Advances in Intelligent Systems and Computing 574

Radek Silhavy
Roman Senkerik
Zuzana Kominkova Oplatkova
Zdenka Prokopova
Petr Silhavy *Editors*

Cybernetics and Mathematics Applications in Intelligent Systems

Proceedings of the 6th Computer Science On-line Conference 2017 (CSOC2017), Vol 2

Springer

Advances in Intelligent Systems and Computing

Volume 574

Series editor

Janusz Kacprzyk, Polish Academy of Sciences, Warsaw, Poland
e-mail: kacprzyk@ibspan.waw.pl

About this Series

The series "Advances in Intelligent Systems and Computing" contains publications on theory, applications, and design methods of Intelligent Systems and Intelligent Computing. Virtually all disciplines such as engineering, natural sciences, computer and information science, ICT, economics, business, e-commerce, environment, healthcare, life science are covered. The list of topics spans all the areas of modern intelligent systems and computing.

The publications within "Advances in Intelligent Systems and Computing" are primarily textbooks and proceedings of important conferences, symposia and congresses. They cover significant recent developments in the field, both of a foundational and applicable character. An important characteristic feature of the series is the short publication time and world-wide distribution. This permits a rapid and broad dissemination of research results.

Advisory Board

More information about this series at http://www.springer.com/series/11156

Radek Silhavy · Roman Senkerik
Zuzana Kominkova Oplatkova
Zdenka Prokopova · Petr Silhavy
Editors

Cybernetics and Mathematics Applications in Intelligent Systems

Proceedings of the 6th Computer Science On-line Conference 2017 (CSOC2017), Vol 2

Springer

Editors
Radek Silhavy
Faculty of Applied Informatics
Tomas Bata University in Zlín
Zlin
Czech Republic

Zdenka Prokopova
Faculty of Applied Informatics
Tomas Bata University in Zlín
Zlin
Czech Republic

Roman Senkerik
Faculty of Applied Informatics
Tomas Bata University in Zlín
Zlin
Czech Republic

Petr Silhavy
Faculty of Applied Informatics
Tomas Bata University in Zlín
Zlin
Czech Republic

Zuzana Kominkova Oplatkova
Faculty of Applied Informatics
Tomas Bata University in Zlín
Zlin
Czech Republic

ISSN 2194-5357 ISSN 2194-5365 (electronic)
Advances in Intelligent Systems and Computing
ISBN 978-3-319-57263-5 ISBN 978-3-319-57264-2 (eBook)
DOI 10.1007/978-3-319-57264-2

Library of Congress Control Number: 2017937149

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

This book constitutes the refereed proceedings of the Cybernetics and Mathematics Applications in Intelligent Systems Section of the 6th Computer Science On-line Conference 2017 (CSOC 2017), held in April 2017.

Particular emphasis is laid on modern trends in mathematical application in intelligent systems, cybernetics, and automation control theory. New algorithms, methods, and applications of intelligent systems in technological systems are also presented.

The volume Cybernetics and Mathematics Applications in Intelligent Systems brings and presents new approaches and methods to real-world problems and exploratory research that describes novel approaches in the defined fields.

CSOC 2017 has received (all sections) 296 submissions, in which 148 of them were accepted for publication. More than 61% of accepted submissions were received from Europe, 34% from Asia, 3% from Africa, and 2% from America. Researches from 27 countries participated in CSOC 2017 conference.

CSOC 2017 conference intends to provide an international forum for the discussion of the latest high-quality research results in all areas related to computer science. The addressed topics are the theoretical aspects and applications of computer science, artificial intelligences, cybernetics, automation control theory, and software engineering.

Computer Science On-line Conference is held online, and modern communication technology which is broadly used improves the traditional concept of scientific conferences. It brings equal opportunity to participate to all researchers around the world.

The editors believe that readers will find the following proceedings interesting and useful for their own research work.

March 2017

Radek Silhavy
Petr Silhavy
Zdenka Prokopova
Roman Senkerik
Zuzana Kominkova Oplatkova

Organization

Program Committee

Program Committee Chairs

Zdenka Prokopova, Ph.D., Associate Professor, Tomas Bata University in Zlin, Faculty of Applied Informatics, email: prokopova@fai.utb.cz

Zuzana Kominkova Oplatkova, Ph.D., Associate Professor, Tomas Bata University in Zlin, Faculty of Applied Informatics, email: kominkovaoplatkova@fai.utb.cz

Roman Senkerik, Ph.D., Associate Professor, Tomas Bata University in Zlin, Faculty of Applied Informatics, email: senkerik@fai.utb.cz

Petr Silhavy, Ph.D., Senior Lecturer, Tomas Bata University in Zlin, Faculty of Applied Informatics, email: psilhavy@fai.utb.cz

Radek Silhavy, Ph.D., Senior Lecturer, Tomas Bata University in Zlin, Faculty of Applied Informatics, email: rsilhavy@fai.utb.cz

Roman Prokop, Ph.D., Professor, Tomas Bata University in Zlin, Faculty of Applied Informatics, email: prokop@fai.utb.cz

Prof. Viacheslav Zelentsov, Doctor of Engineering Sciences, Chief Researcher of St. Petersburg Institute for Informatics and Automation of Russian Academy of Sciences (SPIIRAS).

Program Committee Members

Boguslaw Cyganek, Ph.D., DSc, Department of Computer Science, University of Science and Technology, Krakow, Poland.

Krzysztof Okarma, Ph.D., DSc, Faculty of Electrical Engineering, West Pomeranian University of Technology, Szczecin, Poland.

Monika Bakosova, Ph.D., Associate Professor, Institute of Information Engineering, Automation and Mathematics, Slovak University of Technology, Bratislava, Slovak Republic.

Pavel Vaclavek, Ph.D., Associate Professor, Faculty of Electrical Engineering and Communication, Brno University of Technology, Brno, Czech Republic.

Miroslaw Ochodek, Ph.D., Faculty of Computing, Poznan University of Technology, Poznan, Poland.

Olga Brovkina, Ph.D., Global Change Research Centre Academy of Science of the Czech Republic, Brno, Czech Republic & Mendel University of Brno, Czech Republic.

Elarbi Badidi, Ph.D., College of Information Technology, United Arab Emirates University, Al Ain, United Arab Emirates.

Luis Alberto Morales Rosales, Head of the Master Program in Computer Science, Superior Technological Institute of Misantla, Mexico.

Mariana Lobato Baes, M.Sc., Research-Professor, Superior Technological of Libres, Mexico.

Abdessattar Chaâri, Professor, Laboratory of Sciences and Techniques of Automatic control & Computer engineering, University of Sfax, Tunisian Republic.

Gopal Sakarkar, Shri. Ramdeobaba College of Engineering and Management, Republic of India.

V.V. Krishna Maddinala, Assistant Professor, GD Rungta College of Engineering & Technology, Republic of India.

Anand N. Khobragade, Scientist, Maharashtra Remote Sensing Applications Centre, Republic of India.

Abdallah Handoura, Assistant Prof, Computer and Communication Laboratory, Telecom Bretagne, France

Technical Program Committee Members

Ivo Bukovsky
Miroslaw Ochodek
Bronislav Chramcov
Eric Afful Dazie
Michal Bliznak
Donald Davendra
Radim Farana
Zuzana Kominkova Oplatkova
Martin Kotyrba
Erik Kral
David Malanik
Michal Pluhacek
Zdenka Prokopova
Martin Sysel

Roman Senkerik
Petr Silhavy
Radek Silhavy
Jiri Vojtesek
Eva Volna
Janez Brest
Ales Zamuda
Roman Prokop
Boguslaw Cyganek
Krzysztof Okarma
Monika Bakosova
Pavel Vaclavek
Olga Brovkina
Elarbi Badidi

Organizing Committee Chair

Radek Silhavy, Ph.D., Tomas Bata University in Zlin, Faculty of Applied Informatics, email: rsilhavy@fai.utb.cz

Conference Organizer (Production)

OpenPublish.eu s.r.o.
Web: http://www.openpublish.eu
Email: csoc@openpublish.eu

Conference Website, Call for Papers

http://www.openpublish.eu

Contents

Cost-Effective Computational Modeling of Fault Tolerant Optimization of FinFET-Based SRAM Cells

H. Girish[1(✉)] and D.R. Shashikumar[2]

[1] Department of ECE, J C Bose Centre for Research and Development, Cambridge Institute of Technology, K.R. Puram, Bangalore, India
hgirishphd@gmail.com

[2] Department of Computer Science Engineering, Cambridge Institute of Technology, K.R. Puram, Bangalore, India

Abstract. In the area of computational memory management, energy efficiency and proper utilization of memory cell area is being constantly investigated. However, record of research manuscript in this regards are quite less compared to other related research topic in computer science. We reviewed existing techniques of upgrading the performance of FinFET-based SRAM and found that adoption of computational modeling for optimization is quite a few to find. Hence, we model the problem of leakage power minimization as linear optimization problem and develop a technique that ensures better fault tolerance operation of FinFET-based SRAM using enhanced particle swarm optimization. We minimize the computational complexity of the algorithm compared to conventional evolutionary technique and other performance upgrading system found in recent times. Our algorithm has better control over convergence rate, energy dissipation, and capability to ensure fault tolerance.

Keywords: FinFET · SRAM · Leakage power · Particle swarm optimization · Complexity

1 Introduction

With the advancement of the VLSI and need of prevalent evaluation better computational storage framework, FinFET SRAM has been advanced as a mechanism to offer 10 nm size of transistor configuration. The prime reason of this progressive innovation is because of three dimensional configurations of the gate controls which is bringing down its controlling conditions from ordinary drain and source terminal [1]. In traditional design of transistor, consideration of new components calls for short channel impact, which is totally relieved by present statutory guidelines of FinFET [2]. The prime trade-offs in the design principle of the SRAM are (i) speed vs leakage current, (ii) read vs write stability, and (iii) area vs yield. It is required that an SRAM cell should work faster and should dissipate less leakage power, which unfortunately is still an open end problem [3, 4]. The minimum voltage that a memory cell can use for performing reading operation is called as read voltage. Whereas the write voltage is just the opposite of it i.e. maximum voltages to perform write operation. Hence for better stability during read

R. Silhavy et al. (eds.), *Cybernetics and Mathematics Applications in Intelligent Systems*, Advances in Intelligent Systems and Computing 574, DOI 10.1007/978-3-319-57264-2_1

and write operation, it is required that read and write voltage should be kept minimum and driving strength of AC transistor should be make weaker and stronger during read and write operation respectively. Another problem with existing SRAM is that it has shifted into the large scaled technologies in node design that consider smaller size with minimized level of voltage. This causes narrowing of the difference between the cut-off voltage and the supply voltage. Sometimes, the level of the voltage becomes highly unstable especially in the cache design in CPU where the system design calls for inclusion of transistors with higher reduced size in order to maintain large storage points. It is believed that voltage scaling causes bottlenecks in memory system and in order to address this problem, it is preferred to jointly study FinFET with SRAM. This integrated design principle offers a potential energy efficient feature in storage access design.

This paper presents a fault tolerant optimization technique to upgrade the performance of FinFET based SRAM cells. Section 2 discusses about the existing mechanism of doing so followed by discussion of problems in Sect. 3. Proposed contribution in briefed in Sect. 4 followed by algorithm discussion in Sect. 5. Research methodology is discussed in Sect. 6. Section 7 discusses about the result analysis and finally summary of the paper is done in Sect. 8.

2 Review of Literature

This section discusses about the existing literatures towards improvement in FinFET-based SRAM cells. It also acts as continuation of our prior review work [5]. The design aspects of the SRAM could be significantly improve by focusing on the nanometer area which could be populated with various alternative devices of Field Effect Transistors i.e. FET. The recent reviews performed by Parimaladevi et al. [6] have discussed about the performance factor of the SRAM and has theoretically discussed multiple solutions. The study assists to understand two facts i.e. (i) there are better scope of FET in SRAM for design improvement and (ii) the design of FET itself can be hybridized to attain better objectives. Similar review was also carried out by Bhattacharya and Jha [7]. Discussion on design challenges on FinFET was carried out by Burnett et al. [8]. The most recent study of Zhang et al. [9, 10] have emphasized on low powered applications with FinFET technologies of 7/8 nm. The prototype designed by the author was used to gauge the SRAM with 6 transistors. The study outcome was evaluated with respect to current and voltage. Study towards significance of FinFET on the design improvement was also recently carried out by Lee [11]. The authors presents elaborated discussion towards bulk FinFET and compared its performance over with another type of the FinFET i.e. SOI FinFET. The evaluation was carried out over 14 nm of node and was tested with respect to current-voltage charecteristics. The study has also investigated about the trends of heat dissipation from the 14 nm node to find the temperature reduction capability of 325 K. Study in similar direction was also carried out by Song et al. [12] most recent in 2016. The author has introduced the similar design principle with 10 nM of node with FinFET on SRAM with 128 Mb capacities.

Ansari et al. [13] have presented an elaborated study of design improvement of SRAM cells with 7 transistors. The author has considered a simulation-based study with

HSPICE using multiple number of transistor (20, 16, 14, 10, 7 nm). The outcome of the presented simulation study was found to possess better write speed as well as enhanced stability. The mean static power was also found to be reduced by approximately 57% with existing design of 5T SRAM. A trend of using multiple numbers of transistors involvements was investigated by various researchers. The work carried out by Dani et al. [14] have discussed the charecteristics of 6T SRAM design using FinFET with respect to standby mode, read/write mode, etc. The simulation study outcome was evaluated with respect to power and delay mainly for both read/write operation. Similar trend of study on 6T SRAM was also carried out by Gupta and Roy [15]. Same year (i.e. 2015), Kushwah and Akashe [16] have presented a technique of enhancing the stability of noise margin using SRAM cells with 6 transistors. The study outcome shows better feasibility of stability enhancement during read operation and minimizing the voltage reduction and leakage current. Hence, it can be seen that there are many researchers who choose to implement in SRAM cells with 6 transistors. However, usage of 6 transistors cannot be used to accomplish near-cut-off voltage which is quite important for devices with restricted energy. This problem was addressed by Park et al. [17] where a unique buffer for reading operation was introduced with near cut-off voltage. The outcome shows better write capability with stabilized device operation. Similar direction of the study using SRAM cells with 6 transistors and 22 nm FinFET devices was also investigated by Manju and Kumar [18]. The author have considers access time variation between read and write operation in order to maximize it.

Farkhani et al. [19] have presented a new SRAM design with cell size of 65 nm for incorporating new methods in read/write operations. The technique uses non-positive voltage for enhancing the write charecteristics of SRAM cells. The complete design evaluation was done for SRAM cells with 10 transistors. The simulation outcome of the study was found to possess 82% enhancement to write operation in contrast to conventional SRAM cells with 8 transistor. Shafaei et al. [20] have presented a unique technique to improving performance of FinFET devices. The authors have built a 6T and 8T SRAM cells with 7 nm of FinFET device. The overall study objective was to attain the energy efficient cache memory on FinFET device. Zeinali et al. [21] have presented a study using SRAM cells of 9 transistors with 14 nm FinFET device. The study outcome shows minimization of leakage power by 20% and enhancement of memory access time by 30%. Pal et al. [22] have introduced a double dielectric for enhancing the electrostatic integrity of FinFET in SRAM. Ghai et al. [23] have presented a study that compares the multiple significant parameters for FinFET with respect to analog design. Kerber et al. [24] have developed a double gated FinFET to checks its influence due to strained effects of silicon on static memory. The study outcome shows enhancement in read/write stability in comparison to unstrained FinFET. Villacorta et al. [25] have focused on reliability of SRAM using statistical approach. The next section discusses about the problem identification.

3 Problem Identification

It was observed that majority of the existing mechanism chooses to use either experimental approach or by using hardware-based simulation environment for SRAM and FinFET. Experimental approaches give highly reliable outcomes but none of the studies done till date have actually checked for computational complexity, which makes the approach less applicable in real-time and big-scale commercial usage. Hardware-based approach uses a specific simulation environment which narrows down the scope of computational capability in this. There is a need to develop a computational optimization model in order to enhance the design performance of SRAM FinFET as well as to address the problems of fault tolerance too. There is a need of mathematical optimization principle supported by probability theory for giving better edge to the upgradation of design principles of SRAM based FinFET. There is also a need to focus on the variability factor which has received less attention till date in this field except for few numbers of studies. There are many problems in this topic, but we need to choose such a problem, where we don't have much research work. Hence, some unique problem, which are found to be less addressed in IEEE transaction papers are:

- *Ignorance towards Fault Tolerance*: It is discussed on many papers that gate tunneling and threshold current during read/write process are highly influenced by static leakage current in FinFET SRAM. Usage of computational model of optimization is also less found in literatures.
- *Vague implication of optimization*: Majority of the existing literatures just perform minor improvement of performance parameters and claimed it as optimization. Whereas in real-sense, none of the paper related to FinFET SRAM is found actually implement optimization modeling. However, there are few papers e.g. Wang [26], Lu [27], and Kashfi [28].
- *Tradeoff in research approach and real need of computing*: A closer look into all the IEEE papers on FinFET SRAM will show that their approach is like fine-tuning the technology in order to ensure better customization of transistor charecteristics. However, none of the techniques implemented till date in this can be never considered to be sufficient enough as a transistor will always need to design requirements with respect to system, circuits, and corresponding application. It was because; enough computational modeling is missing from literatures. You can also think of why now-a-days researchers are more inclined towards rapid prototyping in VLSI using computational model.

Along with this, it was also observed that frequently used techniques are more inclined towards memory and size problems, but now we have more problems (but specific) to address i.e. fault tolerance, energy efficiency, and high level optimization, for which we do not have much transaction papers to claim so in FinFET SRAM published between 2010–2016. The next section briefs about the contribution of the proposed system to bridge this research issues.

4 Proposed System

The prime aim of the proposed research work is to develop a computational model for high level of design optimization of FinFET based SRAM cells. The core design objectives is to develop a simple and cost effective fault-tolerant model that can significantly optimize stochastically the design performance of FinFET based SRAM cells. The schematic architecture of the proposed system is shown in Fig. 1.

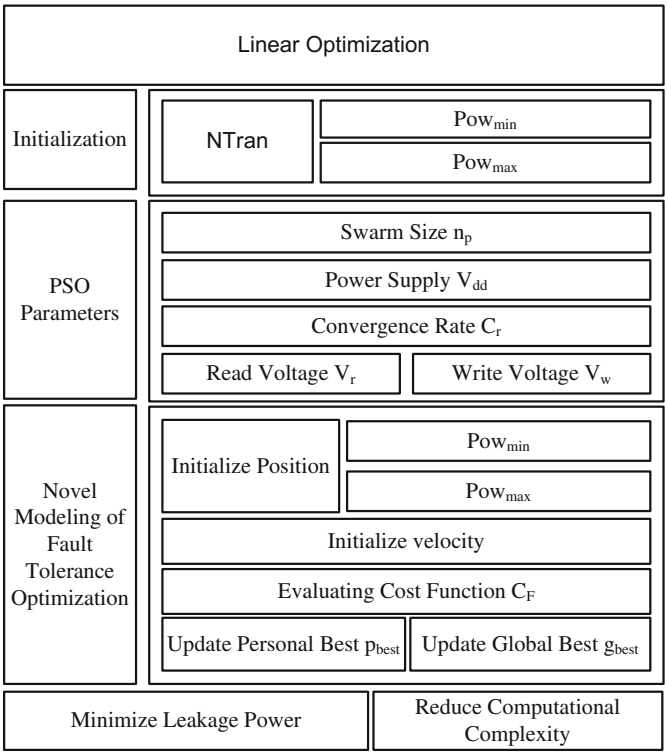


Fig. 1. Schematic architecture of proposed system

Figure 1 shows that the proposed system maps the problem of power optimization as linear optimization problem where the system is initialized by number of transistors with upper and lower limits of power. This mechanism is designed to obtain reduced power level in order to ensure that the system has achieved fault tolerance. This will mean that the outcome of the proposed system (i.e. leakage power) will be very much lower than maximum limit of initialized power of cell array. Apart from power efficiency, we are also interested to incorporate algorithm computational efficiency, which was never considered in any existing techniques of enhancing performance of FinFET based SRAM cells. For this reason, we apply enhanced Particle Swarm Optimization (PSO) with its associated parameters as shown to be Swarm size (i.e. population), power supply, convergence rate, and read/write voltage in Fig. 1. The next part of the study

focused on developing fault tolerance optimization where the focus was laid to obtain better personal best and global best responses from the technique by applying cost function. Finally, the system computes the leakage power and checks for computational complexity. The consecutive sections elaborate about the research methodologies and algorithm implementation.

5 Research Methodology

The research methodology adopted for the proposed approach is purely analytical approach. The primary objective to be achieved in this proposed methodology is to perform optimization of the design of FinFET based SRAM cells. The targets achieved in this objective are to increase the yield of SRAM and make it more faults tolerant. A novel computational framework is developed to understand impact of simple optimization principle on ensuring fault tolerance execution of FinFET SRAM. The model consider the problem of read/write stability of FinFET based SRAM cell arrays and also the model considers the problem of access time, which it will address by investigating/evaluating the read current of SRAM. The investigation of fault tolerance was carried out by considering optimization parametric variation of channel length and thickness of silicon for FinFET using probability theory.

In order to perform optimization, the proposed system adopts evolutionary algorithm where the prime target is to develop a computational model of evolution of elite outcomes for a specific optimization problem. The study considers minimization of leakage power and ensures fault tolerance towards multiple forms of failures (e.g. read, write, access) during individual operations in FinFET based SRAM cell array. The complete design of the proposed optimization was carried out using particle swarm optimization which has both beneficial factor and limiting factor while implementing in circuit designs. The beneficial factors would be consideration of lesser number of dependable components as compared to any optimization (evolutionary) techniques exist and its applicability to larger dimensional problem. The problem of minimization of leakage power with constraint of fault tolerant compliant parameters of read/write/access can be stated as linear optimization problem, which can be suitable modeled and optimized using particle swarm optimization. However, usage of conventional particle swarm optimization is also associated with limitations e.g. the conventional particle swarm optimization is highly recursive in case of finding the particle best solution if the problem space is too large. Hence, we address the problem of recursive charecteristics by incorporating update computation of the power instantly using a linear optimization representation of cost function. This allows reaching the best solution (minimal power dissipation) within two ranges of limit (low, high) without using conventional threshold based mechanism. Our focus is more on achieving algorithm efficiency on 6T SRAM cells; however, we kept it flexible to be changed from 6T to some other numbers of transistors to check for optimized outcome. Till date the techniques implemented using particle swarm optimization and genetic algorithms are more focused on achieving operational efficiency of hardware without even choosing the correct parameters. Our

computational model is designed in such a way that can be used for multiple forms of FinFET-based SRAM configurations.

6 Algorithm Implementation

The primary aim of the proposed algorithm is to ensure that leakage power is minimized. We have developed a mechanism where a simple linear programming is used as a cost function in order to perform optimization of leakage power. An algorithm is developed that takes the input of Number of Transistor and Variable size. The algorithm is designed to compute for multiple numbers of transistors on any FinFET-based SRAM cells. The steps of the algorithm are shown as following:

Algorithm for Minimizing Leakage Power
Input: *nTran*, v_{size}, Pow_{min}, Pow_{max}, V_{dd}, C_r, V_r, V_w, n_p, p_{pos}, p_{vel}, p_{cost}, C_F, g_{best}
Output: Minimized Leakage Power
Start
1. init nTran, v_{size}, Pow_{min}, Pow_{max}, w, C_r, V_r, V_w.
2. Pow_{min} = - (0.1*[Pow_{max}-Pow_{min}])
3. If i=1:n_p
4. p_{pos}→cud(Pow_{min}, Pow_{max})
5. p_{vel}→[]
6. p_{cost}→C_F(p_{pos}(i));
7. p_{best}→[p_{pos}, p_{cost}] & g_{best}→if(p_{best}(p_{cost}<inf))
8. end
9. For i=1:round
10. For j=1:n_p
11. p_{vel}→V_{dd}*p_{vel}+V_r*rand(p_{best}(p_{pos})-p_{pos})+V_w*rand(g_{best}(p_{pos})-p_{pos})
12 opt→[p_{pos}<Pow_{min} | p_{pos}>Pow_{max}]
13. p_{cost}→C_F(p_{pos})
14. If p_{cost}<p_{best}
15. update p_{best}(p_{pos}, p_{cost})
16. if (p_{best}(p_{cost})<g_{best}(p_{cost}))
17. g_{best}←p_{best}
18. end
19. end
20. V_{dd}←V_{dd}*C_r
21. leakage Power←g_{best}(p_{cost})
22. End
23. End
End

We develop cost function C_F as the function representing linear programming i.e. C_F = sum(x^2) where x are data points of integer type. The algorithm is capable of performing simulation study for 1T to XT where X is maximum number of transistors. We considered X to be 6. We apply the optimization technique using upper and lower bounds of power factor. The algorithm than initializes power supply voltage, read voltage and writes voltage. The limits of power are computed by multiplying lowest value of probability with difference of maximum and minimum power (Line-2). The

negative value of it is considered to be lowest power possible by FinFET SRAM cells. The next step is to apply the particle swarm optimization where we add out cost function in it to achieve minimized output of leakage power. The particle position (p_{pos}) is found by using numbers of continuous uniform arbitrary lying between minimum and maximum power of FinFET based SRAM cells (Line-4). We keep the particle velocity as empty matrix in order to obtain better solution (Line-5). This empty matrix is going to be base for all optimized outcomes in future iterations. A cost function is then applies to particle position (p_{pos}), which is none other than power factor itself. In order to minimize the recursive steps of particle swarm optimization to perform update of position and velocity factor, we apply evaluation of particle best and global best value right after this step (Line-7). This step results in minimization of iterations (Line-3) to 50% as compared to conventional particle swarm optimization. We consider p_{best} with respect to particle position and cost (Line-7), while global best g_{best} as any p_{best} value whose cost which is less than infinity (Line-8). This step of the algorithm ensure faster convergence rate and allow not to iterate it for infinite loop.

The strategy towards PSO implementation is to maximize the outcomes of SRAM array for optimizing certain operations e.g. write/read/access time failures. The PSO parameters are Power Supply, Convergence Rate, Read Voltage, and Write Voltage. The next step is to perform optimization for all the number of populations. We calculate particle velocity (p_{vel}) with the equation highlighted in Line-11. The dependable parameter of this equation is power supply voltage (V_{dd}), read voltage (V_r), and write voltage (V_w). This line of algorithm allows the FinFET-based SRAM to be highly fault tolerant. The algorithm allows to access the SRAM during read operation of the cell in lowered voltage only. We perform updating of the particle velocity with respect to limits of fault tolerant Pow_{min} and Pow_{max}. The updating operation is also carried out for particle position. The algorithm than perform concatenation of particle position which is within an exact range of minimum and maximum power. The concatenated value will represent the eligible power or optimized power that will ensure fault tolerance while working with FinFET-based SRAM (Line-12). Cost function is further applied on particle position (Line-13) in order to be used for comparative analysis for elite value of leakage power. If the particle cost is found less than cost of pbest value, its respective position and cost values are recorded and updated (Line-15). In case the cost of pbest is found to be less than the cost of gbest, we consider pbest itself as the gbest value (Line-17). The global best or gbest value is basically the leakage power. The applicability of this algorithm for minimizing leakage power is quite high as PSO is a stochastic method with each particle associated with velocity and position information. It has been already proven in the past that PSO could be easily modeled in analog circuits, RF filter, on-chip spiral inductors, etc. in comparison to other existing optimization techniques. The significant contribution of the proposed system is its total governing of convergence rate in order to ensure fault tolerant design principle of FinFET based SRAM cells. The next section discusses about the outcomes being accomplished.

7 Result Analysis

This section discusses about the results being accomplished from the implementation of an algorithm discussed in previous section. The results were observed considering the maximum simulation rounds of 1000, number of population as 100, supply voltage as 1 v, convergence rate as 0.99, read voltage as 1.5 v and write voltage as 2.0 v. We observe a gradient descent trend of our Leakage power curve just in a matter of 0.02 s. The outcome shown in Fig. 2 is for 6T SRAM with FinFET where after 842 rounds, the leakage power is totally neutralized. For 7T and 8T SRAM, the neutralization of leakage power was observed at 906 and 993 simulation rounds respectively.

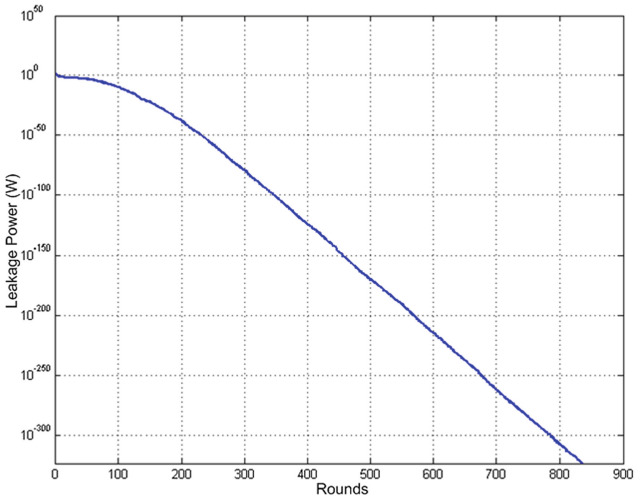


Fig. 2. Trend of leakage power

At present, there are various literatures to prove that there are some attempts of optimization using evolutionary techniques like genetic algorithm and particle swarm optimization. For the purpose of the better benchmarking, the proposed system is compared with the recent work being carried out by Ebrahimi [29] and Tang [30]. Ebrahimi et al. [30] have implemented a technique of statistical optimization to upgrade the performance of FinFET based SRAM cells using particle swarm optimization. The author didn't performed computational analysis on their PSO techniques in their study. Similarly, the work done by Tang [30] considers performing optimization using Genetic Algorithm (GA) for similar cause. Hence, the aim of both the researchers were common i.e. upgrading the performance of FinFET based SRAM cells but their objectives differs as one is achieved by using PSO and other by GA. We also perform statistical verification from the present simulation environment to see that proposed system offer higher failure probability for read operation (0.75), low failure probability for write operation (0.29), and much lower statistical value of failure probability (0.01) for access operation.

The numerical outcome of the computational complexity is tabulated in Table 1.

Table 1. Complexity analysis

n	Proposed (PSO)	Ebrahimi (PSO) [29]	Tang (GA) [30]
	log n	n log n	N^(3/2) log N
100	2.00	200.000	2000.000
200	2.30	460.205	6508.295
300	2.47	743.136	12871.499
400	2.60	1040.823	20816.479
500	2.69	1349.485	30175.401
600	2.77	1666.890	40830.317
700	2.84	1991.568	52691.953
800	2.90	2322.471	65689.427
900	2.95	2658.818	79764.547
1000	3	3000	94868.329

The tabulated outcome clearly indicates the proposed system offers better fault tolerance with much lesser computational complexity as compared to existing mechanism of optimization. By this way, the system ensures a higher degree of resilience against read failures, write failures, and access time failures. The best part of the proposed system is its flexibility to be configured as the way the user wants it to be implemented. The mechanism adopted by Ebrahimi [29] has introduced too many variables of recursive type e.g. on-current, sub-threshold current, read stability, write stability, read current, and sub-threshold leakage power. This results in extra computational complexity. Apart from this the authors have also used yield optimization using back gate voltage on Monte Carlo simulation resulting is exponential growth of complexity by $n \log n$, where n is total number of population. Work carried out by Tang et al. [30] used an optimizer called as GenFin for applying genetic algorithm using TCAD and used a cache model for developing 6T SRAM cell arrays. We replace all of these using simple particle best and global best by cutting down the recursive rounds by 50%. We also use simple mechanism to control fault tolerance by setting the permissible limits of leakage power. Because of such simple inclusion, our mechanism offers similar solutions but in much cheaper computational cost.

8 Conclusion

This paper discusses about the techniques that is meant for enhancing the operations of FinFET-based SRAM. The complete optimization is carried out by enhancing the conventional particle swarm optimization technique by changing the way the pbest and gbest values to be calculated. This feature now allows more intelligence of the cells and blocks of memory to be considered and can be applied on any configuration of transistor i.e. 6T, 7T, 8T etc. The potential contribution of the proposed study is its mechanism to overcome the recursive problems to reduce the computational complexity by roughly 89%. The technique is very simple and completely adheres to specified limits of fault

tolerance i.e. minimum and maximum power. Our future direction of the study will be to further apply more extensive optimization for accomplishing enhanced throughput.

References

1. Reis, R., Cao, Y., Wirth, G.: Circuit Design for Reliability. Springer, New York (2014)
2. Han, W., Wang, Z.M.: Toward Quantum FinFET. Springer Science & Business Media, Switzerland (2013)
3. Shin, C.: Variation-Aware Advanced CMOS Devices and SRAM. Springer, Dordrecht (2016)
4. Prince, B.: Vertical 3D Memory Technologies. Wiley (2014)
5. Girish, H., Shashikumar, D.R.: Insights of performance enhancement techniques on FinFET-based SRAM cells. Commun. Appl. Electr. (CAE) **5**(6), 20–26 (2016). Foundation of Computer Science
6. Parimaladevia, M., Sharmilab, D., Kowsikaa, L.: A survey on the performance analysis of 6t sram cell using novel devices. South Asian J. Eng. Technol. **2**(18), 71–77 (2016)
7. Bhattacharya, D., Jha, N.K.: FinFETs: from devices to architectures. Adv. Electr. (2014). Hindawi Publishing Corporation
8. Burnett, D., Parihar, S., Ramamurthy, H., Balasubramanian, S.: FinFET SRAM design challenges. In: IEEE International Conference on IC Design and Technologies, pp. 1–4 (2014)
9. Zhang, X., Connelly, D., Zheng, P., Takeuchi, H.: Analysis of 7/8-nm bulk-si FinFET technologies for 6T-SRAM scaling. IEEE Trans. Electron Dev. **63**(4), 1502–1507 (2016)
10. Zhang, X.: Simulation-based study of super-steep retrograde doped bulk FinFET technology and 6T-SRAM yield. Doctorial Thesis on University of California at Berkeley (2016)
11. Lee, J.H.: Bulk FinFETs: design at 14 nm node and key characteristics. In: Kyung, C.-M. (ed.) Nano Devices and Circuit Techniques for Low-Energy Applications and Energy Harvesting. KAIST Research Series, pp. 33–64. Springer, Dordrecht (2016)
12. Song, T., Rim, W., Park, S., Kim, Y.: A 10 nm FinFET 128 Mb SRAM with assist adjustment system for power, performance, and area optimization. In: IEEE International Solid-State Circuits Conference (2016)
13. Ansari, M., Kusha, H.A., Ebrahimi, B., Navabi, Z.: A near-threshold 7T SRAM cell with high write and read margins and low write time for sub-20 nm FinFET technologies. J. Integr. VLSI J. **50**, 91–106 (2015). Elsevier
14. Dani, L.M., Singh, G., Kaur, M.: FinFET based 6T SRAM cell for nanoscaled technologies. Int. J. Comput. Appl. **127**(13), 3 (2015)
15. Gupta, S.K., Roy, K.: Low power robust FinFET-based SRAM design in scaled technologies. In: Reis, R., Cao, Y., Wirth, G. (eds.) Circuit Design for Reliability, pp. 223–253. Springer, New York (2015)
16. Kushwah, R.S., Akashe, S.: FinFET-based 6T SRAM cell design: analysis of performance metric, process variation and temperature effect. InderScience Int. J. Sig. Imaging Syst. Eng. **8**(6), 2500–2506 (2015)
17. Park, J., Yang, Y., Jeong, H., Song, S.C., Wang, J.: Design of a 22-nm FinFET-based SRAM with read buffer for near-threshold voltage operation. IEEE Trans. Electron Devices **62**(6), 1698–1704 (2015)
18. Manju, I., Kumar, A.S.: A 22 nm FinFET based 6T-SRAM cell design with scaled supply voltage for increased read access time. Analog Integr. Circ. Sig. Process **84**(1), 119–126 (2015). Springer
19. Farkhani, H., Peiravi, A., Moradi, F.: A new write assist technique for SRAM design in 65 nm CMOS technology. Integr. VLSI J. **50**, 16–27 (2015). Elsevier

20. Shafaei, A., Chen, S., Wang, Y., Pedram, M.: A cross-layer framework for designing and optimizing deeply-scaled FinFET-based cache memories. J. Low Power Electr. Appl. **5**, 165–182 (2015)
21. Zeinali, B., Madsen, J.K., Raghavan, P., Moradi, F.: Sub-threshold SRAM design in 14 nm FinFET technology with improved access time and leakage power. In: IEEE Computer Society Annual Symposium on VLSI (2015)
22. Pal, P.K., Kaushik, B.K., Dasgupta, S.: Design metrics improvement for SRAMs using symmetric dual-k spacer (SymD-k) FinFETs. IEEE Trans. Electron Devices **61**(4), 1123–1130 (2014)
23. Ghai, D., Mohanty, S.P., Thakral, G.: Comparative analysis of double gate FinFET configurations for analog circuit design. In: IEEE International Midwest Symposium on Circuits and Systems, pp. 809–812 (2013)
24. Kerber, P., Kanj, R., Joshi, R.V.: Strained SOI FINFET SRAM design. IEEE Electron Device Lett. **34**(7), 876–878 (2013)
25. Villacorta, H., Champac, V., Bota, S., Segura, J.: FinFET SRAM hardening through design and technology parameters considering process variations. In: IEEE European Conference on Radiation and Its Effects on Components and Systems, pp. 1–7 (2013)
26. Wang, W., Areibi, S., Anis, M.: Modeling leakage power reduction in VLSI as optimization problems. Optim. Eng. **8**(2), 129–162 (2007). Springer
27. Lu, B., Sapatnekar, S.S., Du, D.: Layout Optimization in VLSI Design, vol. 8. Springer, New York (2001)
28. Kashfi: Multi-objective optimization techniques for VLSI circuits (2011)
29. Ebrahimi, B., Rostami, M., A-Kusha, A., Pedram, M.: Statistical design optimization of FinFET SRAM using back-gate voltage. IEEE Trans. Very Large Scale Integr. (VLSI) Syst. **19**(10), 1911–1916 (2011)
30. Tang, A., Gao, X., Chen, L.-Y., Jha, N.K.: Delay/Power modeling and optimization of FinFET circuit modules under PVT variations: observing the trends between the 22 nm and 14 nm technology nodes. ACM J. Emerg. Technol. Comput. Syst. **12**(4), 42 (2016). Article 42

Application of Risk Theory Approach to Fuzzy Abduction

V.N. Tsypyschev(✉)

Moscow Technological University,
78, Vernadsky Avenue, Moscow 119454, Russian Federation
tsypyschev@yandex.ru

Abstract. In this article, learning under the absence or incompleteness of some facts or premises about the problem domain is considered. This task does not fall under semi-supervised learning in the classical sense, because there it is assumed that the target signals are known and correct. The assumption of incompleteness is, however, natural for pattern recognition, e.g. for medical diagnostics.

In such a situation, it is natural to base a learning process on abductive reasoning instead of induction or transduction. It is then important to have a quality criterion for the state of knowledge on the object to be studied.

Previously, to reconstruct missing training data, a fuzzy logical approach to the application of the abductive reasoning method was studied. Now, fuzzy abduction is considered from a risk-theoretical point of view.

As a result, in addition to the fuzzy abduction method, a general algorithm is suggested for finding the true state of the object to be studied in the case when known hypotheses about its state are mutually far from each other.

Keywords: Fuzzy abduction · Fuzzy systems · Logical deduction · Risk theory

1 Introduction

During learning processes, sometimes cases appear which may be characterized by incompleteness, implausibility or just absence of some necessary information. Not only, as in semi-supervised learning, target signals might be absent. It may be that more generally knowledge of the object to be studied is too incomplete to infer from known facts.

In this case, the task of learning should be seen as a task of elicitation, eduction or establishing of causal relationships. Unfortunately, in this case, an application of any model is impossible because modelling, as usual, demands presence and completeness of information about the object to be modelled.

In memory of T.Y. Morozova.

R. Silhavy et al. (eds.), *Cybernetics and Mathematics Applications in Intelligent Systems*, Advances in Intelligent Systems and Computing 574, DOI 10.1007/978-3-319-57264-2_2

It is then natural to apply abduction as a reasoning method to the considered problem. This method may be viewed as a special kind of inference generating supplements. Supplements consist of auxiliary disjunction forms necessary to construct a disjunction form of additional premises, which is essential for successful deductive inference.

For this, an analytical model of learning explanations can be used, and from an algorithmic point of view, one can use a resolution method as developed in [1–3,7,8,10–12,14,16–19].

Abduction may be considered as reverse deduction. In classical deductions, it is assumed [9] that the facts are true and the inferences rules by which conclusions are drawn are known and therefore the conclusions one draws are true. Unlike an application of deduction, an application of abduction is characterized by incomplete knowledge of facts and the necessity to reconstruct the cause of known output. For this, the deductive inference rule is transformed into the new abduction rule, which can be stated as follows: If the conclusion Q is true and P causes Q then this suggests that P might also be true. Previously, this approach was for example studied in [6] and many other works referenced above.

Let us assume that there are reasons to construct a decision support system, wherein it is necessary to infer online from some incomplete set of not entirely trusted facts by applying a particular a priory rule given by an expert, resulting in a conclusion and an according reasoning supporting the conclusion. To give an example: a decision-making person supervising some process must have plausible explanations in an abnormal case. It is then obvious that an application of the abduction method might be fruitful.

This approach mandates quality criteria for explanations supporting the conclusions. For example, the above decision-making person must have quantized evaluations of plausibility of each of the explanations or, at least, an ordering of the explanations according to their plausibility.

With inductive learning one searches for the state of the object in the search space, and the method consists in an application of inductive inference rules to a fixed set of initial patterns. This however falls short of what one wishes to obtain here.

This suggests to apply abductive reasoning to construct the necessary decision support system. But then the following problem arises: The learning process is not a homogeneous process, and the quality of generated explanations is not constant in time and depends crucially on the quantity of absorbed information via the learning process. The decision support system as part of the learning system must be controlled permanently with respect to the quality of what has been learned. We call the current state of the learning system studying an object the *current state of knowledge regarding the object* or, shortly, *state of knowledge*.

This suggestion is necessary because the method of abductive learning is very complicated, not properly investigated, and may be only roughly represented by inductive learning. On the other hand, the decision-making person has to apply the most plausible explanation. Hence, it is necessary to suggest some additions to classical abduction.

2 Methods

To start with, we consider the case that $m+1$ simple hypotheses $H_j, j = \overline{0,m}$, are suggested to determine the true position of an object under learning in the search space. Each hypothesis consists in that: if observing position of object $\bar{x} = (x_1, \ldots, x_n)$ falls to domain X_k of the search space then decision γ_k is adopted and it means that this position $\bar{x}$ corresponds to state of knowledge $S_k, k = \overline{0,m}$.

To construct the decision making rule, we shall use the criterion of minimal average risk [5,15].

An application of any prior established decision making rule involves the possibility of false decision because of the probabilistic nature of the considered object. The observed sample of explanations $\bar{x} = (x_1, \ldots, x_n)$ may fall into the domain X_k corresponding to decision γ_k that the statement S_k is true though indeed this sample corresponds to other state S_j, $j \neq k$. The presence of not only true but also of false decisions in the sequence of decisions is a price for making decisions under conditions of incomplete information. Consequences of false decisions are accounted for by a function (matrix) of losses, which ties with every false decision, i.e. a pair (S_j, γ_k), $j \neq k$, the payment $\ddot{I}_{j,k} = \ddot{I}(S_k, \gamma_j) > 0$, and with right decision the payment $\ddot{I}_{j,j} = \ddot{I}(S_j, \gamma_j) < \ddot{I}_{j,k}, k \neq j$.

An application of a certain decision making rule means nothing less than that the sample space is divided into domains $\{X_k\}$ and corresponding decisions $\{\gamma_k\}$ are given. For a given state S_k, an average value of losses is equal to the average value of losses in the sample space (mathematical expectation):

$$r_j = \sum_{k=0}^{m} \ddot{I}_{j,k} P(\gamma_k / S_j) = \sum_{k=0}^{m} \ddot{I}_{j,k} P(\bar{x} \in S_k / S_j),$$

wherein $P(\gamma_k/S_j)$ is the conditional probability of getting samples to domain X_k under the condition that the true state is S_j. The conditional average value of losses r_j for state S_j is known as a conditional risk [5].

Taking an average conditional risk for all states S_j, we obtain

$$R = \sum_{j=0}^{m} r_j p_j = \sum_{j=0}^{m} \sum_{k=0}^{m} p_j \ddot{I}_{j,k} P(\bar{x} \in X_k / S_j), \tag{1}$$

wherein p_j is a prior probability of S_j.

This value may be seen as a quality criterion of the decision making rule, consisting of partitioning the sample space into m nonintersecting domains and assigning to each of the domains X_k a decision γ_k that the hypothesis H_k is true.

The probability that the observed sample $\bar{x}$ will entail accepting of decision γ_k under condition the hypothesis H_j is true is equal to

$$P(\gamma_k / H_j) = P(\bar{x} \in X_k / S_j) = \int_{X_k} W(\bar{x} / S_j) d\bar{x}. \tag{2}$$

If we substitute (2) into (1), we obtain the value of the average risk

$$R = \sum_{j=0}^{m} r_j p_j = \sum_{j=0}^{m} \sum_{k=0}^{m} p_j \ddot{I}_{j,k} \int_{X_k} W(\bar{x}/S_j) d\bar{x},$$

which depends on the partitioning the sample space into the domains $X_k, k = \overline{0,m}$. This means that the value of R is a quantized measure of quality of the decision rule.

Now a good criterion to determine the optimal selection of rule-making consists in the minimization of the value of the average risk R.

3 Main Results

Before the suggested contribution is given, we explain the method by an example:

Example 1
Consider the matrix of losses for the case of two hypothesis:

$$\ddot{I} = \begin{pmatrix} \ddot{I}_{0,0} & \ddot{I}_{0,1} \\ \ddot{I}_{1,0} & \ddot{I}_{1,1} \end{pmatrix} \tag{3}$$

wherein $\ddot{I}_{1,0} > \ddot{I}_{0,0} \geq 0,\ \ddot{I}_{1,0} > \ddot{I}_{1,1} \geq 0$, the rows correspond to hypothesis H_0, respectively hypothesis H_1 and the columns correspond to decisions $\gamma_k, k = \overline{0,1}$. The right solution costs are located on the main diagonal, losses for the wrong decisions are located on the side diagonal. The average value of losses (average risk) is equal to

$$R = qr_0 + pr_1, \tag{4}$$

wherein

$$r_0 = \ddot{I}_{0,0} P(\gamma_0/H_0) + \ddot{I}_{0,1} P(\gamma_1/H_0) = \ddot{I}_{0,0}(1-\alpha) + \ddot{I}_{0,1}\alpha, \tag{5}$$

$$r_1 = \ddot{I}_{1,0} P(\gamma_0/H_1) + \ddot{I}_{1,1} P(\gamma_1/H_1) = \ddot{I}_{1,0}\beta + \ddot{I}_{1,1}(1-\beta) \tag{6}$$

are conditional risks corresponding to states H_0, H_1 respectively, α is the probability of type I error, i.e. the probability of rejecting a correct hypothesis H_1 and accepting incorrect hypothesis H_0 (false negative), β is a probability to reject a correct hypothesis H_0 and to accept an incorrect hypothesis H_1 (false positive, probability of type II error).

Substituting (5) and (6) into (4), we obtain that

$$R = q\ddot{I}_{0,0} + p\ddot{I}_{1,0} + q(\ddot{I}_{0,1} - \ddot{I}_{0,0})\alpha - p(\ddot{I}_{1,0} - \ddot{I}_{1,1})(1-\beta). \tag{7}$$

The dependence of an average risk on the domain X_1 is expressed via values α and $1-\beta$. Let us substitute them into (1):

$$R = q\ddot{I}_{0,0} + p\ddot{I}_{1,0} - \int_{X_1} \left[p(\ddot{I}_{1,0} - \ddot{I}_{1,1}) W(\bar{x}/S_1) - q(\ddot{I}_{0,1} - \ddot{I}_{0,0}) W(\bar{x}/S_0) \right] d\bar{x}, \tag{8}$$

where $W(\bar{x}/S_0), W(\bar{x}/S_1)$ are likelihood functions.

Since $q\ddot{I}_{0,0} + p\ddot{I}_{1,0}$ is a constant term, an average risk R gets its minimal value under condition:

$$\forall \bar{x} \in X_1 \; p(\ddot{I}_{1,0} - \ddot{I}_{1,1})W(\bar{x}/S_1) \geq q(\ddot{I}_{0,1} - \ddot{I}_{0,0})W(\bar{x}/S_0),$$

That is, the set X_1 may be determined as

$$X_1 = \left\{ \bar{x} \; : \; \frac{W(\bar{x}/S_1)}{W(\bar{x}/S_0)} \geq \frac{q}{p} \cdot \frac{(\ddot{I}_{0,1} - \ddot{I}_{0,0})}{(\ddot{I}_{1,0} - \ddot{I}_{1,1})} \right\}.$$

The function

$$l(\bar{x}) = \frac{W(\bar{x}/S_1)}{W(\bar{x}/S_0)}$$

is a likelihood ratio and represents a non-negative random variable obtained by transformation $z = l(\bar{x})$, i.e. by transformation mapping points of n-dimensional sample space into $\mathbb{R}_+$.

End of Example 1

Using the previous example as a basis and continuing by induction, one obtains a proof of the following Theorem:

Theorem 1. *Let assume that the positions S_k of the object O in a search space X are determined by the sample $\bar{x} = (x_1, \ldots, x_n)$. Let us also assume that it is possible to generate $m+1$ distinguishable simple hypotheses $H_j, j = \overline{0,m}$ with distribution density functions $W(\bar{x}/S_j), j = \overline{0,m}$ and suppose that the true state S of an object under consideration is determined by appropriate decision making.*

Than the decision making rule based on minimization of the average risk R may be constructed by division of the sample space X into $m+1$ non-intersected domains $X_0, X_1, \ldots, X_m$ according to this rule: the domain $X_k, k = \overline{1,m}$, is the set of solutions of m linear inequalities

$$\sum_{i=0}^{m} (\ddot{I}_{i,j} - \ddot{I}_{i,k}) \frac{p_i \cdot W(\bar{x}/S_i)}{p_0 \cdot W(\bar{x}/S_0)} \geq 0, j = \overline{0,m}, j \neq k,$$

$$X_0 = X \setminus \bigcap_{k=1}^{m} X_k,$$

and the state S_k is adopted as true if and only if $\bar{x} \in X_k$.

Theorem 1 may be simplified and, moreover, the decision making rule may be transformed to operating in a sample space of fixed dimension:

Theorem 2. *Under the conditions of Theorem 1, let*

$$y_i = \frac{p_i}{p_0} l_i(\bar{x}) = \frac{p_i \cdot W(\bar{x}/S_i)}{p_0 \cdot W(\bar{x}/S_0)}.$$

Then the set $\tilde{X}_k, k = \overline{1,m}$ *is determined by intersection of planes in* m*-dimensional space*

$$\sum_{i=1}^{m}(\ddot{I}_{i,j} - \ddot{I}_{i,k})y_i \geq \ddot{I}_{0,k} - \ddot{I}_{0,j}, j = \overline{0,m}, j \neq k,$$

$$\tilde{X}_0 = \tilde{X} \setminus \bigcap_{k=1}^{m} \tilde{X}_k,$$

and the state S_k *is adopted as true iff* $\bar{y} \in \tilde{X}_k$.

4 Conclusions

The question of obtaining plausible knowledge about the true position of some object in a learning process under the condition of implausible or uncertain information used in learning was considered. Because usual logical-based reasoning is inapplicable (see, e.g. [4]), a method of quasi-abduction reasoning about the true state of the considered object based not on a logical but on a risk-theoretic approach was suggested. Namely, it was suggested to reduce this problem to the problem of minimizing the risk associated with the decision to be made. This problem was then reformulated as a linear programming problem in a space of fixed dimension. An algorithm for decision-making is provided. The next step of investigation is to provide an algorithm for generating hypotheses about the current state of knowledge regarding an object studied with an abductive learning process.

Author used ideas of [13,20]. Also he's very grateful to Claus Diem for attention paid to this work.

References

1. Belohlavek, R.: Pavelka-style fuzzy logic in retrospect and prospect. Fuzzy Sets Syst. **281**(15), 61–72 (2015)
2. Cheng, C.-L., Lee, R.C.-T.: Symbolic Logic and Mechanical Theorem Proving. Academic Press, New York (1973)
3. Dubois, Didier, Lang, J., Prade, H.: Fuzzy sets in approximate reasoning, Part 2: logical approaches. Fuzzy Sets Syst. **40**(1), 203–244 (1991)
4. Dubois, D., Esteva, F., Godo, L., Prade, H.: Fuzzy-set based logics - an history-oriented presentation of their main developments. In: Handbook of the History of Logic 8. The Many Valued and Nonmonotonic Turn in Logic, pp. 325–449. Elsevier (2007). ISBN 978-0-444-51623-7
5. Roeser, S., Hillerbrand, R., Sandin, P., Peterson, M. (eds.): Handbook of Risk Theory: Epistemology, Decision Theory, Ethics, and Social Implications of Risk. Springer, Dordrecht (2012)
6. Ivanova, I.A., Morozova, T.Yu.: Logical conclusion in indistinct systems and indistinct abduction. In: 8th Open German-Russian Workshop "PATTERN RECOGNITION and IMAGE UNDERSTANDING" OGRW-8, pp. 96–99 (2014)

7. Ivanova, I., Morozova, T., Nikonov, V., Nikolaev, A.: Fuzzy gestures recognition method in development of contact-less interfaces. Int. J. Adv. Stud. **4**(1), 27–31 (2014)
8. Kim, C.S., Kim, D.S., Park, J.S.: A new fuzzy resolution principle based on the antonyms. Fuzzy Sets Syst. **113**, 299–307 (2000)
9. Kleene, S.C.: Mathematical Logic. Wiley, New York (1967)
10. Lee, C.T.: Fuzzy logic and the resolution principle. J. ACM **19**(1), 109–119 (1972)
11. Leonenkov, A.V.: Fuzzy Modeling in MATLAB and fuzzyTECH Environment, 278 p. BHV, Petersburg, Saint-Petersburg (2013)
12. Mendel, J.M., John, R.I., Liu, F.: Interval type-2 fuzzy logic systems made simple. IEEE Trans. Fuzzy Syst. **14**(6), 808–821 (2006)
13. Pankov, V.L.: Stimulation mechanism effectiveness and potential level of satisfaction of the needs of the employee. HERALD of MSTU MIREA **4**(1), 288–291 (2015)
14. Pedrycz, W., Reformat, M.: Evolutionary fuzzy modeling. IEEE Trans. Fuzzy Syst. **11**(5), 652–665 (2003)
15. Sakawa, M., Nishizaki, I., Uemura, Y.: Fuzzy programming and profit and cost allocation for a product and transportation problem. Eur. J. Oper. Res. **131**(1), 1–15 (2001)
16. Zadeh, L.A.: The concept of a linguistic variable and its applications to approximate reasoning, part I. Inf. Sci. **8**, 199–249 (1975)
17. Zadeh, L.A.: The concept of a linguistic variable and its applications to approximate reasoning, part II. Inf. Sci. **8**, 301–357 (1975)
18. Zadeh, L.A.: The concept of a linguistic variable and its applications to approximate reasoning, part III. Inf. Sci. **9**, 43–80 (1975)
19. Zadeh, L.A.: Fuzzy logic and approximate reasoning. Synthese **30**, 407–428 (1975)
20. Tsypyschev, V.N.: Full periodicity of Galois polynomials over nontrivial Galois rings of odd characteristic. J. Math. Sci. **131**(6), 6120–6132 (2005)

Enhanced TDS Stability Analysis Method via Characteristic Quasipolynomial Polynomization

Libor Pekař(✉)

Faculty of Applied Informatics, Tomas Bata University in Zlín,
Nad Stráněmi 4511, 76005 Zlín, Czech Republic
pekar@fai.utb.cz

Abstract. Time delay systems own infinite spectra which cannot be simply analyzed or controlled. A way how to deal with this task consists of an approximation of the characteristic quasipolynomial by a polynomial that can be further handled via conventional tools. This contribution is aimed at an improved extrapolation method transforming a retarded quasipolynomial into a corresponding polynomial. It is equivalent to the finding of a finite-dimensional model related to an infinite-dimensional one describing a time delay system. The approximating polynomial is then used to analyze the dependence of delay values to exponential stability of the system. Two ideas are adopted and compared here; namely, a linear interpolation method via the Regula Falsi method, and the root Newton's method with root tendency. The whole procedure is simply implementable by using standard software tools. To demonstrate this issue, a numerical example performed in MATLAB® & Simulink® environment is given to the reader.

Keywords: Delay dependent stability · Quasipolynomial approximation · Time delay systems

1 Introduction

Time delay systems (TDSs) are representatives of infinite-dimensional systems, i.e. those having infinitely many solution modes or system characteristic values (poles) [1]. They inherently appear and are present throughout various human activities [2–4]; thus, the studying of their properties and the development of philosophies how to steer them have attracted scientists and engineers since the mid of the last century [3, 5–9]. However, due to the infinite-spectrum, these tasks are challenging in their complexity and mostly suffer from advanced and practically hardly implementable mathematics.

The characteristic quasipolynomial of a TDS gives the full information about the system poles loci and thus about its exponential stability (unless distributed delays are included in the dynamics) since its roots coincide with the poles. However, there is no purely analytic method for the calculation of the transcendental roots of general quasipolynomials with non-commensurate delays. Several methods and software packages were developed for direct numerical computation of quasipolynomial roots without the use of a quasipolynomial simplification or approximation, see e.g. [10, 11];

R. Silhavy et al. (eds.), *Cybernetics and Mathematics Applications in Intelligent Systems*,
Advances in Intelligent Systems and Computing 574, DOI 10.1007/978-3-319-57264-2_3

however, they usually require some apriori information about the system spectrum and/or the use of a special software. Another family of methods is aimed at the seeking of the so-called pseudospectrum of the system based interpolation or extrapolation method for the discretization of the state-space formulation [12]. Some ideas of discrete-time (digital) filters designing were used to compute a polynomial approximation of a quasipolynomial in [13]. Then, the approximating spectrum can simply be computed by means of standard mathematical and software tools.

The above introduced methods for the spectrum computation/estimation may be used to determine delay dependent (exponential) stability in terms of the verification of the existence of a pole in the right-half complex plane. Hence, the task of the delay dependent stability verification lies in the determination of delay intervals for which the system remains stable [14, 15]. In [16], we presented a preliminary study to a numerical gridding method for the determination of delays switching the system from/to stability/instability via an iterative polynomial approximation of the characteristic quasipolynomial by means of the extrapolation method followed by the linear interpolation, namely, Taylor's series expansion and the Regula Falsi (RF), respectively. The linear connection of the eventual delay values enables to obtain the stability margin with infinitely many switching delays. The procedure is simple, easily programmable, and applicable to even TDSs with multiple and non-commensurate delays. It enables to estimate positions of stability switching poles with a sufficient precision.

The goal of this contribution lies in an attempt to improve the algorithm such that a more accurate estimation of delay dependent stability windows is obtained, whilst preserving the algorithm simplicity and speed. The innovation is based on the use of the root tendency (RT) expressing the sensitivity of a root loci to a quasipolynomial parameter, instead of RF, followed by the Newtons' method for the zero point computation. One-step and two-step strategies are benchmarked.

The paper is organized as follows: The definition of the TDS and the corresponding (retarded) quasipolynomial along with the introduction of exponential stability are given in the preliminary Sect. 2. Afterward in Sect. 3, the reader is acquainted with an overview of the original gridding approximation-based algorithm to determine stabilizing-delays windows [16]. The algorithm improvements and extensions are suggested in Sect. 4. Prior to conclusions, in Sect. 5, the reader is provided with a numerical example that verifies the proposed ideas and compares them with the original proposition; further suggestions and the limitations are presented as well.

2 TDS, Retarded Quasipolynomial, Exponential Stability

2.1 TDS and its Characteristic Quasipolynomial

Consider a TDS described by the input-output ordinary differential equation with shifted arguments as

$$
\begin{aligned}
&y^{(n)}(t) + \sum_{j=0}^{h_{a,n-1}} a_{n-1,j} y^{(n-1)}\left(t - \vartheta_{a,n-1,j}\right) + \ldots + \sum_{j=0}^{h_{a,1}} a_{1,j} y'\left(t - \vartheta_{a,1,j}\right) \\
&\quad + \sum_{j=0}^{h_{a,0}} a_{0,j} y\left(t - \vartheta_{a,0,j}\right) \\
&= b_m u^{(m)}(t) + \sum_{j=0}^{h_{b,m-1}} b_{m-1,j} u^{(m-1)}\left(t - \vartheta_{b,m-1,j}\right) + \ldots + \sum_{j=0}^{h_{b,1}} b_{1,j} u'\left(t - \vartheta_{b,1,j}\right) \\
&\quad + \sum_{j=0}^{h_{b,0}} b_{0,j} u\left(t - \vartheta_{b,0,j}\right)
\end{aligned}
\tag{1}
$$

where $u(t), y(t)$ stand for system input and output, respectively, $a_{\cdot,\cdot}, b_{\cdot,\cdot}$ are real-valued coefficients, and $\vartheta_{a,\cdot,\cdot}, \vartheta_{b,\cdot,\cdot}$ express general delays where $\vartheta_{a,\cdot,0} = \vartheta_{b,\cdot,0} = 0$. Note that $n \geq m$. Let, moreover,

$$
\vartheta_{\cdot,i,j} = \sum_{k=1}^{L} \lambda_{ij,k} \tau_k \tag{2}
$$

where $\boldsymbol{\tau} = (\tau_1, \tau_2, \ldots, \tau_L)$ are independent delays. If $\vartheta_{\cdot,i,j} = \lambda_{ij}\tau_0, \lambda_{ij} \in \mathrm{N}_+$ for all $\vartheta_{\cdot,i,j}$ and some fixed base delay τ_0, then delays are called *commensurate*. Otherwise, they are non-commensurate.

The transfer function corresponding to (1) reads

$$
G(s) = \frac{b(s)}{a(s)} = \frac{b_m s^m + \sum_{i=0}^{m-1} \sum_{j=1}^{h_{b,i}} b_{ij} s^i \exp\left(-s \sum_{k=1}^{L} \lambda_{b,ij,k} \tau_k\right)}{s^m + \sum_{i=0}^{n-1} \sum_{j=1}^{h_{a,i}} a_{ij} s^i \exp\left(-s \sum_{k=1}^{L} \lambda_{a,ij,k} \tau_k\right)} \tag{3}
$$

where $a(s)$, $b(s)$ are retarded quasipolynomials.

In the further text, it is assumed that are no common roots of $a(s)$, $b(s)$. Under this assumption, $a(s)$ corresponds to the *characteristic quasipolynomial* of the TDS, the roots of which are system *poles* (characteristic values), i.e. the poles constitute the set $\Sigma := \{s : a(s) = 0\}$. In general, $|\Sigma| = \infty$ for a TDS.

2.2 Exponential Stability

The notion of exponential stability for TDSs (1) is analogous to that for finite-dimensional systems, i.e. it expresses that the convergence of the output and all its derivatives are bounded by exponential decays. The stability condition is the same as well, and it is given by the following expression

$$
\alpha < 0 \tag{4}
$$

where α stands for the spectral abscissa defined as $\alpha := \sup \mathrm{Re}\Sigma$. From (4), it is obvious that exponential stability becomes broken whenever $\alpha = 0$, i.e. the rightmost pair of poles cross the imaginary axis (note that this crossing cannot pass off purely in the real axis [1]).

3 Original Algorithm to Detect Stability Switching Delays

3.1 Problem Formulation and Motivation

As introduced above, the system is switched from/to stability/instability when the rightmost pair of poles - the so-called *switching poles*, $\{\bar{s}, \bar{s}'\} = \bar{\beta} \pm \mathrm{j}\bar{\omega}$, are located exactly on the imaginary axis (hereinafter, the complex conjugate $\bar{s}'$ is omitted). The value of α may depend on system parameters, namely on values of $\boldsymbol{\tau}$. Hence, it is possible to determine (estimated) the *switching delays*, $\bar{\tau}$, corresponding to the switching poles.

To study the dependence $\boldsymbol{\tau} \rightarrow \alpha$ or, equivalently, the searching of pairs $\{\bar{s}, \bar{\tau}\}$, consider the characteristic quasipolynomial as the function of $\boldsymbol{\tau}$, i.e. $a(s, \boldsymbol{\tau})$, and the spectral abscissa let be written as $\alpha(\boldsymbol{\tau})$.

3.2 Original Algorithm

The following algorithm based on the polynomial approximation (extrapolation) of the characteristic quasipolynomial intends to give the estimation of the set of pairs $\{\bar{s}, \bar{\tau}\}$ within sets $\bar{\mathrm{T}} := \{\bar{\tau}\}$, $\bar{\Sigma} := \{\bar{s}\}$ for TDSs. Note that symbol $\hat{}$ expresses the estimation, and $f(s|p)$ means a function f of variable s computed for parameter(s) p. The algorithm be formulated, in a concise form, as follows.

Step 1: For the given $a(s, \boldsymbol{\tau})$, define the mesh grid $\tau_{l,j+1} = \tau_{l,j} + \Delta\tau_{l,j}$, $\tau_{l,0} = 0$, $l \in [1, L]$, $j \in [0, N-1]$ for a selected delay range, initialize the counter $i = 0$ and choose $\varepsilon > 0$. Set estimations $\hat{\bar{\mathrm{T}}} = \hat{\bar{\Sigma}} = \emptyset$.

Step 2: Compute $\hat{s}_{0,\ldots,0} = s_{0,\ldots,0} = \max\mathrm{Re}\{s : a(s, \mathbf{0}) = 0\}$.

Step 3: For $(j_1 = 0 \ldots N-1)$, for $(j_2 = 0 \ldots N-1)$, etc. for $(j_L = 0 \ldots N-1)$ do Steps 4 to 9.

Step 4: If $j_l = 0, \forall l$, the inner loop is finished; else, define $M := \max\{l : j_l \neq 0\}$ and set $\boldsymbol{\tau} = \left(\tau_{1,j_1}, \tau_{2,j_2} \ldots, \tau_{L,j_L}\right)$, $\hat{s}_{old} = \hat{s}_0 = \hat{s}_{j_1,\ldots,j_{M-1},j_M-1,0\ldots0}$.

Step 5: Compute the polynomial estimation $\hat{a}(s|\boldsymbol{\tau}, \hat{s}_0)$ of $a(s, \boldsymbol{\tau})$ via the Taylor's series expansion in $\hat{s}_0$ and find its roots, s_k. Calculate $\hat{s}_1 = \arg\min|s_k - \hat{s}_0|$.

Step 6: While $|\hat{s}_1 - \hat{s}_0| \geq \varepsilon$, set $\hat{s}_0 = \hat{s}_1$ and go to Step 5.

Step 7: Set $\hat{s}_{new} = \hat{s}_{j_1,\ldots,j_L} = \hat{s}_{j_1,\ldots,j_{M-1},j_M,0\ldots0} := \hat{s}_1$. If $\mathrm{sgn}(\mathrm{Re}\hat{s}_{new}) = \mathrm{sgn}(\mathrm{Re}\hat{s}_{old})$ the inner loop is finished (see Step 3); else, $i = i + 1$.

Step 8: Calculate the switching delay estimation $\bar{\tau}_M = \bar{\tau}_M\left(\tau_{M,j_M-1}, \hat{s}_{old}, \hat{s}_{new}\right)$ by using the linear interpolation (RF) as

$$\bar{\tau}_M = \bar{\tau}_M\left(\tau_{M,j_M-1}, \hat{s}_{old}, \hat{s}_{new}\right) = \tau_{M,j_M-1} - \mathrm{Re}\hat{s}_{old}\frac{\tau_{M,j_M} - \tau_{M,j_M-1}}{\mathrm{Re}\hat{s}_{new} - \mathrm{Re}\hat{s}_{old}} \tag{5}$$

Step 9: For $l = M-1, \ldots, 1$ do: If $j_l = 0$, set $\bar{\tau}_l = \tau_{l,0}$; else set $\hat{s}_0 = \hat{s}_{new}$ and $\boldsymbol{\tau}_{old} = \left(\tau_{1,j_1}, \ldots, \tau_{l,j_l-1}, \bar{\tau}_{l+1}, \ldots, \bar{\tau}_M, 0, \ldots, 0\right)$, $\boldsymbol{\tau} = \left(\tau_{1,j_1}, \ldots, \tau_{l,j_l}, \bar{\tau}_{l+1}, \ldots, \bar{\tau}_M, 0, \ldots, 0\right)$ and compute the leading (rightmost) root $\hat{s}_1$ from $\hat{a}(s|\boldsymbol{\tau}_{old}, \hat{s}_0)$ as in Steps 5 and 6. Update values $\hat{s}_{old} = \hat{s}_0 := \hat{s}_1$ and find the leading root $\hat{s}_1$ of $\hat{a}(s|\boldsymbol{\tau}, \hat{s}_0)$ and update the value $\hat{s}_{new} = \hat{s}_1$. Then calculate $\bar{\tau}_l = \bar{\tau}_l\left(\tau_{l,j_l-1}, \hat{s}_{old}, \hat{s}_{new}\right)$ via the RF function defined in (5).

Step 10: Consolidate $\bar{\tau}_k = (\bar{\tau}_1, \ldots \bar{\tau}_M, 0, \ldots 0)$, update $\hat{s}_0 = \hat{s}_{new}$, $\hat{\bar{\mathrm{T}}} = \hat{\bar{\mathrm{T}}} \cup \bar{\tau}_i$. Compute iteratively the leading zero $\bar{s}_i = \hat{s}_1$ of $\hat{a}(s|\bar{\tau}_i, \hat{s}_0)$ and set $\hat{\bar{\Sigma}} = \hat{\bar{\Sigma}} \cup \bar{s}_i$.

3.3 Remarks on the Algorithm

Simulation experiments in the MATLAB® & Simulink® environment have proven that this algorithm based on the Taylor's series expansion [16] gives more precise quasipolynomial leading zero estimation compared to the discrete-time idea presented in [13]; however, the computational time is slightly higher due to symbolic computation.

Isolated roots of $a(s, \boldsymbol{\tau})$ behave continuously w.r.t. $\boldsymbol{\tau}$ [1, 19], yet a problem can emerge when seeking the leading pole estimation due to a discontinuity or a non-smooth behavior of $\alpha(\boldsymbol{\tau})$ [17]. Whereas the former case is rare and it can be omitted, the latter one can appear e.g. when there are two or more rightmost poles with the same real part. In such a cases, the estimation of $\hat{s}_1$ (see Steps 5 and 6 of the algorithm) may fail; thus, it is desirable to reset $\hat{s}_0$ (e.g. by means of the QPmR - Quasi-Polynomial mapping based Rootfinder [11] - which is, however, much time consumptive).

Last but not least, the extrapolation yields complex-valued coefficients of $\hat{a}(s|\cdot)$. This i.a. means that root loci are not symmetrical to the real axis, which implies the fact that the rightmost "pair" cannot be determined in the algorithm. Nevertheless, such a property does not make the approximating polynomial pole loci computation worse. Anyway, one may use e.g. a technique introduced in [18] to get real-valued coefficients, yet with a worse estimation.

Infinitely many switching delays can be obtained by the linear interpolation of entries of $\hat{\bar{\mathrm{T}}}$, as in the original version of the algorithm, or by the use of *fit* function of MATLAB®.

4 Suggested Enhancement

The algorithm presented in the preceding section is now attacked in Steps 8 and 9 in order to attempt to get a more accurate switching delays estimation. Specifically, the information about the RT values followed by the Newton's technique rather than RF (as in (5)) are used. Eventually, two different values are obtained; hence, the arithmetical mean of both is taken as the final value. Note that whereas the RF needs two points to be interpolated, the RT is an extrapolation technique, which implies that a single switching delay value can be used to improve the estimation.

The definition of the RT (that expresses the sign of the speed of the real-part position of a characteristic value w.r.t. the corresponding delay value) and the leading idea of its use for the solved problem follows. It is defined as the vector

$$\mathrm{RT} := \mathrm{sgnRe}\{\mathrm{grad}\, s(\boldsymbol{\tau})\} \tag{6}$$

the element of which can be calculated as

$$\mathrm{RT}_l(s,\boldsymbol{\tau}) \approx \mathrm{Re}\left(-\frac{\partial a(s,\boldsymbol{\tau})}{\partial \tau_l}\left(\frac{\mathrm{d}}{ds}a(s,\boldsymbol{\tau})\right)^{-1}\right), l \in [1,L] \tag{7}$$

in a case of poles with multiplicity one.

To introduce the consequent idea of the Newton's method for searching the zero point, consider function $r_{s_k}(\boldsymbol{\tau}) := \boldsymbol{\tau} \mapsto \mathrm{Re}s_k$ for any pole s_k. If s_k is located near the imaginary axis for the particular $\boldsymbol{\tau}$, the zero point $\boldsymbol{\tau}_0$ of $r_{s_k}(\boldsymbol{\tau})$ can be extrapolated as

$$\tau_{0,l} \approx \tau_l - \frac{r_{s_k}(\boldsymbol{\tau})}{\mathrm{RT}_l(s_k,\boldsymbol{\tau})}, l \in [1,L] \tag{8}$$

In Steps 8 and 9 of the above introduced algorithm, two zero point estimations (for $\hat{s}_{old}$, $\hat{s}_{new}$ respectively) are then calculated by means of (8), and the mean value, $\bar{\tau}_l = \tau_{0,l,mean}$, of both is eventually taken as the result. The whole idea yielding $\tau_{0,\mathrm{RT},mean}$ compared to the linear interpolation via the RF $\left(\tau_{0,\mathrm{RF}}\right)$ is depicted in Fig. 1.

As mentioned above, the RT along with the Newton's method can be repeated to obtain a more precise solution. If it is done once again, this two-step strategy is denoted as RT^+ hereinafter.

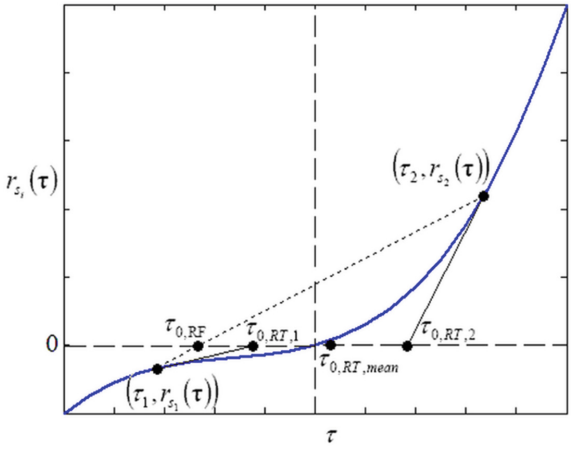


Fig. 1. A schematics comparing the idea of the zero point searching via the RF $\left(\tau_{0,\mathrm{RF}}\right)$ against that of via the value of RT $\left(\tau_{0,RT,mean}\right)$

5 Example

Assume a model of a skater on the remotely controlled swaying bow be considered, already published many times, see e.g. in [13, 16]. Following (1)-(3), it can be expresses by the following equation describing the relation between the horizontal angle deviation remotely driven by the skater and the output angle between the skater and the bow symmetry axis

$$y^{(4)}(t) + a_{2,1}y''\big(t - \vartheta_{a,2,1}\big) = b_0 u\big(t - \vartheta_{b,0,1}\big) \tag{9}$$

where $\vartheta_{a,2,1} = \tau_1 + \tau_2, \vartheta_{b,0,1} = \tau_2$ (τ_1 expresses the skater's reaction time and τ_2 means the servo latency). Denote $a = a_{2,1}, b = b_0$, for the simplicity, hereinafter. The corresponding transfer function then reads

$$G(s) = \frac{b\ \exp(-(\tau_1 + \tau_2)s)}{s^2(s^2 + a\ \exp(-\tau_2 s))} \tag{10}$$

Consider the habitual negative control feedback loop equipped with a finite-dimensional linear controller

$$C(s) = \frac{\sum_{i=0}^{3} q_i s^i}{s^3 + \sum_{i=0}^{2} p_i s^i} \tag{11}$$

where p_i, q_i are real-valued parameters.

Then the characteristic retarded quasipolynomial reads

$$\begin{aligned} a(s,(\tau_1,\tau_2)) &= \mathrm{num}(1 + C(s)G(s)) \\ &= s^2\big(s^2 + a\ \exp(-\tau_2 s)\big)\Big(s^3 + \sum_{i=0}^{2} p_i s^i\Big) + b\ \exp(-(\tau_1 + \tau_2)s)\Big(\sum_{i=0}^{3} q_i s^i\Big) \end{aligned} \tag{12}$$

where $\mathrm{num}(\cdot)$ stands for the numerator quasipolynomial of a meromorphic function. Controller parameters can by optimally tuned e.g. in order to reach the spectral abscissa minimization [19]. Let nominal controlled system parameters and delay values be $a = -1$, $b = 0.2$, $\tau_1 = 0.3$, $\tau_2 = 0.1$ for which the optimized parameters yield the nominal spectral abscissa as $\alpha((0.3, 0.1)) = -1.4454$ (i.e. stable control system) while the delay-free case gives $\alpha((0,0)) = 0.1323$ (i.e. unstable control system). This i.a. implies that there must exist some sets of nonzero delay vectors stabilizing the control feedback loop.

Compare now the use of the original algorithm introduced in Sect. 3.2 applying the RF against the utilization of the RT averaging described in Sect. 4. Let the particular delay region be selected as $R_1 := \tau_1 \times \tau_2 \in [0, 0.8] \times [0, 0.8]$ with $\Delta\tau_\cdot = 0.01$, i.e. $N = 80$, and $\varepsilon = 10^{-6}$. In Fig. 2, the results are given to the reader and compared with a switching delays estimation calculated by the QPmR of a rough delay resolution of $\Delta\tau_\cdot = 0.01$ and the selected precision of 10^{-9}, measured by absolute values of real

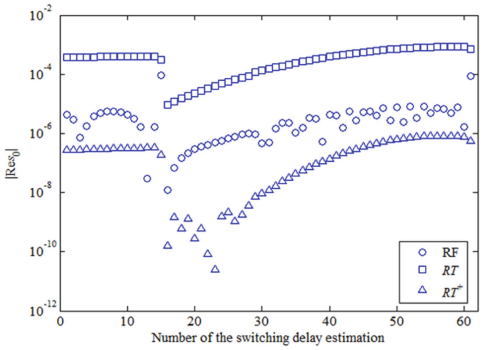


Fig. 2. Switching delay values error in R_1 measured by $|\mathrm{Re}s_0|$ against the result of the QPmR ($\Delta\tau. = 0.01$, $\varepsilon = 10^{-9}$)

parts of dominant poles (s_0). Note that it is not reasonable to use QPmR directly for the switching delays estimation since it requires a rather long lasting computation for a sufficiently high precision and, moreover, the searching region has to be a priori selected. Found switching delays estimations are then joined by the simple linear interpolation.

As can be seen, the simple use of the RT value with consequent averaging does not bring an improvement compared to the RF method; however, the two-step use on the Newton's method (RT^{+}) gives better switching delay estimation.

6 Conclusion

Main objectives of the presented paper have been in improvements and significant extensions of a recently developed gridding multiple stability switching delays seeking algorithm for retarded TDSs. The original procedure can be fitted in a group of frequency-domain direct methods that are based on the effort to find all characteristic roots (poles) located on the stability border, and it can deal with non-commensurate delays more effectively omitting a complex mathematical apparatus. The linear Regula Falsi interpolation has been compared to the use of the root tendency expressing the sensitivity of the leading pole's infinitesimal changes in delays. In addition, one-step and two-step iterative Newton's strategies has been used to enhance the switching delays estimation. Once a finite set of stability switching delays' values is determined, they can be joined e.g. by a linear interpolation procedure. It has been shown by simulations, when controlling a model of a skater on the swaying bow, that the two-step strategy gives better results compared to the one-step one and even to the use of the Regula Falsi. The future research related to this work may lie in an extension of the methodology to neutral TDSs with more complicated dynamics.

Acknowledgments. The work was performed with the financial support by the Ministry of Education, Youth and Sports of the Czech Republic within the National Sustainability Programme project No. LO1303 (MSMT-7778/2014).

References

1. Hale, J.K., Verduyn Lunel, S.M.: Introduction to Functional Differential Equations. Applied Mathematical Sciences, vol. 99. Springer, New York (1993)
2. Chiasson, J., Loiseau, J.J.: Applications of Time Delay Systems. Springer, New York (2007)
3. Sipahi, R., Vyhlídal, T., Niculescu, S.-I., Pepe, P.: Time Delay Systems: Methods, Applications and New Trends. LNCIS, vol. 423. Springer, New York (2012)
4. Han, Q.-L., Liu, Y., Yang, F.: Optimal communication network-based H∞ quantized control with packet dropouts for a class of discrete-time neural networks with distributed time delay. IEEE Trans. Neural Netw. Learn. Syst. **27**, 426–434 (2016)
5. Bellmann, R., Cooke, K.L.: Differential-Difference Equations. Academic Press, New York (1963)
6. Krasovskii, N.N.: Stability of Motion: Applications of Lyapunov's Second Method to Differential Systems and Equations with Delay. Standford University Press, Chicago (1963)
7. Osipov, Y.S.: Stabilization of controlled systems with delay. Differentsial'nye Uravneniya **1**, 605–618 (1965)
8. Richard, J.P.: Time-delay systems: an overview of some recent advances and open problems. Automatica **39**, 1667–1694 (2003)
9. Michiels, W., Niculescu, S.-I.: Stability, Control and Computation of Time-Delay Systems. SIAM Publications, Philadelphia (2014)
10. Engelborghs, K., Luzyanina, T., Samaye, G.: DDE-BIFTOOL v. 2.00: A Matlab Package for Bifurcation Analysis of Delay Differential Equations. Technical report TW-330, Department of Computer Science, K. U. Leuven, Leuven, Belgium (2001)
11. Vyhlídal, T., Zítek, P.: QPmR - quasi-polynomial root-finder: algorithm update and examples. In: Vyhlídal, T., Lafay, J.-F., Sipahi, R. (eds.) Delay Systems: From Theory to Numerics and Applications, pp. 299–312. Springer, New York (2014)
12. Breda, D., Maset, S., Vermiglio, R.: Pseudospectral methods for stability analysis of delayed dynamical systems. Int. J. Dyn. Control **2**, 143–153 (2014)
13. Pekař, L., Navrátil, P.: Polynomial approximation of quasipolynomials based on digital filter design principles. In: Silhavy, R., Senkerik, R., Oplatkova, Z., Silhavy, P., Prokopova, Z. (eds.) Automation Control Theory Perspectives in Intelligent Systems. AISC, vol. 466, pp. 25–34. Springer, Cham (2016)
14. Hertz, D., Jury, E.I., Zeheb, E.: Stability independent and dependent of delay for delay differential systems. J. Franklin Inst. **318**, 143–150 (1984)
15. Sönmez, S., Ayasun, S., Nwankpa, C.O.: An exact method for computing delay margin for stability of load frequency control systems with constant communication delays. IEEE Trans. Power Syst. **31**, 370–377 (2015)
16. Pekař, L., Prokop, R.: On delay (In)dependent stability for TDS. In: 2015 7th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT), Brno, Czech Republic, pp. 73–78. IEEE Press, Brno (2015)

17. Vanbiervliet, T., Verheyden, K., Michiels, W., Vandewalle, S.: A nonsmooth optimization approach for the stabilization of time-delay systems. ESIAM Control Optim. Ca. **14**, 478–493 (2008)
18. Pekař, L., Chalupa P.: A comparison of possible exponential polynomial approximations to get commensurate delays. In: MATEC Web of Conferences, vol. 76, p. 02012 (2016)
19. Pekař, L., Prokop, R.: Algebraic optimal control in RMS ring: a case study. Int. J. Math. Comput. Simul. **7**, 59–68 (2013)

Dissipativity of Multistep Runge–Kutta Methods for Nonlinear Neutral Delay Integro Differential Equations with Constrained Grid

Haiyan Yuan[1(✉)] and Cheng Song[2]

[1] Department of Mathematics, Heilongjiang Institute of Technology, Harbin 150050, China
yhy82_47@163.com
[2] Harbin Institute of Technology, College of Management, Harbin 150001, China

Abstract. This paper is concerned with the numerical dissipativity of multistep Runge-Kutta methods for nonlinear neutral delay-integro-differential equations. We investigate the dissipativity properties of (k,l)-algebraically stable multistep Runge-Kutta methods with constrained grid. The finite-dimensional and infinite-dimensional dissipativity results of (k,l)-algebraically stable multistep Runge-Kutta methods are obtained.

Keywords: Dissipativity · (k,l)-algebraically stability · Nonlinear neutral delay-integro-differential equation · Multistep Runge-Kutta methods

1 Introduction

Many dynamical systems in physics and engineering are characterized by the property of possessing a bounded absorbing set which all trajectories enter in a finite time and thereafter remain inside [1–4]. They are modeled by dissipative dynamical systems. In the study of dissipative systems it is often the asymptotic behavior of the system that is of interest, and so it is important to analyze whether or not numerical methods inherit the dissipativity of the dynamical systems when considering the applicability of numerical methods for these systems.

Humphries and Stuart [3, 4] first studied the dissipativity of Runge–Kutta methods for initial value problems (IVPs) of ordinary differential equations (ODEs) in 1994, and proved that an algebraically stable, irreducible method can inherit the dissipativity of finite-dimensional systems. Later, many results on the dissipativity of numerical methods for ODEs have already been found [5–7]. For the delay differential equations (DDEs) with constant delay, Huang [8] gave a sufficient condition for the dissipativity of theoretical solution, and investigated the dissipativity of – algebraically stable Runge–Kutta methods. Huang and Chen [9] and Huang [10], subsequently, obtained some results about the dissipativity of linear-methods and - algebraically stable one-leg methods, respectively. In addition, Huang [11] further discussed the dissipativity of multistep Runge-Kutta methods, and proved that an algebraically stable, irreducible multistep Runge-Kutta methods with linear interpolation procedure is

R. Silhavy et al. (eds.), *Cybernetics and Mathematics Applications in Intelligent Systems*, Advances in Intelligent Systems and Computing 574, DOI 10.1007/978-3-319-57264-2_4

finite-dimensional dissipative. In 2004, Tian [12] studied the dissipativity of DDEs with a bounded variable lag and the dissipativity of - method. Moreover, Wen [13] discussed the dissipativity of Volterra functional differential equations, and further investigated the dissipativity of DDEs with piecewise delays and a class of linear multistep methods. In recent years, a number of works on the dissipativity of numerical methods have been carried out. Gan [14–16] studied the dissipativity of numerical methods for nonlinear integro differential equations (IDEs), nonlinear delay-integro-differential equations (DIDEs) and nonlinear pantograph equations, respectively. As to nonlinear Volterra delay-integro-differential Equations, it was shown that for [1/2, 1], any linear-method and one-leg method can inherit the dissipativity property, which was obtained by Gan [15]. In addition, Cheng and Huang [19], Wen et al. [20] and Wang et al. [21] considered the dissipativity for nonlinear neutral delay differential equations (NDDEs). Wu and Gan [22] consider the dissipativity for a class of nonlinear neutral delay integro differential equations (NDIDEs). So far we have not seen in literature more dissipativity results for nonlinear NDIDEs.

This paper pursues this, and further investigates the dissipativity of multistep Runge-Kutta methods for nonlinear NDIDEs. The motivations are as follows. Multistep Runge-Kutta methods are a wider class of methods which has as special cases not only one-leg methods, linear multistep methods, and Runge-Kutta methods, but also a wide range of hybrid methods. In particular, there exist algebraically stable multistep Runge-Kutta methods with only real eigenvalues such that they not only possess very good stability, but also can be performed in parallel.

2 The Description of the Problem and Numerical Methods

Let H be a real or complex, finite dimensional or infinite-dimensional Hilbert space with the inner product $\langle \cdot, \cdot \rangle$ and the corresponding induced norm $\|\cdot\|$, and the matrix norm is subordinated to the vector norm. X be a dense continuously imbedded subspace of H. Consider the following initial value problems (IVPs) of nonlinear NDIDEs:

$$\begin{cases} \dfrac{d}{dt}[y(t) - Ny(t-\tau)] = f(y(t), y(t-\tau), \displaystyle\int_{t-\tau}^{t} g(t, \xi, y(\xi))d\xi), t \geq 0, \\ y(t) = \varphi(t), -\tau \leq t \leq 0. \end{cases} \tag{2.1}$$

where τ is a given constant delay, $N \in X \times X$ stands for a constant matrix with $\|N\| < 1$, $\varphi : [-\tau, 0] \to X$ is a continuous function, $f : X \times X \times X \to H$ is a locally Lipschitz continuous function, $g : [0, +\infty) \times [-\tau, +\infty) \times X \to X$ is a continuous function, f and g satisfy the following conditions:

$$\mathrm{Re}\langle u - Nv, f(u, v, w)\rangle \leq \beta_0 + \beta_1\|u\|^2 + \beta_2\|v\|^2 + \beta_3\|w\|^2 \ u, v, w \in X, \tag{2.2}$$

$$\|g(t, s, u)\| \leq \eta\|u\|, \ t \in [0, +\infty), \ s \in [-\tau, +\infty), \ u \in X \tag{2.3}$$

where β_0, β_1, β_2, β_3 and η are real constants.

Throughout this paper, we assume that the problem (2.1) has unique exact solution $y(t)$. For the study of solvability, we refer the reader to [2].

Remark 2.1. When $N = 0$, the problem (2.1) degenerates into an IVP of DIDEs. When the right-hand side function of the problem (2.1) does not possess the integral term, the problem (2.1) degenerates into an IVP of NDDEs. When $N = 0$ and the right-hand side function of the problem (2.1) does not possess the integral term, the problem (2.1) degenerates into an IVP of DDEs. In the above various cases, the number of papers dealing with different aspects of their numerical integration now amounts to several hundreds.

Proposition 2.2 [15]. Condition (2.2) implies that $\beta_0 \geq 0$, $\beta_2 \geq 0$ and $\beta_3 \geq 0$.

Next, let us consider the adaptation of s-stage multistep Runge-Kutta methods for solving problem (2.1) based on the formula

$$\begin{cases} Y_i^{(n)} - N\bar{Y}_i^{(n)} = \sum_{j=1}^{r} a_{ij}(y_{n+j-1} - N\bar{y}_{n+j-1}) + h\sum_{j=1}^{s} b_{ij} f(Y_j^{(n)}, \bar{Y}_j^{(n)}, \bar{G}_j^{(n)}), \\ y_{n+r} - N\bar{y}_{n+r} = \sum_{j=1}^{r} \theta_j(y_{n+j-1} - N\bar{y}_{n+j-1}) + h\sum_{j=1}^{s} \gamma_j f(Y_j^{(n)}, \bar{Y}_j^{(n)}, \bar{G}_j^{(n)}). \end{cases} \tag{2.4}$$

where $h > 0$ is the fixed stepsize, the parameters a_{ij}, b_{ij}, θ_j and γ_j are real constants, $Y_i^{(n)}$ and y_n are approximation to $y(t_n + c_i h)$ and $y(t_n)$, respectively, and $t_n = nh$. The argument $\bar{Y}_i^{(n)}$, $\bar{G}_i^{(n)}$ and $\bar{y}_n$ denotes an approximation to $y(t_n + c_i h - \tau)$, $\int_{t_n + c_i h - \tau}^{t_n + c_i h} g(t_n + c_i h, \xi, y(\xi))d\xi$ and $y(t_n - \tau)$, those are obtained by a specific interpolation procedure using values $Y_i^{(k)}$ and y_{k+r-1} $(k \leq n)$. The initial values $y_n = \varphi(t_n)$, $\bar{y}_n = \varphi(t_n - \tau)$ for $t_n \leq 0$, and $Y_i^{(n)} = \varphi(t_n + c_i h)$, $\bar{Y}_i^{(n)} = \varphi(t_n + c_i h - \tau)$ for $t_n + c_i h \leq 0$. Following the referee's suggestion, we assume that $0 \leq c_i \leq 1$, $i = 1, 2, \cdots, s$.

As to the computation of the delay terms $\bar{Y}_i^{(n)}$ and integral terms $\bar{G}_i^{(n)}$, $i = 1, 2, \cdots, s$, we use the constrained stepsize h satisfying $hm = \tau$ with a positive integer m.

Let

$$\bar{Y}_i^{(n)} = Y_i^{(n-m)}, i = 1, 2, \cdots, s, \tag{2.5a}$$

$$\bar{y}_n = y_{n-m}. \tag{2.5b}$$

and the compound quadrature (CQ) formula for the integral terms:

$$\bar{G}_i^{(n)} = h\sum_{q=0}^{m} v_q g(t_i^{(n)}, t_i^{(n-q)}, Y_i^{(n-q)}),\ i = 1, 2, \cdots, s. \tag{2.5c}$$

where $t_i^{(n)} = t_n + c_i h$, $i = 1, 2, \cdots, s$.

The quadrature formula (2.5c) can be derived from a uniform repeated rule [19]. For our stability analysis we need the rule to satisfy the following condition:

$$h\sqrt{(m+1)\sum_{q=0}^{m}\left|v_q\right|^2}<v, \tag{2.6}$$

with $hm=\tau$ and a positive constant v.

Remark 2.3. We consider the procedure (2.5a, 2.5b and 2.5c) here because, in the case that the order of the method is more than 2, there will be no order reduction if the corresponding quadrature rule is used. But it must be noticed that the stepsize h is limited by $mh=\tau$.

The used values $Y_i^{(n)}$ and y_n with $n<-m<0$ are assumed to be 0. Here, we do not discuss other details.

It is well known that multistep Runge-Kutta methods are a subclass of a general linear methods. Let

$$C_{11}=[b_{ij}]\in R^{s\times s},\quad C_{12}=[a_{ij}]\in R^{s\times r}, \tag{2.7a}$$

$$C_{21}=\begin{bmatrix}0 & \cdots & 0\\ \vdots & & \vdots\\ 0 & \cdots & 0\\ \gamma_1 & \cdots & \gamma_s\end{bmatrix}\in R^{r\times s},\ C_{22}=\begin{bmatrix}0 & 1 & 0 & \cdots & 0\\ 0 & 0 & 1 & \cdots & 0\\ \vdots & \vdots & \ddots & \ddots & \vdots\\ 0 & 0 & \cdots & 0 & 1\\ \theta_1 & \theta_2 & \cdots & \theta_{r-1} & \theta_r\end{bmatrix}\in R^{r\times r}, \tag{2.7b}$$

For any given $k\times l$ real matrix $Q=[q_{ij}]$, we define the corresponding linear operator $Q: X^l\to X^k$,

$$QU=V=(v_1,v_2,\cdots,v_k)\in X^k,\ U=(u_1,u_2,\cdots,u_l)\in X^l,\ u_j\in X,$$

with $v_i=\sum_{j=1}^{l}q_{ij}u_j,\ i=1,2,\cdots,k.$

Then, method (2.3) can be rewritten in the form of general linear method

$$\begin{cases}G^{(n)}=hC_{11}F(Y^{(n)},\bar{Y}^{(n)},\bar{G}^{(n)})+C_{12}g^{(n-1)},\\ g^{(n)}=hC_{21}F(Y^{(n)},\bar{Y}^{(n)},\bar{G}^{(n)})+C_{22}g^{(n-1)}.\end{cases} \tag{2.8}$$

with the following notational conventions:

$$
\begin{aligned}
&Y^{(n)} = (Y_1^{(n)}, Y_2^{(n)}, \cdots, Y_s^{(n)})^{\mathrm{T}},\ \bar{Y}^{(n)} = (\bar{Y}_1^{(n)}, \bar{Y}_2^{(n)}, \cdots, \bar{Y}_s^{(n)})^{\mathrm{T}},\\
&\bar{G}^{(n)} = (\bar{G}_1^{(n)}, \bar{G}_2^{(n)}, \cdots, \bar{G}_s^{(n)})^{\mathrm{T}}\\
&G^{(n)} = (Y_1^{(n)} - N\bar{Y}_1^{(n)}, Y_2^{(n)} - N\bar{Y}_2^{(n)}, \cdots, Y_s^{(n)} - N\bar{Y}_s^{(n)})^{\mathrm{T}},\\
&g^{(n)} = (y_{n+1} - N\bar{y}_{n+1}, y_{n+2} - N\bar{y}_{n+2}, \cdots, y_{n+r} - N\bar{y}_{n+r})^{\mathrm{T}},\\
&F(Y^{(n)}, \bar{Y}^{(n)}, \bar{G}^{(n)}) = (f(Y_1^{(n)}, \bar{Y}_1^{(n)}, \bar{G}_1^{(n)}),\ f(Y_2^{(n)}, \bar{Y}_2^{(n)}, \bar{G}_2^{(n)}),\\
&\qquad\qquad \cdots,\ f(Y_s^{(n)}, \bar{Y}_s^{(n)}, \bar{G}_s^{(n)}))^{\mathrm{T}}.
\end{aligned}
$$

We introduce the following notations for brevity, for any real symmetric $p \times p$ matrix $Q = [q_{ij}]$, and $Q \geq 0 (\ > 0)$ means that Q is nonnegative definite (positive definite). For any $Q \geq 0$, define a pseudo inner product on H^p by

$$
\langle Y, Z\rangle_Q = \sum_{i,j=1}^{p} q_{ij}\langle Y_i, Z_j\rangle,\ Y = (Y_1, Y_2, \cdots, Y_p) \in H^p,\ Z = (Z_1, Z_2, \cdots, Z_p) \in H^p,
$$

and the corresponding pseudo norm on H^p by $\|Y\|_Q = \langle Y, Y\rangle_Q^{1/2}$.

Especially $\|\cdot\|$ is the simplicity for $\|\cdot\|_Q$ when Q is identity matrix.

Definition 2.4. Let k, l be real constants. A multistep Runge-Kutta method (2.4) is said to be (k, l)-algebraically stable if there exists a real symmetric $r \times r$ matrix $G > 0$ and a diagonal matrix $D = diag(d_1, d_2, \cdots, d_s) \geq 0$ such that $M = [M_{ij}] \geq 0$, where

$$
M = \begin{bmatrix} kG - C_{22}^{\mathrm{T}} G C_{22} - 2lC_{12}^{\mathrm{T}} D C_{12} & C_{12}^{\mathrm{T}} D - C_{22}^{\mathrm{T}} G C_{21} - 2lC_{12}^{\mathrm{T}} D C_{11} \\ DC_{12} - C_{21}^{\mathrm{T}} G C_{22} - 2lC_{11}^{\mathrm{T}} D C_{12} & C_{11}^{\mathrm{T}} D + DC_{11} - C_{21}^{\mathrm{T}} G C_{21} - 2lC_{11}^{\mathrm{T}} D C_{11} \end{bmatrix}. \tag{2.9}
$$

As an important special case, a $(1, 0)$-algebraically stable method is called algebraically stable for short.

Definition 2.5. Let l be a real constant, and H be a finite-dimensional (or infinite-dimensional) space. A multistep Runge-Kutta method (2.4) with an interpolation procedure and integral terms are said to be finite-dimensionally (or infinite-dimensionally) $D(l)$ dissipative if, when the method is applied to problem (2.1) in H with stepsize h satisfying

$$
d(\beta_1 h + \beta_2 h + h\beta_3 \eta^2 v^2) < l d_{\min}(1 - \|N\|)^2,\ (\text{or}\, d(\beta_1 h + \beta_2 h + h\beta_3 \eta^2 v^2) < l d(1 - \|N\|)^2)
$$

and constraint $\tau = mh$, where $d_{\min} = \min\limits_{1 \le j \le s} d_j$ and $d = \sum\limits_{j=1}^{s} d_j$, there exists a constant C such that, for any initial values, there exists an n_0, dependent only on initial values, such that

$$\|y_n\| \le C,\ n \ge n_0,$$

holds. As an important special case, a $D(0)$-dissipative method is called D-dissipative for short.

$GD(l)$-and GD-dissipativity are defined by dropping restriction $\tau = mh$.

Definition 2.6 [11]. A multistep Runge-Kutta method (2.4) is said to be stage-reducible if, for some nonempty index set $T \subset \{1, 2, \cdots, s\}$,

$$\begin{aligned} \gamma_j &= 0, \quad for \quad j \in T, \\ b_{ij} &= 0, \quad for \quad j \in T, i \notin T. \end{aligned}$$

Otherwise, it is said to be stage-irreducible.

Definition 2.7. A multistep Runge-Kutta method (2.4) is said to be step-reducible if polynomials $\{\sigma_i(x)\}_{i=0}^{s}$ have common divisor where

$$\begin{aligned} \sigma_0(x) &= x^r - N\bar{x}^r - \sum_{j=1}^{r} \theta_j (x^{j-1} - N\bar{x}^{j-1}), \\ \sigma_i(x) &= \sum_{j=1}^{r} a_{ij} (x^{j-1} - N\bar{x}^{j-1}), i = 1, 2, \cdots, s. \end{aligned}$$

Otherwise, it is said to be step-irreducible.

Definition 2.8 [11]. A multistep Runge-Kutta method (2.4) is said to be reducible if it is stage-reducible or step-reducible.

3 Finite-Dimensional Numerical Dissipativity

In this section, we focus on the dissipativity analysis of (k, l)-algebraically stable multistep Runge-Kutta methods with respect to nonlinear NDIDEs in finite-dimensional spaces. We always assume that $H = X = C^N$.

Lemma 3.1 [11]. Suppose $\{\xi_i(x)\}_{i=1}^{r}$ are a basis of polynomials for P^{r-1}, the space of polynomials of degree strictly less than r and E is the translation operator: $Ey_n = y_{n+1}$. Then there is always a unique solution y_n, y_{n+1}, …, y_{n+r-1} to the system of equations

$$\begin{aligned} &\xi_i(E) y_n = \Delta_i, \\ &\Delta_i \in C^N,\ i = 1, 2, \cdots, r. \end{aligned}$$

and there exists a constant χ, independent of Δ_i, such that

$$\max_{0\le i\le r-1}\|y_{n+i}\|\le\chi\max_{0\le i\le r-1}\|\Delta_i\|.$$

Lemma 3.2 [11]. Suppose that a multistep Runge-Kutta method (2.4) is step-irreducible. Then, there exist real constants $v_i, i=1,2,\cdots,s.$ such that $\sigma_0(x)$ and $\sum_{i=1}^{s} v_i\sigma_i(x)$ have no common divisor.

Now we state and prove the main results.

Theorem 3.3. Assume that a step-irreducible multistep Runge-Kutta method (2.4) is (k,l)-algebraically stable, $D>0$, $l>0$ and $k\le 1$, the problem (2.1) satisfies (2.2) and (2.3) with $d(\beta_1 h+\beta_2 h+h\beta_3\eta^2 v^2)<ld_{\min}(1-\|N\|)^2$. Then the method (2.4) with (2.5a, 2.5b and 2.5c) is finite-dimensionally $D(l)$-dissipative.

Proof. From (2.5a, 2.5b and 2.5c), using Cauchy-Schwarz inequality we can obtain

$$\left\|\bar{Y}^{(j)}\right\|^2=\left\|Y^{(j-m)}\right\|^2 \tag{3.1a}$$

$$\|\bar{y}_n\|=\|y_{n-m}\| \tag{3.1b}$$

$$\begin{aligned}\left\|\bar{G}_i^{(n)}\right\|^2&=\left\|h\sum_{q=0}^{m}v_q g(t_i^{(n)},t_i^{(n-q)},Y_i^{(n-q)})\right\|^2\le h^2\eta^2\sum_{q=0}^{m}|v_q|^2\sum_{q=0}^{m}\left\|Y_i^{(n-q)}\right\|^2\\&\le\frac{\eta^2v^2}{m+1}\sum_{q=0}^{m}\left\|Y_i^{(n-q)}\right\|^2\end{aligned}$$

Therefore,

$$\left\|\bar{G}^{(n)}\right\|^2\le\frac{\eta^2v^2}{m+1}\sum_{q=0}^{m}\left\|Y^{(n-q)}\right\|^2 \tag{3.1c}$$

As in [17] and [15], by means of (k,l)-algebraically stability of the method, we can easily obtain that

$$\begin{aligned}&\|g^{(n)}\|_G^2-k\|g^{(n-1)}\|_G^2-2\mathrm{Re}\left\langle G^{(n)},hF(Y^{(n)},\bar{Y}^{(n)},\bar{G}^{(n)})\right\rangle_D+2l\|G^{(n)}\|_D^2\\&=\left\langle C_{21}hF(Y^{(n)},\bar{Y}^{(n)},\bar{G}^{(n)})+C_{22}g^{(n-1)},G(C_{21}hF(Y^{(n)},\bar{Y}^{(n)},\bar{G}^{(n)})+C_{22}g^{(n-1)})\right\rangle\\&+\left\langle g^{(n-1)},-kGg^{(n-1)})\right\rangle\\&+2\mathrm{Re}\left\langle C_{11}hF(Y^{(n)},\bar{Y}^{(n)},\bar{G}^{(n)})+C_{12}g^{(n-1)},-DhF(Y^{(n)},\bar{Y}^{(n)},\bar{G}^{(n)})\right\rangle\\&+\left\langle C_{11}hF(Y^{(n)},\bar{Y}^{(n)},\bar{G}^{(n)})+C_{12}g^{(n-1)},2lD(C_{11}hF(Y^{(n)},\bar{Y}^{(n)},\bar{G}^{(n)})+C_{12}g^{(n-1)})\right\rangle\\&=-\left\langle\left\langle g^{(n-1)},hF(Y^{(n)},\bar{Y}^{(n)},\bar{G}^{(n)})\right\rangle,M\left\langle g^{(n-1)},hF(Y^{(n)},\bar{Y}^{(n)},\bar{G}^{(n)})\right\rangle\right\rangle\le 0.\end{aligned} \tag{3.2}$$

Considering (2.2), (2.3) and $k \leq 1$, we have

$$\begin{aligned}\left\|g^{(n)}\right\|_G^2 &\leq \left\|g^{(n-1)}\right\|_G^2 + 2hd\beta_0 + 2\beta_1 h\left\|Y^{(n)}\right\|_D^2 + 2h\beta_2\left\|\bar{Y}^{(n)}\right\|_D^2 + 2h\beta_3\left\|\bar{G}^{(n)}\right\|_D^2 - 2l\left\|G^{(n)}\right\|_D^2, \\ &\leq \left\|g^{(n-1)}\right\|_G^2 + 2hd\beta_0 + 2\beta_1 dh\left\|Y^{(n)}\right\|^2 + 2hd\beta_2\left\|\bar{Y}^{(n)}\right\|^2 + 2hd\beta_3\left\|\bar{G}^{(n)}\right\|^2 \\ &\quad - 2ld_{\min}(\left\|Y^{(n)}\right\|^2 + \|N\|^2\left\|\bar{Y}^{(n)}\right\|^2 - 2\|N\|\left\langle Y^{(n)}, \bar{Y}^{(n)}\right\rangle)\end{aligned}$$

Using (3.1a, 3.1b and 3.1c) and Cauchy-Schwarz inequality, we have

$$\begin{aligned}\left\|g^{(n)}\right\|_G^2 &\leq \left\|g^{(n-1)}\right\|_G^2 + 2hd\beta_0 + 2\beta_1 dh\left\|Y^{(n)}\right\|^2 + 2hd\beta_2\left\|\bar{Y}^{(n)}\right\|^2 + 2hd\beta_3\frac{\eta^2\nu^2}{m+1}\sum_{q=0}^{m}\left\|Y^{(n-q)}\right\|^2 \\ &\quad - 2ld_{\min}(\left\|Y^{(n)}\right\|^2 + \|N\|^2\left\|\bar{Y}^{(n)}\right\|^2) + 2ld_{\min}\|N\|(\left\|Y^{(n)}\right\|^2 + \left\|\bar{Y}^{(n)}\right\|^2) \\ &\leq [2\beta_1 dh - 2ld_{\min}(1 - \|N\|)]\left\|Y^{(n)}\right\|^2 + [2hd\beta_2 + 2ld_{\min}\|N\|(1 - \|N\|)]\left\|\bar{Y}^{(n)}\right\|^2 \\ &\quad + \left\|g^{(n-1)}\right\|_G^2 + 2hd\beta_0 + 2hd\beta_3\frac{\eta^2\nu^2}{m+1}\sum_{q=0}^{m}\left\|Y^{(n-q)}\right\|^2\end{aligned} \tag{3.3}$$

Where

$$d_{\min} = \min_{1 \leq j \leq s} d_j,\ d = \sum_{j=1}^{s} d_j. \tag{3.4}$$

By induction, we can easily obtain

$$\begin{aligned}\left\|g^{(n)}\right\|_G^2 &\leq \left\|g^{(-1)}\right\|_G^2 + 2(n+1)hd\beta_0 + [2\beta_1 dh - 2ld_{\min}(1 - \|N\|)]\sum_{j=0}^{n}\left\|Y^{(j)}\right\|^2 \\ &\quad + [2hd\beta_2 + 2ld_{\min}\|N\|(1 - \|N\|)]\sum_{j=0}^{n}\left\|\bar{Y}^{(j)}\right\|^2 \\ &\quad + 2hd\beta_3\frac{\eta^2\nu^2}{m+1}\sum_{j=0}^{n}\sum_{q=0}^{m}\left\|Y^{(j-q)}\right\|^2.\end{aligned} \tag{3.5}$$

When using (2.5a, 2.5b and 2.5c) and (3.1a, 3.1b and 3.1c) on substitution into (3.5) gives

$$
\begin{aligned}
\left\|g^{(n)}\right\|_G^2 &\le \left\|g^{(-1)}\right\|_G^2 + 2(n+1)hd\beta_0 \\
&+ 2hd\beta_3\frac{\eta^2v^2}{m+1}\left[(m+1)\sum_{j=0}^{m}\left\|Y^{(j)}\right\|^2 + \frac{m(m+1)}{2}\max_{-m\le i\le -1}\left\|Y^{(i)}\right\|^2\right] \\
&+ [2d\beta_1 h - 2ld_{\min}(1-\|N\|)]\sum_{j=0}^{n}\left\|Y^{(j)}\right\|^2 \\
&+ [2hd\beta_2 + 2ld_{\min}\|N\|(1-\|N\|)]\sum_{j=0}^{n}\left\|Y^{(j-m)}\right\|^2 \qquad (3.6)\\
&\le \left\|g^{(-1)}\right\|_G^2 + 2(n+1)hd\beta_0 \\
&+ 2\left[d\beta_1 h + d\beta_2 h - ld_{\min}(1-\|N\|)^2 + hd\beta_3\eta^2v^2\right]\sum_{j=0}^{n}\left\|Y^{(j)}\right\|^2 \\
&+ \left[2d\beta_2\tau + 2ld_{\min}\|N\|(1-\|N\|)m + d\beta_3\eta^2v^2\tau\right]\max_{-m\le i\le -1}\left\|Y^{(i)}\right\|^2
\end{aligned}
$$

Let λ_1 denote the maximum eigenvalue of the matrix G,

$$
\begin{aligned}
a_1 &= \max_{0\le i\le r-1}\|y_i\|^2,\ a_2 = \max_{-m\le i\le -1}\left\|Y^{(i)}\right\|^2,\ R_1 = \max(a_1, a_2),\\
\mu &= ld_{\min}(1-\|N\|)^2 - d(\beta_1 h + \beta_2 h + h\beta_3\eta^2v^2).
\end{aligned}
$$

Then, we have $\mu > 0$ and

$$
\begin{aligned}
\left\|g^{(n)}\right\|_G^2 + 2\mu\sum_{j=0}^{n}\left\|Y^{(j)}\right\|_D^2 &\le \left[2d\beta_2\tau + 2ld_{\min}\|N\|(1-\|N\|)m + d\beta_3\eta^2v^2\tau\right]\max_{-m\le i\le -1}\left\|Y^{(i)}\right\|^2 \\
&+ \left\|g^{(-1)}\right\|_G^2 + 2(n+1)hd\beta_0 \\
&\le r\lambda_1(1+\|N\|)a_1 + 2(n+1)hd\beta_0 \\
&+ \left[2d\beta_2\tau + 2ld_{\min}\|N\|(1-\|N\|)m + d\beta_3\eta^2v^2\tau\right]a_2 \\
&\le \left[r\lambda_1(1+\|N\|) + 2d\beta_2\tau + 2ld_{\min}\|N\|(1-\|N\|)m + d\beta_3\eta^2v^2\tau\right]R_1 + 2(n+1)hd\beta_0. \qquad (3.7)
\end{aligned}
$$

When $\beta_0 = 0$, it follows from (3.7) and $\mu > 0$ that

$$
\lim_{n\to\infty}\left\|Y^{(n)}\right\| = 0,
$$

which shows that for any $\varepsilon > 0$, there exists $n_0(R_1, \varepsilon) > 0$, such that

$$
\left\|Y_j^{(n)}\right\| \le \varepsilon,\ \left\|\bar{Y}_j^{(n)}\right\| \le \varepsilon\ j = 1, 2, \cdots, s,\ n \ge n_0. \qquad (3.8)
$$

Hence, (2.4) implies that

$$\|\sigma_i(E)(y_n - N\bar{y}_n)\| = \left\| \sum_{j=1}^{r} a_{ij}(y_{n+j-1} - N\bar{y}_{n+j-1}) \right\| \leq hL \sum_{j=1}^{s} |b_{ij}| + \varepsilon, \quad i = 1, 2, \cdots, s,\ n \geq n_0 \tag{3.9a}$$

$$\|\sigma_0(E)(y_n - N\bar{y}_n)\| = \left\| y_{n+r} - N\bar{y}_{n+r} - \sum_{j=1}^{r} \theta_j(y_{n+j-1} - N\bar{y}_{n+j-1}) \right\| \leq hL \sum_{j=1}^{s} |\gamma_j|, \quad n \geq n_0 \tag{3.9b}$$

where

$$L = \sup_{\|u\| \leq \varepsilon, \|v\| \leq \varepsilon, \|w\| \leq \varepsilon} \|f(u, v, w)\|,\ u, v, w \in X.$$

From Lemma 3.2 it follows that there exist real constants v_i, $i = 1, 2, \cdots, s$, such that $\sigma_0(x)$ and $\sum_{i=1}^{s} v_i \sigma_i(x)$ have no common divisor. Therefore,

$$\left\| \sum_{i=1}^{s} v_i \sigma_i(E)(y_n - N\bar{y}_n) \right\| \leq \sum_{i=1}^{s} |v_i| \left[hL \sum_{j=1}^{s} |b_{ij}| + \varepsilon \right],\ n \geq n_0,$$

which further gives

$$\begin{aligned} &\left\| \left[\sigma_0(E) - \sum_{i=1}^{s} v_i \sigma_i(E) \right] (y_n - N\bar{y}_n) \right\| \\ &\quad \leq hL \sum_{j=1}^{s} |\gamma_j| + \sum_{i=1}^{s} |v_i| \left[hL \sum_{j=1}^{s} |b_{ij}| + \varepsilon \right], \quad n \geq n_0. \end{aligned} \tag{3.10}$$

Since $\sigma_0(x)$ and $\sigma_0(x) - \sum_{i=1}^{s} v_i \sigma_i(x)$ are coprime, and both are of degree r. Hence,

$$\left\{ x^i \sigma_0(x), x^i [\sigma_0(x) - \sum_{i=1}^{s} v_i \sigma_i(x)] : i = 0, 1, \cdots, r-1. \right\},$$

form a basis for P^{2r-1}. Considering (3.9a and 3.9b), (3.10) and Lemma 3.1, we have

$$\|y_n - N\bar{y}_n\| \leq \chi \left[hL \sum_{j=1}^{s} |\gamma_j| + \sum_{i=1}^{s} |v_i| \left(hL \sum_{j=1}^{s} |b_{ij}| + \varepsilon \right) \right],\ \text{for } n \geq n_0 \text{ and } \beta_0 = 0.$$

Therefore,

$$\begin{aligned}\|y_n\| &\le \|N\|\|\bar{y}_n\| + \chi\left[hL\sum_{j=1}^{s}|\gamma_j| + \sum_{i=1}^{s}|v_i|\left(hL\sum_{j=1}^{s}|b_{ij}| + \varepsilon\right)\right] \\ &\le \|N\|\|y_{n-m}\| + \chi\left[hL\sum_{j=1}^{s}|\gamma_j| + \sum_{i=1}^{s}|v_i|\left(hL\sum_{j=1}^{s}|b_{ij}| + \varepsilon\right)\right] \\ &\le \frac{\chi\left[hL\sum_{j=1}^{s}|\gamma_j| + \sum_{i=1}^{s}|v_i|\left(hL\sum_{j=1}^{s}|b_{ij}| + \varepsilon\right)\right]}{1-\|N\|} + \max_{-\tau\le\xi\le 0}\|\varphi(\xi)\|\end{aligned} \tag{3.11}$$

where $n \ge n_0'$, $n_0' = m + n_0$.

When $\beta_0 > 0$, let us take $n = 2(m+r)q - 1$,

$$q = \left\lfloor \frac{[r\lambda_1(1+\|N\|) + 2d\beta_2\tau + 2ld_{\min}\|N\|(1-\|N\|)m + d\beta_3\eta^2v^2\tau]R_1}{4(m+r)hd\beta_0} \right\rfloor + 1$$

where the notation $\lfloor x \rfloor$ means the maximum integer no greater than x, then

$$\left[r\lambda_1(1+\|N\|) + 2d\beta_2\tau + 2ld_{\min}\|N\|(1-\|N\|)m + d\beta_3\eta^2v^2\tau\right]R_1 \le 4(m+r)qhd\beta_0.$$

It follows from (3.7) that

$$\mu\sum_{j=0}^{2(m+r)q-1}\left\|Y^{(i)}\right\|^2 \le 4(m+r)qhd\beta_0,$$

which gives

$$\sum_{i=0}^{2q-1}\sum_{j=(m+r)i}^{(m+r)(i+1)-1}\left\|Y^{(i)}\right\|^2 \le 4(m+r)\frac{qhd\beta_0}{\mu}.$$

Hence, there exists an integer $c \in [q, 2q-1]$ such that

$$\sum_{j=(m+r)c}^{(m+r)(c+1)-1}\left\|Y^{(i)}\right\|^2 \le 4(m+r)\frac{hd\beta_0}{\mu}. \tag{3.12}$$

Let $p = (m+r)c + m$, then for all $j \in [p-m, p+r-1]$,we have

$$\left\|Y^{(i)}\right\|^2 \le a_2', \tag{3.13}$$

where

$$a_2' = 4(m+r)\frac{hd\beta_0}{\mu}.$$

Therefore, by (2.4) and (2.5a, 2.5b and 2.5c), for all $n \in [p, p+r-1]$,

$$\begin{aligned}\|\sigma_i(E)(y_n - N\bar{y}_n)\| &= \left\|\sum_{j=1}^{r} a_{ij}(y_{n+j-1} - N\bar{y}_{n+j-1})\right\| \\ &\le hL_1 \sum_{j=1}^{s} |b_{ij}| + \sqrt{\frac{a_2'}{d_i}}, i = 1, 2, \cdots, s.\end{aligned} \tag{3.14a}$$

$$\begin{aligned}\|\sigma_0(E)(y_n - N\bar{y}_n)\| &= \left\|y_{n+r} - N\bar{y}_{n+r} - \sum_{j=1}^{r} \theta_j(y_{n+j-1} - N\bar{y}_{n+j-1})\right\| \\ &\le hL_1 \sum_{j=1}^{s} |\gamma_j|,\end{aligned} \tag{3.14b}$$

where

$$L_1 = \sup_{\|y\| \le w, \|z\| \le w, \|\omega\| \le w} \|f(y, z, \omega)\|,\ y, z, \omega \in X,$$

where $w = \sqrt{a_2'/d_{\min}}$.

Therefore,

$$\begin{aligned}&\left\|\left[\sigma_0(E) - \sum_{i=1}^{s} v_i\sigma_i(E)\right](y_n - N\bar{y}_n)\right\| \\ &\le hL_1 \sum_{j=1}^{s} |\gamma_j| + \sum_{i=1}^{s} |v_i| \left[hL_1 \sum_{j=1}^{s} |b_{ij}| + \sqrt{\frac{a_2'}{d_i}}\right],\end{aligned} \tag{3.15}$$

with $n \in [p, p+r-1]$.

Considering (3.14a and 3.14b), (3.15), Lemmas 3.1 and 3.2, similar to (3.11), we have

$$\|y_n - N\bar{y}_n\|^2 \le a_1',\ n \in [p, p+r-1]. \tag{3.16}$$

where

$$a_1' = \chi^2 \left[hL_1 \sum_{j=1}^{s} |\gamma_j| + \sum_{i=1}^{s} |v_i| \left(hL_1 \sum_{j=1}^{s} |b_{ij}| + \sqrt{\frac{a_2'}{d_i}}\right)\right]^2.$$

Let

$$R_2 = \max(a_1', a_2').$$

A repetition of the above analysis implies that there exists a p',

$$p' \in [p + (m+r)q' + m, p + (2q'-1)(m+r) + m]$$
$$q' = \left\lfloor \frac{[r\lambda_1(1+\|N\|) + 2d\beta_2\tau + 2ld_{\min}\|N\|(1-\|N\|)m + d\beta_3\eta^2 v^2\tau]R_2}{4(m+r)hd\beta_0} \right\rfloor + 1,$$

such that

$$\left\|Y^{(n)}\right\|^2 \le a_2',\ n \in [p'-m, p'+r-1], \tag{3.17}$$

$$\|y_n - N\bar{y}_n\|^2 \le a_1',\ n \in [p', p'+r-1]. \tag{3.18}$$

Similar to (3.3), (3.5) and (3.7), for $n \in [p, p']$, we can obtain

$$\begin{aligned}\left\|g^{(n-1)}\right\|_G^2 &\le \left\|g^{(p-1)}\right\|_G^2 + 2(n-p)hd\beta_0 \\ &\quad + \left[2d\beta_2\tau + 2ld_{\min}\|N\|(1-\|N\|)m + d\beta_3\eta^2v^2\tau\right] \max_{p-m\le i\le p-1}\left\|Y^{(i)}\right\|^2 \\ &\le 2\left(\left[2d\beta_2\tau + 2ld_{\min}\|N\|(1-\|N\|)m + d\beta_3\eta^2v^2\tau\right] + r\lambda_1(1+\|N\|)\right)R_2 \\ &\quad + 2(2m+r)hd\beta_0\end{aligned} \tag{3.19}$$

Similar to (3.11), we can obtain

$$\begin{aligned}\|y_n\| &\le \|N\|\|\bar{y}_n\| \\ &\quad + \sqrt{2([2d\beta_2\tau + 2ld_{\min}\|N\|(1-\|N\|)m + d\beta_3\eta^2v^2\tau] + r\lambda_1(1+\|N\|))R_2 + 2(2m+r)hd\beta_0} \\ &\le \|N\|\|y_{n-m}\| \\ &\quad + \sqrt{2([2d\beta_2\tau + 2ld_{\min}\|N\|(1-\|N\|)m + d\beta_3\eta^2v^2\tau] + r\lambda_1(1+\|N\|))R_2 + 2(2m+r)hd\beta_0}\end{aligned}$$

Hence, by induction, we have

$$\begin{aligned}\|y_n\| &\le \frac{\sqrt{2([2d\beta_2\tau + 2ld_{\min}\|N\|(1-\|N\|)m + d\beta_3\eta^2v^2\tau] + r\lambda_1(1+\|N\|))R_2 + 2(2m+r)hd\beta_0}}{1-\|N\|} \\ &\quad + \max_{-\tau\le\xi\le 0}\|\varphi(\xi)\|\end{aligned} \tag{3.20}$$

for

$$n \geq \frac{([2d\beta_2\tau + 2ld_{\min}\|N\|(1-\|N\|)m + d\beta_3\eta^2v^2\tau] + r\lambda_1(1+\|N\|))R_2}{2hd\beta_0} + 2m + r$$

and $\beta_0 > 0$.

A combination of (3.11) and (3.20) shows that the method is finite-dimensionally $D(l)$-dissipative.

Theorem 3.4. Assume that a step-irreducible multistep Runge-Kutta method (2.4) is (k,l)-algebraically stable, $D > 0$, $l < 0$ and $k \leq 1$, the problem (2.1) satisfies (2.2) and (2.3) with $d(\beta_1 h + \beta_2 h + h\beta_3\eta^2v^2) < ld_{\min}(1-\|N\|)^2$. Then the method (2.4) with (2.5a, 2.5b and 2.5c) is finite-dimensionally $D(l)$-dissipative.

Proof. In the proof of Theorem 3.3, change all $d_{\min}$ into d, we can get the proof of Theorem 3.4.

Theorem 3.5. Assume that a method (2.4) is irreducible and algebraically stable, the problem (2.1) satisfies (2.2) and (2.3) with $\beta_1 h + \beta_2 h + \beta_3 h < 0$. Then, the method (2.4) with (2.5a, 2.5b and 2.5c) is finite-dimensionally D-dissipative.

Proof As in [18], we can prove that, if a stage-irreducible method (2.4) is algebraically stable for the matrices G and D, then $D > 0$, therefore, use the proof of Theorem 3.3 for $k = 1, l = 0$, we prove this theorem.

4 Comparison with Existing Results

(1) When $N = 0$, the problem (2.1) degenerates into an IVP of DIDEs

$$\begin{cases} y'(t) = f(y(t), y(t-\tau), \int_{t-\tau}^{t} g(t,\xi,y(\xi))d\xi), t \geq 0, \\ y(t) = \varphi(t), -\tau \leq t \leq 0. \end{cases} \tag{4.1}$$

The conditions (2.2) and (2.3) degenerate into

$$\mathrm{Re}\langle u, f(u,v,w)\rangle \leq \beta_0 + \beta_1\|u\|^2 + \beta_2\|v\|^2 + \beta_3\|w\|^2,\ u,v,w \in X, \tag{4.2}$$

and

$$\|g(t,s,u)\| \leq \eta\|u\|, \quad t \in [0,+\infty), \quad s \in [-\tau,+\infty), \quad u \in X \tag{4.3}$$

Gan [15] studies the dissipativity of θ-methods for DIDEs (4.1), Qi et al. [23] study the dissipativity of multistep Runge-Kutta methods for nonlinear VDIDEs. So far we have not seen in literature other numerical dissipativity results for nonlinear DIDEs. But Theorems 3.3 and 3.4 in this paper can be applied to this class of problem directly, and we can obtain the following Corollaries.

Corollary 4.1. Assume that a step-irreducible multistep Runge-Kutta method (2.4) is (k,l)-algebraically stable, $D>0$, $l>0$ and $k\leq 1$, the problem (4.1) satisfies (4.2), (4.3) with $d(\beta_1 h+\beta_2 h+h\beta_3\eta^2 v^2)<ld_{\min}$. Then the method (2.4) with (2.5a, 2.5b and 2.5c) for DIDEs is finite-dimensionally $D(l)$-dissipative.

Corollary 4.2. Assume that a step-irreducible multistep Runge-Kutta method (2.4) is (k,l)-algebraically stable, $D>0$, $l>0$ and $k<1$, the problem (4.1) satisfies (4.2), (4.3) with $d(\beta_1 h+\beta_2 h+h\beta_3\eta^2 v^2)<ld_{\min}$. Then the method (2.4) with (2.5a, 2.5b and 2.5c) for DIDEs is infinite-dimensionally $D(l)$-dissipative.

(2) When the right-hand side function of the problem (2.2) does not possess the integral term, the problem (2.1) degenerates into an IVP of NDDEs

$$\begin{cases} \frac{d}{dt}[y(t)-Ny(t-\tau)]=f(y(t),y(t-\tau)),t\geq 0, \\ y(t)=\varphi(t),-\tau\leq t\leq 0. \end{cases} \tag{4.4}$$

and the fourth term of the right side of condition (2.2) vanishes and thus (2.2) degenerates into

$$\mathrm{Re}\langle u-Nv,f(u,v)\rangle\leq\beta_0+\beta_1\|u\|^2+\beta_2\|v\|^2,\ u,v\in X \tag{4.5}$$

Wen [20] studies the dissipativity of θ-methods for nonlinear NDDEs (4.4), Wang [21] studies the dissipativity of Runge-Kutta methods for NDDEs with piecewise constant delay. So far we have not seen in literature other numerical dissipativity results for nonlinear NDDEs. Theorems 3.3, 3.4, 4.1 and 4.2 in this paper can be applied to this class of problem directly, and we can obtain the following Corollaries.

Corollary 5.3. Assume that a step-irreducible multistep Runge-Kutta method (2.4) is (k,l)-algebraically stable, $D>0$, $l>0$ and $k\leq 1$, the problem (4.4) satisfies (4.5) with $d(\beta_1 h+\beta_2 h)<ld_{\min}$. Then the method (2.4) with (2.5a, 2.5b and 2.5c) for NDDEs is finite-dimensionally $D(l)$-dissipative.

Corollary 5.4. Assume that a step-irreducible multistep Runge-Kutta method (2.4) is (k,l)-algebraically stable, $D>0$, $l>0$ and $k<1$, the problem (4.4) satisfies (4.5) with $d(\beta_1 h+\beta_2 h)<ld_{\min}$. Then the method (2.4) with (2.5a, b and c) for NDDEs is infinite-dimensionally $D(l)$-dissipative.

(3) When $N=0$ and the right-hand side function of the problem (2.1) does not possess the integral term, the problem (2.1) degenerates into an IVP of DDEs. Therefore, the results of Theorems 3.3, 3.4, 4.1 and 4.2 in this paper partially cover the numerical dissipativity of multistep Runge-Kutta for DDEs which is given by Huang in [11].

Acknowledgements. This work were supported by the Creative Talent Project Foundation of Heilongjiang Province Education Department (UNPYSCT-2015102).

References

1. Bocharov, G.A., Rihan, F.A.: Numerical modelling in biosciences with delay differential equations. J. Comput. Appl. Math. **125**, 183–199 (2000)
2. Xiao, A.G.: On the solvability of general linear methods for dissipative dynamical systems. J. Comput. Math. **18**, 633–638 (2000)
3. Humphries, A.R., Stuart, A.M.: Model problems in numerical stability theory for initial value problems. SIAM Rev. **36**, 226–257 (1994)
4. Humphries, A.R., Stuart, A.M.: Runge-Kutta methods for dissipative and gradient dynamical systems. SIAM J. Numer. Anal. **31**, 1452–1485 (1994)
5. Hill, A.T.: Global dissipativity for A-stable methods. SIAM J. Numer. Anal. **34**, 119–142 (1997)
6. Hill, A.T.: Dissipativity of Runge-Kutta methods in Hilbert spaces. BIT **37**, 37–42 (1997)
7. Xiao, A.G.: Dissipativity of general linear methods for dissipative dynamical systems in Hilbert spaces. Math. Numer. Sin. **22**, 429–436 (2000). (in Chinese)
8. Huang, C.M.: Dissipativity of Runge-Kutta methods for dissipative systems with delays. IMA J. Numer. Anal. **20**, 153–166 (2000)
9. Huang, C.M., Chen, G.: Dissipativity of linear-methods for delay dynamical systems. Math. Numer. Sin. **22**, 501–506 (2000). (in Chinese)
10. Huang, C.M.: Dissipativity of one-leg methods for dissipative systems with delays. Appl. Numer. Math. **35**, 11–22 (2000)
11. Huang, C.M.: Dissipativity of multistep Runge-Kutta methods for dynamical systems with delays. Math. Comput. Model. **40**, 1285–1296 (2004)
12. Tian, H.J.: Numerical and analytic dissipativity of the θ-method for delay differential equations with a bounded lag. Int. J. Bifurcation Chaos, 1839–1845 (2004)
13. Wen, L.P.: Numerical stability analysis for nonlinear Volterra functional differential equations in abstract spaces, Ph.D. thesis, Xiangtan University (2005). (in Chinese)
14. Gan, S.Q.: Dissipativity of linear θ-methods for integro-differential equations. Comput. Math Appl. **52**, 449–458 (2006)
15. Gan, S.Q.: Dissipativity of θ-methods for nonlinear Volterra delay-integro-differential equations. J. Comput. Appl. Math. **206**, 898–907 (2007)
16. Gan, S.Q.: Exact and discretized dissipativity of the pantograph equation. J. Comput. Math. **25**(1), 81–88 (2007)
17. Burrage, K., Butcher, J.C.: Nonlinear stability of a general class of differential equation methods. BIT **20**, 185–203 (1980)
18. Li, S.F.: Theory of computational methods for stiff differential equation. Huan Science and Technology Publisher, Changsha (1997)
19. Cheng, Z., Huang, C.M.: Dissipativity for nonlinear neutral delay differential equations. J. Syst. Simul. **19**(14), 3184–3187 (2007)
20. Wen, L.P., Wang, W.S., Yu, Y.X.: Dissipativity of θ-methods for a class of nonlinear neutral delay differential equations. Appl. Math. Comput. **202**(2), 780–786 (2008)
21. Wang, W.S., Li, S.F.: Dissipativity of Runge-Kutta methods for neutral delay differential equations with piecewise constant delay. Appl. Math. Lett. **21**(9), 983–991 (2008)
22. Wu, S.F., Gan, S.Q.: Analytical and numerical stability of neutral delay integro-differential equations and neutral delay partial differential equations. Comput. Math Appl. **55**, 2426–2443 (2008)
23. Qi, R., Zhang, C.J., Zhang, Y.J.: Dissipativity of multistep Runge-Kutta methods for nonlinear Volterra delay-integro-differential equations. Acta Math. Applicatae Sinica **28**(2), 225–236 (2012). English Series

Evaluation of Uncertainties of ITS-90 by Monte Carlo Method

Peter Sopkuliak[1(✉)], Rudolf Palenčár[1], Jakub Palenčár[1], Emil Suroviak[1], and Jaromír Markovič[2]

[1] Slovak University of Technology in Bratislava, Námestie Slobody 17, 812 31 Bratislava, Slovak Republic
{xsopkuliak, xsuroviak}@is.stuba.sk, {rudolf.palencar, jakub.palencar}@stuba.sk
[2] Slovak Legal Metrology, Hviezdoslavova 31, 974 01 Banská Bystrica, Slovak Republic
markovic@slm.sk

Abstract. The article briefly describes the approach of evaluating calibration using the adaptive method of Monte Carlo and the subsequent validation by the law of uncertainties when applied on the primary realization of the temperature scale, with emphasis on measurement with standard platinum resistance thermometer (SPRT) illustrated by the range (0 ÷ 660) °C of the international temperature scale (ITS-90).

Keywords: Uncertainty · Correlation · Monte Carlo method

1 Introduction

One of the most measured quantities, which undoubtedly has a dominant influence in all sectors of the national economy, is the temperature. The measurement of temperature has a significant effect on the quality and efficiency of each production process and ultimately an accuracy of temperature measurement plays more important role in this process. Temperature measurement can be carried out with devices with different levels of accuracy. For listed reasons, as well as taking into account the requirements of the practice it is essential to improve temperature measurement at the international level. Many tasks today without a deepening degree of knowledge in science, in technology and in the development of living standards could not be addressed without a fundamental knowledge of the theory of measurement, measuring equipment, physical principle of sensors, their attributes and metrological characteristics.

2 Scope

This article presents a method based on the propagation of distributions by MCM. The procedure is based on the generation of pseudo-random numbers of input variables of multi-dimensional distribution. Multi-dimensional distribution is used because it takes into account correlation between the SPRT resistances from calibration as well as the

R. Silhavy et al. (eds.), *Cybernetics and Mathematics Applications in Intelligent Systems*, Advances in Intelligent Systems and Computing 574, DOI 10.1007/978-3-319-57264-2_5

SPRT resistances in temperature measurement. In cases with uncorrelated resistances it is sufficient to generate input variables only from the one-dimensional distribution.

In our case it is necessary to know the probability distributions of input quantities and in the case of correlated input quantities relevant multivariate distribution function. We can assume normal distribution for all input SPRT resistances and therefore multivariate normal distribution for correlated resistances. This assumption is based on the central limit theorem because at the measurement there are several sources of uncertainties e.g. self-heating effect of the SPRT, chemical impurities of the substance in the DFPs, immersion effect of the SPRT, hydrostatic-head effect, etc.

The aim of this study is

1. the presentation of the procedure of MCM for uncertainty evaluation of the international temperature scale ITS-90 by using SPRT calibrated at DFPs;
2. the validation of the process by using the law of propagation of uncertainty according to the GUM for specific conditions.

The procedure was designed to take into account the correlation between the SPRT resistances calibrated at the DFPs as well as in temperature measurements. Influences such as fluctuations, drifts, temperature gradients and further, are not analyzed. Also, the uncertainties caused by non-uniqueness and consistency subranges are not included. These questions are presented e.g. in [1].

3 Current Status of the Issue

The vast majority of published papers submits an approach based on the law of propagation of uncertainty GUM [2] and its supplements [3, 4]. Smaller amount of articles are based on the orthogonal polynomials. The overview of the approaches is presented in [3, 6]. Propagation of distributions using MCM based on Supplement 1 [3] to the GUM [2] occurs only in a few isolated cases [7]. In general the authors predict uncorrelated resistance between the defining fixed points (DFP) and neglect the impact of correlations. Omitting the correlations between resistances in the DFP does not always correspond to reality and in accordance [6] they can have a significant impact. Articles based on the Monte Carlo method, which follows the recommendations and procedures listed in [3] appear sporadically

4 Theoretical Bases of ITS-90

International Temperature scale of 1990 defines the temperature from the inverse function

$$T = f(W_r) \tag{1}$$

Corresponding subranges of the ITS-90 from 0 °C for the functions (1) are stated in [8]. Function W_r is given by

$$W_r - W = \sum_{i=1}^{N} a_i f_i(W) \tag{2}$$

where

$$W = \frac{R}{R_{\mathrm{TPW}}} \tag{3}$$

while R is the SPRT resistance at temperature T and R_{TPW} is the SPRT resistance at TPW, $f_i(W)$ are functions of the individual subranges stated in [8], a_i are the coefficients of deviation function from the calibration of SPRT at DFPs and $a_i = g_1(R_{\mathrm{TPW1}}, \ldots, R_{\mathrm{TPW}N}, R_{\mathrm{DFP1}}, \ldots, R_{DFPN})$ or $a_i = g_2(W_{\mathrm{DFP1}}, W_{\mathrm{DFP2}}, \ldots, W_{DFPN})$. Matrix notation for the calculation of the coefficients of deviation function can be used. If the relationship (2) is applied to N fixed points, then

$$\begin{pmatrix} \Delta W_{\mathrm{DFP1}} \\ \vdots \\ \Delta W_{\mathrm{DFP}N} \end{pmatrix} = \begin{pmatrix} f_1(W_{\mathrm{DFP1}}) & \cdots & f_N(W_{\mathrm{DFP1}}) \\ \vdots & \ddots & \vdots \\ f_1(W_{\mathrm{DFP}N}) & \cdots & f_N(W_{\mathrm{DFP}N}) \end{pmatrix} \begin{pmatrix} a_1 \\ \vdots \\ a_N \end{pmatrix} \tag{4}$$

where $\Delta W_{\mathrm{DFP}i} = W_{\mathrm{r,DFP}i} - W_{\mathrm{DFP}i}$, $W_{\mathrm{DFP}i}$ are resistance ratios for corresponding DFPi and $W_{\mathrm{r,DFP}i}$ are defined in [8]. Equation (4) is written in the form

$$\Delta \boldsymbol{W}_{\mathrm{DFP}} = \boldsymbol{M}_{\mathrm{DFP}}\, \boldsymbol{a} \tag{5}$$

Because there exists $\boldsymbol{M}_{\mathrm{DFP}}^{-1}$ the coefficients of deviation function are given by

$$\boldsymbol{a} = \boldsymbol{M}_{\mathrm{DFP}}^{-1} \Delta \boldsymbol{W}_{\mathrm{DFP}} \tag{6}$$

Then the Eq. (2) can be rewritten as follow

$$W_r = W + \boldsymbol{a}^T \boldsymbol{f}(\boldsymbol{W}) \tag{7}$$

5 Application of Monte Carlo Method

5.1 Procedure of Calculation

In the process of calculation the temperature and its standard uncertainty by using Monte Carlo method based on propagation of distributions is schematically illustrated on Fig. 1.

5.2 Used Software and Generating of Pseudo-Random Numbers

When selecting a suitable programming software environment, it was crucial to create an application which would be easy to use and portable. It was also necessary to consider the efficiency of the calculation of final application and the way of

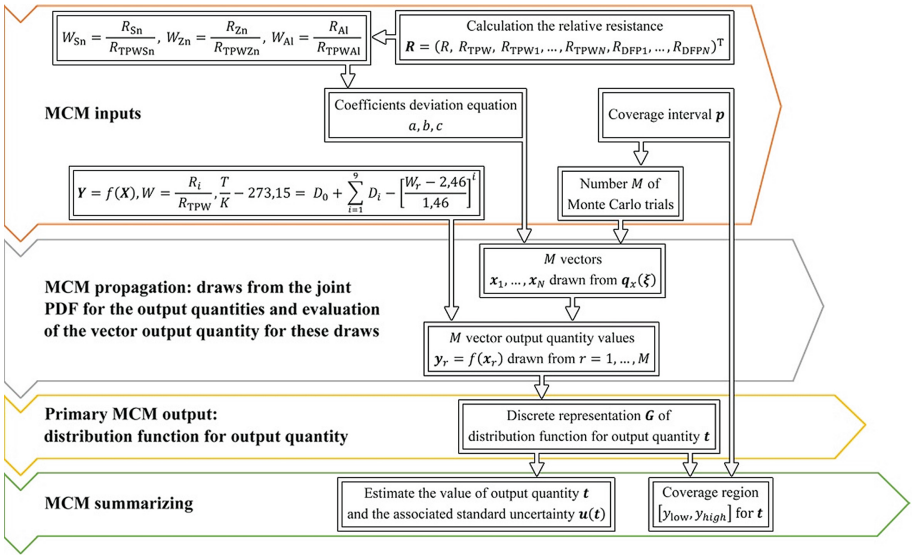


Fig. 1. Computing phase of calibration using Monte Carlo method

implementation of Marsenne Twister generator. For these reasons Microsoft Visual Basic .NET programming environment was chosen and 32 bit version of operating system has been used due to direct compatibility with newer 64-bit operating systems.

Visual Basic does not integrate the generation of pseudo-random numbers with Marsenne Twister (MT) generator which is currently the best rated algorithm which has undergone a large number of experiments for testing pseudo-random numbers. For this reason, the final algorithm uses the original MT source code translated for VB .NET framework. Since MT generator generates numbers from uniform distribution, it was necessary to use the Box-Muller transformation method. This method allows to transform uniformly distributed random variables to the Gaussian distribution.

6 Evaluation of Uncertainty and Experimental Data

In order to compare the results of both methods of calibration, it is necessary for the evaluation based on the Monte Carlo method to apply it on the data obtained by the calibration in our case by the Slovak Institute of Metrology (SMU). Results from the calibration were carried out by the SMU. These values will be used as inputs for evaluating the calibration EOST based on the Monte Carlo method and results will be compared.

6.1 Inputs and Considered Calibration Case

The input data for example calculated in this section are from SMU. Temperature has been measured outside the calibration laboratory. We used one triple point water cell for SPRT calibration. SPRT resistance of temperature measurement were considered uncorrelated with the other SPRT resistances. The SPRT resistances at defining fixed point are considered uncorrelated and calibration was made outside the laboratory. Correlation and covariance matrix of vector of input SPRT resistances R for simulated case is listed below.

Using SPRT outside calibration laboratory, resistances at DFPs are uncorrelated

$$\boldsymbol{R}_R = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix},$$

$$\boldsymbol{V}_R = 10^{-10} \begin{pmatrix} 1.61 & 1.61 & 1.61 & 0 & 0 & 0 & 0 & 0 \\ 1.61 & 1.61 & 1.61 & 0 & 0 & 0 & 0 & 0 \\ 1.61 & 1.61 & 1.61 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 14.8 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 24.8 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 39.9 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1.61 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 4.00 \end{pmatrix}$$

Whereby $\boldsymbol{V}_R = \boldsymbol{P}_R \boldsymbol{R}_R \boldsymbol{P}_R^{\mathrm{T}}$ and $\boldsymbol{P}_R$ is a diagonal matrix of dimension 8×8 with diagonal elements $u(R_{\mathrm{TPWSn}})$, $u(R_{\mathrm{TPWZn}})$, $u(R_{\mathrm{TPWAl}})$, $u(R_{\mathrm{Sn}})$, $u(R_{\mathrm{Zn}})$, $u(R_{\mathrm{Al}})$, $u(R_{\mathrm{TPW}})$, $u(R)$. The results of simulation by Monte Carlo method and GUM are presented in Table 1. The graphical comparison of both methods for 66 calibration points in the range $(0 \div 660)$ °C of the ITS-90 is illustrated on the Fig. 9. We consider $M = 10^5$ Monte Carlo trials. For the output quantity T we consider the reference probability $p = 0,95$ and the number of significant decimals digits $n_{\mathrm{dig}} = 2$, from standard uncertainty $u(T)$. Histogram for the resistance R_{Al} from the example given in 6.1, from $N(\boldsymbol{\mu}, \boldsymbol{V})$ distribution is shown in Fig. 2.

The a, b, c coefficients of deviation function can be determined from Eq. (6) for pseudo-random generated values for the input quantities from which we get a_1 to a_{10^5} samples. The procedure for calculation of coefficients a, b, c of deviation function is shown in Fig. 4. Histograms of coefficients of deviation function have similar shape as the coefficient a presented in Fig. 3.

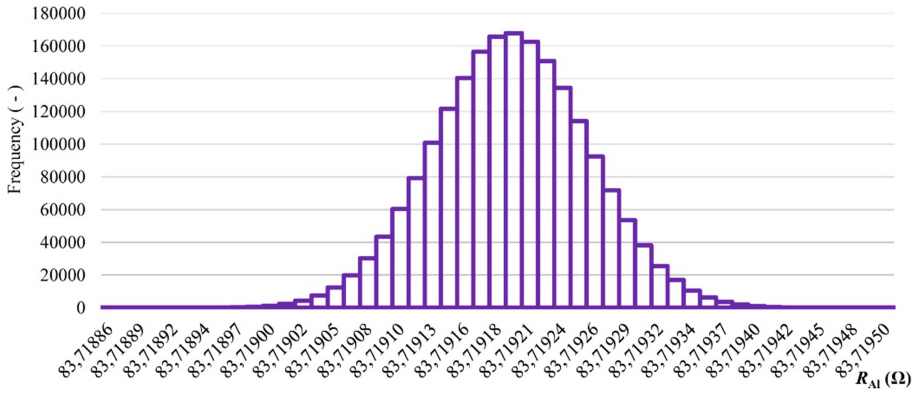


Fig. 2. Histogram for the input resistance

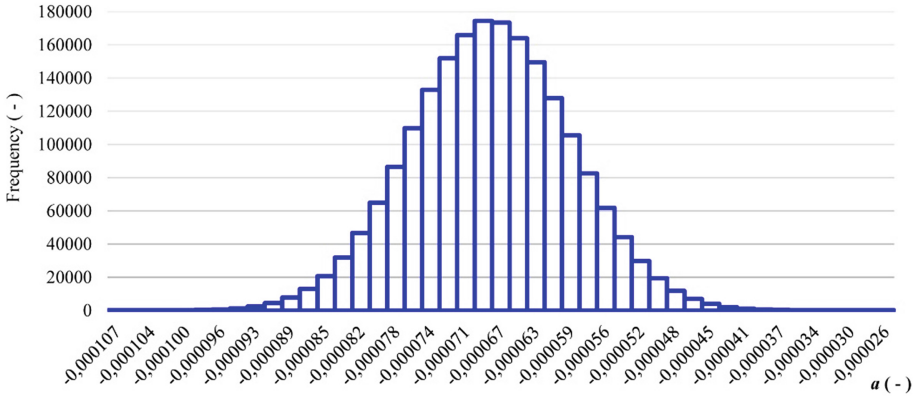


Fig. 3. Histogram of the deviation equation for coefficient a

The coefficients of deviation function a, b, c were calculated in vector form by Monte Carlo method in the previous chapter in order to calculate the estimate of the unknown temperature, associated standard uncertainty and confidence interval.

The measured resistance R at temperature T, mentioned above, is considered to be uncorrelated with other resistances and it may be generated from the one-dimensional probability distribution. In our analysis it is supposed (to be constant) as a constant (the uncertainty is equal to zero), so the problem can be analyzed without influencing the measurement uncertainty of the resistance for temperature measurements. For specific uncertainty measurement we use a specific type of distribution. The calculation procedure is shown in Fig. 5, where $f_8 = \frac{R}{R_{\mathrm{TPW}}}$, from Eq. (2) is $f_9 = a(W-1) + b(W-1)^2 + c(W-1)^3$ and from Eq. (7) is $f_{10} = W - a(W-1) + b(W-1)^2 + c(W-1)^3$. Histogram of estimated temperature t is presented in Fig. 7.

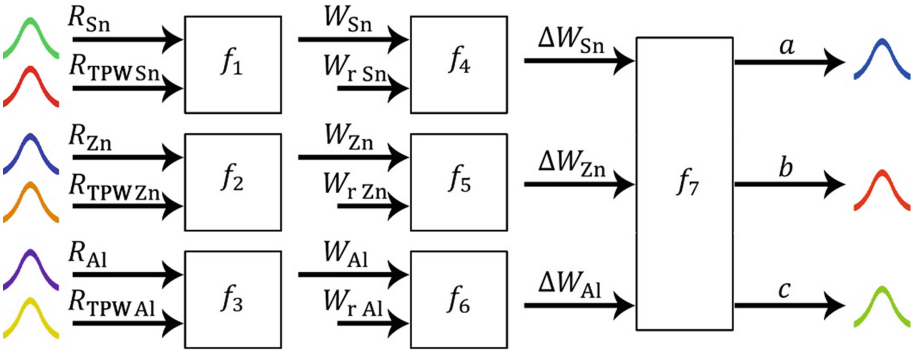


Fig. 4. Sub-model of calculation the coefficients of deviation function

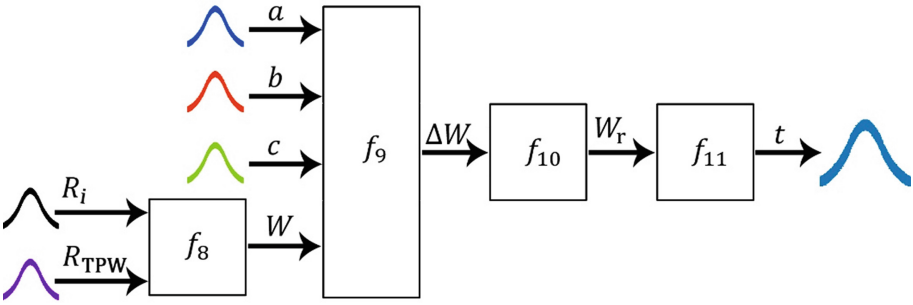


Fig. 5. Model of temperature calculation and its associated standard uncertainty

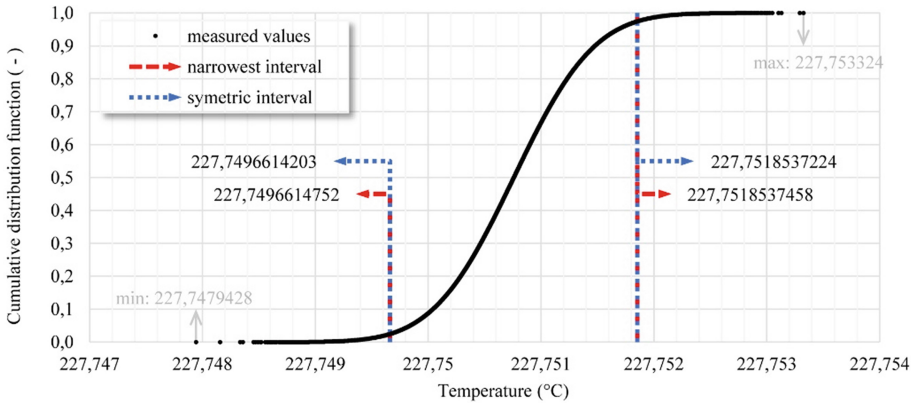


Fig. 6. The distribution function of the output temperature

6.2 Example of Calculation of Output Characteristics for Selected Case

Correlation and covariance matrix for the following example is listed above and results obtained by the law of uncertainty for specific resistance are $R_i = 46.55489887\ \Omega$, $t_i = 227.75076\ °C$, $u(t_i) = 5.59864 \cdot 10^{-4}\ °C$. Based on the generated input values for resistors and using appropriate relationship we get for our case with $h = 20$ and $M = 10^5$ estimate of the temperature t (°C). Determining the symmetrical reference interval with the specified probability for the estimating output quantity t is obtained from its generated discrete representation (distribution function shown on Fig. 6) by arranging

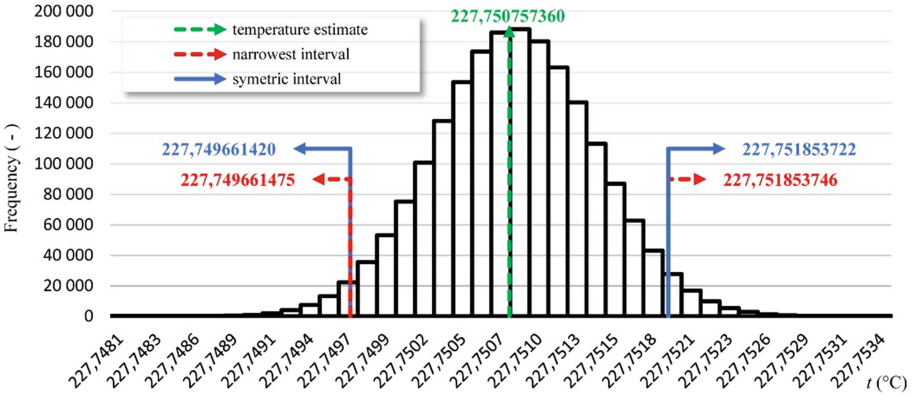


Fig. 7. Histogram of output quantity t

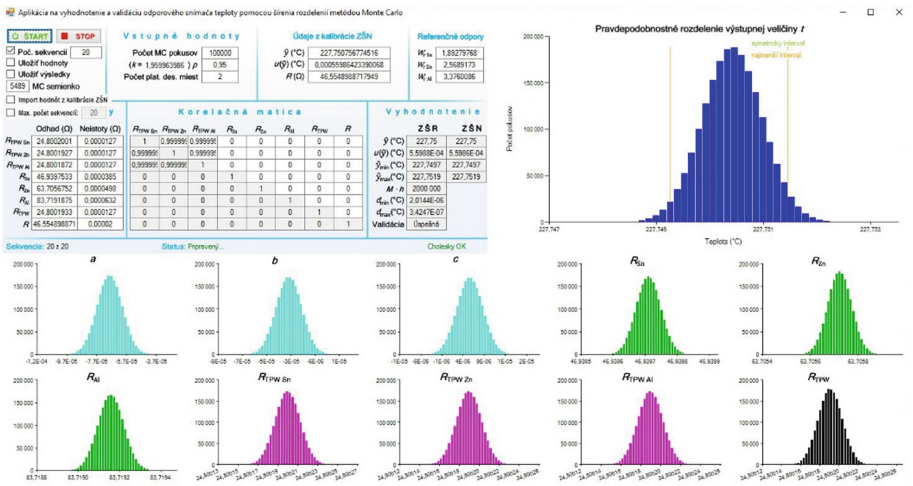


Fig. 8. User interface of the created application at the end of calculation

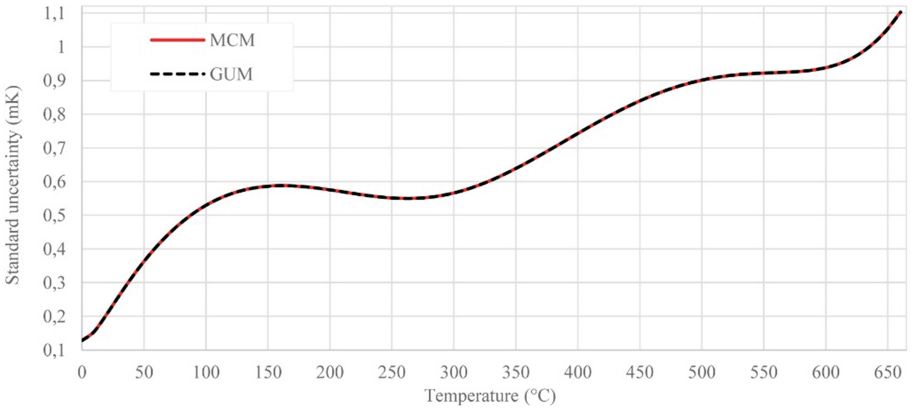


Fig. 9. Comparison of MCM and GUM for given cases illustrated on subrange (0 ÷ 660) °C of ITS-90

Table 1. Comparison and verification of the law of uncertainty propagation by the law of distribution propagation using MCM for one value t for specific case

Method	M	t(°C)	$u(t)$(°C)	95% coverage interval	GUF validated
GUF	–	227, 750757	0, 000560	[227, 749659; 227, 751854]	–
MCM shortest	1×10^5	227, 750758	0, 000560	[227, 749672; 227, 751863]	✗
MCM shortest	20×10^5	227, 750757	0, 000560	[227, 749661; 227, 751854]	✓
MCM symmetric				[227, 749661; 227, 751854]	✓

GUM uncertainty framework (GUF) [3], each sequence of MCM consists of $M = 1 \times 10^5$ trials

values t at non-decreasing sequence and rules listed in [3]. Using the above we get the symmetric and the narrowest confidence interval shown in Fig. 6.

Statistical characteristics of the resulting estimates are calculated from the partial estimates in each sequence $y^{(h)}$, $u\left(y^{(h)}\right)$. After carrying out the last sequence h it is possible to calculate the resulting parameters for estimation. Stabilization criterion determines whether it is necessary to increase the current value of the sequence $h > 2$ calculation of Monte Carlo method to the next sequence, in the case if one of the values $2s_{\hat{y}}$, $2s_{\hat{u}(y)}$, $2s_{\hat{y}_{\min}}$ and $2s_{\hat{y}_{\max}}$ becomes greater than the value of δ. After the successful fulfillment of the conditions of the stabilization criteria $y^{(h\times M)}$, $u\left(y^{(h\times M)}\right)$ and $\left[y_{\min}^{(h\times M)}; y_{\max}^{(h\times M)}\right]$ are final statistical characteristics determined from all the generated values. In the given example, the results are obtained by the Monte Carlo method following: estimation of the temperature $t = 227,75075736$°C, standard uncertainty $u(t) = 5,59884 \times 10^{-4}$ °C, 95% narrowest interval and symmetric interval as well as the histogram of output variable t obtained by the adaptive Monte Carlo method are shown in Fig. 7.

6.3 Validation of Uncertainty Propagation Law

Whether the reference interval obtained by uncertainty propagation and by Monte Carlo method are identical in certain numerical tolerance can be found out by the calculation. This numerical tolerance is rated from the point of view of limit values of the reference intervals and it gives as an expression of standard uncertainty $u(y)$ to existing number of decimal places. Numerical expression of tolerance δ with associated standard uncertainty $u(y)$ as described in [3] is as follows:

$$\delta = \frac{1}{2} \times 10^{r}, u(y) = 56 \times 10^{-5}\,^{\circ}\mathrm{C}\ , a = 56,\ r = -5 \Rightarrow \delta = \frac{1}{2} \times 10^{-5}$$

Absolute differences of limit values both confidence intervals are determined from the relations: $d_{\min} = \left|y_{\mathrm{GUM}} - U_{0,95(\mathrm{GUM})} - y_{\min(\mathrm{MCM})}\right|$ and $d_{\max} = |y_{\mathrm{GUM}} + U_{0,95(\mathrm{GUM})} - y_{\max(\mathrm{MCM})}|$. Table 1 shows detailed result of calibration for calibration case listed above for one selected point. From the calculation of absolute differences of limit values d_{low} and d_{high} results that are not larger than δ and verification of uncertainty propagation law is successful. Of course, it's not to be applied for other conditions.

The user interface of created application for the evaluation and validation of the standard platinum resistance thermometer by the Monte Carlo method is shown in the figure below (Fig. 8).

7 Conclusions

This paper presents a procedure for determination of uncertainties of temperature measurement by Monte Carlo method. This procedure is based on generating of pseudo-random numbers as the input of standard platinum resistance thermometer resistances at DFPs and at TPW. To be able to take into account the correlations between DFPs the approach of generating pseudo-random numbers from multivariate distributions was used.

We assumed an eight-dimensional Gaussian probability distribution. The assumption of Gaussian distribution is quite acceptable, because of the several sources of uncertainty of SPRT resistances at DFPs. If the correlations among the SPRT resistances at DFPs are negligible, it is possible to adapt the model in such a way that the input resistances are uncorrelated and we can use one-dimensional distributions for each input resistance. The calculation procedure has been prepared in accordance with the document [5, 8, 9] whereas in the algorithm strictly integrates general approach for evaluating the measurements listed in annex 1 and 2 GUM [3, 4].

Attention was also paid to validation of the use of uncertainty propagation law in accordance with the GUM for particular conditions. For the case of uncorrelated input SPRT resistances at defining fixed points the validation is successful and the results obtained by using uncertainty propagation law and by using MCM were consistent.

Acknowledgments. Authors would also like to thank the Slovak University of Technology in Bratislava, the grant agency VEGA projects number 1/0604/15, 1/0748/15 and the agency KEGA project number 014STU-4/2015 for their support.

References

1. BIPM. Uncertainties in the Realization of the SPRT Sub-ranges of the ITS-90. BIPM, Paris, Sèvres (2009). http://www.bipm.org
2. JCGM 100: Evaluation of Measurement Data – Guide to the Expression of Uncertainty in Measurement, (BIPM) (2008)
3. JCGM 101: Evaluation of measurement data – Supplement 1: Guide to the expression of uncertainty in measurement – Propagation of distributions using a Monte Carlo method (2008)
4. JCGM 102: Evaluation of Measurement Data – Supplement 2: Guide to the Expression of Uncertainty in Measurement – Extension to any number of output quantities, (BIPM) (2011)
5. Rosenkranz, P.: Uncertainty propagation for platinum resistance thermometers calibrated according to ITS-90. Int. J. Thermophys. **32**, 106–119 (2010)
6. Palenčár, R., Ďuriš, S., Brokeš, V.: Neistoty pri realizácii teplotnej stupnice. Bratislava: Vydavateľstvo STU, p. 159 (2015). ISBN 978-80-227-4286-3
7. Ribeiro, S., Alves, J., Oliveira, C., Pimenta, M., Cox, M.G.: Uncertainty evaluation and validation of a comparison methodology to perform in-house calibration of platinum resistance thermometers using a Monte Carlo method. Int. J. Thermophys. **29**, 902–914 (2008)
8. The International Temperature Scale of 1990. BIPM, Sévres, France (1989)
9. Palenčár, R.: Postup výpočtu kovariancií. Metrológia a skúšobníctvo **2**, 22–25 (2000)

Exploiting Model Continuity in Agent-Based Cyber-Physical Systems

Domenico L. Carní[1], Franco Cicirelli[2], Domenico Grimaldi[1], Libero Nigro[1(✉)], and Paolo F. Sciammarella[1]

[1] Engineering Department of Informatics Modelling Electronics and Systems Science, University of Calabria, 87036 Rende, CS, Italy
l.nigro@unical.it, {dlcarni,d.grimaldi,p.sciammarella}@dimes.unical.it
[2] CNR - National Research Council of Italy, Institute for High Performance Computing and Networking (ICAR), 87036 Rende, CS, Italy
f.cicirelli@dimes.unical.it

Abstract. This work develops an agent and control based approach for modeling, analysis and implementation of Cyber-Physical Systems (CPSs). Novel in this software engineering approach is a support to model continuity, that is the possibility of transitioning a same model from property analysis based on simulation, down to design, implementation and real-time execution. The paper introduces the basic concepts of the methodology, illustrates some implementation issues and presents a case study concerned with power management in a smart micro-grid.

Keywords: Control-centric approach · Multi-agent systems · Model continuity · Simulation · Real-time · Arduino · Power management · Smart micro-grid

1 Introduction

Currently many efforts are directed to an effective design of Cyber-Physical Systems (CPSs) [1], whose major goal is the real-time control of a physical plant, interfaced by suitable sensor/actuator devices, by a cyber software component linked to the physical plant through an interconnection network.

In this work the agent paradigm [2] is experimented for CPS development. As shown in [3], the agent paradigm proves to be a natural and powerful one for general distributed control applications, where agents can operate cooperatively and can dynamically learn from, and adapt to, their influencing environment. A multi-agent system design is well-suited, e.g., for power management in a smart home automation system [4].

An original agent and control based software engineering architecture [5,6] is adopted in this paper, which fosters software modularity while providing abstraction mechanisms for managing the timing aspects. The computational model is based on actors, asynchronous messages and time-sensitive actions. Scheduling and processing of both messages and actions are responsibility of a

R. Silhavy et al. (eds.), *Cybernetics and Mathematics Applications in Intelligent Systems*, Advances in Intelligent Systems and Computing 574, DOI 10.1007/978-3-319-57264-2_6

control structure selected from a library, which manages a particular notion of time (simulated or real-time). A unique feature of the adopted framework is a support to *model continuity* [7,8], that is the possibility of using a same model both for property analysis by simulation, and for implementation and real-time execution. Seamless transition from a development phase to the next one relies on a particular interpretation of agent actions and associated processing units and on the adoption of a particular control structure. No changes are required to the model, i.e., to actor behaviors and message exchanges.

Authors' current work is focussed on an exploitation of model continuity in the domain of CPSs. A *Gateway* component is introduced which interfaces the cyber part with the physical part. The gateway requires to be re-interpreted when moving from the analysis to the implementation phase. During the analysis phase the component can also take in to account causal-effect relations tied to operations carried out on the environment. The contribution of this paper consists in experimenting the agent methodology in the development of a concrete system devoted to power management in a smart micro-grid.

The paper is structured as follows. Section 2 introduces the basic concepts of the agent architecture. Section 3 focuses on an exploitation of the agent framework for CPSs. Section 4 describes the case study. Section 5 concludes the paper with an indication of on-going and future work.

2 Agent Architecture

The adopted control-sensitive agent framework is founded on the notions of *actors* (agents) and *actions*. Details about the developed APIs, i.e., class hierarchies for actors, actions and control structures, can be found in [5].

2.1 Actors Concepts

Actors are thread-less agents whose behavior is modeled by a finite state machine. Actors communicate to one another by asynchronous message passing. An actor is at rest until a message arrives. An incoming message is processed by the *handler()* method whose execution causes in general (a) an update to local variables of the actor, (b) a change to the current state of the state machine, (c) new messages to be sent to known actors (acquaintances), and (d) one or more actions to be created and submitted.

A Logical Process (LP) is a subsystem of actors allocated for execution on a computing node. Local actors of an LP are managed by a *control machine*, which transparently buffers exchanged messages into one or more message queues and ultimately deliver messages to destination actors, according to a time-sensitive *control strategy*, tailored to simulation or real-time execution. Message processing in an LP represents the *unit of scheduling* and *dispatching*. As a consequence, messages are ultimately handled one at a time and in an interleaved way, thus enabling a cooperative (not overkilling) concurrency schema. True actor concurrency exists among actors belonging to different LPs of a distributed system.

2.2 Actions Concepts

Message exchanges promote sociality among actors and capture the occurrence of events. Besides messages, the actor framework also provides the notion of *actions*, i.e., activities which consume time and require *processing units* (PUs) for them to be executed. Actions are executed in parallel, depending on the availability of PUs. In general an action, after its submission by an actor, can run to completion or it can be suspended/resumed or aborted. Each action is accompanied by a *list of input parameters* and a *list of output parameters*. Purposely, actions have no visibility to the internal data variables of the submitter actor. As a consequence, no interference can ever occur from the action parallel execution. When an action terminates, it can inform the submitter by an *action completion message*.

Actions can be reified in different ways. Simulated actions are pure time consuming activities, used to advance the simulated time. Real or effective actions have a concrete instruction body (algorithm) whose execution increases the real time. Pseudo real actions advance the real time too but have no concrete algorithm to execute. They can be useful for the *preliminary real-time execution* of a model [6], which is a key to check how the overhead introduced by message exchanges and message processing affects the system timing constraints.

Action execution can be atomic or it can be preempted [5]. An action can return to its submitter a (partial) result at selected time points (e.g., spaced uniformly for a periodic behavior) belonging to an assigned time window.

The different types of actions are managed by corresponding *action schedulers*. An action scheduler administers local processing units and can store actions which find no available PU in pending action queues, waiting for some specific or unspecific PU to be ready to accept a new action execution. A PU can be either a physical core or it can be realized by a Java thread, or it can be a fake object in the case of simulated actions.

Novel in the actor control framework is the possibility of transitioning, without changes, a model from property analysis to real time execution (*model continuity* [7,8]). Switching from simulation to real execution needs replacing the control machine, the time notion and the nature of actions. The developer has to substitute simulated actions with real actions and associated action schedulers. All the remaining details of the model, that is actor behaviors and message passing, remain exactly the same during the transition.

3 Exploiting the Agent Framework for CPS

A Gateway is a boundary component, which requires to be reified when switching from the analysis phase to real system implementation. It abstracts the physical devices which are used to sense and monitoring a controlled environment. During analysis, the Gateway simulates such physical devices, during real execution the goal is hiding communication protocols and allowing both agents and actions to exploit hardware functionalities in a standard and uniform way.

3.1 Gateway Design Guidelines

A Gateway exposes a simple API allowing to carry out the basic read/write operations to sensors/actuators selected by a unique id, and abstracts the way such operations are ultimately handled. The read/write operations are typically invoked by submitted actions which, in a real execution scenario, are carried out on dedicated Java threads. The Gateway maintains a collection of data variables, which correspond to sensor/actuator devices. An In/Out layer controls the communication with physical devices and the update to data variables. The In/Out layer is composed of input/output Java threads, which interface the communication channels with a number of i/o hardware components, e.g., Arduino [9] or similar. Sensor/actuator devices are physically linked to the i/o hardware. To simplify configuration and operation, each i/o hardware can be specialized to managing a disjoint subset of sensors or actuators. The communication channels between the Gateway and the i/o hardware can be based on a serial connection [10] or on a wireless connection. The experiments described in this paper were carried out using the serial connection managed by the RXTX.jar Java library. Separate concurrent hash maps are used for handling the input (sensor) data variables and the output (actuators) commands and data.

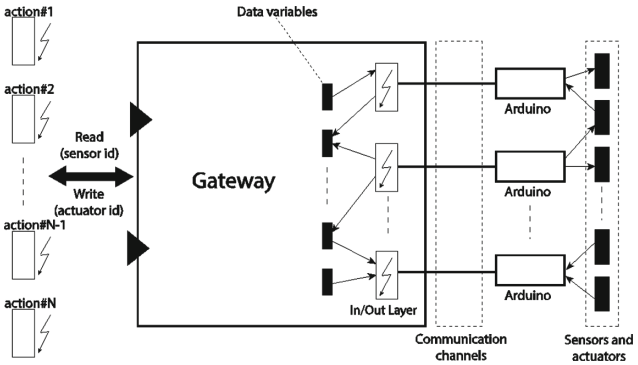


Fig. 1. Organization of a Gateway component

A design issue of the Gateway is concerned with the adoption of an anticipation schema. The i/o hardware components are supposed to be programmed so as to repeatedly reading the sensors and providing the data to the Gateway. From time to time, the values of the data variables represent the most fresh data values. For generality concerns, an action which requires some sensor data can specify a *filter* object at the request time. The filter provides a guard method (a *rule*) which must be satisfied by the involved data variables to cause the data values to be actually returned.

The anticipation schema proves effective for coping with exceptional events. Let's consider, for example, the situation where the cyber part has to react to a

crash event in the external environment. As soon as the physical event is sensed it gets captured in a data variable where it remains available for an action to read, without any risk of losing it for timing problems of action execution.

3.2 Specializing the Gateway to Work with Arduino

Figure 1 reports the case in which the Gateway was concretely interfaced with Arduino [9], which can be flexibly configured and programmed, using a C-like language, for carrying out the i/o operations. Arduino configuration is accomplished in the **setup()** function, where details of pin connections with physical devices are established. **setup()** is executed only once following a reset of the device. After that, Arduino enters its main **loop()**. The instructions of the loop can be directed to reading from sensors and to put the data, e.g., after some A/D conversions, onto the communication channels towards the Gateway (see Fig. 1). At the end of the loop, a delay statement is executed which defers the starting of the next loop iteration. During the loop operation, the Arduino can also receive and process interrupt signals. For instance, a *serialEvent interrupt* which is raised whenever new data arrive through the hardware serial communication link (RX), can be heard and processed only at the end of each loop iteration. The mechanism can be exploited to dynamically change the amount of the loop delay.

An Arduino devoted to controlling only actuators has an empty loop and the output operations are all delegated to serial interrupt handling.

3.3 Using the Gateway in Simulation

A different concretization of the Gateway is required for simulation. In this phase the In/Out and i/o hardware layers are transparently replaced by software agents which can provide, in simulation time, pre-generated input data to the Gateway or simply consume output commands.

Since an actuation induces, in general, a change in the controlled environment which must then be perceivable by the cyber part through the Gateway, an EnvAgent can be used which (possibly) through a mathematical model, fuzzy logic etc., can reproduce the necessary changes in the environmental variables monitored by the Gateway.

4 A Case Study Using Power Management

The developed case study is devoted to power management in a smart micro-grid [3] acting as a home automation system [4], where a generated power signal of a renewable energy source like a photovoltaic panel, represents a reference to which user power loads are required to adapt dynamically. The goal is to control the power system so as to consume to the maximum extent the available generated power, thus minimizing costs of recourse to the external energy provider.

Differently from [4] where the power management system was prototyped only in simulation, the presented application depends on the distributed agent infrastructure, discussed in the previous section, and exploits model continuity from model simulation (for debugging and analysis purposes) to real-time execution.

The control algorithm is implemented using agents and rests on an *adaptation mechanism* based on *resource scheduling* and *shedding* [11].

User power loads, or equipments, are supposed to be ranked by priority and, in some cases, to be modularly organized thus enabling control actions.

4.1 Concrete System Example

For demonstration purposes a real case study was assembled in the context of an academic laboratory, with the production power signal pre-generated according to real observed behavior. A group of physical and modular loads (lamps) are controlled by corresponding load agents, which in turn receive control commands from a SchedulerAgent which is in charge of monitoring the reference input power signal and to adapt the power loads accordingly.

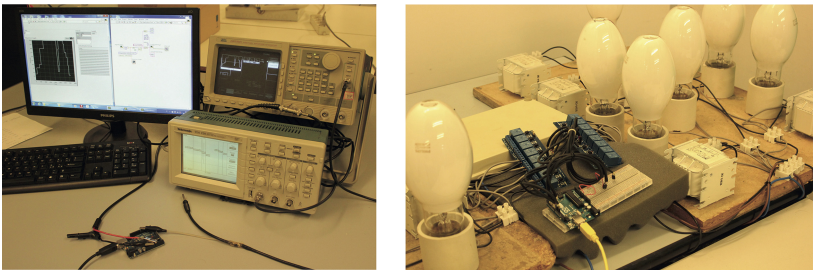


Fig. 2. Measurement stand and controlled loads (lamps)

Two Arduino One are used: the first one is devoted to acquiring the samples of the input available power signal generated by an Arbitrary Waveform Generator (Sony/Tektronix AWG2021); the second Arduino is dedicated to commanding the physical loads (see Fig. 2).

4.2 Loads Subsystem

The load subsystem is composed of one Arduino One board, 2 relay board with 8 high voltage channels characterized by rating of 10 A at 250 and 125 V AC and 10 A at 30 and 28 V DC, managed by means of 8 digital pins adapted to work with the Arduino output operating voltage, and 12 Standard High Pressure Mercury lamps Philips HPL-N 250 W and 12 ballasts. Each ballast is connected to a lamp in order to regulate the current to the lamps and provides sufficient voltage to

start the lamp. The loads are obtained by connection of a pre-established number of lamps. The digital pins of the Arduino board feed the digital pins of the relay modules, permitting the turn on and off of the loads.

4.3 Communication Protocol

The interaction between the cyber and physical subsystems is based on the exchange of character strings. Strings can represent commands to execute on the loads, or information that Arduino collects from sensors, ultimately destined to agents. The following format is adopted: *sensorId/actuatorId # message*, where the content of the token message is variable. Information about sampling of the available power signal is sent from Arduino as: *watt # powerLevel* where powerLevel is expressed as a double number. The command used to change the sampling period, can be sent by agents through the format: *arduinoId # samplingTimeInMillis*. To drive a relay it is necessary communicate to Arduino the id of the load and the type of command to execute. The format is: *loadId # command powerLevel* where the type of *command* can be:

- A, for *activate*. It is used to (re)connect a load with a specified power level;
- C, for *change load power*. It is used to change the power level of an already connected load;
- D, for *deactivate*. It is used to disconnect a load (in this case no power level is specified).

4.4 Multi-agent Design

The cyber system naturally mirrors the physical plant needs. Distinct agents are introduced for controlling respectively the loads and for the sample acquisition of the generated power signal. Details of input/output devices are confined within the corresponding agents. This modular structure simplifies the design of the scheduler agent, whose behavior can focus on the logic of the power management and optimization, and not on such physical aspects like the ids of devices and the particular commands for interacting with them.

The multi-agent system purposely exploits a multi time point action. Such an action naturally copes both with the data acquisition of the available produced power and with the monitoring requirements of the loads.

For the purpose of the case study the scheduler agent is designed so as to admit a *predictive* mechanism and a *reactive* mechanism, both operating at the level of the basic time unit of the reference power signal. The predictive mechanism relies on predicted data about the available power, established, e.g., with the help of a weather forecast model or simply coincident with the true values observed the day before. Such data are assumed to be already known to the scheduler. The reactive mechanism, instead, adapts the scheduling according to real data. At the end of the elementary period t the predictive mechanism proposes a scheduling plan for the loads for the next period $t+1$ on the basis of available predicted data. The next sample of the reference power signal arrived at

the beginning of the period $t+1$, causes the scheduling plan to be regenerated in the case the real sample differs from the predicted one. In the case the real power sample agrees with the predicted one, the proposed plan is instead executed by sending commands to the load agents. It is worth noting that a scheduler message for activating a load, can actually be ignored by the relevant load agent if the current state of the load is already compliant with the activation request.

The multi-agent model is characterized by its openness and flexibility. For instance, a new load can dynamically announce its willingness to execute by sending a notify message to the scheduler agent. The message carries all the load parameters such as priority, time to start and so forth. In a similar way, a load can detach from the scheduler by stating, through a message, that it is terminated.

4.5 Data Configuration

Each load (see Table 1) is characterized by:

- a priority level π, ranging from 1 to 4, where 4 is the highest level;
- a power consumption temporal behavior, specified by a stepwise function;
- the number of lamps, for each step, necessary to set up the power consumption in the real experiment.

The power level of a modular load can be reduced by a scaling factor so as to adapt to the actual power level available. For example, an increase in the generated power signal implies that the scheduler reconsiders each load consumption starting from its maximum value, and reduces it according to the new reference scenario.

The execution of the highest priority load is supposed to be cyclic. Its time varying power level is a matter of load requirement and does not depend on a scheduler decision.

Table 1 also furnishes the details about the connections used to drive the lamps through a minimal number of relays. From the different power levels

Table 1. Loads parameters

Priority π	Duration (t.u.)/PowerRequest /ActiveLamps	Scaling factor	Cyclic load	Relays $\times$ Lamp
4	10/500/2 30/1250/5 10/1000/4 10/750/3	1	Yes	(1×2) (1×1) (1×2)
3	50/500/2	2	No	(1×1) (1×1)
2	300/1000/4	2 or 4	No	(1×1) (1×1) (1×2)
1	220/250/1	1	No	(1×1)

required by each load, some lamps can purposely be connected in parallel so as to be driven by a single relay. All of this affects the number of pins needed to perform the test, and makes it possible to use only one Arduino board to turn on and off the actuators.

For example, the maximum power needed by the load with $\pi = 2$ is 1000 W (4 lamps each of 250 W) and it can be regulated (due to the scaling factor being 2 or 4) so as to consume 1000 W, or 500 W or 250 W. For this modularization, two lamps are connected in parallel (1 controlling pin) to achieve 500 W, and the other two are separately controlled (other 2 pins).

A predicted reference power signal for the energy production of a photovoltaic panel, is shown in Table 2. Each row specifies the duration of a step, expressed in minutes, and the relative power level.

Table 2. Predicted reference power steps

Duration (t.u.)	Power (W)	Duration (t.u.)	Power (W)
20	400	5	2000
30	800	50	2500
20	1000	5	2300
40	2000	15	1800
130	2500	10	1500
10	600	5	1000
5	1000	15	400

4.6 Execution Experiments

The prototyped agent model was preliminarily investigated in simulation, and subsequently executed in real-time.

Property Analysis by Simulation. The simulation study was mainly devoted to checking the scheduling algorithm. Simulated versions of the *acquisition action* which periodically reads the reference power samples from the Gateway, and of the *load actions* which reproduce the power consumption durations, are used. A GeneratorAgent supplies, in simulation time, the reference power samples to the Gateway. Actuation commands are simply turned to log commands. The EnvAgent modeling the external environment reduces to the use of booleans for recording the active/inactive state of each load (lamp).

Figure 3 shows a scenario with 3 loads l_1 l_2 l_3 and associated priorities $\pi(l_i)$ along with a time dependent consumption plan. For simplicity, predicted and real power signals are assumed to coincide at each instant. The sequence of scheduling operations at selected change time points 0, 2, 3 and 4 (see Fig. 3b) together with the associated sequence of load commands are collected in Fig. 4. For brevity, scheduling is not shown in other time points where it coincides with

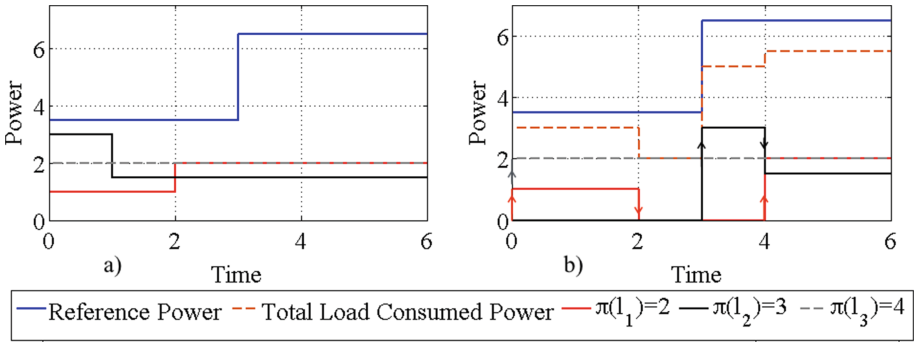


Fig. 3. A power scenario (a) and some scheduling points (b)

```
TIME: 0.0
LOADS TO MANAGE
(ID, PRIORITY, POWER CONSUMPTION)
L3 π=4,4,2.0
L1 π=2,2,1.0

COMMAND SEQUENCE:
L3#A2.0
L1#A1.0
```

```
TIME: 2.0
LOADS TO MANAGE
(ID, PRIORITY, POWER CONSUMPTION)
L3 π=4,4,2.0

COMMAND SEQUENCE:
L1#D
```

```
TIME: 3.0
LOADS TO MANAGE
(ID, PRIORITY, POWER CONSUMPTION)
L3 π=4,4,2.0
L2 π=3,3,3.0

COMMAND SEQUENCE:
L2#A3.0
```

```
TIME: 4.0
LOADS TO MANAGE
(ID, PRIORITY, POWER CONSUMPTION)
L3 π=4,4,2.0
L2 π=3,3,1.5
L1 π=2,2,2.0

COMMAND SEQUENCE:
L2#C1.5
L1#A2.0
```

Fig. 4. Scheduler decisions and scheduled load commands for the example scenario of Fig. 3

the last computed one. Time advancement in simulation is caused by the periodicity of the acquisition action and by the duration of load action consumptions. Activating a load for a given duration δ, implies the action completion message gets scheduled with the timestamp $now + \delta$. The scheduler agent only transmits activation/deactivation messages to the load agents. The activation duration δ is known to the load agent.

Simulation experiments of the case study are depicted in Figs. 5 and 6, where the total load consumption and the available power are shown vs. time. The power levels are normalized to the generated power signal.

In Fig. 5, predicted and real generated power are assumed to coincide at each instant. Figure 6 illustrates a scenario where the reactive scheduling replaces the predictive one. In particular, the available power signal is supposed to have a sudden decrease at time 130 with the lower level which is held for 50 time units.

Only the load $\pi = 2$, remoduled to operate at 500 W, is allowed to be active during the reactive time interval.

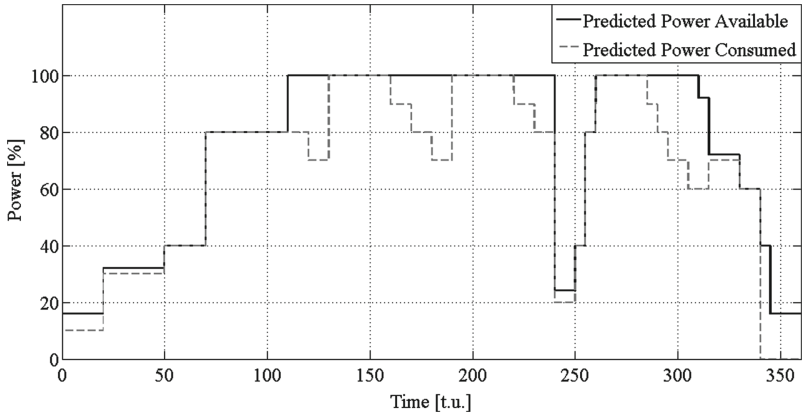


Fig. 5. Power monitoring and control - pure predictive case

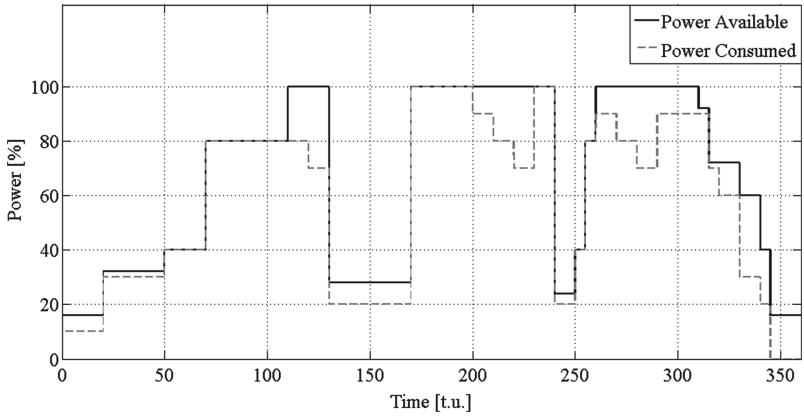


Fig. 6. Power monitoring and control - predictive and reactive case

Real-Time Execution. The case study was then studied in real-time mode, by replacing the control machine and turning actions from the simulated versions to their effective versions. Now the input power samples are acquired by the combination $< Gateway, inputArduino >$. In addition, the physical loads are directly controlled through the Gateway and the output Arduino which executes the output command. It is worth noting that in real-time the action duration sent by the scheduler agent to a load agent is redundant, because the physical load remains active/deactive until the next scheduled command is actuated.

The LabView software loads the predicted power signal samples (Fig. 2) in the AWG2021 memory and configures the AWG2021 so as to generate the signal with a sampling frequency equal to 10 Hz and an amplitude in the range $[0 : 5]$ V. With this configuration, the signal is compatible with the acquisition amplitude

range of the Arduino board, and it has a total duration of 360 s. In the set up the Arduino configures its analog input channel and sends a synchronization trigger signal to the AWG2021 that starts the generation of the power signal. After the set up configuration phase, the Arduino board acquires the AWG2021 output with a sampling frequency of 1 Hz and sends back the acquired data to the Gateway from which it is ultimately acquired by the scheduler agent.

The observed real-time behavior of reference power vs. the total load consumption coincides with that reported in Fig. 5, with 1 s being the time unit, except for an initial offset of 7 s needed by the Arduino set up and the opening of the associated serial ports.

5 Conclusions

This paper focuses on the use of a control-centric agent architecture [5,6] for CPS design, which fosters *model continuity* [7,8]. The approach is applied to a case study concerned with power management in a smart micro-grid home automation system [3,4]. Prosecution of the research aims to:

- improving/extending the interconnection between the cyber and the physical parts currently based on the mediation of Arduino [9], with more powerful hardware also in the presence of wireless communications and protocols;
- experimenting with use of the agent and control based approach in general smart environments by using IoT;
- optimizing the runtime infrastructure of agents through a direct support of the theatre architecture [12].

References

1. Lee, E.A.: Cyber physical systems: design challenges. In: Proceedings of International Symposium on Object/Component/Service Oriented Real-Time Distributed Computing (ISORC), pp. 363–369 (2008)
2. Wooldridge, M.: An Introduction to Multi-agent Systems, 2nd edn. Wiley, Chichester (2009)
3. Rohbogner, G., Hahnel, U.J.J., Benoit, P., Fey, S.: Multi-agent systems' asset for smart grid applications. Comput. Sci. Inf. Sys. (ComSIS) **10**(4), 1799–1822 (2013)
4. Abras, S., Ploix, S., Pesty, S., Jacomino, M.: A multi-agent home automation system for power management. In: Cetto, J.A., Ferrier, J.-L., Costa dias Pereira, J.M., Filipe, J. (eds.) Informatics in Control Automation and Robotics. LNEE, vol. 16, pp. 59–68. Springer, Heidelberg (2008). ISBN: 978-3-540-97141-6
5. Cicirelli, F., Nigro, L.: Control centric framework for model continuity in time-dependent multi-agent systems. Concurrency Comput. Pract. Exp. **28**(12), 3333–3356 (2016)
6. Cicirelli, F., Nigro, L., Sciammarella, P.F.: Agents+Control: a methodology for CPSs. In: IEEE/ACM 20th International Symposium on Distributed Simulation and Real Time Applications (DSRT), pp. 45–52. IEEE Computer Society (2016)

7. Hu, X., Zeigler, B.P.: Model continuity to support software development for distributed robotic systems: a team formation example. J. Intell. Robot. Syst. **39**(1), 71–87 (2004)
8. Hu, X., Zeigler, B.P.: A simulation-based virtual environment to study cooperative robotic system. Integr. Comput.-Aided Eng. **12**, 353–367 (2005)
9. Arduino on-line. https://www.arduino.cc
10. Arduino serial-link. http://playground.arduino.cc/Interfacing/Java
11. Cicirelli, F., Furfaro, A., Grimaldi, D., Nigro, L., Pupo, F.: MADAMS: a software architecture for the management of networked measurement services. Comput. Stand. Interfaces **28**(4), 396–411 (2006)
12. Cicirelli, F., Giordano, A., Nigro, L.: Efficient environment management for distributed simulation of large-scale situated multi-agent systems. Concurrency Comput. Pract. Exp. **27**(3), 610–632 (2015)

Design of Processor in Memory with RISC-modified Memory-Centric Architecture

Danijela Efnusheva(✉) and Aristotel Tentov

Computer Science and Engineering Department, Faculty of Electrical Engineering and Information Technologies, Skopje, Republic of Macedonia
{danijela,toto}@feit.ukim.edu.mk

Abstract. The technological developments in the areas of computer hardware and software resulted in a wide range of fast and cheap single- or multi-core processors, compilers, operating systems and programming languages, each with its own benefits and drawbacks, but with the ultimate goal to increase overall computer system performances. Although the number of transistors on a chip continues to double roughly every two years, there is still difficult to improve the performance of sequential processors, and even of the parallel multi-core and multi-processor shared-memory systems. The main reason for this resides in the ever-increasing gap between processor and memory speeds in the classical Von Neumann's computer model. Therefore in this paper we propose a novel memory-centric approach of computing in a RISC-modified processor core that includes on-chip memory, which can be directly accessed, without the use of general-purpose registers (GPRs) and cache memory. Considering that the proposed RISC-modified core allows for a high on-chip memory bandwidth and low latency, we examine its performances in applications with different arithmetical intensity (dense matrix multiplication, Fast Fourier Transform - FFT, Partial Differential Equations - PDEs), according to the Roofline model. The results show that the proposed memory-centric RISC-modified core outperforms the initial RISC-based MIPS processor core for problems with medium or large arithmetical intensity.

Keywords: Intelligent RAM · Memory-centric computing · RISC architecture · Processing in/near memory · Von Neumann bottleneck

1 Introduction

Standard computer systems use a processor-centric approach of computing, which means that their processing and memory resources are strictly separated [1]. In such systems the processor has the central role, executing operations on the data which need to be constantly moved from and to the main memory. In order to overcome the bottleneck problem [2], during the memory access and to

R. Silhavy et al. (eds.), *Cybernetics and Mathematics Applications in Intelligent Systems*, Advances in Intelligent Systems and Computing 574, DOI 10.1007/978-3-319-57264-2_7

approach data to the processor, today's modern computer systems usually utilize cache memory organized in several layers [3], which is basically faster, but smaller than the main memory. For instance, up to 40% of the die area in Intel processors [4,5], is occupied by caches, used solely for hiding memory latency.

The cache memory presents solely a redundant copy of the main memory data that would not be necessary if the main memory had kept up with the processor speed. Although cache memory positively affects the memory access time reduction, it demands constant movement and copying of data, which contributes to an increase in energy consumption in the system [6]. Besides that, the implementation of cache memory introduces extra hardware resources in the system and as well complex mechanisms for maintaining memory consistency.

Other latency tolerance techniques [7], include combining large caches with some form of out-of-order execution and speculation. These methods also increase the chip area and complexity. Some powerful processor architectures, like wide superscalar, VLIW (very long instruction word) and EPIC (explicitly parallel instruction computing) suffer from low utilization of resources, implementation complexity, and immature compiler technology [8,9]. On the other hand, the integration of multiple processors on a single die brings even greater demands on the memory system, increasing the number of slow off-chip memory accesses.

Contrary to the standard model of processor-centric architecture [10], some computer designers have investigated alternative approaches of memory-centric computing, which suggests integrating or placing the memory near to the processor [11]. This research have resulted with several proposals: register-less processor that performs all the operations directly with the cache memory, organized in several layers (on-chip and off-chip) [12], use of Scratchpad memory as a small high-speed on-chip memory that maps into the processors address space at a predefined address range [13], and variety of memories that include processing capabilities, known as computational RAM, intelligent RAM, processing in memory chips, intelligent memory systems [14–22] etc. The given merged memory/logic chips usually integrate on-chip DRAM memory (instead of SRAM) which allows high internal bandwidth, low latency and high power efficiency, eliminating the need for expensive, high speed inter-chip interconnects [11]. These characteristics make the smart memory chips applicable for performing computations which require high memory bandwidth and stride memory accesses, such as the fast Fourier transform, multimedia processing, network processing etc.

The aim of this paper is to propose a RISC-modified processor core, that will achieve stronger match between processing and memory elements in the system. Accordingly, we suggest a memory-centric approach that implements on-chip memory into the processor core to store and provide direct access to data. Contrary to the other memory/logic merged chips, the proposed RISC-modified processor core excludes the use of GPRs and cache memory, which significantly simplifies the memory accesses, since it avoids unnecessary copying of data and decreases the number of data movements. Actually, the proposed RISC-modified processor core implements standard instruction set, but with addressing modes that define direct access of each operand, placed into the

on-chip memory. Additionally, the proposed RISC-modified processor core operates in pipeline mode, allowing every (memory, jump or arithmetic) instruction to be completed in single tact cycle.

The rest of this paper is organized as follows: Sect. 2 presents the current state of research, discussing the methods for decreasing memory access time in processor-centric systems and presenting other alternative solutions, which use memory-centric approach of computing. Section 3 clarifies the general idea of modifying a RISC-based MIPS processor and proposing a processor in memory with memory-centric architecture. This section discusses the novel elements of the proposed RISC-modified processor core, including: memory segmentation concept, on-chip memory access, pipelining, operand addressing, instruction's formatting etc. After that, Sect. 4 presents the results of program's simulations with different arithmetical intensity (dense matrix multiplication with various sizes, FFTs with different number of points, PDEs), and provides a performance comparison of the initial MIPS processor and the proposed RISC-modified processor core. Section 5 gives a summary of the research from this paper.

2 Current State

The technological development over the last decades has caused dramatic improvements in processor performances, causing significant speed-up of processor working frequency and increased amount of instructions which can be processed in parallel in single cycle. According to the Moore's law [23], the advances in integrated circuits production technology doubled the number of transistors on chip every 18 months, which resulted in the creation of multi-core processors over the last decade. The given development of processor technology has brought performance improvements on computer systems, but not for all the types of applications. The reason for such divergence is due to the bottleneck problem between the processor and main memory (which is located out of the processor), caused by the growing disparity of memory and processor speeds. Therefore, we can say that not long ago, off-chip memory was able to supply the processor with data at an adequate rate. Today, with processor performance increasing at a rate of about 70% per year and memory latency improving by just 7% per year [1], it takes a dozens of cycles for data to travel between the processor and the main memory.

Several approaches [7], have been proposed to help alleviate this bottleneck, including branch predictor algorithms, techniques for speculative and re-order instructions execution, wider and faster memory connections and multi-level cache memory. These techniques can serve as a temporal solution, but are still not able to completely solve the problem of increased memory latency. Even the development of a powerful superscalar, VLIW and EPIC processor [9], (which are capable of executing several millions of instructions per second), didn't cause significant performance improvement as a result of the memory stalls, during the slow memory accesses. Other techniques [24], which allow the processor to execute other operations while a memory request is being processed include:

multithreading [3], pre-fetch [1], non-blocking cache [25]. Generally, the research has shown that the usage of the previous techniques contributes to reducing memory latency, but causes increased memory traffic (higher instruction rate and need for operands). As a result of the limited bandwidth of the memory interface, additional latency is caused. This introduces more intensive work with the memory resources, causing a bottleneck in the system again.

In order to provide simultaneous improvement of the memory latency and bandwidth computer architects have proposed several memory-centric approaches of integrating or placing the memory near to the processing elements. This research includes several proposals: register-less processor, Scratch-pad memory as an addition to the on-chip cache memory, chips that integrate processing and memory into the same die (32R/D, Terasys, DIVA, intelligent RAM, parallel processing RAM and DataScalar) and the active pages model which adds reconfigurable logic blocks to every virtual page in the memory [12–22]. Within these memory-centric system, the processor can be implemented as some sophisticated standard superscalar processor and may contain a vector unit, as is the case with the IRAM [18]. The integrated on-chip memory is realized as SRAM or embedded DRAM, which is mostly accessed through the processor's cache memory. Although the processing in/near memory brings latency and bandwidth improvement, still the system has to perform unnecessary copying and movement of data between the on-chip memory, caches and GPRs. Besides that, the processing speed, the on-chip memory size, and the chip cost are limited due to technological constrains and the production process. Additionally, it is an even greater challenge to develop suitable compiler support, which will recognize the program parallelism and will enable effective utilization of the internal memory bandwidth (ex. wide datapath in DIVA system [16]).

Having in mind that modern processors are lately dealing with both technical and physical limitations, while the memory capacity is constantly increasing, it seems that now is the right moment to reinvestigate the idea of placing the processor in or near to the memory in order to overcome their speed difference [22]. The latest research in that field is held from the Hewlett Packard international company, which suggests novel computer architecture - the Machine [26], that utilizes non-volatile memory as a true DRAM replacement. Our research will continue into the direction of developing novel memory-centric architecture similar to PERL, which will provide direct access to the memory that is integrated into the processor chip (without the use of GPRs and cache memory).

3 Design of RISC-modified Memory-Centric Processor Architecture

In order to design the proposed RISC-modified memory-centric processor core, we decided to use a MIPS implementation of a single-cycle pipelined RISC architecture, given by D. A. Patterson and J. L. Hennessy in [1]. The appropriate MIPS processor is characterized with: fix-length straightforward decoded

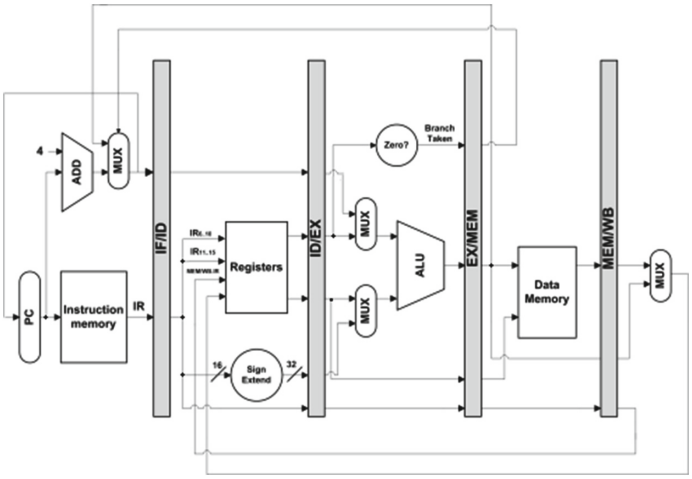


Fig. 1. Pipelined MIPS processor architecture

instruction format, memory accesses limited to load and store instructions, hardwired control unit, large GPR file and pipeline operation in five stages (fetch, decode, execute, memory access and write back), as shown on Fig. 1.

In our proposal, we are augmenting and extending the original MIPS processor architecture in order to involve some hardware improvements that will provide easier access and manipulation with memory data. In fact, the suggested RISC-modified processor core implements separated on-chip instruction and data memories, allowing direct access to the on-chip data memory with standard instructions. This approach avoids unnecessary copying of data and decreases the number of data movements, because the GPRs and the cache memory are excluded from the memory hierarchy. Besides that, the number of pipeline stages is reduced by one, excluding the MEM phase, which in the MIPS processor is responsible to transfer (load or store) data between the cache memory and the general-purpose registers. As can be seen in Fig. 2, the proposed memory-centric processor directly addresses two sources and one result operand in each instruction. These addresses are applied to a specialized memory address selector unit that is responsible to select the data from the on-chip data memory.

In our approach, the memory system is observed as a set of blocks, quite similar to the virtual memory concept. According to that, the RISC-modified memory-centric processor core is associated with one instruction block and up to three data memory blocks at a given moment: two for the sources, and one for the result. The selection of the proper blocks is performed via special SetMBS instruction that affects the content of the memory block selector units (for op1, op2 and result). Once the memory block selectors are set, the RISC-modified memory-centric processor core performs page table lookup and selects the appropriate physical blocks from the on-chip data memory, as shown on Fig. 3.

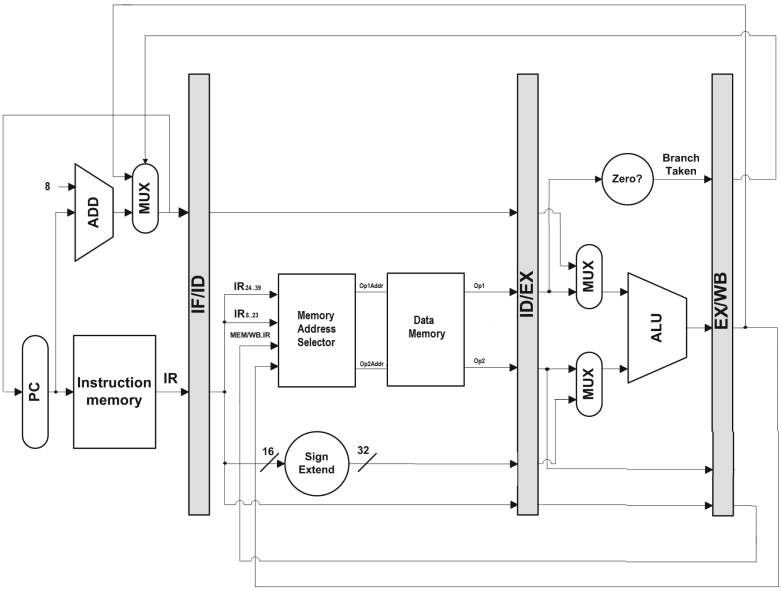


Fig. 2. Pipelined RISC-modified memory-centric architecture

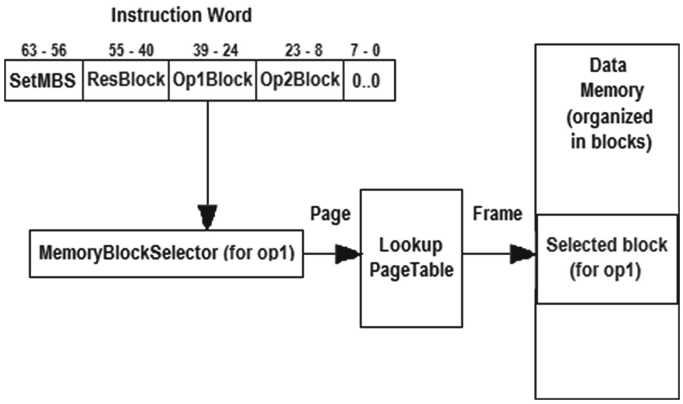


Fig. 3. Block selection for the first operand from the on-chip data memory of the proposed RISC-modified memory-centric processor core

The memory segmentation concept results with some simplifications in instructions formatting, since the data memory addressing is performed in two steps. First the memory blocks are selected via special SetMBS instruction, and then the operands are specified as address offsets to the given blocks (instead of complete addresses) into the following instructions. As shown on Fig. 4, each instruction is 64-bits wide, typically specifying the operation code and the operands, given as direct address offsets or immediate values. Additionally,

Format	Bits					
	63-56	55-40	39-24	23-8	7 - 3	2 - 0
M-type	op. code	Result Offset	Op1 Offset	Op2 Offset	Shift Amount	Base Address Mode
I-type	op. code	Result Offset/ Imm16	Op1 Offset/ Imm16	Op2 Offset /Imm16	Shift Amount	Base Address Mode
J-type	op. code	Result Offset	DestImm32		Zeros	Base Address Mode

Fig. 4. Instruction formats of RISC-modified memory-centric processor core

the instructions have two extra fields for: shift amount and base address mode. The shift amount field specifies a constant value for performing a shift operation of the second operand, during the execution of some other ALU operation. As a result, the second operand is also called a flexible second operand.

Besides direct and immediate addressing modes, the proposed RISC-modified memory-centric processor core supports base addressing, which use is specified with the base address mode instruction field. This three-bit field indicates which of the operands (op1, op2 or res) use base addressing. In order to provide this address mode, the RISC-modified memory-centric processor core implements three separate base address units (for op1, op2 and res), and a specialized instruction SetBR (with similar syntax to SetMBS) that is intended to set the initial address values into the base address units. If some operand has its base address mode field bit set to one, then its address is computed by addition of its base address unit value and its address offset specified into the instruction, as shown on Fig. 5.

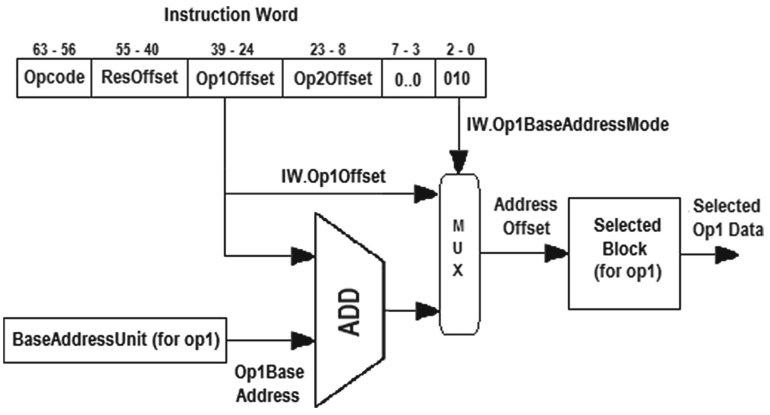


Fig. 5. Selection of the first operand from the appropriate on-chip data memory block of the proposed RISC-modified processor core

Basically, the instruction set architecture of the proposed memory-centric processor core is RISC-like, and includes three types of instructions (-type, I-type and -type), organized in four different groups. The arithmetical-logical group of instructions consists of: addition with overflow detection, subtraction, multiplication, integer division (div), modulo division (mod) and AND, OR, XOR and NOT logical bit-wise operations. The shifting group of instruction consists of: left and right logical and arithmetical shifts and rotations. The branching group consists of instructions for conditional and unconditional change of the program flow. The last group is the auxiliary group, consisting of control instructions for program termination and system halt, some SET instructions that update the value of the base address units and the memory block selectors, memory instructions for loading of 8-bit, half word or 32-bit constants and IN/OUT instructions for communication with external hardware units.

4 Performance Evaluation

The performance evaluation is presented by analysing the behaviour of the initial MIPS processor and the proposed RISC-modified memory-centric processor, while executing programs with different arithmetical intensity. We consider that the proposed RISC-modified memory-centric processor includes on-chip memory that is segmented into 64 KB physical blocks and is with total capacity that equals to the amount of on-chip cache memory into the MIPS processor (128 KB L1 and 2 M L2 cache). The both processors make use of pipelining, allowing each instruction to be executed in single cycle, excluding the MIPS's memory instructions that generate a miss in L1 or L2 cache. In that case, we consider that the MIPS processor use associative cache with with 128-word blocks, which L1 hit time is 1 cycle, L2 hit time is 21 cycles and L2 miss penalty is 272 cycles.

The performance analysis of the suggested RISC-modified memory-centric processor was realized by an instruction-level compiler and simulator, which were specially developed for that purpose. On the other hand, MIPS processor behaviour (GPRs and cache change) was simulated with the well known MARS simulator [27]. This study includes simulation of three different groups of algorithms, which according to the Roofline model are characterized with different arithmetical intensity. By comparing the execution time of the given programs for the both processors we can determine the area of application, where the suggested RISC-modified memory-centric processor achieves improvement in speed.

The first program is characterized with the largest arithmetical intensity and includes multiplication of dense matrices with different dimensions (8×8, 16×16, 32×32, 64×64, 128×128, accordingly). This program executes many repetitive arithmetical operations over a huge amount of data (768 B, 3072 B, 12288 B, 49152 B, 196608 B, accordingly). The results shown on Fig. 6 present the percentual speedup in execution time for matrix multiplications with different sizes (8×8, 16×16, 32×32, 64×64, 128×128, accordingly) for the proposed RISC-modified memory-centric processor, in comparison with the MIPS processor. According to that, we can notice that as the problem size is increased,

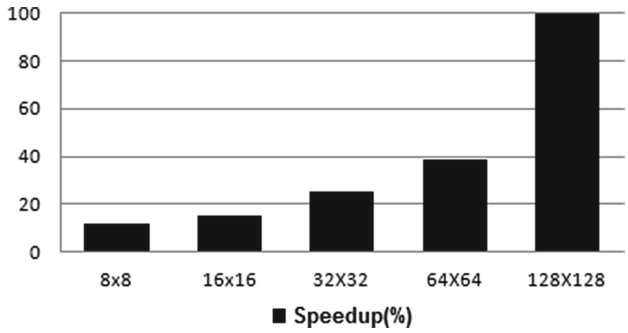


Fig. 6. Matrix multiplication speedup of the proposed RISC-modified processor

the proposed RISC-modified memory-centric processor accelerates the program execution by 12%, 15%, 25%, 39%, 99%, accordingly. This is due to the constant increase of number of misses in the MIPS processor's cache, which introduce further delay.

The second program is characterized with medium arithmetical intensity and represents a spectral method problem. This program implements FFT calculations, according to the well known CooleyTukey algorithm. The results shown on Fig. 7 present the percentual speedup in execution time of FFT computations for different number of points: 8, 16, 32, 64, 128, 256, 512, 1024, 2048, 4096, 8192 and 16384, achieved by the proposed RISC-modified memory-centric processor, in comparison with the MIPS processor. According to that, we can notice that there is a decrease, and then an increase of the total speedup which is achieved by the proposed RISC-modified memory-centric processor. This is because the proposed RISC-modified memory-centric processor raises the number of instructions which are executed by increasing the extent of the problem for FFT calculations, in relation to MIPS. Still, as a result of the increased amount of cache misses that occur into the MIPS processor when calculating FFT with over 500 points, the proposed RISC-modified memory-centric processor instead of a decrease, introduces an increase in speed.

The third program is characterized with small arithmetical intensity and implements PDE solver, which is an important problem from the structure grids area. This program includes small data set (128B) and performs repetitive operations over this data set by calling two functions. As a result, the number of MIPS's cache misses is only 2, causing two blocks to be transferred from the main memory to the two-level cache memory. The results shown on Fig. 8 present the PDE execution time, measured in cycles, for the both processors. According to the results we can notice that the suggested RISC-modified memory-centric processor slows down the program execution time by 14,7%, in comparison with MIPS. This is due to the increased amount of instructions (ex. SETBR) which is introduced by the proposed RISC-modified processor. In the future we could improve this, by optimizing the RISC-modified processor's compiler.

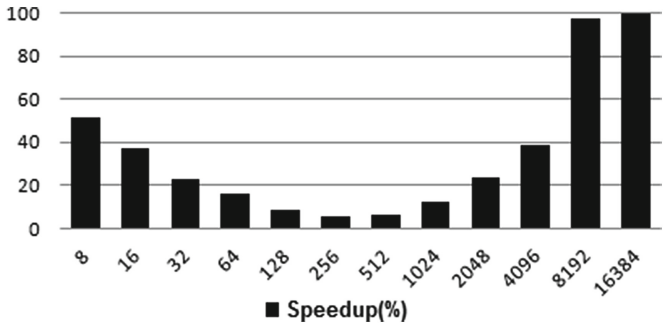


Fig. 7. FFT speedup of the proposed RISC-modified processor

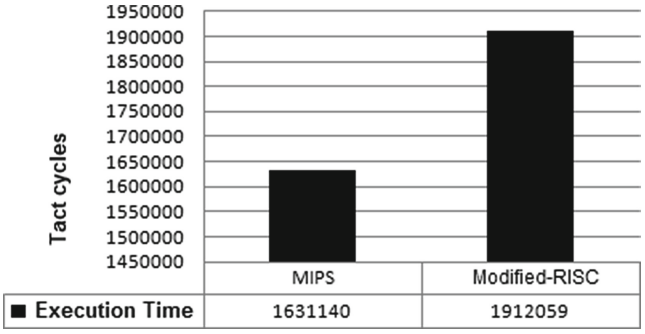


Fig. 8. PDE execution time comparison between MIPS and modified-RISC processor

5 Conclusion

In this paper, we have proposed a memory-centric processor core that is based on a standard MIPS implementation of RISC architecture, which is further improved with several hardware accelerators, reorganized memory architecture, and adjusted instruction set for direct operation with the on-chip data memory. The basic modification of the initial MIPS processor is presented by removal of GPRs and cache memory, and addition of separate on-chip instruction and data memory into the processing core. This memory-centric approach avoids unnecessary copying of data between GPRs and cache memory, and thus decreases the number of data movements, and allows 4-stage pipeline processing with direct access to the on-chip memory.

In order to estimate the performances of the proposed RISC-modified memory-centric processor core, we have designed a VHDL model of the proposed processor in Virtex7 VC709 FPGA, and a dedicated compiler and instruction-level simulator as well. The previously given performance estimation shows that the proposed RISC-modified memory-centric processor achieves significant improvement while executing programs with large arithmetical intensity and

partial improvement while executing programs with medium arithmetical intensity, in relation to a standard MIPS processor. On the other hand, the proposed RISC-modified memory-centric processor does not bring improvements while executing programs with small arithmetical intensity, so in these cases it is better to use MIPS processor. Considering that the analysed programs were sequential, in a very near future we would explore the possibility of designing a homogeneous multi-core processor and parallelising the examined programs for that extension.

References

1. Patterson, D.A., Hennessy, J.L.: Computer Organization and Design: The Hardware/Software Interface. Elsevier, USA (2014)
2. Wulf, W.A., McKee, S.A.: Hitting the memory wall: implications of the obvious. ACM SIGARCH Comput. Archit. News **23**(1), 20–24 (1995)
3. Hennessy, J.L., Patterson, D.A.: Computer Architecture: A Quantitative Approach. Morgan Kaufmann, USA (2012)
4. Borkar, S., Chien, A.A.: The future of microprocessors. Commun. ACM **54**(5), 67–77 (2011)
5. Intel Corporation: New Microarchitecture for 4th Gen. Intel Core Processor Platforms. Product Brief (2013)
6. Carvalho, C.: The gap between processor and memory speeds. In: ICCA 2002, Portugal (2002)
7. Machanick, P.: Approaches to addressing the memory wall. Technical report. University of Queensland Brisbane, Australia (2002)
8. Smotherman, M.: Understanding EPIC architectures and implementations. In: ACM Southeast Conference, Atlanta (2002)
9. Jakimovska, D., et al.: Modern processor architectures overview. In: XVIII ICEST Conference, Bulgaria, pp. 239–242 (2012)
10. Eigenmann, R., et al.: Von Neumann computers. Wiley Encyclopedia Electr. Electron. Eng. **23**, 387–400 (1998)
11. Saulsbury, A., Pong, F., Nowatzyk, A.: Missing the memory wall: the case for processor/memory integration. In: 23rd International Symposium on Computer Architecture, USA (1996)
12. Suresh, P.: PERL - a register-less processor. Ph.D. Thesis. Indian Institute of Technology, Kanpur (2004)
13. Wang, P.: Designing scratchpad memory architecture with emerging STT-RAM memory technologies. In: IEEE International Symposium on Circuits and Systems (2013)
14. Cojocaru, C.: Computational RAM: implementation and bit-parallel architecture. Master Thesis. Carletorn University, Ottawa (1995)
15. Tsubota, H., et al.: The M32R/D, a 32b RISC microprocessor with 16Mb embedded DRAM. Technical report (1996)
16. Draper, J., et al.: A prototype processing-in-memory (PIM) chip for the data-intensive architecture (DIVA) system. J. VLSI Sig. Process. Syst. **40**(1), 73–84 (2005)
17. Gokhale, M., Holmes, B., Jobst, K.: Processing in memory: the Terasys massively parallel PIM array. IEEE Comput. J. **28**(4), 23–31 (1995)

18. Gebis, J., et al.: VIRAM1: a mediaoriented vector processor with embedded DRAM. In: 41st Design Automation Student Design Contest, San Diego (2004)
19. Murakami, K., Shirakawa, S., Miyajima, H.: Parallel processing RAM chip with 256 Mb DRAM and quad processors. In: Solid-State Circuits Conference (1997)
20. Kaxiras, S., Burger, D., Goodman, R.: DataScalar: a memory-centric approach to computing. J. Syst. Archit. **45**, 1001–1022 (1999)
21. Oskin, M., Chong, F.T., Sherwood, T.: Active pages a computation model for intelligent memory. In: 25th Annual International Symposium on Computer Architecture, pp. 192–203 (1998)
22. Azarkhish, E., Rossi, D., Loi, I., Benini, L.: Design and evaluation of a processing-in-memory architecture for the smart memory cube. In: 29th International Conference Architecture of Computing Systems, Germany (2016)
23. Moore's Law is dead - long live Moore's Law. IEEE Spectr. Mag. (2015)
24. Bakshi, A., et al.: Memory latency: to tolerate or to reduce? In: 12th Symposium on Computer Architecture and High Performance Computing (2000)
25. Li, S., et al.: Performance impacts of non-blocking caches in out-of-order processors. Technical paper (2011)
26. Hewlett Packard Labs: The Machine: The future of technology. Technical Paper (2016)
27. Vollmar, K., Sanderson, P.: MARS: an education-oriented MIPS assembly language simulator. In: 37th SIGCSE Technical Symposium on Computer Science Education, USA (2007)

CARIC: A Novel Modeling of Combinatorial Approach for Radiological Image Compression

M. Lakshminarayana[1(✉)] and Mrinal Sarvagya[2]

[1] Department of ECE, Visvesvaraya Technological University, Belgaum, India
lakshminarayana.m.2015@ieee.org
[2] School of ECE, REVA University, Bangalore, India
mrinalsarvagya@gmail.com

Abstract. The contribution of several compression algorithms plays a significant role in minimizing the size of multiple radiological images from last decade. However, a closer look into existing work will show that there is a big trade-off between compression performance and data quality during the reconstruction process. We review the existing research work being carried out and briefs such problems and trade-off. This paper presents a framework called as CARIC (Combinatorial Approach for Radiological Image Compression) that uses a combinatorial approach of both lossy and lossless compression schemes unique in any radiological image. Using maximum numbers and modalities of different radiological images, we also compare CARIC with some recent and relevant work of compression to find that CARIC offers better image compression ratio along with a great balance among quality of the reconstructed image and faster response time.

Keywords: Compressive sensing · Compression ratio · Discrete Wavelet Transform · Medical images · Peak Signal-to-Noise Ratio · Radiological image

1 Introduction

Radiological imaging has a huge effect on diagnosis of various diseases and also assists in the surgical process. Also, the storage of significant radiological data and transmission is also a vital dilemma due to its insufficient memory as well as the limited bandwidth. Existing radiological imaging modalities produces a digital form of diagnosed medical images [1]. There occurs a necessity for image compression for storage and data transmission in a communicating channel. The traditional lossless image compression methods provide large compression rates and a significant quality of the medical image. The most important a problem arises in teleradiology is the difficulty of transferring huge volume of medical data in a limited bandwidth [2]. Image compression methods have been increased the feasibility by minimizing the bandwidth constraint along with cost-effective delivery of medical imaging for primary diagnosis [3]. In medicine, it is essential to have a good perceptual quality of medical data in the diagnostically critical area. In other words, two different image compression methods should apply to both region and non-region of sectors. The basic idea is to accomplish a superior image quality

R. Silhavy et al. (eds.), *Cybernetics and Mathematics Applications in Intelligent Systems*,
Advances in Intelligent Systems and Computing 574, DOI 10.1007/978-3-319-57264-2_8

in diagnostically relevant areas, whereas encoding Non-ROI area's general practitioner be able to visualize the important regions in the original image.

The overall idea of proposed article is to reconstruct a picture with the help of measuring a minimum number of samples using compressive sensing with no loss in the region of interest. The projected manuscript discuss a combinatorial modeling of the traditional lossless compression scheme in a critical area of the medical image with better compression rate, compressive sensing method in other parts of medical images. Section 2 discusses the existing research work towards compression followed by problem identification in Sect. 3. Segment Sect. 4 talks about proposed methodology followed by algorithm implementation in Sect. 5. Comparative analysis of finished result has discussed under Sect. 6 followed by conclusion in Sect. 7.

2 Related Work

Since the development of medical imaging schemes, numerous researchers have been endeavored to implement an image compression techniques. The primary concern was a storage and transmission of a vast number of medical images obtained from many imaging devices. In addition to this, it is imperatively expected to get highest possible data compression speed and improved image quality. Our previous review work has previously deliberated several existing schemes to carry out compressive sensing over multimedia communication in medical image processing [4] through an illustrative explanation of the research gap. This division, we evaluate specific relevant investigation work in the direction of hybrid compression using compressive sensing given more complications.

Perumal et al. [5] proposed the performance analysis of various medical images using hybrid of both wavelets transform and neural network-based back propagation scheme for achieving an efficient image compression and reconstruction. Yadav et al. [6] presented a scheme to improve the compression rate along with reducing the computational complexity and excellently removed image quality using compressive sensing method. A significant enhancement in medical image quality and the better compression ratio is achieved using a hybrid technique such as Log-discrete wavelets transform and logarithmic number system is presented by Ibraheem et al. [7]. Raju et al. [8] demonstrated a hybrid of both lossy and lossless traditional compression techniques to acquire a finer quality of decompressed image and get a better compression ratio. Wang et al. [9] proposed a hybrid compressive sensing scheme; it successfully incorporates both $\ell 0$-norm and $\ell 1$-normalization of medical image gradient using presenting a threshold to achieve better image quality. The similar direction of work is carried out by Bhattacharjee et al. [10] to achieve better compression ratio and restoration of therapeutic images utilizing block compressive sensing method. Khalid et al. [11] have studied the conceptions of CS and its requirements on models using deterministic CS matrix designed with BCH code vectors. The authors also performed a real time based medical image reconstruction using compressed sensing method. Malczewski [12] have presented a new PET (Positron Emission Tomography)-based image compression and reconstructing method using compressive sensing scheme. It improves the resolution of clinical PET scanners. Chen et al. [13]

encompass an innovative real-time system of multipliers based optimization algorithm to reverse the linear framework containing two normalization terms to reconstructs ultrasound medical images. The study toward the compressed sensing is done by Nan et al. [14] for a biomedical image to achieve better-restored image and diminish the computational difficulty of compressive sensing technique. The reconstruction of MRI (Magnetic Resonance Imaging) image using the f-MRI technique based on compressive sensing scheme is proposed by Hirabayashi et al. [15] where the authors have carried out a slice similarity using sparsity matrix. Chernyakova et al. [16] have investigated an ultrasound device using sub-Nyquist sampling and dispensation where authors presented compressive sensing methods to improve the compression ratio, device size, less power consumption and also computational cost of the system. Another unique application of CS scheme was implemented by Imtiaz et al. [17] where the compression scheme his adopted in wearable wireless sensor nodes. This system uses less power consumption in wireless sensor nodes. Liu et al. [18] have proposed a novel framework to recovery a biomedical data using CS system where the authors have used a real life biomedical signal for the analysis purposes. The newly implemented technique accomplishes enhanced reconstruction precision performance regarding *l1* and *l2* errors. The hybrid compression scheme combines the DCT (Discrete Cosine Transform) and Huffman encoding technique to perform medical image compression. It also reduces the bandwidth communication channel as seen in the work of Padmavati et al. [19]. Lee et al. [20] have proposed a hybrid compression scheme for hyperspectral images based on PCA (Principle Component Analysis). The experimental outcome provides an outstanding compression ratio and superior reconstructed image with excellent accuracy than traditional image compression techniques.

Hence, we can observe that where many compression techniques should present in recent years for evolving up with novel approaches of compressive sensing. All the methods discussed have a significant favorable point of study and approve while related with restrictions and constraints too. The issues relating to current studies are discussed concisely in a subsequent segment.

3 Problem Description

This section briefs about the issues has identified after reviewing the existing research techniques related to compressive sensing. A closer look at existing studies shows that compressive sensing is used more for compression purpose on critical image section or the overall image even after knowing that it is lossy compression scheme. Hence, actual usage of compressive sensing is not for diagnosis of critical or complex disease condition from radiological images. It has found that focus on faster response time was totally ignored while using compressive sensing. Delay in response time directly represents that existing techniques uses higher recursive techniques resulting in less image quality and more algorithm time. These problems have addressed in proposed model. The next section discusses the adopted research methodology to overcome the problems.

4 Proposed Methodology

Our prior study has presented a simple and yet robust framework of compression towards medical images using compressive sensing [21–23]. This paper has presented a combinatorial approach CARIC that combines both lossy and lossless compression techniques to perform cost-effective compression of radiological images. The core objectives of proposed study are (i) To present a non-iterative method for compressing the radiological image, (ii) To introduce a modeling of the combinatorial approach to accomplish higher compression performance. CARIC has designed on the concept of implementing lossy compression scheme of compressive sensing on non-ROI while it applies lossless compression schemes on ROI portion of the image. This methodology has undertaken to ensure a better equilibrium between compression performance and image quality of output signal. The schematic representation of the proposed CARIC is as shown in Fig. 1.

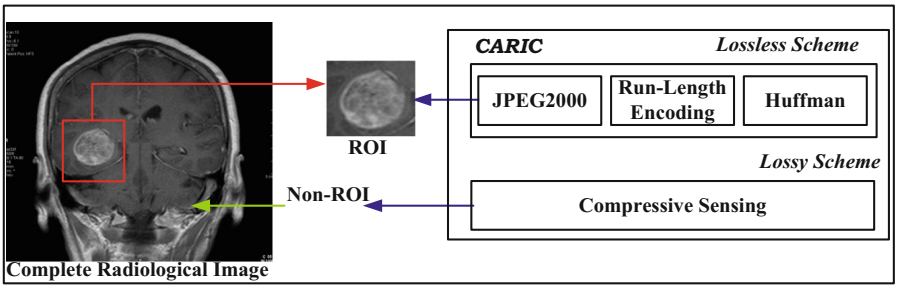


Fig. 1. Schematic representation of CARIC

5 Algorithm Implementation

The algorithm design of the CARIC is carried out in three stages i.e. (i) preparing the input radiological images, (ii) applying compression techniques, and (iii) performing decompression techniques. The Matlab-based *imcrop* method has used for extraction region of interest. The ROI is selected based on the visual analysis by the user who is assumed to be a physician. The CARIC algorithm considers input of I_1 (region of interest) and I_2 (non-region of interest), L (length of the coefficient for Compressed Sensing), m being many rows, m being many columns, δ is ratio of image compression. The algorithm defines a length of factor L as 150 for compressive sensing that has followed by using a new attribute σ. This attribute selects the coefficient randomly by rows m (*Line-3*). An attribute ϕ (*Line-2*) has generated by product of α and ε_1, where $\varepsilon_1 = \sum \left[\theta(I).\psi.\psi^T\right]$. The parameter ψ is a 2D coefficient that is equivalent to m rows, while θ is the sparse matrix of image I. This process will generate an ultimate outcome of ϕ (*Line-2*). Initialization of bits is done by the unsigned integer of 8 bits from ϕ as well as implying JPEG2000 standard, and the ratio of compression ratio δ. An attribute of compressive sensing ϕ is again generated once again by generated bits as well as maximizing its precision to double. Finally, a computation of output (or reconstructed)

image *rec* (*Line-4*) is carried out by applying compressive sensing function on ϕ, α and *m* rows. ROI encoding should initiate by evaluating error *Err* (*Line-6*). A new parameter *mx* is used for representing non-iterative elements in image *I* (*Line-6*). The resultant *o* should use for subjecting it to next process of run-length encoding to accomplish Δ (data) accompanied by Huffman encoding to enhance the quality of a radiological image (*Line-7*). In the algorithm f_1, f_2, f_3 is a function for JPEG2000, compressive sensing, and Run-Length Encoding (*Line-8*). The important thing to understand is ROI image was subjected to a lossless compression scheme (*JPEG2000, Run-Length Encoding, and Huffman*) while non-ROI image has subjected to lossy compression scheme of compressive sensing.

Combinatorial Algorithm for Radiological Image Compression

Input: *I_1(region of interest), I_2 (non-region of interest), L (length of the coefficient for Compressed Sensing), m (number of rows), n (number of columns), δ (ratio of compression)*

Output: *I_3 (Reconstructed image)*

Start

1. $I = [I_1\ I_2]$
2. $\phi \rightarrow ((\ \phi \rightarrow 255*[(w)/max(w)]), f_1, \delta)$, *where* $w = \alpha * \varepsilon_1$ & $\alpha \rightarrow rand(L, m)$
3. ***For*** $p=1: m$
4. $rec \rightarrow f_2(\phi, \alpha, m)$
5. ***End***
6. $o \rightarrow (Err(mx))$
7. $\Delta \rightarrow f_3\ (o)$
8. $I_3 \rightarrow [b\ d] \rightarrow hc(\Delta)$

End

6 Results and Discussion

The proposed scheme uses standard medical datasets of Cornell University [24]. For better forms of assessment, the assessment of the proposed study is carried out considering multiple types of radiological images for strong validation of outcomes. The outcome was accessed using Peak Signal-to-Noise Ratio (PSNR) and algorithm processing time. The resultant of the study has compared with the most recent work being carried out by Kathirvalavakumar [25], where author have used Self-Organizing Map (SOM), Discrete Wavelet Transform, and Arithmetic encoding. We do minor editing in the approach by using Self-Organizing Map and Discrete Wavelet Transform for encoding non-region of interest and Arithmetic encoding for encoding region-of-interest. This amendment is carried out to compare our combinatorial approach. Similarly, we also found that Kourav [26] and Kumar et al. [27] have used Wavelet Difference

Reduction and Arithmetic method for compressing the medical signal. We apply similar amendment in this approach by using Wavelet Difference Reduction to be used for encoding non-region of interest and Arithmetic method for encoding region of interest. Table 1 highlights the changes has made to suit the existing approach should compare with proposed combinatorial approach to assessing the compression performance.

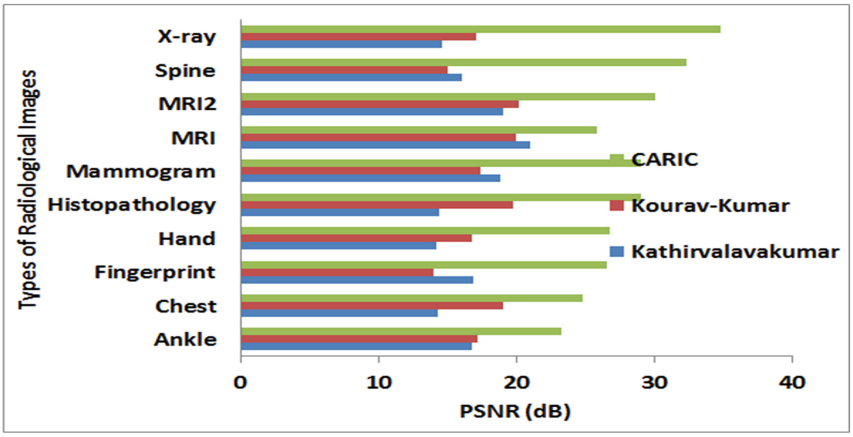
Table 1. Comparative performance analysis

Scheme	Adopted encoding scheme	
	Non-ROI	ROI
Kathirvalavakumar [25]	Self-organizing map, discrete wavelet	Arithmetic
Kourav [26], Kumar [27]	Wavelet difference reduction	Arithmetic
CARIC	Compressive sensing	JPEG2000, Huffman, run-length encoding

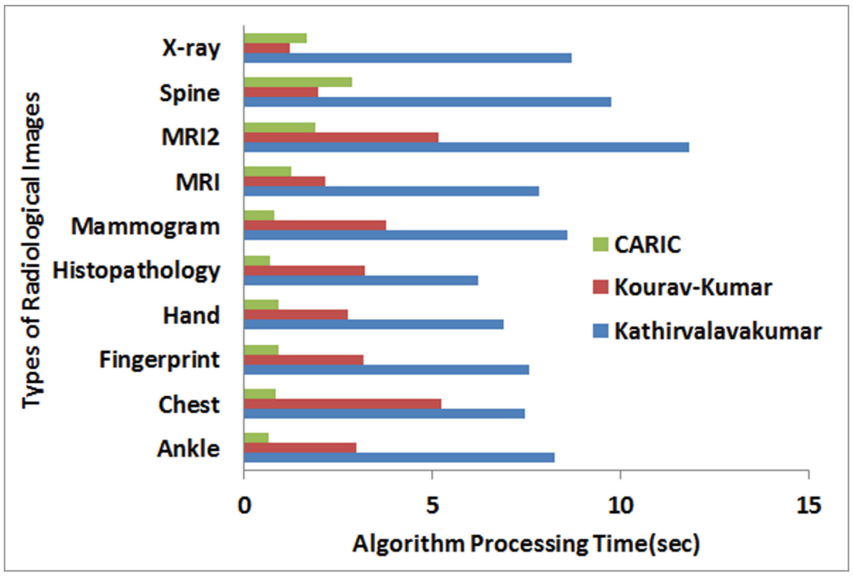
Figure 2 highlights the similar outcome of proposed system CARIC with the existing system with PSNR (Fig. 2(a)) and Algorithm processing time (Fig. 2(b)). The PSNR outcome accomplishment of CARIC has found in the range of 23–35 dB for various forms of radiological images. The PSNR results of Kourav [26] and Kumar et al. [27] has found in the range of 10–20 dB, and that of Kathirvalavakumar [25] is found to be 15–20 dB. A similar analysis of the algorithm processing time is found to be lower for proposed system CARIC within a range of 0.15 s–1.5 s. The processing time for Kourav [26] and Kumar [27] is found within the range of 0.9–5.2 s while that of Kathirvalavakumar [25] is in the range of 9–13.4 s.

The prime reasons behind the outcomes are as follows: the approach of Kathirvalavakumar [25] uses highly recursive operations of training using Self-Organizing map that results in superior version of encoding performance in terms of PSNR, but it is not able to offer better processing time as its clusters has constructed from higher amount of data, which negatively affects processing time consumption. The finished result of the study exhibits that proposed system (CARIC) offers highly enhanced the balance between retention of maximum image quality and the compression performance for the reconstructed image. The algorithm has also proven with faster response time evaluated on various processors Core i3-i7 with the standard windows machine, and thereby it shows its direct applicability over medical image compression.

Figure 3 shows that offered CARIC offer higher compression performance in comparison to the existing techniques. The outcome of this analysis should accomplish after using ten different modalities of datasets i.e. Spine, Fingerprint, Chest, MRI2, Ankle, Mammogram, MRI, Histopathology, Hand, X-ray. Our outcome on higher ranges of radiological images shows superior compression ratio.



(a)



(b)

Fig. 2. Comparative analysis of CARIC (a) PSNR, (b) Processing time

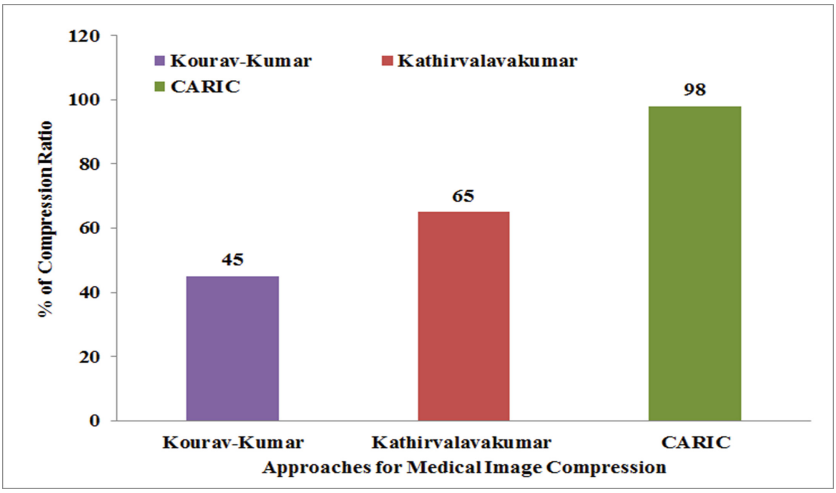


Fig. 3. Comparative study of CARIC method (Compression ratio)

7 Conclusion

The prime agenda of this paper is to check the applicability of the combinatorial approaches of compression over various radiological images. It has seen that compressive sensing has an excellent reputation for offering optimal compression of different signals but usage towards vast forms of radiological images are not seen much in existing studies. This principle also provides network-based efficiency regarding channel capacity and memory to support faster response time. A significant research gap has observed as less research work has emphasized over accomplishing efficiency towards computational potential and was more stressed on using compression. The proposed paper makes a balanced use of both lossy and lossless compression scheme which has directed for using over the entire radiological image but in a very discrete manner. The proposed CARIC applies lossy compression scheme e.g. compressive sensing over non-ROI section which is maximum in number while lossless compression scheme over ROI section which has required for higher content of information. The study outcome of the proposed CARIC was found to perform superior in contrast to existing algorithms e.g. self-organizing map, discrete wavelet transformation, Arithmetic and Huffman encoding on compression ratio, processing time, and PSNR.

References

1. Majumdar, A.: Compressed Sensing for Magnetic Resonance Image Reconstruction. Cambridge University Press, Cambridge (2015). Computers
2. Carmi, A.Y., Mihaylova, L., Godsill, S.J.: Compressed Sensing & Sparse Filtering. Springer Science & Business Media. Technology & Engineering, Heidelberg (2013)

3. Boche, H., Calderbank, R., Kutyniok, G., Vybíral, J.: Compressed Sensing and Its Applications. Birkhäuser, Boston (2015). Mathematics
4. Lakshminarayana, M., Sarvagya, M.: Scaling the effectiveness of existing compressive sensing in multimedia contents. Int. J. Comput. Appl. **115**(9), 16–26 (2015)
5. Perumal, B., Rajasekaran, M.P.: A hybrid discrete wavelet transform with neural network back propagation approach for efficient medical image compression. In: International Conference on Emerging Trends in Engineering, Technology and Science (ICETETS), Pudukkottai, pp. 1–5 (2016)
6. Yadav, V., Verma, M., Kaushik, V.D.: A hybrid image compression technique for medical images. In: International Conference on Computational Intelligence and Communication Networks (CICN), Jabalpur, pp. 222–227 (2015)
7. Ibraheem, M.S., Ahmed, S.Z., Hachicha, K., Hochberg, S., Garda, P.: Medical images compression with clinical diagnostic quality using logarithmic DWT. In: IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI), Las Vegas, NV, pp. 402–405 (2016)
8. Raju, C.S., Reddy, D.V.R.K.: On compression characteristics of white band and narrow band images using hybrid DCT and DWT. In: 2nd International Conference on Electronics and Communication Systems (ICECS), Coimbatore, pp. 827–830 (2015)
9. Wang, Y., Liang, D., Chang, Y., Ying, L.: A hybrid total-variation minimization approach to compressed sensing. In: 9th IEEE International Symposium on Biomedical Imaging (ISBI), Barcelona, pp. 74–77 (2012)
10. Bhattacharjee, S., Choudhury, S.K., Das, S., Pramanik, A.: DPCM block-based Compressed sensing with frequency domain filtering and Lempel-Ziv-Welch compression. In: International Conference on Advances in Computing, Communications and Informatics (ICACCI), Kochi, pp. 1244–1249 (2015)
11. Khalid, S., Khan, S.: Application of compressed sensing on images via BCH measurement matrices. In: International Conference on Robotics and Emerging Allied Technologies in engineering (iCREATE), Islamabad, pp. 78–81 (2014)
12. Malczewski, K.: PET image reconstruction using compressed sensing. In: Proceedings of Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA), Poznan, pp. 176–181 (2013)
13. Chen, Z., Basarab, A., Kouamé, D.: Reconstruction of enhanced ultrasound images from compressed measurements using simultaneous direction method of multipliers. IEEE Trans. Ultrason. Ferroelectr. Freq. Control **63**(10), 1525–1534 (2016)
14. Nan, Y., Yi, Z., Bingxia, C.: Review of compressed sensing for biomedical imaging. In: 7th International Conference on Information Technology in Medicine and Education (ITME), Huangshan, pp. 225–228 (2015)
15. Hirabayashi, A., Inamuro, N., Mimura, K., Kurihara, T., Homma, T.: Compressed sensing MRI using sparsity induced from adjacent slice similarity. In: International Conference on Sampling Theory and Applications (SampTA), Washington DC, pp. 287–291 (2015)
16. Chernyakova, T., Eldar, Y.C.: Fourier-domain beamforming: the path to compressed ultrasound imaging. IEEE Trans. Ultrason. Ferroelectr. Freq. Control **61**(8), 1252–1267 (2014)
17. Imtiaz, S.A., Casson, A.J., Rodriguez-Villegas, E.: Compression in wearable sensor nodes: impacts of node topology. IEEE Trans. Biomed. Eng. **61**(4), 1080–1090 (2014)
18. Liu, Y., De Vos, M., Gligorijevic, I., Matic, V., Li, Y., Huffel, S.V.: Multi-structural signal recovery for biomedical compressive sensing. IEEE Trans. Biomed. Eng. **60**(10), 2794–2805 (2013)

19. Padmavati, S., Mesharam, V.: DCT combined with fractal quadtree decomposition and Huffman coding for image compression. In: International Conference on Condition Assessment Techniques in Electrical Systems (CATCON), Bangalore, pp. 28–33 (2015)
20. Lee, C., Youn, S., Jeong, T., Lee, E., Serra-Sagristà, J.: Hybrid compression of hyper-spectral images based on PCA with pre-encoding discriminant information. IEEE Geosci. Remote Sens. Lett. **12**(7), 1491–1495 (2015)
21. Lakshminarayana, M., Sarvagya, M.: Algorithm to balance compression and signal quality using novel compressive sensing in medical images. In: Silhavy, R., Senkerik, R., Oplatkova, Z.K., Silhavy, P., Prokopova, Z. (eds.) Software Engineering Perspectives and Application in Intelligent Systems. AISC, vol. 465, pp. 317–327. Springer, Cham (2016). doi: 10.1007/978-3-319-33622-0_29
22. Lakshminarayana, M., Sarvagya, M.: Random sample measurement and reconstruction of medical image signal using compressive sensing. In: IEEE-International Conference on Computing and Network Communications (CoCoNet), Trivandrum, pp. 255–262 (2015)
23. Lakshminarayana, M., Sarvagya, M.: Lossless compression of medical image to overcome network congestion constraints. In: Shetty, N.R., Prasad, N.H., Nalini, N. (eds.) Emerging Research in Computing, Information, Communication and Applications, ERCICA 2015, vol. 1, pp. 305–311. Springer, New Delhi (2015). doi:10.1007/978-81-322-2550-8_30
24. "Finding Articles, Databases and Images". Cornell University Library. https://www.library.cornell.edu/research/introduction/articles. Accessed 3rd Dec 2016)
25. Kathirvalavakumar, T., Ponmalar, E.: Self-organizing map and wavelet based image compression. Int. J. Mach. Learn. Cybern. **4**(4), 319–326 (2013). doi:10.1007/s13042-012-0099-3. Springer
26. Kourav, A., Sharma, A.: Comparative analysis of wavelet transforms algorithms for image compression. In: IEEE-International Conference on Communication and Signal Processing (2014)
27. Kumar, R., Kumar, A., Singh, G.K.: Electrocardiogram signal compression using singular coefficient truncation and wavelet coefficient coding. IET Sci. Meas. Technol. **10**(4), 266–274 (2016). IEEE

Torque Characteristics of Antagonistic Pneumatic Muscle Actuator with an Oval Cam

Mária Tóthová(✉) and Alena Vagaská

Department of Mathematics, Informatics and Cybernetics, Faculty of Manufacturing Technologies, Technical University of Košice, Prešov, Slovakia
{maria.tothova,alena.vagaska}@tuke.sk

Abstract. The current equipment for generating of the rotational motion by antagonistic pneumatic muscle actuator is standardly solved with a circular pulley, on which a flexible strip is strung and its ends are connected with artificial muscles. However, in this solution, torque and stiffness of such actuator decrease with increasing rotation of the actuator arm. This is due to the nonlinear decrease of muscles forces according to their contraction. By application of the oval cam, a smaller decrease in torque and thus the greater and symmetrical stiffness of the actuator with pneumatic artificial muscles can be obtained.

Keywords: Torque characteristics · Pneumatic muscles actuator · Oval cam

1 Introduction

Pneumatic artificial muscles (PAMs) are very powerful actuators that work very similar to a human muscle, they have a phenomenal ratio force to weight and they can exert a force up to 400 times their own weight [1, 2]. They will work when twisted or bent and can work under water. They're also easy and cheap.

The pneumatic muscle actuators (PMAs) are designed for generating a rotary motion using two pneumatic artificial muscles in antagonistic connection (Fig. 1). The principle of PMA based on pneumatic muscles is described in more details for example in [3–5].

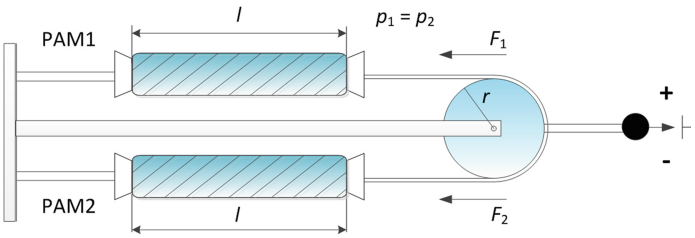


Fig. 1. Principle diagram of the antagonistic PMA based on pneumatic muscles

R. Silhavy et al. (eds.), *Cybernetics and Mathematics Applications in Intelligent Systems*, Advances in Intelligent Systems and Computing 574, DOI 10.1007/978-3-319-57264-2_9

2 Force Characteristics of Antagonistic PMA

The most important properties of the pneumatic artificial muscles are the static characteristics showing the dependence the generated force F on the muscle contraction κ under a constant muscle pressure p regardless the time factor [6–8]. These characteristics can be specified by the manufacturer of muscles and we used characteristics which are recommended by FESTO catalogue [9]. Figure 2 shows, that the range of an angle of actuator arm rotation increase by increasing the initial muscle contraction but the resulting generated force of the muscle decrease [10]. This initial muscle contraction has also an effect on the overall dynamics of the antagonistic PMA based on pneumatic muscles.

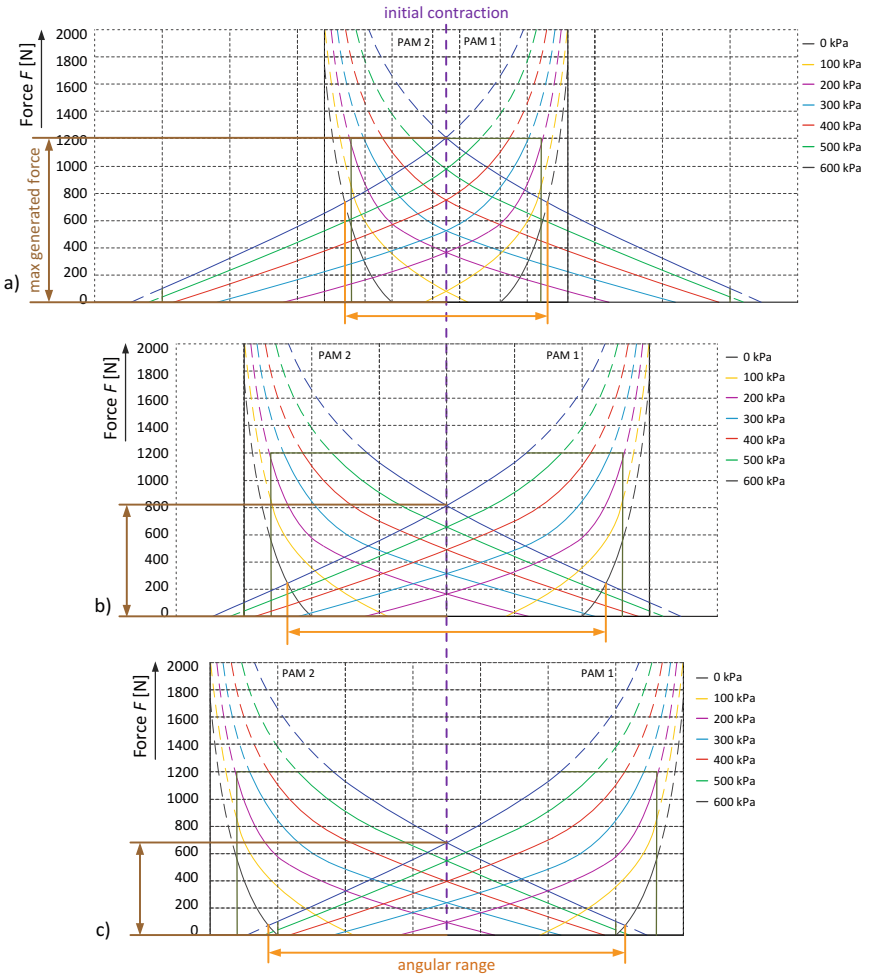


Fig. 2. Correlation between initial muscle contraction: (a) $\kappa_0 = 4\%$, (b) $\kappa_0 = 10\%$, (c) $\kappa_0 = 12.5\%$, the resulting generated force of the muscle and the range of an angle of actuator arm rotation [10]

3 Antagonistic PMA with an Oval Cam

Antagonistic PMA based on pneumatic muscles with an oval cam [11, 12] is shown in Fig. 3 and it has also a similar design as a PMA with circular pulley in [13, 14] (Fig. 1), the difference is only in using an oval cam instead of classical circular pulley. Assumption was that an oval cam is centrally symmetric and rotating around in its geometric center. The principle of the movement of the PMA based on pneumatic muscles with an oval cam is shown in Fig. 4.

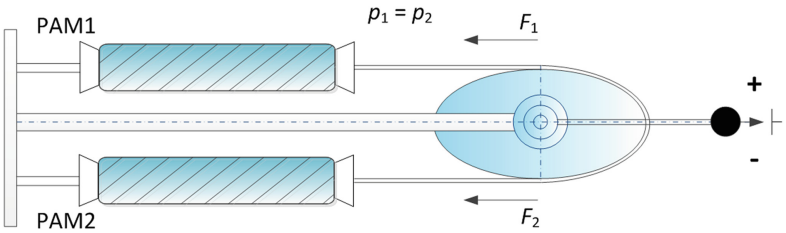


Fig. 3. Antagonistic PMA based on pneumatic muscles with an oval cam

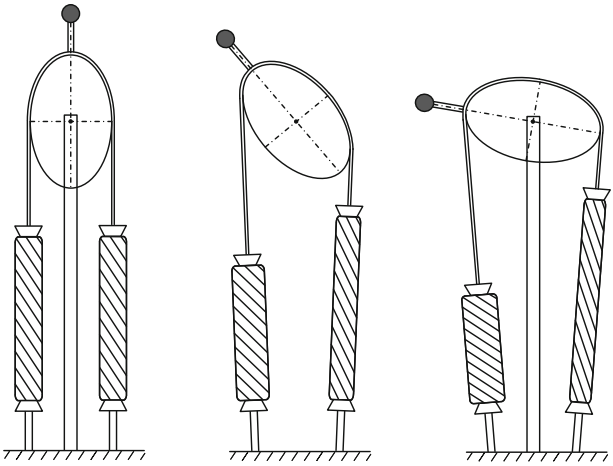


Fig. 4. The principle of the movement of the antagonistic PMA based on pneumatic muscles with an oval cam

Basic parameters for geometric description of the antagonistic PMA based on pneumatic muscles with an oval cam (Fig. 5):

- the major axis a of oval cam [m],
- the minor axis b of oval cam [m],
- the height h of the antagonistic PMA from a base plate to the axis of rotation of oval cam (to the point S) [m].

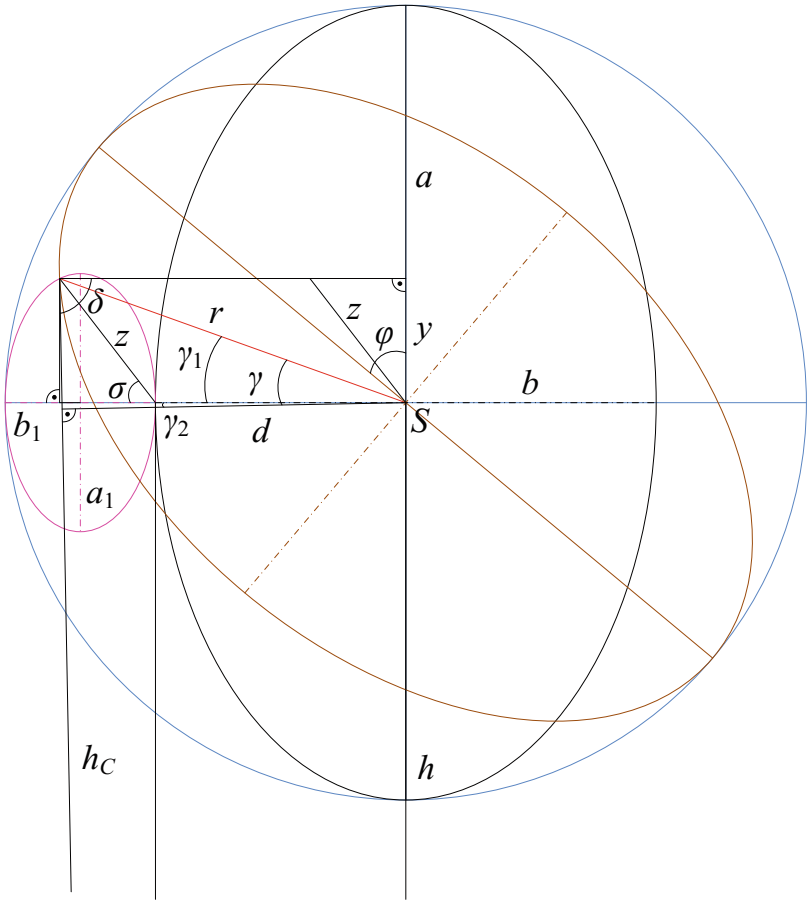


Fig. 5. Correlation between parameters of the oval cam

The length of the arm r on which muscle force acts (i.e. the distance of the point of muscle force action from the axis of rotation of the oval cam) is possible to obtain from the relation for the diagonal of the parallelogram as follows:

$$r = \sqrt{b^2 + z^2 + 2b \cdot z \cdot \cos \sigma}, \tag{1}$$

where b is the minor axis of oval cam [m], z is the auxiliary variable [m] and σ is the auxiliary angle between b and z [°].

The auxiliary variable z can be expressed (using the equation for auxiliary oval cam with constant a_1, b_1 and mathematical adjustment) as:

$$z = \frac{3(a-b) \cdot \cos \sigma}{\cos(2\sigma) + 2}. \tag{2}$$

Dependence of auxiliary angle σ (Fig. 5) to the angle φ of actuator arm rotation with a constant parameter a (value of the minor axis was $b = 0.05$ m) is shown in Fig. 6, [15].

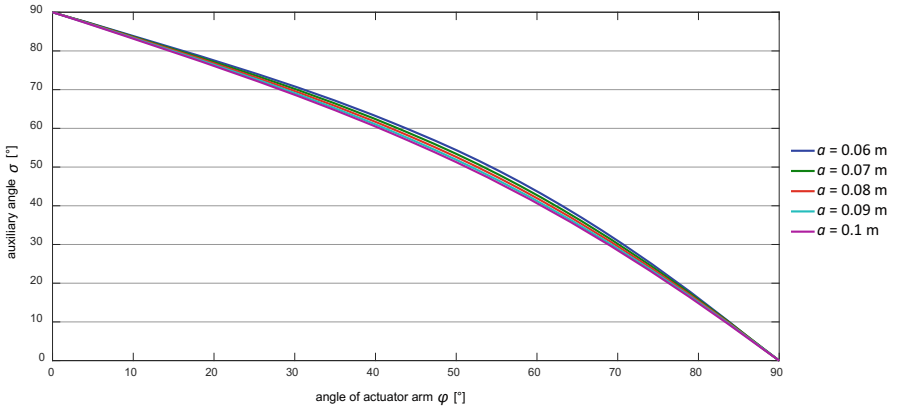


Fig. 6. Dependence of the angle φ of actuator arm rotation to the auxiliary angle σ

This dependence was approximated by a polynomial function and in the case of achieving the best possible approximation was chosen third degree polynomial with nine coefficients:

$$\sigma(\phi, a) = p_{00} + p_{10} \cdot \phi + p_{01} \cdot a + p_{20} \cdot \phi^2 + p_{11} \cdot \phi \cdot a + p_{02} \cdot a^2 + p_{30} \cdot \phi^3 + p_{21} \cdot \phi^2 \cdot a + p_{12} \cdot \phi \cdot a^2, \tag{3}$$

where φ is an angle of actuator arm rotation [°], a is the major axis of oval cam and p_{00}, p_{10}, p_{01}, p_{20}, p_{11}, p_{02} p_{30}, p_{21} p_{12} are unknown coefficients which values are found using Matlab Curve Fitting Toolbox [−] and they are presented in Table 1.

Table 1. The values of coefficients from (3)

Coefficients	Values
p_{00}	89.6800000
p_{10}	−0.3111000
p_{01}	−0.0011960
p_{20}	−0.0033640
p_{11}	−0.0004092
p_{02}	1.681E−06
p_{30}	−4.381E−05
p_{21}	3.808E−06
p_{12}	2.624E−08

Fit results of the static characteristics of dependence of auxiliary angle σ to the angle φ of actuator arm rotation with a constant parameter *a,* approximated according to (3), shows Table 2.

Table 2. Fit results for the approximation (3) using a third degree polynomial function

SSE	R-square	Adj R-square	RMSE	Coeff
7.977	0.9999	0.9999	0.19	9

The angle γ in Fig. 5 between d (the perpendicular distance of muscle forces from the axis of the oval cam) and r (the length of the arm on which muscle force acts) can be expressed as follows:

$$\gamma = \gamma_1 + \gamma_2 \tag{4}$$

and

$$\sin\gamma_1 = \frac{y}{r} = \frac{z \cdot \sin\sigma}{\sqrt{b^2 + z^2 + 2b \cdot z \cdot \cos\sigma}}, \tag{5}$$

$$\gamma_2 = \frac{\pi}{2} - \delta, \tag{6}$$

where δ is the auxiliary angle between b_1 (the minor axis of the auxiliary ellipse) and h_C (the height of the antagonistic actuator based on PAMs from a base plate to the point C) [°].

The height h_C can be expressed from Pythagorean theorem as follows:

$$h_C = \sqrt{h^2 + z^2 + 2h \cdot z \cdot \sin\sigma}. \tag{7}$$

For the auxiliary angle δ using (7) apply:

$$\sin\delta = \frac{h + y}{h_C} \Rightarrow \delta = \arcsin\frac{h + z \cdot \sin\sigma}{\sqrt{h^2 + z^2 + 2h \cdot z \cdot \sin\sigma}}. \tag{8}$$

Then the angle γ can be expressed using (5), (6) and (8) as follows:

$$\gamma = \frac{\pi}{2} + \arcsin\frac{z \cdot \sin\sigma}{\sqrt{b^2 + z^2 + 2b \cdot z \cdot \cos\sigma}} - \arcsin\frac{h + z \cdot \sin\sigma}{\sqrt{h^2 + z^2 + 2h \cdot z \cdot \sin\sigma}}. \tag{9}$$

4 Torque Characteristics of the Antagonistic PMA with an Oval Cam

For the torque M of antagonistic PMA based on pneumatic muscles with an oval cam is valid:

$$M = F \cdot d = F \cdot r \cdot \cos\gamma, \tag{10}$$

where F is the muscle tensile force [N], d is the perpendicular distance of muscle forces from the axis of the oval cam [m], r is the length of the arm on which muscle force acts (i.e. the distance of the point of muscle force action from the axis of rotation of the oval cam) [m] and α is an angle between d and r [°].

Then for torque M of the antagonistic PMA based on pneumatic muscles with an oval cam can be written:

$$M = F \cdot \sqrt{b^2 + z^2 + 2b \cdot z \cdot \cos\sigma} \cdot \cos\left(\begin{array}{c} \frac{\pi}{2} + \arcsin\frac{z \cdot \sin\sigma}{\sqrt{b^2 + z^2 + 2b \cdot z \cdot \cos\sigma}} \\ -\arcsin\frac{h + z \cdot \sin\sigma}{\sqrt{h^2 + z^2 + 2h \cdot z \cdot \sin\sigma}} \end{array}\right). \tag{11}$$

For calculation of the torque of the antagonistic PMA based on pneumatic muscles it is necessary to know some parameters and geometric constants. The basic parameters and constants are the initial muscle length ($l_0 = 0.264$ m), the initial muscle contraction ($\kappa_0 = 12.5\%$), the minor axis of the oval cam ($b = 0.05$ m) and the height of actuator ($h = 1$ m).

The torque characteristics of the antagonistic PMA based on pneumatic muscles with an oval cam dependent of the angle φ of the actuator arm for the different major axis a of oval cam are shown in Fig. 7.

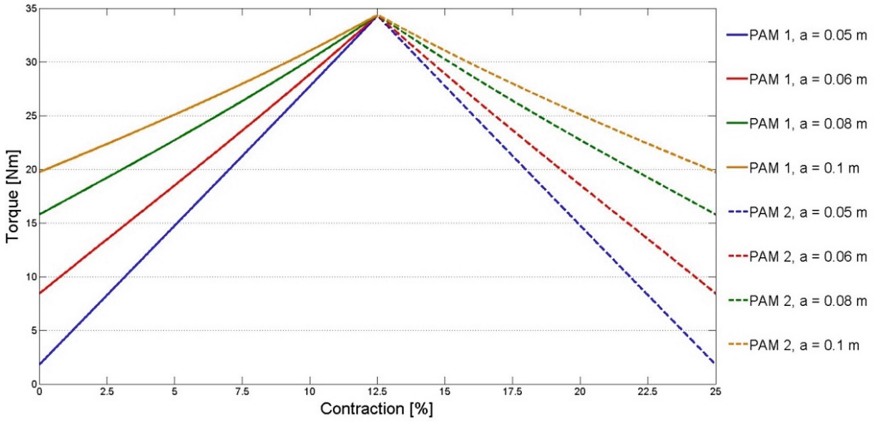


Fig. 7. Torque characteristics of the antagonistic PMA based on pneumatic muscles with an oval cam

5 Conclusion

The torque characteristics in Fig. 7 show that the torque of the antagonistic PMA based on pneumatic muscles with an oval cam decrease is the less and it is given by the value of the major axis a and the minor axis b of oval cam. It is therefore possible to conclude

that by using of an oval cam it is possible to achieve even higher stiffness of the antagonistic PMA based on pneumatic muscles. As it can be seen from Fig. 7 the disadvantage of using an oval cam is the need to increase muscle contraction with increasing rotation angle of the actuator arm (using long muscles), which is limited by maximum contraction.

Acknowledgments. The research work is supported by the project KEGA 026TUKE-4/2016 "Implementation of Modern Information and Communication Technologies in Education of Natural Science and Technical Subjects at Technical Faculties".

References

1. Chou, C.P., Hannaford, B.: Static and dynamic characteristics of McKibben Pneumatic artificial muscles. In: Proceedings of 1994 IEEE International Conference on Robotics and Automation, pp. 281–286, San Diego, USA (1994)
2. Daerden, F.: Conception and realization of Pleated Pneumatic artificial muscles and their use as compliant actuation elements, p. 176, Vrije Universiteit Brussel (1999)
3. Tondu, B., Lopez, P.: Modeling and control of McKibben artificial muscle robot actuators. Control Syst. Mag. **20**(2), 15–38 (2000)
4. Boržíková, J., Piteľ, J.: Nonlinearity of static characteristics of the antagonistic system. In: Proceedings of XXI International Educational Conference MMTT-21, pp. 196–197, Saratov State Technical University (2008)
5. Boržíková, J.: The determination of analytic dependence of static characteristic of PAM-based Antagonistic actuator. Acta Mechanica Slovaca **12**(1-A), 227–230 (2008)
6. Sárosi, J., Piteľ, J., Šeminský, J.: Static force model-based stiffness model for Pneumatic muscle actuators. Int. J. Eng. Res. Afr. **18**, 207–214 (2015)
7. Boržíková, J.: Non-linear approximation of the static characteristic $F = f(p,\kappa)$ of Antagonistic system. In: Proceeding of ARTEP 2008, pp. 4-1–5, TU Kosice (2008)
8. Tóthová, M., Piteľ, J., Hošovský, A., Sárosi, J.: Numerical approximation of static characteristics of McKibben Pneumatic artificial muscle. Int. J. Math. Comput. Simul. **9**, 228–233 (2015)
9. Fluid muscles DMSP/MAS Actuators with Special Functions 2004/10 (2016). https://www.festo.com/cat/sk_sk/data/doc_sk/PDF/SK/MAS_SK.PDF
10. Kerscher, T., Albiez, J., Zollner, J.M., Dillmann, R.: Evaluation of the dynamic model of fluidic muscles using quick-release. In: Proceeding of International Conference on Biomedical Robotics and Biomechatronics (BioRob 2006), pp. 637–642, Pisa (2006)
11. Balara, M., Piteľ, J., Tóthová, M., Vagaská, A.: Actuator with artificial muscle V. Patent No. 288296 (2015)
12. Tóthová, M., Balara, M.: Torque characteristics of actuator with Pneumatic artificial muscles and oval cam. Eng. EXTRA **19**(12), 88–89 (2015)
13. Balara, M., Vagaská, A.: The torque moment of rotary actuator with artificial muscles. In: Proceeding of ARTEP 2013, pp. 31-1–10, TU Kosice (2013)
14. Piteľ, J., et al.: Torque characteristics of Pneumatic muscle actuator with eccentric pulley. Int. J. Mech. **8**, 276–281 (2014)
15. Hrehová, S., Fechová, E.: Gain knowledge of selected properties of artificial muscles using tools of matlab. In: Proceedings of the 2014 15th International Carpathian Control Conference (ICCC 2014), pp. 164–167. IEEE, Danvers (2014)

Adaptive Control System of a Robot Manipulator Based on a Decentralized Position-Dependent PID Controller

Jan Cvejn(✉) and Jiří Tvrdík

Faculty of Electrotechnics and Informatics, University of Pardubice, Studentská 95, 532 10 Pardubice, Czech Republic
jan.cvejn@upce.cz

Abstract. The paper describes an approach to adaptive feedback control of a robot manipulator, based on partitioning of the joint space into segments. Within each segment the robot is controlled as a decoupled linear system by means of conventional PID controllers. To achieve continuity of control variables the segments are represented as fuzzy sets. The controller settings are adapted by online identification from past measurements of position and control signals.

Keywords: Robot manipulator · Motion control · PID controller · Fuzzy modeling

1 Introduction

Control of robot manipulators is difficult in general, because robot dynamics is usually strongly non-linear. Although the influence of non-linear terms in the arm motion equations can be suppressed by using high-ratio gears in the actuators, is such cases the friction in the gears and bearings often degrades the actuator performance for high-velocity motions. Especially in the case of light-weight, high-velocity robot arms for pure manipulating purposes, the arm dynamics cannot be neglected to achieve optimal performance.

In this paper the problem of tracking a trajectory, provided by the motion planning layer of the robot control system, is discussed. It is assumed that the trajectory, defined in the robot operational space, is transformed into the robot joint space by the algorithm of inverse kinematics [1], before the motion task is performed. The motion control layer then works with the information on the robot joint positions, i.e. the relative positions of the robot links.

Multiple approaches to design of the feedback control system of robot manipulator are described in literature [1, 2]. The simplest approach, suitable only for low-velocity motions, works with the actuators as with velocity generators and the effects of the robot dynamics are considered as unknown disturbances [2]. The feedback control can be then based on PI or PID controllers. To enhance the performance, cascade configuration with additional velocity or even acceleration feedback can be used. It is also possible to use a partial feed-forward compensation of non-linearities, if a partial knowledge of the robot mathematical model is available [1].

R. Silhavy et al. (eds.), *Cybernetics and Mathematics Applications in Intelligent Systems*, Advances in Intelligent Systems and Computing 574, DOI 10.1007/978-3-319-57264-2_10

More advanced robot control architectures use actuators as torque generators [1, 2]. This approach is utilized in centralized control systems, viewing the robot dynamics in full complexity as a high-order, coupled and non-linear one. The centralized methods utilize some special features of the robot dynamics. In particular, dynamic inversion method transforms the controller design problem into a linear one by means of additional interior loop. However, applicability of this approach depends on precision of the mathematical model available, which often cannot be guaranteed due to unknown influences, such as backslashes and flexibilities in the gears or saturations of the action forces. Therefore, practical usability requires some extensions, guaranteeing at least closed-loop asymptotical stability [1]. Among alternative approaches e.g. the non-linear PID control based on Lyapunov stability theory or passive systems theory can be mentioned [1, 2].

An advantage of the decentralized control approach is that for the controller tuning only rough robot mathematical model is sufficient, describing approximate inertial effect on individual axes and damping effects. If we assume that terms in the robot model depend only on position, it is possible to improve the performance by dividing the joint space into segments with constant controller settings. The controller parameters can be changed during motion when the trajectory goes across the segment boundary. Within each segment the robot then can be controlled as a decoupled linear system by means of conventional PID controllers. It is possible to use initially the same settings in all the segments and to adapt the controller parameters in each segment automatically by processing past measurements of the kinematic and control variables. Such an algorithm adapts the controller parameters also when the robot manipulated load changes.

In this paper an extension of the decentralized robot control approach is proposed, based on the idea outlined above, including some additional enhancements. Practical implementation is indeed more complex than in the case of conventional decentralized control. Partitioning of the joint space into segments brings increased requirements on the control system hardware, as regards both performance and memory capacity. However, it reveals that these requirements can be fulfilled by using current 32-bit microcontroller-based platforms.

2 The Robot Mathematical Model

The mathematical model of a robot arm consisting of n links in an open kinematic chain, moving freely in the operation space, can be considered in the form

$$\mathbf{B}(\mathbf{q})\ddot{\mathbf{q}} + \mathbf{C}(\mathbf{q}, \dot{\mathbf{q}})\dot{\mathbf{q}} + \mathbf{g}(\mathbf{q}) = \mathbf{f} \tag{1}$$

where $\mathbf{q}$ is the vector of joint positions and $\mathbf{f}$ the vector of total force effects of actuators. If K and P denote the total kinetic and potential energy, $\mathbf{B}(\mathbf{q}) = \partial^2 K / \partial \dot{\mathbf{q}}^2$ is a positive definite position-dependent inertia matrix,

$$\mathbf{C}(\mathbf{q}, \dot{\mathbf{q}})\dot{\mathbf{q}} = \frac{\partial^2 K}{\partial \mathbf{q} \partial \dot{\mathbf{q}}}\dot{\mathbf{q}} - \frac{\partial K}{\partial \mathbf{q}} \tag{2}$$

is a non-linear function corresponding to the effects of centrifugal and Coriolis forces and $\mathbf{g}(\mathbf{q}) = \partial P/\partial \mathbf{q}$ is the vector function corresponding to the gravity-force effects [1]. If we assume only electrical DC actuators, by neglecting the winding inductance and mechanical friction, we obtain

$$\mathbf{M} \approx \mathbf{K}_u \mathbf{u} - \mathbf{K}_v \boldsymbol{\omega} \tag{3}$$

where $\mathbf{M}$ is the vector of motor output torques, $\mathbf{u}$ the input voltages, $\boldsymbol{\omega}$ the vector of motor angular velocities and $\mathbf{K}_u > 0$, $\mathbf{K}_v > 0$ constant diagonal matrices. In principle, (3) allows using the motors as velocity generators, where the connected load is represented as a disturbance. Alternatively, the motor can play the role of a torque generator, where the term $\mathbf{K}_v\boldsymbol{\omega}$, corresponding to induced voltage in winding, is considered as electromagnetic friction. In this case, the motor is usually equipped with inner current feedback, which reduces the effect of $\mathbf{K}_v\boldsymbol{\omega}$ and protects from overload [1]. Although this paper is based on the idea of decentralized control, the motors are considered as torque generators, like at most centralized control methods. In this case

$$\mathbf{f} = \mathbf{K}_r\left(\mathbf{K}_u \mathbf{u} - \mathbf{K}_v \boldsymbol{\omega}\right) - \mathbf{F}\dot{\mathbf{q}} \tag{4}$$

where $\mathbf{K}_r > 0$ and $\mathbf{F} > 0$ are diagonal matrices. The term $\mathbf{F}\dot{\mathbf{q}}$ corresponds to viscous friction in bearings and gears and $\mathbf{K}_r$ is the mechanical gear ratio. Coulomb friction is not considered. Since $\dot{\mathbf{q}} = \mathbf{K}_r^{-1}\boldsymbol{\omega}$,

$$\mathbf{f} = \mathbf{K}_r\mathbf{K}_u\mathbf{u} - \mathbf{F}_r\dot{\mathbf{q}},\ \mathbf{F}_r = \mathbf{K}_r\mathbf{K}_v\mathbf{K}_r + \mathbf{F}. \tag{5}$$

The Eq. (1) then can be rewritten as

$$\mathbf{B}(\mathbf{q})\ddot{\mathbf{q}} + \left(\mathbf{C}(\mathbf{q}, \dot{\mathbf{q}}) + \mathbf{F}_r\right)\dot{\mathbf{q}} + \mathbf{g}(\mathbf{q}) = \mathbf{K}\mathbf{u} \tag{6}$$

where $\mathbf{K} = \mathbf{K}_r\mathbf{K}_u$. Since the dependence of $\mathbf{F}_r$ on $\mathbf{K}_r$ is quadratic, the term $\mathbf{C}(\mathbf{q}, \dot{\mathbf{q}})$ has low influence in the case of higher mechanical gear ratios and the model depends predominantly only on $\mathbf{q}$.

3 The Decentralized Control with Partial Knowledge of B(q)

One possible version of the decentralized robot control algorithm uses partial knowledge of the inertia matrix $\mathbf{B}(\mathbf{q})$, which is decomposed as

$$\mathbf{B}(\mathbf{q}) = \bar{\mathbf{B}} + \Delta\mathbf{B}(\mathbf{q}) \tag{7}$$

where $\bar{\mathbf{B}}$ is a constant diagonal positive definite matrix, corresponding to approximate average inertial effects on individual axes [1]. The robot model for the controller design is in the form

$$\bar{\mathbf{B}}\ddot{\mathbf{q}} + \mathbf{F}_r\dot{\mathbf{q}} = \mathbf{K}\mathbf{u} - \mathbf{d} \tag{8}$$

where

$$\mathbf{d} = \Delta\mathbf{B}(\mathbf{q})\ddot{\mathbf{q}} + \mathbf{C}(\mathbf{q}, \dot{\mathbf{q}})\dot{\mathbf{q}} + \mathbf{g}(\mathbf{q}) \tag{9}$$

is disturbance. Since all matrices in (8) are diagonal, it is possible to write (8) as

$$b_i\ddot{q}_i + f_{ri}\dot{q} = k_i u_i - d_i, \;\; i = 1, .., n \tag{10}$$

where b_i, f_{ri} and k_i denote the diagonal terms of $\bar{\mathbf{B}}$, $\mathbf{F}_r$ and $\mathbf{K}$, respectively. Equation (10) can be rewritten as

$$T_i\ddot{q}_i + \dot{q} = K_i u_i - \delta_i \tag{11}$$

where $T_i = b_i/f_{ri}$, $K_i = k_i/f_{ri}$ and $\delta_i = d_i/f_{ri}$. The corresponding axis transfer function is in the form

$$F_i(s) = \frac{K_i}{s\left(T_i s + 1\right)}. \tag{12}$$

To compensate the effect of persistent input-type disturbance d_i of the system (10) and to achieve zero tracking error in the case of ramp reference trajectory, a controller with additional zero pole is needed. If we use the PID controller with the transfer function in Laplace transform

$$R_i(s) = r_i\left(1 + \frac{1}{T_{Ii}s} + T_{Di}s\right) = \frac{r_i}{T_{Ii}s}\left(T_{Ii}T_{Di}s^2 + T_{Ii}s + 1\right) \tag{13}$$

for generating the control signals $u_i(t)$, the characteristic polynomial of the i-th axis is in the form

$$Q_i(s) = \frac{T_{Ii}s^2\left(T_i s + 1\right)}{r_i K_i} + \left(T_{Ii}T_{Di}s^2 + T_{Ii}s + 1\right) = \frac{T_{Ii}T_i}{r_i K_i}s^3 + \frac{T_{Ii} + r_i K_i T_{Di}T_{Ii}}{r_i K_i}s^2 + T_{Ii}s + 1. \tag{14}$$

The PID controller parameters can be determined so that poles of $Q_i(s)$ are placed at desired locations [6]. If we assume $Q_i(s) = \left(T_{1i}s + 1\right)\left(T_{2i}^2 s^2 + 2\xi_i T_{2i}s + 1\right)$, where T_{1i}, T_{2i} are chosen closed-loop response time constants and ξ_i is the relative damping ratio, by comparison of the coefficients we obtain

$$r_i = \frac{\left(T_{1i} + 2\xi_i T_{2i}\right)T_i}{T_{1i}T_{2i}^2 K_i}, \; T_{Ii} = T_{1i} + 2\xi_i T_{2i}, \; T_{Di} = \frac{2\xi_i T_{1i}T_{2i} + T_{2i}^2}{T_{Ii}} - \frac{1}{r_i K_i}. \tag{15}$$

If the disturbance d_i is partially known, its effect can be compensated in part by adding $K_i^{-1}\delta_i$ to the i-th axis controller output, where $\delta_i = d_i/f_{ri}$.

4 The Extended Decentralized Robot Control Algorithm

Consider that the robot joint space S, i.e. the space of all possible joint positions $\mathbf{q} = \left[q_1, \ldots, q_n\right]^T$, is partitioned into m segments S_k, $k = 1, .., m$, such that

$$\bigcup_{k=1}^{m} S_k = S, \bigcap_{k=1}^{m} S_k = \emptyset. \tag{16}$$

Since the diagonal parts of $\mathbf{B}(\mathbf{q})$ and $\mathbf{g}(\mathbf{q})$ are position-dependent, it is possible to approximate the robot dynamics in each segment by a different linear model. Then the tracking performance can be enhanced if to each segment there correspond different controller settings.

A basic approach is that the controller parameters are rewritten during motion when the trajectory goes across the segment boundary. Within each segment the robot is controlled as a decoupled linear system by means of conventional PID controllers. This extension is rather straightforward and can be efficiently implemented, although a sufficient amount of memory in the control system hardware is needed. If a space of each generalized coordinate is divided into d sub-intervals, the joint space will be divided into $m = d^n$ segments. To each segment S_k there corresponds a matrix of the parameters $\mathbf{P}_k = \left[\mathbf{K}_k, \mathbf{T}_k, \boldsymbol{\delta}_k\right]$ describing the plant dynamics as described in the previous section. The columns in $\mathbf{P}_k$ are vectors of n components, e.g. $\mathbf{K}_k = \left[K_{k1}, \ldots, K_{kn}\right]^T$. The corresponding PID controlled settings can be obtained directly by substitution into (15).

However, this concept has an important drawback, consisting in discontinuity of the action variables $u_i(t)$ caused by changes of the controller settings at the segments boundaries. Such discontinuities are undesirable, since they can lead to oscillations of the mechanical structure.

One possible solution is replacing the segments S_k by the fuzzy sets $\tilde{S}_k = \left\{R^n, \mu_k(\mathbf{q})\right\}$, where the membership functions $\mu_k(\mathbf{q})$ are chosen continuous and so that $\mu_k(\mathbf{q}) \in [0, 1]$ and $\mu_k\left(\mathbf{c}_k\right) = 1$, where $\mathbf{c}_k$ denotes the centre of the segment S_k. The matrix of the plant parameters is then computed at each control step as

$$\mathbf{P} = \sum_{k=1}^{m} \mu_k(\mathbf{q})\mathbf{P}_k / \sum_{k=1}^{m} \mu_k(\mathbf{q}). \tag{17}$$

The expression (17) is usually used as an inference rule in Takagi-Sugeno-type fuzzy modeling [4]. However, the computation of (17) at each control step can be time-consuming due to rather large number of segments m. It can be considered that the segments are for given parameter h defined by means of their centers $\mathbf{c}_k$ as

$$S_k = \left\{\mathbf{q} \mid \left\|\mathbf{q} - \mathbf{c}_k\right\|_\infty \leq h\right\}, \tag{18}$$

where $\|\mathbf{x}\|_\infty = \max|x_i|$ denotes the L_∞-norm. Then it is possible to define

$$\mu_k(\mathbf{q}) = \mu\big(\mathbf{q} - \mathbf{c}_k\big) \tag{19}$$

where $\mu(\mathbf{q}) \in [0, 1]$ is a chosen continuous function, such that $\mu(\mathbf{0}) = 1$. Computation of (17) can be made much more efficient if $\mu(\mathbf{q})$ is chosen as a function with compact support, see e.g. [5]. A simple possibility is to choose

$$\mu(\mathbf{q}) = \max\big\{1 - \|\mathbf{q}\|_\infty/((1+\lambda)h),\ 0\big\} \tag{20}$$

where $\lambda > 0$, typically $\lambda \in [0.5, 2]$. In this case the values of $\mu_k(\mathbf{q})$ are zero, except for the k-the segment, where $\mu_k(\mathbf{q})$ has the largest value, and several neighboring segments. This fact can be utilized for efficient implementation of the controller, although such a realization is more complex. A disadvantage of the choice (20) is that this function is not smooth, which will produce non-smooth histories of the control signals. Therefore, it might be preferable to construct $\mu(\mathbf{q})$ as at least continuously differentiable. The compact-support choice of $\mu(\mathbf{q})$ brings the risk of unbounded values in (17) for trajectories exceeding the boundaries of S, but this problem can be easily avoided, e.g. by increasing λ in the cases when $\sum_{k=1}^{m} \mu_k(\mathbf{q}) = 0$.

5 Adaptation of the Controller Settings

Since the number of segments can be rather large, it is necessary that the controller settings are computed automatically. Initially, the settings in all the segments are set to the same values corresponding to the decentralized PID controller design described in Sect. 3. During the robot operation the settings in the segments can be adapted by processing the measured values of $\mathbf{q}\big(t_k\big)$ and $\mathbf{u}\big(t_k\big)$, at the instants $t_k = k\Delta$, where Δ is the identification scan period. During motion in each segment it is needed to estimate the parameters K_i, T_i and δ_i. The index of segment is omitted below for simplicity, i.e. e.g. K_i should be written as K_{ki} in the k-th segment to be precise.

The continuous transfer function (12) has the corresponding transfer function in Z-transform

$$\begin{aligned} F_i(z) &= \mathcal{Z}\big\{K_i T_i\big(-1 + t/T_i + e^{-t/T_i}\big);\ t = k\Delta\big\} \\ &= K_i T_i\left(-1 + \frac{\Delta}{T_i}\frac{1}{z-1} + \frac{z-1}{z - e^{-\Delta/T}}\right) = K_i\left(\frac{\Delta}{z-1} + T_i\frac{e^{-\Delta/T_i} - 1}{z - e^{-\Delta/T_i}}\right). \end{aligned} \tag{21}$$

If Δ/T_i is sufficiently low, $e^{-\Delta/T_i} \approx 1 - \Delta/T_i$, so

$$F_i(z) \approx K_i\Delta\left(\frac{1}{z-1} - \frac{1}{z - e^{-\Delta/T_i}}\right) = K_i\Delta\frac{1-\alpha_i}{z^2 - \big(1+\alpha_i\big)z + \alpha_i} \tag{22}$$

where $\alpha_i = e^{-\Delta/T_i}$. Equation (22) corresponds to the data model

$$q_i^{[k+2]} - \big(1+\alpha_i\big)q_i^{[k+1]} + \alpha_i q_i^{[k]} = K_i\Delta\big(1-\alpha_i\big)\,u_i^{[k]} + \varepsilon^{[k]} \tag{23}$$

where $q_i^{[k]}$ denotes the value of q_i at k-th instant and $\varepsilon^{[k]}$ is the error process. After including the disturbance δ_i, defined by (11), (23) can be rewritten as

$$\Delta^{-1}\left(q_i^{[k]} - q_i^{[k+1]}\right)\alpha_i - u_i^{[k]}\tilde{K}_i + \tilde{\delta}_i = \Delta^{-1}\left(q_i^{[k+1]} - q_i^{[k+2]}\right) + \varepsilon^{[k]} \tag{24}$$

where $\tilde{K}_i = K_i\left(1-\alpha_i\right)$ and $\tilde{\delta}_i = \delta_i\left(1-\alpha_i\right)$. From (24) the value of $\boldsymbol{\theta}_i = \left[\alpha_i, \tilde{K}_i, \tilde{\delta}_i\right]^T$ can be estimated by means of the least-squares method. Then, α_i determines the value of T_i. The parameters K_i and δ_i can be computed directly from $\boldsymbol{\theta}_i$.

Note that to each segment there corresponds one data model (24) and a corresponding data structure have to exist in the control system for storing the measurements. It is advantageous to use the recursive version of the LS estimator [3], which need not store all the data, but only a 3×3 matrix and the vector $\boldsymbol{\theta}_i$ in each segment and for each axis. It is assumed that the settings are updated after a single task is performed, but the same approach can be used when the controller is updated during motion. Let

$$\mathbf{x}_i^{[k]} = w_j\left(\mathbf{q}^{[k]}\right)\left[\Delta^{-1}\left(q_i^{[k]} - q_i^{[k+1]}\right), -u_i^{[k]}, 1\right]^T,\ y_i^{[k]} = w_j\left(\mathbf{q}^{[k]}\right)\Delta^{-1}\left(q_i^{[k+1]} - q_i^{[k+2]}\right) \tag{25}$$

where $w_j(\mathbf{q})$ is the weight of the measurement in the j-th segment, computed as

$$w_j(\mathbf{q}) = \mu_j(\mathbf{q}) \,/\, \sum_{l=1}^{m} \mu_l(\mathbf{q}). \tag{26}$$

The $(k+1)$-th estimate of $\boldsymbol{\theta}_i$ for the j-th segment is obtained as follows:

$$\boldsymbol{\theta}_i^{[k+1]} = \boldsymbol{\theta}_i^{[k]} + \frac{\mathbf{C}_i^{[k]}\mathbf{x}_i^{[k]}}{\gamma + \mathbf{x}_i^{[k]T}\mathbf{C}_i^{[k]}\mathbf{x}_i^{[k]}}\left(y_i^{[k]} - \mathbf{x}_i^{[k]T}\boldsymbol{\theta}_i^{[k]}\right) \tag{27}$$

$$\mathbf{C}_i^{[k+1]} = \frac{1}{\gamma}\left(\mathbf{I} - \frac{\mathbf{C}_i^{[k]}\mathbf{x}_i^{[k]}\mathbf{x}_i^{[k]T}}{\gamma + \mathbf{x}_i^{[k]T}\mathbf{C}_i^{[k]}\mathbf{x}_i^{[k]}}\right)\mathbf{C}_i^{[k]} \tag{28}$$

where γ is the forgetting coefficient, equal or very close to 1. In the considered case it seems to be necessary to require that $T_i > T_{i\min} > 0$ and $K_i > K_{i\min} > 0$, where the constants $T_{i\min}, K_{i\min}$ are suitably chosen. The recursive form of the estimator (27) enables to keep the values of the components of $\boldsymbol{\theta}_i$ in the corresponding range by updating only with feasible values. The matrix $\mathbf{C}_i^{[0]}$ is set as $\mathbf{C}_i^{[0]} = \mathrm{diag}\left(\sigma_1, \sigma_2, \sigma_3\right)$, where $\sigma_k > 0$ are chosen. Total memory requirements can be estimated as $12\,m\,n$ real numbers, which can occupy from tens to hundreds of kbytes of the control system memory.

6 Simulated Results

Consider the 3-DOF anthropomorphic robot arm approximate model in Fig. 1, where $m_1 = m_2 = 1$ kg, $l_1 = l_2 = 0.5$ m and $h = 1$ m. The terms of the diagonal matrices $\mathbf{K}_u$, $\mathbf{K}_v$, $\mathbf{K}_r$ and $\mathbf{F}$ were chosen as $k_{ui} = 1$, $k_{vi} = 0$, $k_{ri} = 10$ and $f_i = 3$, $i = 1, .., 3$.

The mathematical model for simulation, which is strongly non-linear, was obtained by expressing the terms $\mathbf{B}(\mathbf{q})$, $\mathbf{C}(\mathbf{q}, \dot{\mathbf{q}})\dot{\mathbf{q}}$ and $\mathbf{g}(\mathbf{q})$ in (1), where $\mathbf{q} = [\varphi, \psi, \vartheta]^T$, from the expressions for kinetic and potential energy. Since the mathematical model is rather complex, the details had to be omitted due to paper length limitations.

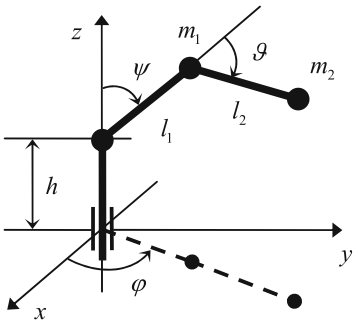


Fig. 1. The robot arm approximate model

First, the fixed axes PID controllers (13) were used. The diagonal matrix $\bar{\mathbf{B}}$ for setting-up the PID controllers, with the meaning of rough estimate of the inertia matrix, was chosen as

$$\bar{\mathbf{B}} = \operatorname{diag}\left(\frac{m_1 l_1^2 + m_2 \left(l_1 + l_2\right)^2}{2},\ \left(m_1 + m_2\right) l_1^2,\ m_2 l_2^2 \right). \tag{29}$$

The reference trajectory was chosen as the step function, which can be viewed as the worst-case situation, since the robot will usually track a continuous trajectory. The controller parameters were computed so that $T_{1i} = T_{2i} = 0.075$ s and $\xi_i = 0.8$ in (15). The joint initial and target positions are considered as

$$\mathbf{q}_0 = [-1.5,\ 2,\ 3],\ \mathbf{q}_f = [1.5,\ -2,\ -1]. \tag{30}$$

Figure 2 shows the corresponding histories of the robot joint positions.

Further, the proposed adaptive controller has been used. The joint space has been divided into $m = 8^3$ segments, $\lambda = 1$ has been chosen in (20). The controller has been initialized to the same settings as in the previous case and executed 20 times the same trajectory with the scan period $\Delta = 0.002$ s to adapt. The desired closed-loop settings $T_{1i} = T_{2i} = 0.075$ s and $\xi_i = 0.8$ were preserved. The matrices $\mathbf{C}_i$ have been initialized as $\mathbf{C}_i^{[0]} = 0.01 \times \operatorname{diag}(1, 0.3, 1)$ and $\gamma = 0.9^{\Delta}$, $T_{i\min} = 0.03$ and $K_{i\min} = 0.05$ has been chosen. Figure 3 shows the final simulated histories. It can be seen that significant enhancement has been achieved. Similar results were obtained also for lower values of m. In these cases the convergence was faster, but it seems that it is necessary to use higher values of $T_{i\min}$ and $K_{i\min}$ and the responses are a little slower.

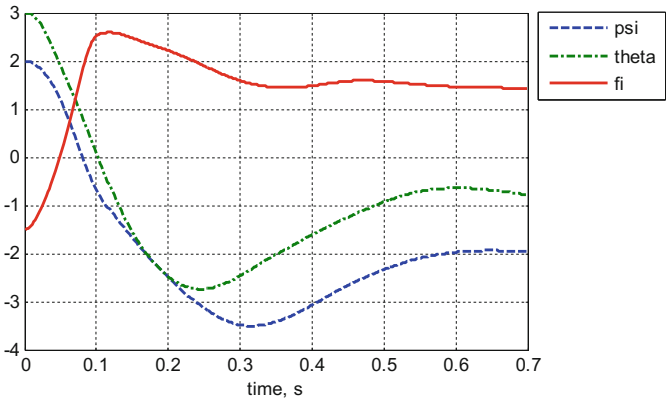


Fig. 2. The joint step responses - fixed axes PID controllers

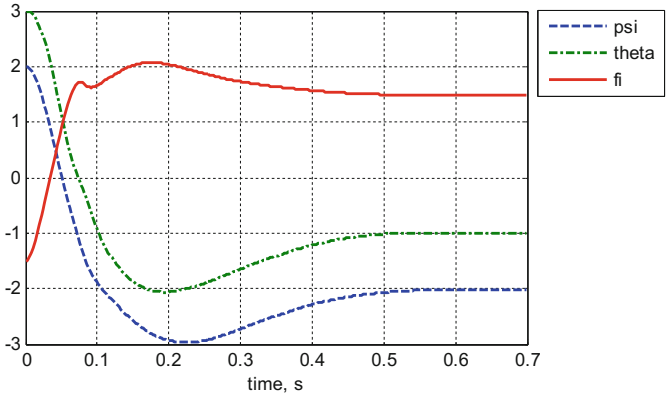


Fig. 3. The joint step responses - adaptive control system, $m = 8^3$

7 Conclusions

The proposed adaptive control system is based on the principles of the PID controller-based decentralized control, where the joint space is divided into segments with different controller settings. To ensure continuity of the control signal at the segment boundaries, the segments are represented as fuzzy sets with a special choice of the membership function. Simulated results show that the approach can be used in the cases when the conventional decentralized control fails to produce good responses, although such situations seem to occur mainly in the cases of long-range and high-velocity movements. The control system memory requirements are large in comparison with conventional control algorithms, but can be fulfilled by using current 32-bit microcontrollers. Thus the worst problem from practical point of view seems to be proper initial settings of $\mathbf{C}_i$, $T_{i\min}$, $K_{i\min}$ and other parameters that influence convergence of the sequence of estimates (27).

References

1. Sicilliano, B., Sciavicco, L., Oriollo, G.: Robotics: Modelling, Planning and Control. Springer, London (2009)
2. Siciliano, B., Khatib, O. (eds.): Springer Handbook of Robotics. Springer, Heidelberg (2008)
3. Goodwin, G.C., Payne, R.L.: Dynamic System Identification. Experimental Design and Data Analysis. Academic Press, London (1977)
4. Sugeno, M.: Industrial Applications of Fuzzy Control. Elsevier Science Pub. Co., New York (1985)
5. Rektorys, K.: Variational Methods in Mathematics, Science and Engineering, 2nd edn. Reidel, Dordrecht (1980)
6. Kiong, T.K., Quing-Guo, W., Chieh, H.C., Hägglund, T.J.: Advances in PID Control. Springer, Heidelberg (1999)

Possibilities of Process Modeling in Pedagogical Cybernetics Based on Control-System-Theory Approaches

Tomas Barot(✉)

Pedagogical Faculty, Department of Mathematics with Didactics, University of Ostrava, Mlynska 5, 701 03 Ostrava, Czech Republic
Tomas.Barot@osu.cz

Abstract. This paper tries to extent the connection between the technical and pedagogical cybernetics. Particularly, the process-modeling possibilities from the control-system theory are applied to the pedagogical-research area in this paper. In the pedagogy, the cybernetics processes are not further mathematically modeled, because the classical approaches are not widely based on mathematical background of control-system theory. The models are usually presented in a schematic form. In the other case, the measured variables can be described using a set of statistical parameters in the statistical research. The feedback-control model is established in the pedagogical cybernetics. For the purposes of pedagogical research, this paper presents the possibilities of process modeling using control-system-theory approaches. The verification of the presented approach can be provided using statistical methods.

Keywords: Pedagogical cybernetics · Control-system theory · Hypothesis testing · Normality testing · Mann-Whitney U-test · Kruskal Wallis test

1 Introduction

The pedagogy is a humanistic area, which describes and researches the aspects of the education in the social context [1–3]. Its particular processes can be studied using approaches, which are part of the pedagogical cybernetics [1–3]. The cybernetics has many areas and its principles are generally applicable; however, the technical cybernetics has the concrete mathematical background in comparison to the humanistic based cybernetics. In the technical areas, the control-system theory [4, 5] can be suitable used. The humanistic researches are in general processed using quantitative statistical methods [6–13].

The quantitative research is based on the statistical methods, which can be in general divided in two categories. At first, the descriptive methods can be used for the whole set of population. Further, the methods of statistical induction include the hypothesis testing according to the significance level α. The p-value approach is suitable for these purposes of hypothesis verification. The particular methods are depended on the normality of data [6–13].

R. Silhavy et al. (eds.), *Cybernetics and Mathematics Applications in Intelligent Systems*, Advances in Intelligent Systems and Computing 574, DOI 10.1007/978-3-319-57264-2_11

The pedagogical cybernetics is a cybernetics discipline, which is based on humanistic problems; however, the general cybernetics principles can be fulfilled. Many various types of educational processes can be described in the pedagogical cybernetics. The feed-back-control model [2] is established there. The advanced possibility of this theory is a definition of an artificial teacher [3], which can be considered in the e-learning systems. In the classical theory, modeling-possibilities of the pedagogical processes using the mathematical background are not widely described [1–3].

The control-system theory [4, 5] is closely bounded with the technical cybernetics. All processes and analyses can be provided using mathematical methods. The analyses of system can be realized with respect to the further synthesis of controller [5]. The processes can be modeled using numerical simulation methods. The main circuit of control is based on feed-back principle. There are many variations of the feed-back-control strategy. For example, the adaptive control reflects the time-variant systems using the online identification. However, these principles can be suitable utilized in the pedagogical cases [4, 5].

In this paper, the unused control-system-theory mathematical background [4, 5] is extended and presented in the pedagogical cybernetics [1–3]. The results of proposed approach should have the similar statistical properties that can be verified using quantitative statistical methods [6–13]. The prerequisite for this application is the existence of the general process model [2] in the pedagogical cybernetics for purposes of modeling of educational processes.

2 Methods for Process Analysis in Pedagogical Cybernetics

The pedagogical cybernetics [1–3] determines a feed-back model of educational process [2], where the controller can be modeled as a teacher and the controlled object can be represented by a student. However, the included particularly models are not widely mathematically described in this area. For example, a student skills-level is in general analyzed using statistical methods (elementary statistics or hypothesis testing) [1–3].

For count of n samples x_i of random variable X, properties of this data can be characterized using elementary statistics: sample average (1) and sample standard deviation (2) [6–13].

$$\hat{\bar{x}} = n^{-1} \sum_{i=1}^{n} x_i \tag{1}$$

$$\hat{\sigma}^2 = (n-1)^{-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{2}$$

Normality test [7] includes a testing of hypotheses (3) and (4). In the null hypothesis (3), an origin of the random variable X is detected on its normal probability distribution N. An alternative hypothesis (4) expresses the independency on this probability distribution. For results of this hypothesis testing, the significance p-value is compared with the significance level α. For $p < \alpha$, the null hypothesis is rejected, in favor of the alternative hypothesis on the significance level α. In the other case, the null

hypothesis is not rejected on the same significance level. In the paper, the Shapiro-Wilk test [7] on data-normality is performing [7].

$$H_0 : X \sim N \tag{3}$$

$$H : X \nsim N \tag{4}$$

The random variable X can be described using parameters: mean value μ and standard deviation σ^2, which are corresponding with the probability distribution of X [6–13].

For comparison of means values between more than two random variables X_1, …, X_m, the hypothesis testing is performed using the ANOVA test [10] for normal probability distribution of all these variables. In the case of failure of this requirement, the Kruskal-Wallis test [13] can be applied. A null hypothesis consists of equality statement (5). For the significance p-value $p < \alpha$, the null hypothesis is rejected, in favor of the alternative hypothesis (6) on the significance level α. For the significance p-value $p > \alpha$, the null hypothesis is not rejected on the significance level α [6, 10, 13].

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_m \tag{5}$$

$$H_1 : \mu_1 \neq \mu_2 \neq \cdots \neq \mu_m \tag{6}$$

In the case of number of two random variables X_i and X_j, the T-test [8] is selected as a method for testing of hypotheses (7) and (8). The normal probability distribution of random variable is required. However; the F-test [9], which is described as a further method, should be performed at first for the purposes of T-test evaluation. For non-normal distributed random variables X_i and X_j, the Mann-Whitney U-test [11] is recommended. For $p < \alpha$, the null hypothesis (7) is rejected, in favor of the alternative hypothesis (8) on α. In opposite case, the null hypothesis is not rejected on α [6, 8, 9, 11].

$$H_0 : \mu_i = \mu_j \tag{7}$$

$$H_1 : \mu_i \neq \mu_j \tag{8}$$

For the hypotheses testing (9) and (10) on equalities of standard deviations, the F-test is recommended for two normal-distributed variables X_i and X_j. The Kolmogorov-Smirnov test [12] is applied as an alternative to the F-test for non-normal distributed random variables X_i and X_j. For $p < \alpha$, the null hypothesis (9) is rejected, in favor of the alternative hypothesis (10) on significance level α. In opposite case, the null hypothesis is not rejected on significance level α [6, 9, 12].

$$H_0 : \sigma_i^2 = \sigma_j^2 \tag{9}$$

$$H_1 : \sigma_i^2 \neq \sigma_j^2 \tag{10}$$

3 Possibilities of Process Modeling in Control-System Theory

In control-system theory, the approximation of real system can be realized using transfer function [4] with variable s in the complex plane. This mathematical representation corresponds with the differential equations using Laplace transformation [5]. For purposes in this paper, the model (11)–(13) of linear continuous dynamic systems of the first order is utilized. In the control-system theory, the roots of a polynom in the denominator of transfer function should be negative values in case of a stable system [4, 5].

$$G(s) = \frac{b_0}{s + a_0} = \frac{K}{Ts + 1} \tag{11}$$

$$K = \frac{b_0}{a_0} \tag{12}$$

$$T = \frac{1}{a_0} \tag{13}$$

A mathematical model of the system (11) can be described by an alternative form using continuous step function $h(t)$ [4, 5]. In case of particular model (11), expression of a time-based step function can be described as (14). Both modeling possibilities (11) and (14) can suitable express the dynamic behaviour of the described real system. Concretely, the step function is an output-response of the model for an input signal $\eta(t)$ in form of unit step ($t < 0$: $\eta(t) = 0$; $t > 1$: $\eta(t) = 1$) [4, 5].

$$h(t) = \frac{b_0}{a_0}(1 - e^{-a_0 t}) = K\left(1 - e^{-\frac{t}{T}}\right) \tag{14}$$

4 Proposed Approach for Process Modeling in Pedagogy

In the simulation theory, the real system should be modeled using a mathematical form [5], which includes any suitable declared parameters. This approximation should represent the original behaviour of the system dynamics. For fulfilling of this base purpose, the new structure of dynamic model is present for the pedagogy area in this paper. The results of classical research methods (statistics methods) can be further compared with dynamics responds performed using the proposed models. The proposed approach is based on extended possibilities of simulation of any education process using the mathematical background of control-system theory.

The structure of a proposed dynamics model (15) is based on the statistics properties of analyzed data: sample average and sample standard deviation. The corresponding realization of sample average is presented using a parameter K (16). Belong to this recommendation; the sample standard deviation is transform to the parameter T using the rule (17). For this structure of proposed model, the parameters a_0 and b_0

have a form (18) and (19). The alternative mathematical possibility - the step function can be realized using equation in time plane as (20) and can be suitable displayed in a plot. The root π (21) of the polynom in denominator can be only negative value for any sample standard deviation that fulfills the requirements for the stability [5] of model.

$$G(s) = \frac{\hat{\bar{x}}}{\hat{\sigma}^2 s + 1} \tag{15}$$

$$K = \hat{\bar{x}} \tag{16}$$

$$T = \hat{\sigma}^2 \tag{17}$$

$$a_0 = \frac{1}{\hat{\sigma}^2} \tag{18}$$

$$b_0 = \frac{\hat{\bar{x}}}{\hat{\sigma}^2} \tag{19}$$

$$h(t) = \hat{\bar{x}}\left(1 - e^{-\frac{t}{\hat{\sigma}^2}}\right) \tag{20}$$

$$\pi = -\frac{1}{T} = -a_0 = -\frac{1}{\hat{\sigma}^2} \tag{21}$$

5 Results

In the practical part of this paper, the statistical characteristics and presented simulation techniques are compared for the data variables with experimentally declared values. These values were generated with respect on fluctuation about any mean value. Two data sets A–B were experimentally determined in MS Excel. The normal probability distribution of these variables is not required. Values of all these declared variables can demonstrate any measurable results in pedagogy (e.g. the student-knowledge level). For the demonstrating purposes, the population of respondents is divided into four general categories: "Category 1"–"Category 4" in each data set.

For each category in data sets A–B, the elementary statistics: sample average (1) and sample standard deviation (2) were computed using MS Excel, as can be seen in Table 1. For the elementary statistics between particularly categories in the concrete data set, the similarities of these values are highlighted. The similarities should be detected using hypothesis testing on the equality of mean values or standard deviations in the verification part of this chapter.

As can be seen in Table 2, the mathematical models $G_1(s)$–$G_4(s)$ were defined in corresponding with the proposed model structure (15) for each category in data sets A–B.

For the verification purposes, the concrete values from models $G_1(s)$–$G_4(s)$ are required using particularly step functions in Table 3. Their graphical interpretation can be displayed in Fig. 1 (for data set A) and in Fig. 2 (for data set B).

Table 1. Statistics for categories in data sets

Data set	Statistics	Cat. 1	Cat. 2	Cat. 3	Cat. 4
A	$\hat{\bar{x}}$	**20.49**	**20.40**	30.57	43.38
	$\hat{\sigma}^2$	**0.10**	**0.11**	0.06	0.05
B	$\hat{\bar{x}}$	**20.55**	**20.52**	**20.58**	**20.53**
	$\hat{\sigma}^2$	**0.13**	**0.14**	**0.13**	**0.13**

Table 2. Modeled categories of data sets using transfer functions $G_1(s)$–$G_4(s)$

Data set	Model 1 - $G_1(s)$ (Cat. 1)	Model 2 - $G_2(s)$ (Cat. 2)	Model 3 - $G_3(s)$ (Cat. 3)	Model 4 - $G_4(s)$ (Cat. 4)
A	$\frac{20.49}{0.1s+1}$	$\frac{20.4}{0.11s+1}$	$\frac{30.57}{0.06s+1}$	$\frac{43.38}{0.05s+1}$
B	$\frac{20.55}{0.13s+1}$	$\frac{20.52}{0.14s+1}$	$\frac{20.58}{0.13s+1}$	$\frac{20.53}{0.13s+1}$

Table 3. Modeled categories of data sets using step functions $h_1(t)$ – $h_4(t)$

Model/data set	Data set A	Data set B
Model 1 - $h_1(t)$ (Cat. 1)	$20.49\,(1-e^{-\frac{t}{0.1}})$	$20.55\,(1-e^{-\frac{t}{0.13}})$
Model 2 - $h_2(t)$ (Cat. 2)	$20.4\,(1-e^{-\frac{t}{0.11}})$	$20.52\,(1-e^{-\frac{t}{0.14}})$
Model 3 - $h_3(t)$ (Cat. 3)	$30.57(1-e^{-\frac{t}{0.06}})$	$20.58\,(1-e^{-\frac{t}{0.13}})$
Model 4 - $h_4(t)$ (Cat. 4)	$43.38\,(1-e^{-\frac{t}{0.05}})$	$20.53(1-e^{-\frac{t}{0.13}})$

As can bee seen in Tables 4 and 5, the models approximations for each category in all data sets can be verified using hypotheses testing. The similar conclusions of this hypotheses testing are required in comparison to the conclusions of the hypothesis testing on original data.

The statistical hypothesis (3) and (4) of normality, and hypothesis of equality of mean values (5)–(8) or standard deviations (9) and (10) are applied on the particularly values of each categories and values of the modeled step functions for these category.

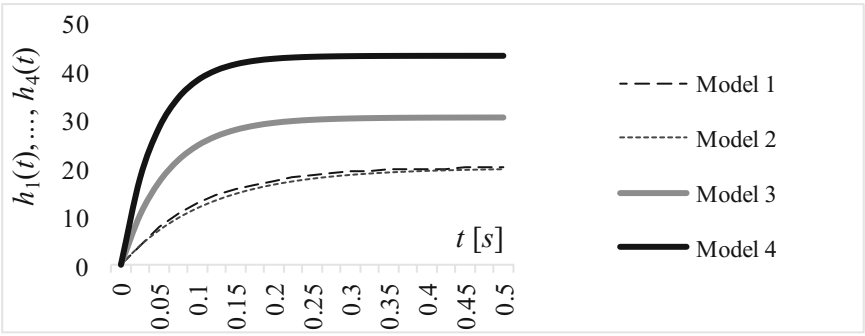


Fig. 1. Step functions for particular models in data set A

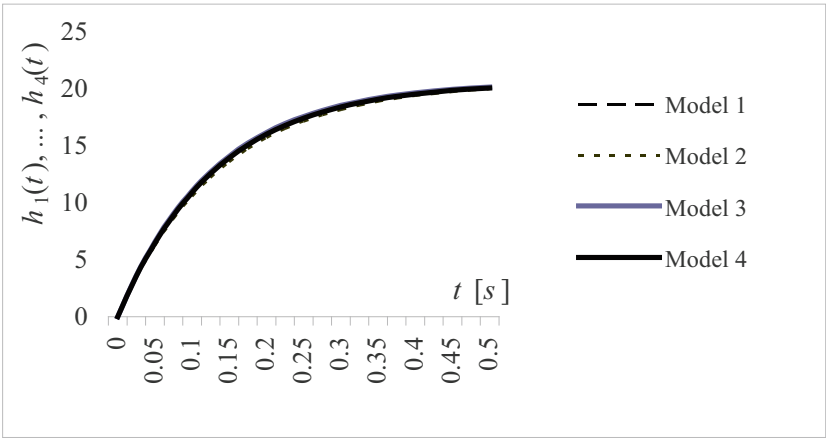


Fig. 2. Step functions for particular models in data set B

Table 4. Hypotheses testing for data set **A** (p-value: p_d) and for modeled values of particular step functions (p-value: p_m).

	Cat. 1	*Cat. 2*	*Cat. 3*	*Cat. 4*
Cat. 1	$H_0 : X_1 \sim N$ $\boldsymbol{H_1 : X_1 \not\sim N}$ $p_d = 9.2 \times 10^{-5}$ $p_m = 1.7 \times 10^{-4}$	$\boldsymbol{H_0 : \mu_1 = \mu_2}$ $H_1 : \mu_1 \neq \mu_2$ $\boldsymbol{p_d = 0.976 > 0.05}$ $\boldsymbol{p_m = 0.466 > 0.05}$	$H_0 : \mu_1 = \mu_3$ $\boldsymbol{H_1 : \mu_1 \neq \mu_3}$ $p_d = 3.4 \times 10^{-6}$ $p_m = 2.2 \times 10^{-5}$	$H_0 : \mu_1 = \mu_4$ $\boldsymbol{H_1 : \mu_1 \neq \mu_4}$ $p_d = 1.1 \times 10^{-6}$ $p_m = 2.4 \times 10^{-6}$
Cat. 2	$\boldsymbol{H_0 : \sigma_1^2 = \sigma_2^2}$ $H_1 : \sigma_1^2 \neq \sigma_2^2$ $\boldsymbol{p_d = 1 > 0.05}$ $\boldsymbol{p_m = 0.797 > 0.05}$	$H_0 : X_2 \sim N$ $\boldsymbol{H_1 : X_2 \not\sim N}$ $p_d = 2.3 \times 10^{-4}$ $p_m = 4.2 \times 10^{-4}$	$H_0 : \mu_2 = \mu_3$ $\boldsymbol{H_1 : \mu_2 \neq \mu_3}$ $p_d = 8.1 \times 10^{-6}$ $p_m = 1.8 \times 10^{-5}$	$H_0 : \mu_2 = \mu_4$ $\boldsymbol{H_1 : \mu_2 \neq \mu_4}$ $p_d = 2.9 \times 10^{-6}$ $p_m = 2.1 \times 10^{-6}$
Cat. 3	$H_0 : \sigma_1^2 = \sigma_3^2$ $\boldsymbol{H_1 : \sigma_1^2 \neq \sigma_3^2}$ $p_d = 2.9 \times 10^{-7}$ $p_m = 8.9 \times 10^{-8}$	$H_0 : \sigma_2^2 = \sigma_3^2$ $\boldsymbol{H_1 : \sigma_2^2 \neq \sigma_3^2}$ $p_d = 8.8 \times 10^{-7}$ $p_m = 8.9 \times 10^{-8}$	$H_0 : X_3 \sim N$ $\boldsymbol{H_1 : X_3 \not\sim N}$ $p_d = 5.7 \times 10^{-5}$ $p_m = 2.8 \times 10^{-6}$	$H_0 : \mu_3 = \mu_4$ $\boldsymbol{H_1 : \mu_3 \neq \mu_4}$ $p_d = 1.1 \times 10^{-6}$ $p_m = 2.5 \times 10^{-5}$
Cat. 4	$H_0 : \sigma_1^2 = \sigma_4^2$ $\boldsymbol{H_1 : \sigma_1^2 \neq \sigma_4^2}$ $p_d = 1.1 \times 10^{-7}$ $p_m = 1.3 \times 10^{-8}$	$H_0 : \sigma_2^2 = \sigma_4^2$ $\boldsymbol{H_1 : \sigma_2^2 \neq \sigma_4^2}$ $p_d = 3.9 \times 10^{-7}$ $p_m = 1.3 \times 10^{-8}$	$H_0 : \sigma_3^2 = \sigma_4^2$ $\boldsymbol{H_1 : \sigma_3^2 \neq \sigma_4^2}$ $p_d = 1.1 \times 10^{-7}$ $p_m = 8.9 \times 10^{-8}$	$H_0 : X_4 \sim N$ $\boldsymbol{H_1 : X_4 \not\sim N}$ $p_d = 5.3 \times 10^{-6}$ $p_m = 4.9 \times 10^{-7}$
Cat. 1 - 4	$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$; $\boldsymbol{H_1 : \mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4}$ $p_d = 1.1 \times 10^{-10}$; $p_m = 3.7 \times 10^{-10}$			

Table 5. Comparison of hypotheses testing for data set **B** (p-value: p_d) and for modeled values of particular step functions (p-value: p_m).

	Cat. 1	*Cat. 2*	*Cat. 3*	*Cat. 4*
Cat. 1	$H_0 : X_1 \sim N$ $\boldsymbol{H_1 : X_1 \nsim N}$ $p_d = 1.6 \times 10^{-5}$ $p_m = 1.3 \times 10^{-3}$	$\boldsymbol{H_0 : \mu_1 = \mu_2}$ $H_1 : \mu_1 \neq \mu_2$ **p_d =0.987>0.05** **p_m =0.715>0.05**	$\boldsymbol{H_0 : \mu_1 = \mu_3}$ $H_1 : \mu_1 \neq \mu_3$ **p_d =0.863>0.05** **p_m =0.81>0.05**	$\boldsymbol{H_0 : \mu_1 = \mu_4}$ $H_1 : \mu_1 \neq \mu_4$ **p_d =0.858>0.05** **p_m =0.81>0.05**
Cat. 2	$\boldsymbol{H_0 : \sigma_1^2 = \sigma_2^2}$ $H_1 : \sigma_1^2 \neq \sigma_2^2$ **p_d =1>0.05** **p_m =0.999>0.05**	$H_0 : X_2 \sim N$ $\boldsymbol{H_1 : X_2 \nsim N}$ $p_d = 7.4 \times 10^{-6}$ $p_m = 2 \times 10^{-3}$	$\boldsymbol{H_0 : \mu_1 = \mu_2}$ $H_1 : \mu_1 \neq \mu_2$ **p_d =0.831>0.05** **p_m =0.68>0.05**	$\boldsymbol{H_0 : \mu_1 = \mu_2}$ $H_1 : \mu_1 \neq \mu_2$ **p_d =0.88>0.05** **p_m =0.79>0.05**
Cat. 3	$\boldsymbol{H_0 : \sigma_1^2 = \sigma_3^2}$ $H_1 : \sigma_1^2 \neq \sigma_3^2$ **p_d =1>0.05** **p_m =1>0.05**	$\boldsymbol{H_0 : \sigma_2^2 = \sigma_3^2}$ $H_1 : \sigma_2^2 \neq \sigma_3^2$ **p_d =1>0.05** **p_m =0.999>0.05**	$H_0 : X_3 \sim N$ $\boldsymbol{H_1 : X_3 \nsim N}$ $p_d = 2.1 \times 10^{-5}$ $p_m = 1.2 \times 10^{-3}$	$\boldsymbol{H_0 : \mu_1 = \mu_2}$ $H_1 : \mu_1 \neq \mu_2$ **p_d =0.708>0.05** **p_m =0.81>0.05**
Cat. 4	$\boldsymbol{H_0 : \sigma_1^2 = \sigma_4^2}$ $H_1 : \sigma_1^2 \neq \sigma_4^2$ **p_d =1>0.05** **p_m =1>0.05**	$\boldsymbol{H_0 : \sigma_2^2 = \sigma_4^2}$ $H_1 : \sigma_2^2 \neq \sigma_4^2$ **p_d =1>0.05** **p_m =1 >0.05**	$\boldsymbol{H_0 : \sigma_3^2 = \sigma_4^2}$ $H_1 : \sigma_3^2 \neq \sigma_4^2$ **p_d =1>0.05** **p_m =1>0.05**	$H_0 : X_4 \sim N$ $\boldsymbol{H_1 : X_4 \nsim N}$ $p_d = 5.4 \times 10^{-6}$ $p_m = 1.4 \times 10^{-3}$
Cat. 1 - 4	$\boldsymbol{H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4}$; $H_1 : \mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4$ **p_d = 0.984>0.05**; **p_m = 0.963>0.05**			

The hypothesis testing should provide the similar conclusions. All these calculations were provided using PAST v2.17 software [14].

For the original values of each category in concrete data set, the significance value is assigned as p_d in the hypotheses tests. The significance value p_m is used for models in form of particular step function.

In Tables 4 and 5, a gray background of cells corresponds with highlighting of similar elementary statistics in Table 1.

As can be seen in Table 4, values of the original data and values of particular step functions has not a normal probability distribution on the significance level $\alpha = 0.05$. For the means values testing for original data and step functions, the Mann-Whitney U-test resp. Kruskal-Wallis test were used with conclusions in this table. The non-rejected hypotheses are highlighted using bold font. Using Kolmogorov-Smirnov test, the hypothesis testing on the similarity of standard deviations is presented in the same form. All conclusions of hypotheses testing are stated on the significance level $\alpha = 0.05$.

These conclusions respect the similarities of the elementary statistics with highlighting in Table 1. The same results were achieved for the original data of categories and for its modeled step-function values.

As can be seen in Table 5, values of original data and values of particular step functions have not a normal probability distribution on the significance level $\alpha = 0.05$. For the testing of means values, the Mann-Whitney U-test resp. Kruskal-Wallis test were applied. The hypothesis testing on the equality of standard deviations was provided using Kolmogorov-Smirnov test. All conclusions of hypotheses testing are stated on the significance level $\alpha = 0.05$.

For all data sets, the differences between statistical properties of declared variables and their assigned modeled behaviour were compared. Classical statistical analyses and proposed simulations of model dynamics achieved the same characteristic properties verified using statistical methods for hypothesis testing. The same results of normality testing and same conclusions of hypotheses testing for comparison of mean values or standard deviations were experimentally demonstrated.

6 Conclusion

In this paper, the simulation conclusions verified, that mathematical models in form of continuous transfer function can be used for process modeling in the pedagogical cybernetics. The connection between the technical and pedagogical cybernetics was presented using the approach of process modeling based on control-system theory. In the classical approaches in the pedagogy, the research data can be analyzed using the statistical methods based on hypotheses testing. In this paper, an alternative possibility for modeling of pedagogical process was presented with respect to the statistical properties of its original researched data. The main parameters for model construction were sample average and sample standard deviation of the declared variable in the pedagogical research. The comparisons between statistical properties of declared variable and its assigned modeled behaviour were verified. Particularly, data of declared variables and values of step responses of its dynamic models were statistically compared using normality testing, hypotheses testing on equality of mean values or standard deviations. Proposed approach achieved the required statistical properties and hypotheses conclusions in the experimentally example of process modeling in the pedagogical-cybernetics research. In this area, the proposed methodology is not widely applied.

References

1. Gushchin, A., Divakova, M.: Trend of e-education in the context of cybernetics laws. Procedia Soc. Behav. Sci. **214**, 890–896 (2015). Elsevier, ISSN 1877-0428
2. Cevik, Y.D., Haslaman, T., Celik, S.: The effect of peer assessment on problem solving skills of prospective teachers supported by online learning activities. Stud. Educ. Eval. **44**, 23–35 (2015). Elsevier, ISSN 0191-491X

3. Granic, A., Mifsud, Ch., Cukusic, M.: Design, implementation and validation of a Europe-wide pedagogical framework for e-Learning. Comput. Educ. **53**, 1052–1081 (2009). Elsevier, ISSN 0360-1315
4. Kucera, V.: Analysis and Design of Discrete Linear Control Systems. Nakladatelstvi Ceskoslovenske akademie ved, Praha (1991). ISBN 80-200-0252-9
5. Datta, B.N.: Numerical methods for linear control systems: design and analysis. Elsevier Academic Press, Amsterdam (2004). ISBN 0-12-203590-9
6. Cortes, J., Casals, M., Langohr, M., et al.: Importance of statistical power and hypothesis in P value. Med. Clin. **146**(4), 178–181 (2016). Elsevier, ISSN 0025-7753
7. Alizadeh Noughabi, H.: Two powerful tests for normality. Ann. Data Sci. **3**(2), 225–234 (2016). Springer, ISSN 2198-5812
8. Rasch, D.: The two-sample t test: pre-testing its assumptions does not pay off. Stat. Pap. **52** (1), 219–231 (2011). Springer, ISSN 1613-9798
9. Mewhort, D.J.K.: A comparison of the randomization test with the F test when error is skewed. Behav. Res. Methods **37**(3), 426–435 (2005). Springer, ISSN 1554-3528
10. Lazic, S.E.: Why we should use simpler models if the data allow this: relevance for ANOVA designs in experimental biology. BMC Physiol. **8**(18), 16 (2008). BioMed Central, ISSN 1472-6793
11. Fischer, D., Oja, H.: Mann-Whitney type tests for microarray experiments: the R package gMWT. J. Stat. Softw. **65**(1), 1–19 (2015). Foundation for Open Access Statistics, ISSN 1548-7660
12. Bradley, D.R., Senko, M.W., Stewart, F.A.: Statistical simulation on microcomputers. Behav. Res. Methods Instruments Comput. **22**(2), 236–246 (1990). Springer, ISSN 1532-5970
13. Kitchenham, B., Madeyski, L., Budgen, D., et al.: Robust statistical methods for empirical software engineering. Empir. Softw. Eng., 1–52 (2016). Springer, ISSN 1573-7616
14. Hammer, O., Harper, D.A.T., Ryan, P.D.: PAST: Paleontological statistics software package for education and data analysis. Palaeontologia Electronica **4**(1) (2001). Coquina Press, ISSN 1094-8074. http://palaeo-electronica.org/2001_1/past/issue1_01.htm

Calibration of Low-Cost Three Axis Magnetometer with Differential Evolution

Ales Kuncar(✉), Martin Sysel, and Tomas Urbanek

Faculty of Applied Informatics, Tomas Bata University in Zlin,
Namesti T.G. Masaryka 5555, 76001 Zlin, Czech Republic
{kuncar,sysel,turbanek}@fai.utb.cz

Abstract. The magnetometers are used in wide range of engineering applications. However, the accuracy of magnetometer readings is influenced by many factors such as sensor errors (scale factors, non-orthogonality, and offsets), and magnetic deviations (soft-iron and hard-iron interference); therefore, the magnetic calibration of magnetometer is necessary before its use in specific applications. This research paper describes calibration method for three axis low-cost MEMS (Micro-Electro-Mechanical Systems) magnetometer. The calibration method uses differential evolution (DE) algorithm for the determination of the transformation matrix (scale factor, misalignment error, and soft iron interference) and bias offset (hard-iron interference). The performance of this method is analysed in experiment on three axis low-cost magnetometer LSM303DLHC and then compared to the traditional method (least square ellipsoid fitting method). The magnetometer readings were obtained while rotating the sensor around arbitrary rotations. The experimental results show that the calibration error is least using DE.

Keywords: Calibration · Differential Evolution · Magnetometer · MEMS

1 Introduction

In last decade, the advances in Micro-Electro-Mechanical System (MEMS) technologies have made a great role in many engineering applications. Magnetometers have been widely used in many areas such as geophysical research [1], military defence, mineral resources, drilling, mining practice [2], navigation, and localization [3–5]. The magnetometers provide information about the strength and direction of the local magnetic field. The measured magnetic field is a combination of the Earth's geomagnetic field and a magnetic field generated by nearby objects. Nevertheless, the main problem is errors such as zero deviation, scale factors, non-orthogonality, measurement noise, misalignment error, and hard-iron and soft-iron interferences. The biggest effect on the magnetometer readings comes from soft-iron and hard-iron interferences in the vicinity of the sensors. These errors lead to deviation between the true and measured value. In absence

R. Silhavy et al. (eds.), *Cybernetics and Mathematics Applications in Intelligent Systems*,
Advances in Intelligent Systems and Computing 574, DOI 10.1007/978-3-319-57264-2_12

of magnetic interferences, the magnetometer will measure only the three components of the geomagnetic field. The vector magnitude is equal to the magnitude of geomagnetic field at a different orientation. The locus of the magnetometer readings lies on the surface of the sphere centred in zero field and the radius is equal to the magnitude of the geomagnetic field.

The hard-iron interference adds fixed offset in each axis to all measurements. That results in displacement of the sphere centred to the hard-iron offset; however, the geomagnetic field strength is still the same. The locus of the magnetometer measurements is distorted in each axis differently to a 3D ellipsoid in presence of soft-iron interference (Fig. 1). Before each application, a calibration and compensation of such errors needs to be conducted.

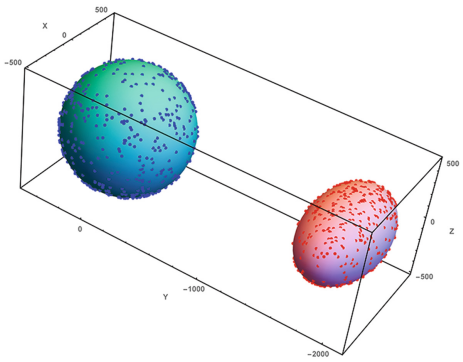


Fig. 1. Calibrated (sphere) and raw magnetometer (ellipsoid) data.

Many researchers deals with the calibration of low-cost inertial magnetometers. Guo et al. [6] used Extended Kalman filter to compensate the soft-iron and hard-iron interference. Crassidis et al. [7] then compared performance of Extended Kalman filter and proposed Unscented filter based on Kalman filter. The results showed that the Unscented filter has better performance than the Extended KF. Kok et al. [8] presented a calibration algorithm using a maximum likelihood method and a additional inertial sensor. Renaudin et al. [9] applied an adaptive least square estimator to ellipsoid fitting problem. Tabatabaei et al. [10] proposed a novel iterative calibration algorithm for three axis magnetometers which outperforms conventional ellipsoid fitting method in accuracy and reliability. Liu et al. [11] describes calibration method using turntable and ellipsoid fitting method. Cheuk et al. [12] used an evolutionary algorithm to minimize misalignment, scale and bias errors. Similar research were conducted on calibration of 9DOF sensor board by Sarcevic et al. [13].

These calibration methods can be divided up into these groups:

- The method involves coil system (Helmholtz coils) for precise calibration. The calibration parameters are determined by measuring the current in the coils.
- Specific highly controlled equipment like turntable, accelerometers, gyroscopes or GPS (Global Positioning System) is used to accurately control the direction of magn etometer sensitive axis.
- Calibration method known as swinging method. This method collects magnetometer readings while rotating the sensor around known sensitive axis.
- IGRF (International Geomagnetic Reference Field) model provides magnetic field magnitude.

The aim of this paper is to use differential evolution algorithm as a calibration method for low-cost three axis magnetometer. This method is furthermore compared with the traditional least square ellipsoid fitting method which will account magnetometer sensor errors and magnetic interferences.

The reminder of this paper is organized as follows. In Sect. 2, the differential evolution is introduced. The magnetometer error model and hardware and software for data collection are briefly described in Sects. 3 and 4, respectively. The experimental results of magnetometer calibration are mentioned in Sects. 5 and 6.

2 Differential Evolution

Differential evolution is an optimization algorithm for heuristic search of function minimums introduced by R. Storn and K. Price in 1995 [14]. This optimization method is an evolutionary algorithm based on population, mutation and recombination. Differential evolution uses only four parameters which need to be set; therefore, it can be easily implemented. The parameters are Generations, NP, F and CR [15].

- **Generations** (Number of iterations) specifies the number of evolutionary cycles (generations) during which the entire population develops.
- **NP** (Number of population members) is a parameter which gives the size of the population. The value of this parameter cannot be lower than 4 because it is the minimum size at which the differential evolution algorithm still works.
- **CR** (Crossover probability). This parameter is a small value in range from 0 to 1. In case of separable function, this value is set close to 0 (clean copy of the fourth parent). Otherwise, it is set to the values close to 1 (random search).
- **F** (Mutation constant) is the last control parameter for differential evolution and its value ranges from 0 to 2.

The flow chart of differential evolution algorithm for the magnetometer calibration parameters estimation is shown in Fig. 2.

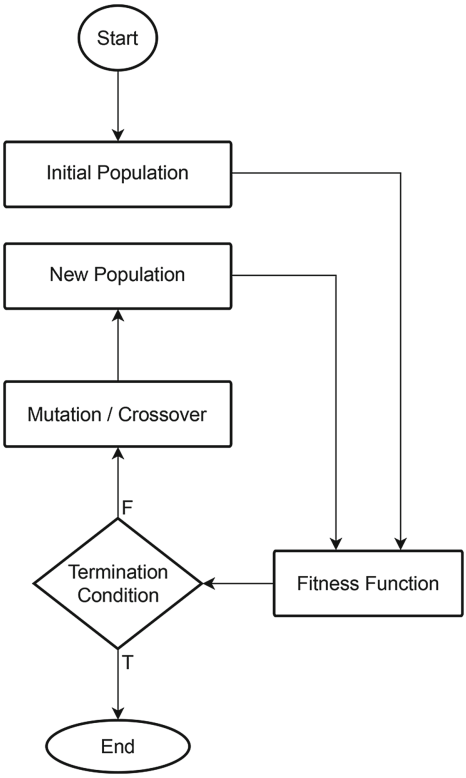


Fig. 2. Flow-chart of DE for parameters estimation.

3 Magnetometer Error Model

The magnetometer readings are influenced by many sources of error like wide-band measurement noise, stochastic biases, installation errors and magnetic interferences in the vicinity of the sensors. These magnetic interferences can be divided up into two groups: soft-iron and hard-iron interference. The hard-iron interference is caused by the presence of magnets or materials generating fixed or slightly time-varying magnetic field. The second type, soft-iron interference, occurs when a ferromagnetic materials is in the vicinity of the sensor or it can be even generated by the device itself. This will cause the distortion of magnetic field. The traditional method for compensation of such errors is equivalent to transforming the 3-D ellipsoid to the centre oriented sphere.

The magnetometer error model is,

$$R = M_m \cdot S \cdot SI \cdot (M + O + n) \tag{1}$$

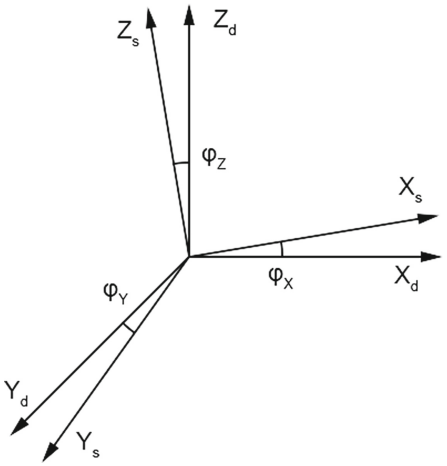


Fig. 3. Misalignment error

In this model, the variables M_m, S and SI are matrices which interpret misalignment errors, scale factors and soft-iron biases, respectively. O and n are vectors representing hard-iron biases and wideband noise which distorts the true magnetic field measurements M (Figs. 3 and 4).

1. **Misalignment error** is defined as angles between the magnetometer axis X_s, Y_s, Z_s and the device body axis X_d, Y_d, Z_d. This caused by imperfect mounting of sensor on the PCB (printed circuit board).

$$M_m = \begin{bmatrix} 1 & m_{xy} & m_{xz} \\ m_{yx} & 1 & m_{yz} \\ m_{zx} & m_{zy} & 1 \end{bmatrix} \tag{2}$$

2. **Scale factor error** corresponds to constants of proportional relationship between the input and output of the magnetometer. The scale factor can be modelled as

$$S = diag\begin{pmatrix} s_x & s_y & s_z \end{pmatrix} \tag{3}$$

3. **Soft-Iron** error can be modelled as

$$SI = \begin{bmatrix} SI_{xx} & SI_{xy} & SI_{xz} \\ SI_{yx} & SI_{yy} & SI_{yz} \\ SI_{zx} & SI_{zy} & SI_{zz} \end{bmatrix} \tag{4}$$

4. **Hard-Iron** is equivalent to a bias and can be represented as

$$O = \begin{bmatrix} O_x & O_y & O_z \end{bmatrix}^T \tag{5}$$

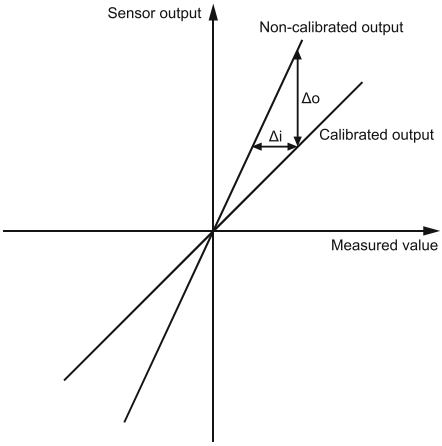


Fig. 4. Scale factor error

4 Equipment

The experimental measurement chain (Fig. 5) includes control unit, inertial measurement unit (IMU) and software for data collection.

The control unit STEVAL-MKI109V2 is built up to provide platform for the evaluation of the MEMS modules. These modules can be connected via 24-pin expansion connector.

Furthermore, the unit consists high-performance 32-bit microcontroller STM32F103RET6, which is based on ARM technology, with 512 kB flash memory functioning as a bridge between the MEMS modules and a graphical user interface (GUI) or dedicated software routines for customized applications.

To provide measurements, 10 axis inertial measurement unit STEVAL-MKI124V1 is connected to the control unit. The IMU includes three axis gyroscope with internal thermometer (L3GD20), three axis accelerometer and three axis magnetometer (LSM303DLHC), and barometer (LPS331AP). All these sensors are based on MEMS technology and they are factory tested, and trimmed. However, this factory calibration is appropriate only for basic applications. Advanced calibration had to be provided for application such as navigation systems.

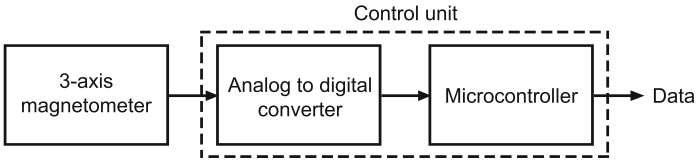


Fig. 5. Measurement chain.

Table 1. Magnetometer characteristics

Parameters	Values
Full scale	±1.3–±8.1 gauss
Sensitivity	230–1100 LSB/gauss
Cross-axis	±1 %FS/gauss
Output data rate	0.75 Hz–220 Hz

Several different configurations allow for settings regarding specific usage. Sensor specifications are given in Table 1 and [16].

To collect measured data, a PC is connected to the control unit using virtual serial port. On the PC, the drivers for interaction and configuration of sensors are installed. This software is called Unico STSW-MKI109W.

The collected data are processed in Wolfram Mathematica 10 and then the differential evolution algorithm, programmed in Lua language, is applied.

5 Magnetometer Calibration

In order to determine the calibration parameters of low-cost three axis magnetometer, sufficient number of samples is needed in different directions. Therefore the data were measured while rotating the sensor around each sensitive axis and also in arbitrary rotations. Duplicate readings were deleted from the dataset.

The average of squared errors have been used as the fitness function where the error is the difference between the calculated output from current parameters and the true value. Therefore, the fitness function of the DE algorithm is modelled as

$$F = \sum_{i=1}^{n} \left(\sqrt{(X)^2 + (Y)^2 + (Z)^2} - R \right)^2 \tag{6}$$

where X, Y, and Z are calibrated values and R is the true scalar value of geomagnetic field intensity taken from IGRF [17] due to the absence of proton magnetometer.

The calibrated values account for bias offset (O_X, O_Y, O_Z), scale factor (S_X, S_Y, S_Z), and misalignment error (α, β, γ). The equations for calculations of such errors are showed in (7), (8) and (9)

$$X = (R_X - O_X) \cdot S_X \tag{7}$$

$$Y = (R_X - O_X) \cdot \alpha + (R_Y - O_Y) \cdot S_Y \tag{8}$$

$$Z = (R_X - O_X) \cdot \beta + (R_Y - O_Y) \cdot \gamma + (R_Z - O_Z) \cdot S_Z \tag{9}$$

The model for calculation of misalignment errors is depicted in Fig. 6.

Table 2 shows the set-up of differential evolution. The best set-up of DE is the subject of further research.

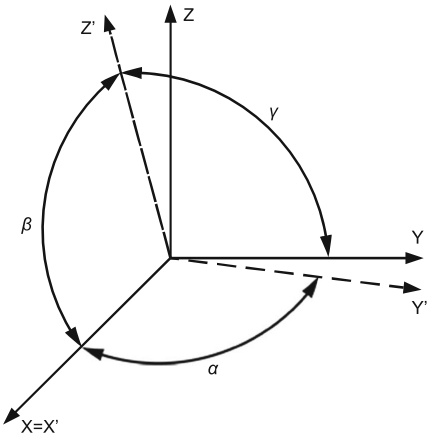


Fig. 6. Model for misalignment error.

Table 2. Set-up of differential evolution.

Parameter	Value
NP	90
Generations	100
F	0.3
CR	0.5

6 Experimental Results

The goal of this experiment is to analyse the performance of calibration method using DE algorithm and than compare it to traditional method. The true magnitude of the geomagnetic field in Zlin, Czech Republic (49.2306827° N, 17.6566617° E) is 48,996 nT which is 489.96 mGauss (taken from IGRF).

As you can see in Fig. 7, the measured data before calibration shows signs of presence of the hard-iron interference. The effect of the soft-iron interference is very small so it can be ignored; however, it was also accounted for.

Figure 8 shows the results after using DE as calibration technique. The magnitude of magnetic field is approximately 490 mGauss because we set this value in one of the parameters in the fitness function. The transformation matrix and bias offset is

$$\begin{bmatrix} 1.02 & 0 & 0 \\ -0.036 & 1.103 & 0 \\ -0.012 & -0.008 & 1.06 \end{bmatrix} \tag{10}$$

The offset caused by the presence of metal shielding of the IMU is

$$\begin{bmatrix} -47.664 & -229.297 & -65.051 \end{bmatrix}^T \tag{11}$$

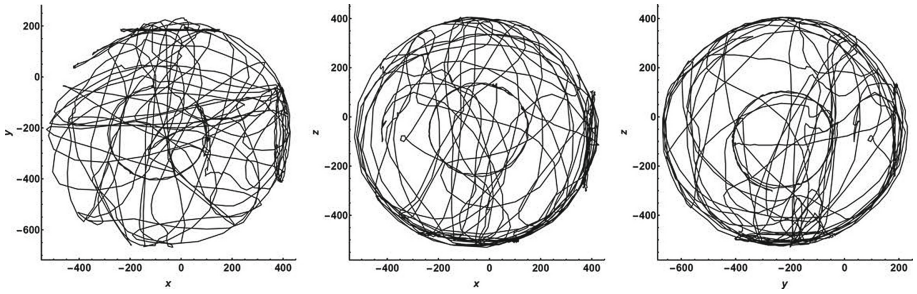


Fig. 7. Slices of raw magnetometer measurements.

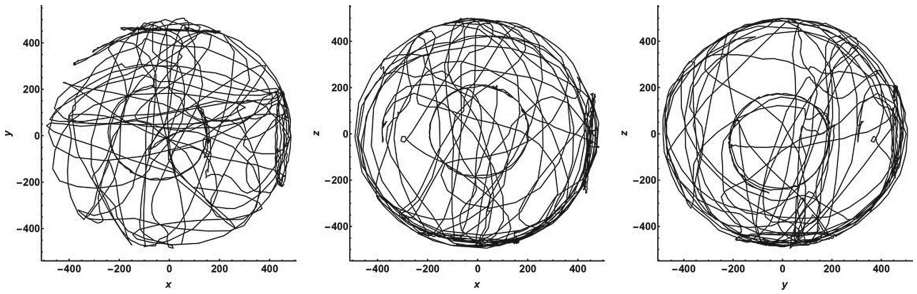


Fig. 8. Slices of calibrated magnetometer measurements after the use of DE.

7 Discussion

For the evaluation of performance, the root mean square error (RMSE) was used. The smaller is value of RMSE, the better is performance of calibration method.

$$RMSE = \sqrt{\frac{1}{N} \cdot \sum_{i=1}^{N} (x_i - \hat{x_i})^2}, \tag{12}$$

where N is equal to the number of samples, $\hat{x_i}$ is measured magnitude of magnetic field, and x_i is true magnitude of magnetic field.

The scalar error is listed in Table 3. The RMSE is the least with DE and it is two times better than ellipsoid fitting method.

Table 3. RMSE using least square ellipsoid fitting method and DE.

Before calibration	Ellipsoid fitting method	DE
30.13	17.91	8.24

Furthermore, Welch's t-test was conducted to provide statistical evidence that the calibration method with DE has lower error than the ellipsoid fitting method.

Welch's t-test
data: EllipsoidFitting and DE
t = 162.0209, df = 36543.34, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
14.71383 15.07419
sample estimates:
mean of x mean of y
505.1649 490.2709

As can be seen, the p-value ot the t-test is lower than $2.2 \cdot 10^{-16}$. Therefore, the null hypothesis that the ellipsoid fitting method and differential evolution have the same true difference in means is rejected. That means, that the alternative hypothesis is accepted; therefore, the DE could have lower error than the ellipsoid fitting method.

8 Conclusion

In applications, the effects of the hard-iron and soft-iron interference will distort the local magnetic field especially in low-cost magnetometers. These interferences needs to be accounted for and removed from the magnetometer measurements. Therefore, this paper describes calibration methods for low-cost three axis magnetometer. This method does not require any additional equipment for data measurement. The proposed calibration method uses DE algorithm for estimation of calibration parameters. The performance is then compared with traditional ellipsoid fitting method. The comparison shows that the DE algorithm has lower error.

In future work, the performance of DE algorithm will be tested for different data sets and also the best set-up of its parameters.

Acknowledgments. This work was supported by Internal Grant Agency of Tomas Bata University in Zlin under the project No. IGA/FAI/2017/007.

References

1. Kawai, J., Uehara, G., Kohrin, T., Ogata, H., Kado, H.: Three axis SQUID magnetometer for low-frequency geophysical applications. IEEE Trans. Magn. **35**(5), 3974–3976 (1999)
2. Haverinen, J., Kemppainen, A.: A geomagnetic field based positioning technique for underground mines. In: 2011 IEEE International Symposium on Robotic and Sensors Environments (ROSE), pp. 7–12. IEEE, September 2011

3. Ashkar, R., Romanovas, M., Goridko, V., Schwaab, M., Traechtler, M., Manoli, Y.: A low-cost shoe-mounted inertial navigation system with magnetic disturbance compensation. In: 2013 International Conference on Indoor Positioning and Indoor Navigation, IPIN 2013 (2013)
4. Bird, J., Arden, D.: Indoor navigation with foot-mounted strapdown inertial navigation and magnetic sensors [emerging opportunities for localization and tracking]. IEEE Wireless Commun. **18**, 28–35 (2011)
5. Glanzer, G., Walder, U.: Self-contained indoor pedestrian navigation by means of human motion analysis and magnetic field mapping. In: Proceedings of the 2010 7th Workshop on Positioning, Navigation and Communication, WPNC 2010, pp. 303–307 (2010)
6. Guo, P., Qiu, H., Yang, Y., Ren, Z.: The soft iron and hard iron calibration method using extended kalman filter for attitude and heading reference system. In: 2008 IEEE/ION Position, Location and Navigation Symposium, pp. 1167–1174. IEEE (2008)
7. Crassidis, J.L., Lai, K.-L., Harman, R.R.: Real-time attitude-independent three-axis magnetometer calibration. J. Guidance, Control Dyn. **28**(1), 115–120 (2005)
8. Kok, M., Hol, J., Schon, T., Gustafsson, F., Luinge, H.: Calibration of a magnetometer in combination with inertial sensors. In: 2012 15th International Conference on Information Fusion (FUSION), pp. 787–793 (2012)
9. Renaudin, V., Afzal, M.H., Lachapelle, G.: Complete triaxis magnetometer calibration in the magnetic domain. J. Sens. **2010**, 1–10 (2010)
10. Tabatabaei, S.A.H., Gluhak, A., Tafazolli, R.: A fast calibration method for triaxial magnetometers. IEEE Trans. Instrum. Meas. **62**(11), 2929–2937 (2013)
11. Liu, Y., Li, X., Zhang, X., Feng, Y.: Novel calibration algorithm for a three-axis strapdown magnetometer. Sensors **14**(5), 8485–8504 (2014)
12. Cheuk, C.M., Lau, T.K., Lin, K.W., Liu, Y.: Automatic calibration for inertial measurement unit. In: 2012 12th International Conference on Control Automation Robotics and Vision (ICARCV), pp. 1341–1346. IEEE, December 2012
13. Sarcevic, P., Pletl, S., Kincses, Z.: Evolutionary algorithm based 9DOF sensor board calibration. In: 2014 IEEE 12th International Symposium on Intelligent Systems and Informatics (SISY), pp. 187–192. IEEE, September 2014
14. Storn, R., Price, K.: Differential evolution - a simple and efficient adaptive scheme for global optimization over continuous spaces, Technical report (1995)
15. Storn, R.: On the usage of differential evolution for function optimization. In: Proceedings of North American Fuzzy Information Processing, pp. 519–523. IEEE (1996)
16. STMicroelectronics, Data brief: STEVAL-MKI124V1, p. 4 (2013)
17. National Centers of Environmental Information (2016)

The Technique of Multi-criteria Decision-Making in the Study of Semi-structured Problems

Alexander N. Pavlov[1,2(✉)], Dmitry A. Pavlov[2], Alexey A. Pavlov[2], and Alexey A. Slin'ko[2]

[1] Volga State University of Technology, Yoshkar-Ola, Russia
Pavlov62@list.ru
[2] Mozhaisky Military Space Academy, St. Petersburg, Russia

Abstract. In the article it is proposed to use additional information from the decision maker (DM) for removing the criteria of uncertainty when making decisions in the framework of semi-structured problems, which is characterized by incomplete information, numerous qualitative and the quantitative selection criteria. This information is represented by the production models and processed by using the methods of the experiment planning theory and parametric fuzzy measures. The essence of the proposed methodology consists of sharing the ideas of verbal analysis of the decisions (simple and complex basic situation of a survey) and procedures of bringing data qualitative indicators to the quantitative ones, which is based on using the mathematical apparatus of the theory of fuzzy sets, relations and measures, and the theory of experiment planning. A parametric fuzzy measure has been constructed in order to reduce the number of calls to the DM in the process of the expert survey and the consistency control of his statements in the set of the production rules that represent basic situation of the survey. This parametric fuzzy measure allows computing the DM's preferences on criteria for achieving the goal set for making the management decisions.

Keywords: Criteria uncertainty · Production model · Reference situation of survey · Theory of experiment planning · Fuzzy measure

1 Introduction

The study of semi-structured problems is carried out in the conditions of incomplete information, lack of knowledge about the behavior of a complex object, multi-purpose, multi-functional nature of the management tasks for the object, the impact of external and internal factors and other reasons. When making decisions in these situations it is necessary to analyze alternative solutions and take into account various factors of uncertainty and incompleteness of available information. It should be noted that it is impractical for a number of criteria in obtaining the accurate quantitative descriptions because of limited financial and time resources, the uniqueness of the solvable problem. Therefore it is used quality estimates obtained from the experts. The real problems of multi-criteria decision-making [1–3], which arising in practice, are extremely diverse, but they all share a general scheme of finding a solution, the essence consists in the creating a

R. Silhavy et al. (eds.), *Cybernetics and Mathematics Applications in Intelligent Systems*,
Advances in Intelligent Systems and Computing 574, DOI 10.1007/978-3-319-57264-2_13

set of procedures carried out on the set of alternatives, which is produced many rational solutions.

It is known [6, 14], that the problem of specification (which is detailed by bringing more qualitative and quantitative information about the properties of criterion functions, on alternatives, for the optimality principles, etc.) is the basis of multi-criteria choice methods. The main source of additional information when searching for the best alternatives are the experts, who know a given subject area, and decision-makers, who is trying to pursue a particular objective (objectives), in order to achieve that and solved the problem. To date a wide variety of methods for solving problems of multi-criteria choice [1, 4–9, 14] are developed. Various principles and features can be offered to the classification of these methods. For example, in the paper [9] it is proposed to distinguish between classes of a priori, a posteriori, and adaptive methods and models of multi-criteria optimization. There is a polynomial (scalarization) among the priori methods of construction methods. We differentiate between heuristic and axiomatic convolution. Another group of the priori methods for solving multi-criteria problems is based on the building components of the resulting preference relations. We differentiate between pareto's, lexicographical, resulting majority preference relation, the first two of which are divided into classic, interval and threshold, and most on without an interval and interval resulting relationship preferences. Each alternative from a finite set of the non-dominated solutions with the relevant decision-makers' preferences can be the best. However lexicographical optimization and Pareto`s dominance is characterized by a minimum volume of the DM preferences, which entails the use of an incomplete selection criteria and the construction of only a partial order on the set of feasible alternatives [1]. To overcome these shortcomings should increase the DM preferences. It should be noted that it is need take into account the possibilities and limitations of human information processing when developing of regulatory decision-making methods.

The normative methods of solving multi-criteria choice on types of information collected and used by the DM, when evaluating alternatives, can be classified as follows:

- the methods based on quantitative dimensions [3, 5, 14];
- the methods based on primary quantitative measurements, the results of which immediately transferred in quantitative form [10, 11];
- the methods based on quantitative measurements, but using several indicators when comparing alternatives [9, 14];
- the methods based on qualitative measurements without any transition to quantitative variables [4, 6].

When choosing specific methods of multi-criteria evaluation [6, 9], it is advisable to comply with the following requirements:

Requirement 1. Completeness and acyclic (transitive) relationship on the set of multi-criterion alternatives.

Requirement 2. The requirement in the methods of decision-making must be provided to verify the information of decision-makers and experts on the consistency. Low sensitivity to human error.

Requirement 3. Any assumptions about the kind of decision rule must be mathematically and psychologically justified.

Requirement 4. Decision-making methods must be used only such means of obtaining information from DM and experts, which correspond to the possibilities of human information processing system.

The analysis of [6] matching the four groups of standard methods, the first two requirements have shown that these methods are not, in general, can simultaneously ensure the completeness of comparisons of alternatives, to provide a linear order, to be rational and insensitive to human measurement errors, leading to not of the transitive relations on the set of the compared solutions.

In this paper the problem of multi-criteria decision making in situations where alternatives are not known or partially known at the time of decision-making, and can also appear in the decision-making process. When the alternatives are evaluated linguistically (verbally) the specified performance indicators characterized the task as not structured. There are two ways to deal with such poorly structured tasks. The first way is to describe the qualitative indicators with quantitative indicators constructed in a special way (ratings, fuzzy numbers, and linguistic variables). It is believed that the use of entirely new mathematical techniques such as fuzzy sets theory, relations and actions, fuzzy integration allows you to effectively formalize and solve semi-structured tasks. The second way is to use methods of verbal (ordinal) decision analysis (ZAPROS I, II, III, and a few others) [6], which are based on the unified scale of changes in quality on the set of values of all criteria and application of so-called anchor (utopian or perfect solution, and the opposite of the decision). It is believed that any procedure qualitative to quantitative information is incorrect and there is no reason to relies on quantitative results.

The methodology of multi-criteria decision-making, which is based on a production model of the decision maker's preference for describing simple and complex reference situations of the survey, data processing knowledge methods of the theory of fuzzy measures [10, 11, 15] and the theory of experiment planning [12, 13] and the verification of statements of decision-maker for consistency.

2 The Technique of Multi-criteria Decision-Making

Let a set of alternatives evaluated a set of performance indicators $F = \{F_1, F_2, \ldots, F_m\}$, each of which represents a linguistic variable. For example, the linguistic variable F_i = «Payback Period» can take values from the set of simple and compound terms $T(F_i) = \{$«low», «below average», «average», «above average», «high»$\}$. For qualitative interpretation of the resulting index will be used the linguistic variable "Effective Solutions", which can take the values $T(F_{res}) = \{$«bad», «below average», «average», «above average», «good»$\}$. In the most general form of knowledge of decision-makers on the relationship private performance $F = \{F_1, F_2, \ldots, F_m\}$ with the resulting index F_{res} can be represented production models of the form:

P_j: «IF $F_1 = A_{1j}$ and $F_2 = A_{2j}$ and … and $F_m = A_{mj}$, THEN $F_{res} = A_{jres}$», where $A_{ij} \in T(F_i)$, $A_{jres} \in T(F_{res})$ are the values of the respective linguistic variables. As a common scale for all values of the indicators used by the bipolar scale $[-1, 0, +1]$, and the values can be set using fuzzy numbers (L-R) like (Fig. 1).

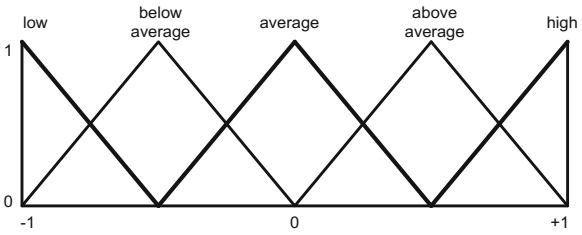


Fig. 1. The values of the linguistic variable in scale [−1, +1]

In accordance with the method of solving the problem of multi-criteria evaluation is proposed in the works [12, 13], extreme («minimum» and «maximum») values of the linguistic variable F_i scale labeled «−1» and «+1» and to build the result indicator, in accordance with the provisions of the theory of planning experiment, form the orthogonal plan of the expert survey, elements of which are extreme marked private values of the performance indicators $\{F_1, F_2, \ldots, F_m\}$. An example of the orthogonal plan of the expert survey for three private performance indicators is presented in Table 1.

Table 1. The orthogonal plan of the expert survey

F_0	F_1	F_2	F_3	F_1F_2	F_1F_3	F_2F_3	$F_1F_2F_3$	F_{res}
1	−1	−1	−1	1	1	1	−1	A_{1res}
1	1	−1	−1	−1	−1	1	1	A_{2res}
1	−1	1	−1	−1	1	−1	1	A_{1res}
1	1	1	−1	1	−1	−1	−1	A_{3res}
1	−1	−1	1	1	−1	−1	1	A_{2res}
1	1	−1	1	−1	1	−1	−1	A_{4res}
1	−1	1	1	−1	−1	1	−1	A_{3res}
1	1	1	1	1	1	1	1	A_{5res}
λ_0	λ_1	λ_2	λ_3	λ_{12}	λ_{13}	λ_{23}	λ_{123}	

In Table 1 the values of the terms of the linguistic variable F_{res} result performance indicator can be represented by triangular fuzzy numbers (Fig. 2). Then, for example, in the second row of the table is presented following expert judgment “If the indicator F_1 is set to «high», the indicator F_2 is set to «low», the indicator F_3 is set to «low», the resulting indicator F_{res} is estimated as «below average». And production itself generally is seen as supporting the situation when holding the expert survey.

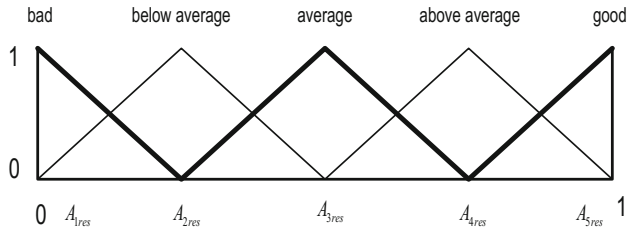


Fig. 2. The scale of the resulting indicator

The calculation of the resultant index coefficients $F_{res} = \lambda_0 + \sum_{i=1}^{m} \lambda_i F_i + \sum_{i=1}^{m} \sum_{\substack{j=1 \\ j \neq i}}^{m} \lambda_{ij} F_i F_j + \ldots + \lambda_{12\ldots m} F_1 F_2 \ldots F_m$ that take into account the effect of a private individual indicators and the impact of a set of two, three and so on indicators is carried out according to the rules which are adopted in the experiment planning theory. For this calculated averaged scalar product of the corresponding columns of an orthogonal matrix (Table 1) by the vector of the resulting performance indicator values. For example, the coefficient λ_2 value is calculated as follows:

$$\lambda_2 = \frac{-A_{1res} - A_{2res} + A_{1res} + A_{3res} - A_{2res} - A_{4res} + A_{3res} + A_{5res}}{8}.$$

Thus, the proposed technique of multi-criteria decision making consists of the following steps.

Step 1. The formation of many linguistic scales for each of the partial indicators and the resulting index of efficiency of the decisions. Transfer individual results to the scale $[-1, +1]$.

Step 2. The construction of orthogonal plan of the expert survey and the equal survey (answers to the questions of production rules).

Step 3. The build result indicator of the effectiveness of the decisions.

Among these the most important steps and responsible ***step 2*** is associated with obtaining expert answers to the questions which are contained in the production rules. On the one hand, this is because, for example, when the number of a particular performance number of questions asked 4 increases and becomes more than 16, which usually leads to inconsistencies in statements mean expert features of human thinking. These characteristics are reflected in the patterns derived by J. Miller; the essence of which lies in the fact that short-term human memory of the expert cannot remember and repeat more than 7 ± 2 elements. On the other hand, according to the Ellsberg paradox, a person (expert) does not think additive that requires evaluation of his answers do not use an additive (fuzzy) measures [10, 11, 15].

It is invited to in ***step 2*** the equal survey proceeds to resolve arisen difficulties as follows.

For example, let us suppose you set private indicators $F_i \quad (i = 1, \ldots, m)$ that evaluate the effectiveness of the decisions. To conduct the expert survey in ***step 2*** is required to draw up 2^m production rules type P_j: «IF $F_1 = A_{1j}$ and $F_2 = A_{2j}$ and … and $F_m = A_{mj}$,

THEN $F_{res} = A_{jres}$», where $A_{ij} \in \{-1_{F_i}, +1_{F_i}\}$ - «low» or «high» value of F_i, $A_{jres} \in T(F_{res})$ - the value of the linguistic variable performance indicator result.

The rules, where all performance indicators except one take «low» values, we call simple rules of expert`s survey or simple support situations. The number of these situations corresponds to the number of partial indicators of efficiency. We assume that the rules $P_1, P_2, \ldots, P_m$ are simple, where the corresponding indicators $F_1, F_2, \ldots, F_m$ take «high» values.

The complex (compound) rule (advanced reference situation) can be described using a simple reference situations in the following way. A rule P_j: «IF $F_1 = A_{1j}$ and $F_2 = A_{2j}$ and … and $F_m = A_{mj}$, THEN $F_{res} = A_{jres}$», where the indicators with the indices $\{i_1, i_2, \ldots, i_k\} \subseteq \{1, 2, \ldots, m\}$ take high values, can be written as $P_j = P_{i_1} \cup P_{i_2} \cup \ldots \cup P_{i_k}$.

An evaluation of result indicator A_{ires} of the simple rules, we denote $g_i = E(a_i, \alpha_i, \beta_i) = E(A_{ires})$, $i = 1, \ldots, m$, where $E(\bullet)$ is the operation defuzzification of triangular fuzzy number (for example, $E(a_i, \alpha_i, \beta_i) = a_i + \dfrac{\beta_i - \alpha_i}{3}$).

The calculation of the resulting index ratings in the complex situations, the reference is invited to be carried out by building a constructive parameter λ-fuzzy measure Sugeno [10, 11, 15] on a finite set of the simple support situations P_i, $i \in \Gamma = \{1, 2, \ldots, m\}$, where g_i - a density of distribution of the fuzzy measures. Sugeno measure reflects an assessment of the resulting figure in a complicated rule $P_j = P_{i_1} \cup P_{i_2} \cup \ldots \cup P_{i_k}$ and it is as follows:

$$G_\lambda(P_j = P_{i_1} \cup P_{i_2} \cup \ldots \cup P_{i_k}) = \left[\prod_{l=1}^{k}\left(1 + \lambda g_{i_l}\right) - 1\right]/\lambda.$$

For the construction of λ-fuzzy measure Sugeno, characterizing the rating of the resulting figure in the complex is generally required to find the root λ^* of the interval $(-1, \infty)$ is the following polynomial of (m−1) order [10, 11, 15] $\left[\prod_{i=1}^{m}\left(1 + \lambda g_i\right) - 1\right]/\lambda = 1, \ -1 < \lambda < \infty.$

It should be noted that the polynomials have exactly one root in the interval $(-1, \infty)$, what is stating in the theorem [11].

The obtained estimates complex rules are used to check the consistency of the DM's remarks. For example, if you reply to a complex rule P_j the evaluation result indicator will be equal A_{jres} and relative deviation of the result from the value $G_{\lambda^*}(P_j)$ will be greater than the specified error value $0 \le \gamma \le 1$ (i.e. $\dfrac{\left|G_{\lambda^*}(P_j) - E(A_{jres})\right|}{G_{\lambda^*}(P_j)} > \gamma$), it is considered that the expert gave a wrong answer. The identified contradictions are facing the DM to analyze and resolve them.

3 Illustrative Example

To illustrate the proposed method we give a small example of computing. Let the effectiveness of the decisions made by three private estimated indicators $F = \{F_i,\ i = 1,\ 2,\ 3\}$ (criteria, $F_i \to \max\ \ i = 1,\ 2,\ 3$). According to the proposed approach for the construction of an indicator of efficiency of the resulting matrix should fill the expert survey in the extreme values (-1_{F_i} - «low», $+1_{F_i}$ - «high») indicators F_i (Table 2). The estimates of expert support for the simple situations 2, 3, 5 are shown in Table 2.

Table 2. The results of a survey expert on the simple rules

Rules	F_1	F_2	F_3	F_{res}
1	«low»	«low»	«low»	0
2	«high»	«low»	«low»	0, 2
3	«low»	«high»	«low»	0, 6
4	«high»	«high»	«low»	?
5	«low»	«low»	«high»	0, 4
6	«high»	«low»	«high»	?
7	«low»	«high»	«high»	?
8	«high»	«high»	«high»	1

To determine the result indicator assessments in the complex situations, taking into account the views of the supporting expert in the simple situations will make a calculation parameter λ^* of fuzzy measures Sugeno, solving the equation $\frac{(1+0.2\lambda)(1+0.6\lambda)(1+0.4\lambda)-1}{\lambda} = 1 \Rightarrow 0.048\lambda^2 + 0.44\lambda + 0.2 = 0.$

The roots of the equation are equal, respectively $\lambda_1^* \approx -0.48$, $\lambda_2^* \approx -8.69$. The second solution does not satisfy the condition $-1 < \lambda < \infty$; therefore, $\lambda^* \approx -0.48$.

Then the expert evaluation result indicator in the 4th situation will be equal $G_{\lambda^*}(P_4) = \frac{(1+0.2\lambda^*)(1+0.6\lambda^*)-1}{\lambda^*} \approx 0.742$, in the 6th situation will be equal $G_{\lambda^*}(P_6) = \frac{(1+0.2\lambda^*)(1+0.4\lambda^*)-1}{\lambda^*} \approx 0.5616$, in the 7th situation will be equal $G_{\lambda^*}(P_7) = \frac{(1+0.6\lambda^*)(1+0.4\lambda^*)-1}{\lambda^*} \approx 0.8848.$

Then the evaluated averaged scalar works corresponding columns of the orthogonal matrix (Table 1) to the vector result performance indicator values. The resulting performance indicator will be look like

$$F_{res} = 0,549 + 0,077F_1 + 0,258F_2 + 0,163F_3 - 0,013F_1F_2 - \\ - 0,0081F_1F_3 - 0,0274F_2F_3 - 0,0014F_1F_2F_3.$$

Evaluate the effectiveness of solutions $\{w_1,\ w_2,\ w_3\}$, the evaluation results are given in Table 3.

Table 3. Private estimates of solutions

Indicators	w_1	w_2	w_3
F_1	**0.33**	**0.44**	**0.22**
F_2	**0.33**	**0.17**	**0.5**
F_3	**0.14**	**0.28**	**0.56**

Let us translate this evaluations, which are shown in Table 3, in a scale of $[-1, +1]$ (Table 4).

Table 4. Scaling results private assessments

Indicators	w_1	w_2	w_3
F_1	**−0.34**	**−0.12**	**−0.56**
F_2	**−0.34**	**−0.66**	**0**
F_3	**−0.72**	**−0.44**	**0.12**

The calculated values of the result indicator of the effectiveness of the solutions $\{w_1, w_2, w_3\}$ are equal $F_{res}(w_1) = 0.307$, $F_{res}(w_2) = 0.288$, $F_{res}(w_3) = 0.525$. The obtained results allow us to evaluate and rank the proposed solutions.

The feature of this example is the fact that the assessment of the expert survey DM result indicators was given the real numbers. For the implementation of the proposed method in the article should be used both the arithmetic fuzzy numbers and fuzzy quantities ranging codes [11] in the case where the evaluation data will be variable linguistic value (such as trapezoidal fuzzy numbers).

4 Conclusions

It is suggested the technique which can reduce the number of appeals to the DM in the process of the expert survey in order to construct the resulting figure of solving problems of multi-criteria decision making under the linguistic given partial indicators. What is the DM survey is carried out using production rules, which take only partial indicators of the extreme (best or worst) values, and the processing of these statements is carried out by methods of the experiment planning theory. However, the limited ability of the human system of information processing even in these conditions will be able to lead to a violation of the rationality of choice, errors and contradictions. To ensure the consistency of the DM preferences in the technique it is provided for a procedure to eliminate operational errors in the answers of the DM. It lies in the fact that on the set of the production rules survey (reference cases) using the DM's answers for simple reference situations based parametric fuzzy measure which is used for check a conflicts with the DM's answers to complex questions.

Acknowledgements. The research described in this paper is partially supported by the Russian Foundation for Basic Research (grants 15-07-08391, 15-08-08459, 16-07-00779, 16-08-00510, 16-08-01277,16-29-09482-ofi-i), grant 074-U01 (ITMO University), project 6.1.1 (Peter the

Great St. Petersburg Polytechnic University) supported by Government of Russian Federation, Program STC of Union State "Monitoring-SG" (project 1.4.1-1), state order of the Ministry of Education and Science of Russian Federation № 2. 3135. 2017, State research 0073–2014–0009, 0073–2015–0007.

References

1. Mikoni, S.V.: Teorija prinjatija upravlencheskih reshenij. Uchebnoe posobie [Theory of administrative decision-making. Tutorial]. SPb.: Lan'. 448 p. (2015). (in Russian)
2. Mattila, V., Virtanen, K.: Ranking and selection for multiple performance measures using incomplete preference information. Eur. J. Oper. Res. **242**(2), 568–579 (2015)
3. Korhonen, P.J., Silvennoinen, K., Wallenius, J., Öörni, A.: Can a linear value function explain choices? An experimental study. Eur. J. Oper. Res. **219**(2). 360–367 (2012)
4. Petrovskij, A.B., Rojzenzon, G.V., Tihonov, I.P., Balyshev, A.V.: Retrospektivnyj analiz rezul'tativnosti nauchnyh proektov [A retrospective analysis of the performance of research projects]. Int. J. Inf. Models Anal. **1**(4), 349–356 (2012). (in Russian)
5. Podinovski V.V.: Decision making under uncertainty with unknown utility function and rank-ordered probabilities. Eur. J. Oper. Res. **239**(2), 537–541 (2014)
6. Larichev, O.I.: Verbal'nyj analiz reshenij [Verbal decision analysis]. M.: Nauka, 181 p. (2006). (in Russian)
7. Mikoni, S.V.: System analysis of multi-criteria optimization methods on a finite set of alternatives. In: Trudy SPIIRAN – SPIIRAS Proceedings, vol. 4, no. 41, pp. 180–199 (2015). (in Russian)
8. Mikoni, S.V.: Axioms of multicriteria optimization methods on a finite set of alternatives. In: Trudy SPIIRAN – SPIIRAS Proceedings, vol. 1, no. 44. pp. 198–214 (2016). (in Russian)
9. Sokolov, B.V., Moskvin, B.V., Pavlov, A.N., et al.: Voennaja sistemotehnika i sistemnyj analiz. Modeli i metody prinjatija reshenij v slozhnyh organizacionno–tehnicheskih kompleksah v uslovijah neopredeljonnosti i mnogokriterial'nosti: uchebnik [Military systems engineering and systems analysis. Models and methods of decision-making in complex technical–organizational systems in conditions of uncertainty and multicriteria]./Pod red. B.V. Sokolova. SPb.: VIKKU imeni A. F. Mozhajskogo, 496 p. (1999). (in Russian)
10. Nechetkie mnozhestva v modeljah upravlenija i iskusstvennogo intellekta [Fuzzy sets in management models and artificial intelligence]/Pod red. D.A. Pospelova. M.: Nauka, 312 p. (1986). (in Russian)
11. Pavlov, A.N., Sokolov, B.V.: Prinjatie reshenij v uslovijah nechetkoj informacii: ucheb. Posobie [Decision-making in conditions of fuzzy information: tutorial]. SPb.: GUAP, 72 p. (2006). (in Russian)
12. Zelentsov, V.A., Pavlov, A.N.: Multi-criteria analysis of the influence of individual elements on the performance of complex systems. Informacionno-upravljajushhie sistemy – Inf. Contr. Syst. **6**(49), 7–12 (2010). (in Russian)
13. Pavlov, A., Sokolov, B., Pashchenko, A., Shalyto, A., Maklakov, G.: Models and methods for multicriteria situational flexible reassignment of control functions in man-machine systems. In: Proceedings of the 2016 IEEE 8th International Conference on Intelligent Systems, pp. 402–408 (2016)

14. Nogin, V.D.: Prinjatie reshenij v mnogokriterial'noj srede: kolichestvennyj podhod [Decision making in multicriteria environment: a quantitative approach]. M.: FIZMATLIT, 176 p. (2005). (in Russian)
15. Pyt'ev Ju, P.: Vozmozhnost' kak al'ternativa verojatnosti. Matematicheskie i jempiricheskie osnovy, primenenie [The possibility alternatively probability. Mathematical and empirical basis, application]. M.: FIZMATLIT, 464 p. (2007). (in Russian)

AnyLogic-Based Discrete Event Simulation Model of Railway Junction

Alexander Lyubchenko(✉), Stanislav Bartosh, Evgeny Kopytov, Alexander Shiler, and Askar Kildibekov

Omsk State Transport University, Omsk, Russia
allyubchenko@gmail.com

Abstract. Nowadays, increase of competitive ability and effectiveness of railway transportation network of the Russian Federation is an important problem, which solution requires the modernization of system elements such as railway junctions. Modern computer technologies allow assisting in correct project decision-making, however, there is a necessity of development of appropriate software instruments for analysis. In this paper, mathematical model of Ekaterinburg railway junction, developed by means of the instrument of simulation modeling AnyLogic, is presented. The model was constructed using discrete event approach and queuing network technique that give an opportunity to estimate indices of railroad operation and discover bottlenecks in structure of the junction. Calculation of the estimation errors of output parameters was conducted on the basis of the simulation experimental results, which analysis allowed making conclusion about the adequacy of the model.

Keywords: Railway junction · Simulation · Queueing network · AnyLogic

1 Introduction

Railroads as a basis of the transportation system of the Russian Federation have extremely important national, economic, social and defense significance. In 2008, the Russian Federation government approved the Strategy for Developing Rail Transport up to 2030 year, supposing the modernization of railway districts for the purpose of increase of the operational effectiveness and track capacity improvement. It is obvious that achievement of this goal is hard without proper modernization of such parts of the transportation network as railway junctions including stations and linked rail tracks between them.

One of the largest russian railway junctions is Ekaterinburg transport railway interchange, which is situated on the principal track of Trans-Siberian railway. Improvement of this rail centre is a strategic task. Application of the modern simulation tools allows analyzing of operation performance, assists in discovering of bottlenecks in junction structure for decision-making support concerning its modernization. Moreover, simulation modeling possesses by a number of advantages, namely, "it permits promptly taking into account all changes in the project, and also obtaining more precise values of the optimal parameters [1]". Therefore, at present, "there are a lot of developments using

R. Silhavy et al. (eds.), *Cybernetics and Mathematics Applications in Intelligent Systems*,
Advances in Intelligent Systems and Computing 574, DOI 10.1007/978-3-319-57264-2_14

simulation techniques in native science and practice for calculation and design of railway stations, districts and transport interchanges [1]". This type of modeling finds a use for the estimation of carrying capacity of tracks [2, 3], determination of the most effective alternatives of traffic flow service [4, 5], traffic estimation [6], analysis of shunting operations on stations [7] and determination of scheduled maintenance work intervals [8]. In [9], the research of analytical and simulation models for track capacity optimization and improvement of timetable quality was performed. Description of modeling methods and problem-oriented software for study and analysis of railroad networks is presented in [10]. "Therefore, the relevance of the simulation method to study the stages of modernization programs is still high [8]".

A mathematical model of the abovementioned railway junction based on the system dynamics approach was developed using Matlab and suggested in [11]. Nevertheless, nowadays, AnyLogic software, which combines "all three modern paradigms of simulation models construction" [7], gains more and more popularity among domestic instruments for simulation modeling. Foreign researches in the field of transportation pay attention to the advantages of this software for wide range of real task research as well [12]. AnyLogic is utilized for modeling of transport vehicles timetable [13], analysis of railroad operations of marshalling yards [14] and estimation of effectiveness of intermodal freight short-distance traffic [15].

Based on the character of railway junction operation, in this paper the discrete event approach is proposed for mathematical description. Moreover, this methodology has been well studied, and as it is shown in [16], the approach is still relevant and has high demand on solving modern simulation problems.

Thus, the problem of a discrete event simulation model construction for Ekaterinburg railway junction using AnyLogic software was formulated. The model is supposed to be utilized for the estimation of rehandling of a train on the stations and calculation of train flows on the blocks.

2 Theory

2.1 Concept Description of Ekaterinburg Railway Junction

Generally, according to railroad operation of Ekaterinburg railway junction, it relates to a transit interchange with high volume of classification and local freight operations. Movement of large quantity of passenger and suburban trains is carried out in this junction.

The borders (approaches) of the junction are stations: Khrustalnaya st., Reshety st., Kosulino st., Aramil st., Shuvakish st., Apparatnaya st., Sysert st., where branching of the main tracks originates.

The junction includes 23 stations: 2 off-grade stations Ekaterinburg-Sortirovochnyi and Ekaterinburg-Passazhirskiy, division terminal Sedelnikovo, 13 freight yards and 8 through depots.

2.2 Discrete Event Simulation Approach

The simulation instrument AnyLogic was chosen for the development of railway junction model using discrete event technique realized by means of the embedded Enterprise library. Enterprise library is a high-level interface for the fast construction of discrete event models using block-diagrams. Consequently, the model was developed on the basis of the queuing network (QN) technique. This approach represents an aggregate of final quantity of servers, where requests circulate according to the traffic routing matrix from one server to another. Server always is an open-loop queue system (QS).

Whereupon, separate QSs represent functionally independent parts of a real system, connections between QSs correspond to the structure of the system, and requests circulating in QN refer to constituent of the material flows.

The proposed model represents a linear network, as the requests are not lost and do not multiply. The model is also an open-loop network, where the requests arrive from the external medium and leave the network after operation. Peculiarity of an open-loop QN is availability of one or several independent external sources, which generate requests entering into the network regardless of the current quantity of requests in QN. At any one time, there is a random quantity of requests in such network.

In accordance to the chosen simulation technique the following components of the model were developed: trains, approaches to the junction, railway stations and blocks. Then the mentioned elements were used for model structure construction in line with the master plan for the development of Ekaterinburg junction [17]. Structure chart of the model is illustrated in Fig. 1.

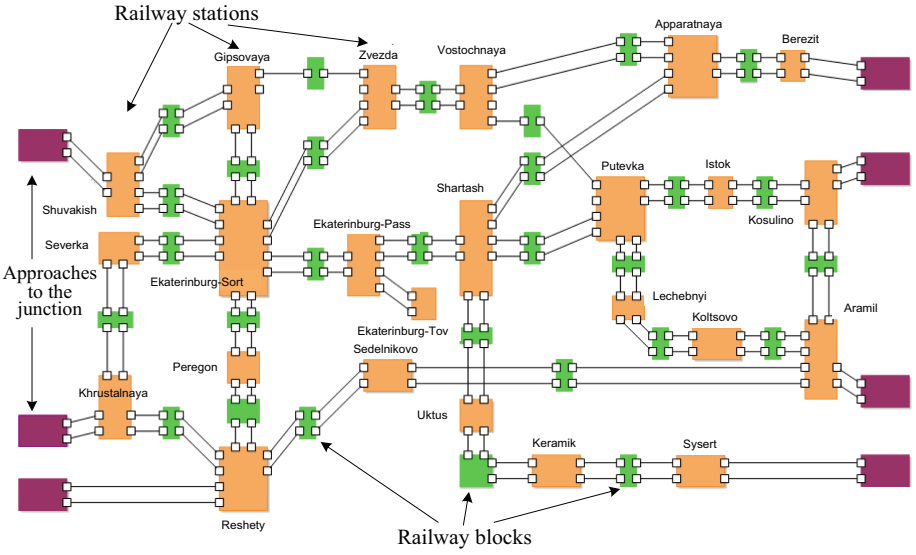


Fig. 1. Structure chart of the Ekaterinburg railway junction model.

Rolling stock plying between stations was represented by means of objects "Message" transmitted between models of the stations. According to the types of rolling stock taking part in rail movement four groups of messages were anticipated: through and division trains; pick-up, transfer and clean-up trains; passenger trains and suburban trains.

The approaches to the junction were built by a way of 4 message generators regarding to the abovementioned types of train flows. Messages simulate request flows arriving at QN. The requests simulating trains in the model come in the railway junction with the time interval subject to Poisson distribution.

The models of railway stations were proposed using active objects of AnyLogic simulation software that gives an opportunity to specify individual functioning logic for each model, and, therefore, to provide design flexibility. For each station of the simulated junction, a separate active object (model) was allocated considering the peculiarities of the station upon implementing operations with arriving trains. Different types of trains pass through during various time slots. Depending on the train flow types the following operations are considered: receiving, departure and turn-over of train; splitting-up and making-up of train; replenishment and setting out of wagons; non-stop proceeding; train processing with crew/locomotive change. The models of all stations anticipate carrying out of overtaking of a train on stations. In this case, another trailing express train, which usually corresponds to the passenger traffic, overtakes arriving train and another one refers to the freight flows.

The messages arriving at the station inputs are distributed between outputs after the performed operations. Definition of the train flow directions is implemented using matrix with N rows and 4N columns, where N is a quantity of approaches to the station. The distribution matrixes are generated on the basis of the data of train flow chart of Ekaterinburg railway junction for 2007 [17].

Blocks between railway stations are also represented by means of active objects of AnyLogic tool. Structure of a railway block provides an opportunity to consider various running time within the block depending on the train flow types. An example of the active object structure for the block between stations Apparatnaya and Berezit is depicted in Fig. 2.

There are two train movement directions: up and down. Active objects of railway blocks are realized with support of the both directions independently. Each active object is characterized by 3 parameters: running time of passenger and freight trains, actual carrying capacity.

3 Experimental Results

Two types of simulation experiments were implemented in this work:

Simulation modeling of railway junction operation in order to analyze the adequacy of the proposed model.

Analysis of time intervals of arriving trains at the approaches of the Ekaterinburg-Sortirovochnyi station.

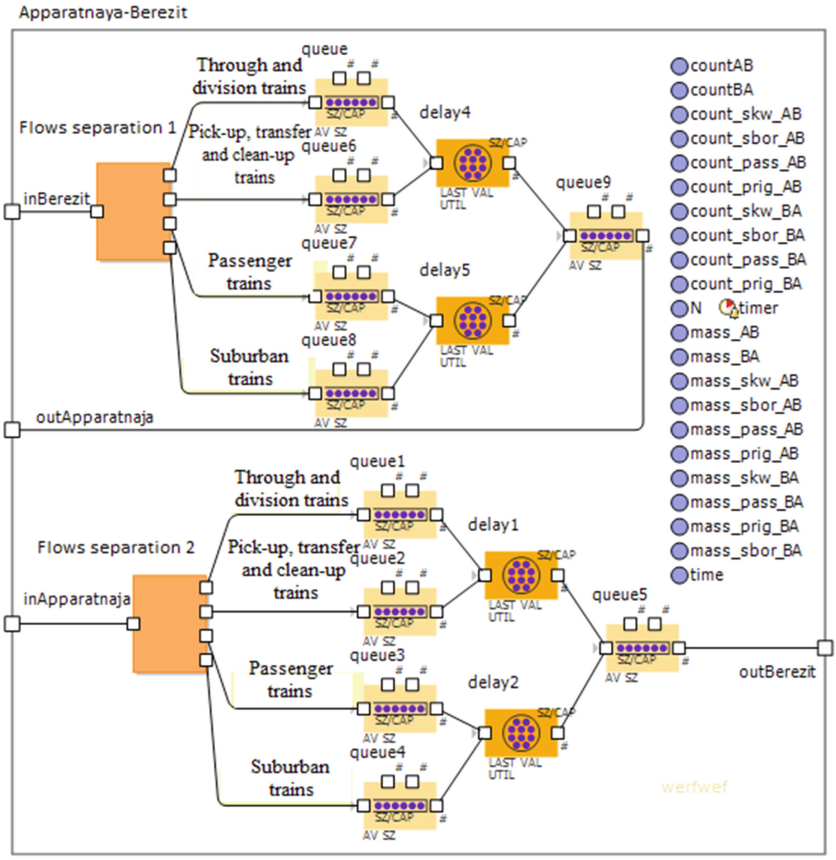


Fig. 2. Model of the railroad block Apparatnaya-Berezit.

Values of the input parameters were specified according to the master plan for development of Ekaterinburg railway junction [17].

Adequacy analysis of the model was performed by comparison of estimation results of train flows and train rehandling with the real experimental data for railway blocks and stations. The relative errors for output values were calculated considering empirical data for each type of the train flow: through and division trains (TDT); pick-up, transfer and clean-up trains (PTCT); passenger trains (PT) and suburban trains (ST). Relative error results of the train rehandling for all 23 stations of Ekaterinburg junction are demonstrated in Table 1.

A question of great interest in railway junction structure is marshalling yard Ekaterinburg-Sortirovochnyi situated on the main rail track of the junction. Therefore, the analysis of the distribution law of train arriving time intervals was implemented by means of the developed model for six connected lines. Six simulation experiments were performed in order to collect statistical data with posterior statistical manipulation and

Table 1. Relative error estimation results

Station	Relative error, %			
	TDT	*PTCT*	*PT*	*ST*
Ekaterinburg-Pass	2.8	−0.4	0	0.1
Putevka	2.2	−1.2	0	−0.2
Ekaterinburg-Sort	0	−0.3	0	0
Shartash	2.8	0	0	−0.1
Apparatnaya	0	2.5	−3.3	−0.6
Aramil	0.7	5	−1.1	0
Berezit	0	0	−3.3	0
Gipsovaya	0	0	0	0
Khrustalnaya	−0.2	0	0.2	1.4
Istok	2.5	0	0.3	0
Keramik	0	−1.6	1.3	0
Koltsovo	1.7	5.0	−0.9	−0.6
Kosulino	1.7	0	0.1	0.3
Lechebnyi	1.7	5.0	−0.9	−0.6
Peregon	0,6	0	0,3	−0.5
Reshety	0	0	0	−0.5
Sedelnikovo	−0.5	0	0	2.0
Severka	0.2	0	0.2	1.4
Shuvakish	0	−1.3	0	−0.3
Sysert	0	0	1.3	0
Uktus	0	−2.5	0	0
Vostochnaya	0	3.3	0	0
Zvezda	0	2.9	0	0

checking of the suggested hypothesis using Pearson criterion. Histogram of time intervals distribution of arriving trains from Ekaterinburg-Passazhirskiy station is presented in Fig. 3.

4 Discussion of Results

The results of relative error estimation of train rehandling operation on the stations of the Ekaterinburg railway junction demonstrated in Table 1 indicate conformity of modeling values to the real experimental data. The error did not exceed 5% threshold for all stations and all types of train flows that corresponds to the adequate simulation of the analyzed operation index. Furthermore, the same situation was observed for train flow estimations on the railway blocks, consequently, the obtained results provide an opportunity to conclude in whole about adequate behavior of the proposed simulation model.

In accordance to the results of time intervals analysis of train arrivals at Ekaterinburg-Sortirovochnyi station the following fact was determined that the time intervals distribution is subject to the exponential law with the confidence probability 0.95 for 4 of 6 approaches of the junction from the direction of the stations: Ekaterinburg-Passazhirskiy, Peregon, Severka and Shuvakish. Arriving time intervals distribution for the rest

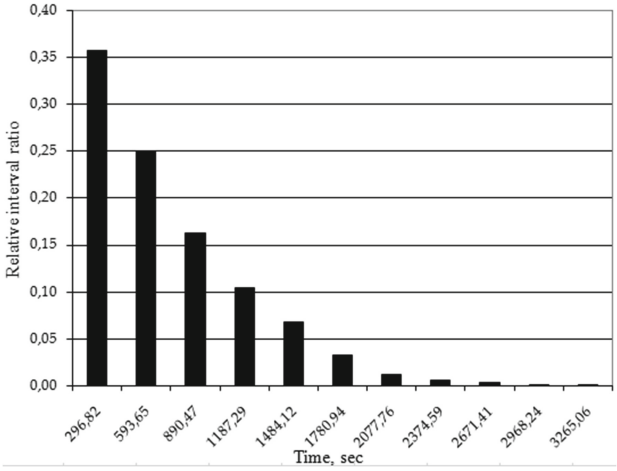


Fig. 3. Histogram of distribution law of time intervals between trains arriving at Ekaterinburg-Sortirovochnyi station from Ekaterinburg-Passazhirskiy station.

2 approaches from the direction of the stations Gibsovaya and Zvezda is characterized by significant variation of the obtained values and impossibility to suggest a hypothesis about the distribution law.

5 Conclusion

Thus, in this work, a simulation model of Ekaterinburg railway junction was proposed, which was constructed in AnyLogic modeling instrument using discrete event approach and queuing network technique. The model allows estimation of such railway operation indices as train rehandling on a station and train flows on blocks of the junction.

Estimation error of the abovementioned operation indices was calculated with reference to the real experimental data. The error did not exceed 5% threshold that gave an opportunity to make a conclusion about adequacy of the developed simulation model. In addition, the model was used for analysis of distribution law of arriving time intervals of trains at Ekaterinburg-Sortirovochnyi station by 6 lines from the direction of the connected stations. The exponential distribution law was verified for 4 of 6 stations: Ekaterinburg-Passazhirskiy, Peregon, Severka and Shuvakish.

The simulation model allows research of influence of the distribution law of the input flows on railway operation indices, analyzing blocks carrying capacity utilization that, consequently, gives a possibility to discover bottlenecks in structure of the junction. Hence, the model can be used as an automated decision-making support instrument for analysis of modernization variants of the transport junction. Moreover, the set of the developed objects of the stations and blocks can be easily utilized for simulation modeling of railway junctions with another topology that emphasizes a flexibility of the proposed model.

Acknowledgments. This work has been supported in part by projects ERANET-Plus (European Commission) and 141 K-2010-176/10, Mathematical modeling of Ekaterinburg railway junction operation (JSC Russian railways).

References

1. Timchenko, V.: Prospects of domestic experience application of railway stations, sites and transport knots calculation by imitating modeling at development of railway the Crimean peninsula infrastructure. Internet J. "Mir-nauki" **4**, 1–9 (2014)
2. Kokurin, I.M., Kudryavtsev, V.A.: Estimation of railway line capacity based on imitation modeling the transportation processes. In: Proceedings of Petersburg Transport University, vol. 2, pp. 18–22 (2012)
3. Alekseev, S.I., Berezhnoy, V.V., Soroka, R.I.: Imitated-animated simulation of the basic technological processes the Murmansk transport hub. In: Proceedings of the International Conference "Eurasian Area: Priorities of Socio-economic Development", Moscow, vol. 1, pp. 28–36, 12 May 2011
4. Maksimey, I.V., Sukach, E.I., Giruts, P.V., Erofeeva, E.A.: Simulation modeling of probabilistic characteristic of railway network operation. Math. Mach. Syst. **4**, 147–153 (2008)
5. Sukach, E.I.: Automatization of research processes of organization variants for transport flows movement of railroad network. Math. Mach. Syst. **4**, 161–168 (2009)
6. Kibzun, A.I., Naumov, A.V., Ivanov, S.V.: Bilevel optimization problem for railway transport hub planning. Large Syst. Manage. Collected Pap. **38**, 140–159 (2012)
7. Rakhmangulov, A.N., Mishkurov, P.N.: Special aspects of railroad station working method simulation model development within AnyLogic system. Mod. Prob. Ways Solution Sci. Transp. Prod. Educ. **4**(2), 7–13 (2012)
8. Kokurin, I.M., Kattsyn, D.V., Timchenko, V.S.: Determining parameters of scheduled maintenance work-intervals within the assessment of future transportation capacity. World Transp. Transp. **13**(2), 142–153 (2015)
9. Hansen, I.A.: State-of-the-art of railway operations research. In: Timetable Planning and Information Quality, pp. 35–47. WIT Press, Boston (2010)
10. Marinov, M., Şahin, I., Ricci, S., Vasic-Franklin, G.: Railway operations, time-tabling and control. Res. Transp. Econ. **41**(1), 59–75 (2013). doi:10.1016/j.retrec.2012.10.003
11. Smirnov, V.A.: Simulation of transportation process and rolling stock service systems on railway transport. Sci. Transp. Prob. Siberia Far East **2**, 13–18 (2012)
12. Möller, D.P.F.: Simulation tools in transportation. In: Möller, D.P.F. (ed.) Introduction to Transportation Analysis, Modeling and Simulation. SFMA, pp. 195–228. Springer, London (2014). doi:10.1007/978-1-4471-5637-6_5
13. Merkuryeva, G., Bolshakovs, V.: Vehicle schedule simulation with AnyLogic. In: 12th International Conference on Computer Modelling and Simulation, pp. 169–174. IEEE (2010)
14. Baugher, R.W.: Simulation of yard and terminal operations. In: Patty, B.W. (ed.) Handbook of Operations Research Applications at Railroads, vol. 222, pp. 219–242. Springer, Boston (2015). doi:10.1007/978-1-4899-7571-3_9
15. Reis, V.: Analysis of mode choice variables in short-distance intermodal freight transport using an agent-based model. Transp. Res. Part A Policy and Practice **61**, 100–120 (2014). doi:10.1016/j.tra.2014.01.002

16. Brailsford, S.: Discrete-event simulation is alive and kicking! J. Simul. **8**, 1–8 (2014). doi: 10.1057/jos.2013.13
17. Development of Ekaterinburg railway junction. General plan. Explanatory note. JSC "Uralgiprotrans", 252 p. (2010)

The Parameters List for Multihop Wireless Networks Cross-Layer Routing Metric

I.O. Datyev(✉), A.A. Pavlov, and M.G. Shishaev

Institute for Informatics and Mathematical Modelling of Technological Processes of the Kola Science Center RAS, 184209 Apatity, Russia
{datyev,pavlov,shishaev}@iimm.ru

Abstract. Multihop wireless networks are the promising direction of communication networks. The main problem of such networks due to links' instability is to find the best route. Different parameters are used to route estimation. The paper presents an attempt to estimate the different parameters influence on wireless multihop networks performance. The parameters list is formed by different authors past experience generalization of cross-layer routing metrics development. We provide the results of Ns-3 experiments for estimation of different parameters influence on network performance. In particular, parameters list that are planned to be considered during the design of routing metrics is proposed.

Keywords: Multihop wireless networks · Network performance · Cross-layer routing metric

1 Introduction

Nowadays, wireless multihop networks are studied by scientists around the world. Link instability due to the constant nodes' mobility complicates data transmission.

Identifying the best data packet transmission path is one of the main problems of wireless multihop networks. The routing metric is intended to solve this problem by routing path estimation. Various parameters can be taken into account. The most common are: hop count, bandwidth of the channel, cost of data transmission over a channel, reliability, delay and etc.

Many routing protocols, for example AODV, OLSR, DSR, DSDV, use the hop count metric for route selection as default (or only) routing metric. However, the more parameters will be used in the evaluation of the route, the more likely choice is really the best of them.

The traditional architecture' protocols are strictly layers-grained, where adjacent levels can communicate only with each other [13].

The work was supported by RFBR grant № 16-29-12878 - The development of the identification methods of the dynamic models with random parameters and their use in Eurasia migration forecasting.

R. Silhavy et al. (eds.), *Cybernetics and Mathematics Applications in Intelligent Systems*, Advances in Intelligent Systems and Computing 574, DOI 10.1007/978-3-319-57264-2_15

Under such approach, the routing protocol developer focuses on a specific level, not taking into account the parameters of another stack layers. On the other hand, cross-layer mechanisms can be employed to make the different parameters' value available at all layers [4]. Cross-layer architecture is a complementary scheme for the layered protocol stack. By weakening the strict functional separation of protocols, networking performance can improve [1].

The goal of our work is to identify the parameters that are used in existing cross-layer metrics and estimate their potential influence on the network performance.

The paper is organized as follows. In Sect. 2, we review the previous works on cross-layer routing metrics for wireless multihop networks and discuss their features, such as used parameters and layers. Section 3 provides our NS-3 experiments for estimation of different parameters influence on network performance. Conclusions and future work are presented in Sect. 4.

2 Related Works

In this section, we overview some cross-layer routing metrics proposed for wireless multihop networks. Currently a unified classification of cross-layer metrics not yet defined.

In paper [1], a cross-layer connectionless routing is proposed based on Dynamic Virtual Router (DVR). In this algorithm, virtual route discovery process is controlled by restricting the request packets' broadcast to the relatively slow speed, and low loaded nodes located in suitably crowded areas. Each destination decides to choose or discard the found route based on several cross-layer metrics, which are joined into a single cross-layer metric. Network layer information is the number of hops, MAC- layer information is node workload, PHY-layer information is the number of node neighbors. This mixed multidimensional criterion is expressed as (1) below:

$$\text{Mix\_Metr} = \text{Hop\_Metr}_{ij} \times (1 - \text{Node\_Load}_{ij}) \times \text{Node\_Neigh}_{ij} \times (1 - \text{Mob\_Metr}_{ij}) \tag{1}$$

$$\text{Mob\_Metr}_{ij} = \text{avr(node)} \times \sqrt[3]{\text{var (node)}} \tag{2}$$

where:

Hop_Metr_{ij}: The number of hops from Source i to Destination j

Node_Load_{ij}: Maximum load metric of the nodes on the route from Source i to Destination j

Node_Neigh_{ij}: Maximum neighboring metric of the nodes on the route from Source i to Destination j

Mob_Metr_{ij}: Mobility metric between source node i and destination node j in terms of average and variance of the mobility

The simulation results [1] showed that the proposed algorithm achieved better performance compared to DVR, in terms of average end-to-end delay and packet delivery ratio.

Another work [2] via predicting the duration of the interference imposed by the neighbors at every hop along the route, a new routing metric is presented which guarantees that the established routes will not break frequently while having the minimum interference. The metric is calculated by (3):

$$\text{Metric} = \min_{\text{Pi}} \left(\sum_{h_i^j} \frac{\dfrac{\text{Interference}_j}{\text{number}^j_{\text{interfer}}}}{t^{(j-1,j)}_{\text{broken}} - t_{\text{current}}} \right) \tag{3}$$

where:

$\text{number}^j_{\text{interfer}}$ – total number of the interference nodes surrounding node j,

$\dfrac{\text{Interference}_j}{\text{number}^j_{\text{interfer}}}$ – is mean duration of interference imposed on node j,

$t^{(j-1,j)}_{\text{broken}} - t_{\text{current}}$ – connectivity duration of current communication link.

Thus, the information of three layers is used: Network, MAC and Physical. Simulation results [2] show that Minimum Interference cross-layer Routing protocol (MIR) can significantly improve the network performance.

Proposed by the authors [3] Multipath Routing Protocol using Cross-layer based QoS Metrics at first calculates multiple disjoint paths and then estimates the Combined Cost (CC) metric based on the metrics such as: Traffic Contention Time (TCT), Average Transmission Delay (Adelay) and Signal Fading Value (SFV). Finally, the path with minimum cost (i.e., with minimum CC) is chosen as the best path from the multiple disjoint paths and the data is sent through. The metric is calculated as shown in (4):

$$(\text{CC}) = \frac{a \times \text{TCT} + b \times A_{\text{delay}}}{c \times \text{SFV}} \tag{4}$$

where a, b, and c are normalization or smoothing constants.

The simulation results [3] demonstrate that the proposed Multi-path routing protocol helps in achieving better delivery ratio and throughput with reduced delay.

Authors [4] modify ad hoc on-demand distance vector routing protocol using cross-layer approach which uses three parameters, namely Signal to Noise Ratio (SNR), node lifetime and delay to improve the performance of routing protocol in mobile ad-hoc network and to avoid routing through bad quality links.

Link Cost of a node has the following three components: (1) Signal to Noise Ratio cost (2) Delay cost (3) Power cost. Link cost of a node is weighted sum of these three costs:

$$\text{LC} = \text{SNR} + \text{delay} + \text{node lifetime} \tag{5}$$

where, LC is Link Cost at receiving node, SNR is Signal to Noise Ratio.

The performance is improved in terms of average end-to-end delay, throughput, packet delivery ratio and normalized routing load for constant bitrate traffic pattern.

The paper [5] proposes a design approach, deviating from the traditional network design, toward enhancing the cross-layer interaction among different layers, namely physical, MAC and network. The Cross-Layer design approach for Power control (CLPC) would help to enhance the transmission power by averaging the receiving signal strength (RSS) values and to find an effective route between the source and the destination.

In another proposed algorithm [6] three input variables are taken and one output variable. Three input variables are Battery power, Received signal strength and Speed of a node.

The authors [7] propose modified AODV algorithm designed to save nodes energy. That is, the authors calculate the total energy consumption of the packet transmission.

Work [8] is used information on the number of hops and average packet transmission time. Overall, authors conclude that their mechanism demonstrates significant benefits at high and unstable traffic scenarios.

In another work [9], the authors use the information on the battery and the availability of the route. Cross layer based approach for link availability prediction (DPCPLP) increases network lifetime and capacity by combining the effect of optimum transmit power in transmitting RTS, CTS, DATA and ACK packets and estimation of link availability time and further, formation of the path prior to the link break to support the Quality of Service (QoS) requirements of applications.

Paper [10] authors to improve the energy awareness of the wireless network, have shared the parameters of MAC and network layers. In Cross layer AODV (CAODV) algorithm the information about the energy of each node is analyzed so that during routing, path with the highest residual energy is selected.

AODV-BER-QoS [11] has been proposed, where the route discovery process of Ad-Hoc on Demand Distance Vector routing (AODV) has been modified to obtain the stable route. The modified route discovery message obtains Bit Error Rate (BER) information from physical layer through cross-layer.

The paper [12] proposes a novel routing technique called Adaptive Link-Weight (ALW) routing protocol. ALW adaptively selects an optimum route on the basis of available bandwidth, low delay and long route lifetime. The technique adapts a cross-layer framework where the ALW is integrated with application and physical layer.

Summarizing the past authors experience (partly presented above) it is possible to create a list of parameters (or estimations of parameters) used in cross-layer routing metrics: hop count, load of the nodes, node neighboring, end-to-end delay, MAC layer delay, link throughput, link rate, link availability, signal fading, Signal to Noise Ratio (SNR), Received Signal Strength (RSS), interference, propagation range, bit error rate, frame error rate, battery power, medium utilization, transmit power.

3 Simulation Results and Analysis

Section 2 demonstrates that different authors most widely use following parameters: number of hops, interference, battery level, time of data transmission. It has been proved that these parameters inclusion helps to achieve more efficient data routing in wireless multihop networks. But there is still a large set of important parameters for wireless networks.

At the initial stage we have selected a few, in our opinion, important parameters. In order to prove that the selected parameters also have a significant impact, we conducted a series of simulation experiments. We have tested the following parameters: packet size, nodes velocity, transmitter power, wireless standard, source data rate, the number of nodes in the network.

Simulation conducted in the network simulator NS-3. Firstly, we have developed the general model with the following default settings, as shown in Table 1.

Table 1. Default settings

Simulation environment	Ns-3
Wi-Fi standard	802.11b
Packet size	256 Bytes
Transmitter power	10 dBm
PropagationDelay	ConstantSpeedPropagationDelayModel
PropagationLoss	FriisPropagationLossModel
Traffic pattern	CBR (Constant Bit Rate)
Source data rate	512 Kb/s
Routing Protocol	OLSR
Nodes velocity	5 m/s
Simulation duration	100 s
Data transmission start time	0 s
Number of nodes	25
Source/sink data pairs	1
Simulation area	1.5 km × 1.5 km
Mobility model	Random waypoint

Further we change the value of one selected parameter and conduct an experiment. This was done in order to identify the impact on data transmission quality, namely, the mean delay and packet loss ratio. We tried to use the range of values that are most likely to meet in real life. Some of our simulation results we provided below.

Simulation results of general model with default settings (OLSR protocol, Hop count metric) are: mean delay–225 ms and packet loss ratio–47%.

With such delay and packet loss ratio values (one source/sink data pair) is impossible to establish normal communication in the network, as given in Table 2.

Table 2. Allowed values for different transmission data types

Transmission data type	Allowed value	
	Delay (ms)	Packet loss ratio (%)
Video	300	1
Voice	400	5
Text	1000	15

It should be noted that in our experiments the data transmission start time equal to 0 s, and the routing protocol takes some time to configure the network.

In the first experiments series we changed the value of transmitter power, as shown in Table 3. We used following values: 7.5, 10, 15, 20.

Table 3. Transmitter power

Transmitter power (dBm)	Source/sink data pairs (source->destination)	Mean delay (ms)	Packet loss ratio (%)
7.5	1 (5->9)	412	63
10	1 (5->9)	225	46
15	1 (5->9)	228	24
20	1 (5->9)	0.4	0
10	2 (5->9) (7->11)	486 267	66 34
15	2 (5->9) (7->11)	278 2,5	43 0

The results of simulation can be judged that this parameter greatly affects the communication quality. Under the high transmitters power almost all nodes are within one step from each other. Hence, it defeats the need to retranslate data and as a result are significantly reduced delays and packet loss ratio.

The choice of wireless standard is important too. Modern standards, as a rule, have a higher data rate. There are specific standards designed for mesh and manet networks, but they are not widespread. Accordingly, there is no implementation of such standards in NS-3. Because of this we have tested different WiFi standards, as shown in Table 4.

Table 4. WiFi standard

WiFi standard	Source/sink data pairs (source->destination)	Mean delay (ms)	Packet loss ratio (%)
802.11b	1 (5->9)	225	46
802.11 g	1 (5->9)	227	47
802.11n	1 (5->9)	58	11
802.11n	2 (5->9) (7->11)	88 41	12 6

Network nodes can support different WiFi standards. It is obvious that when using a node as a relay, it is preferable to choose a route through nodes operating on the standard that supports higher bandwidth.

It should be separately noted impact on data transmission quality in wireless networks provided by the packet size, as shown in Table 5.

Table 5. Packet size

Packet size (KB)	Source/sink data pairs (source->destination)	Mean delay (ms)	Packet loss ratio (%)
256	1 (5->9)	225	46
64	1 (5->9)	910	76
128	1 (5->9)	657	64
512	1 (5->9)	88	50
512	2 (5->9) (7->11)	240 148	64 49

You can dynamically change the packet size depending on the network state or to use it in the routing metric.

The source data rate value can be correlated with the type of transmitted data, as shown in Table 6. For example, small data rate value enough for comfortable communication of subscribers via text message. Audio and video traffic has a higher data rate, lower delay and data packet loss ratio (Table 2).

Table 6. Source data rate

Source data rate (Kb/s)	Source/sink data pairs (source->destination)	Mean delay (ms)	Packet loss ratio (%)
512	1 (5->9)	225	46
128	1 (5->9)	56	36
256	1 (5->9)	148	41
1000	1 (5->9)	676	72
512	2 (5->9) (7->11)	486 267	66 34
128	2 (5->9) (7->11)	89 52	28 13

The nodes velocity has an impact mainly on data transmission delay of due to a network connectivity violation, as shown in Table 7.

Table 7. Nodes velocity

Nodes velocity (m/s)	Source/sink data pairs (source->destination)	Mean delay (ms)	Packet loss ratio (%)
5	1 (5->9)	225	46
2.7	1 (5->9)	155	43
7.5	1 (5->9)	299	50
10	1 (5->9)	236	40
2.7	2 (5->9) (7->11)	499 332	60 46
10	2 (5->9) (7->11)	248 95	67 10

However, too little velocity (at low density of nodes) can lead to the violation of the network connectivity, especially in case when nodes are unevenly distributed across the area.

The same can be said about the number of nodes. At low values of this parameter are not always guaranteed the network connectivity, as shown in Table 8. On the other hand, high value, specifically in the OLSR and other table-driven protocols, is increases service traffic. For large number of concurrent connections and high nodes' mobility the network will be overloaded and stable data transfer will be difficult.

Table 8. Number of nodes

Number of nodes	Source/sink data pairs (source->destination)	Mean delay (ms)	Packet loss ratio (%)
25	1 (5->9)	225	46
15	1 (5->9)	398	35
40	1 (5->9)	278	64
60	1 (5->9)	235	29
15	2 (5->9) (7->11)	127 16	20 4
60	2 (5->9) (7->11)	125 25	19 5

In addition, the modulation type has a significant influence on the network perfomance, as shown in Table 9.

Table 9. Modulation/Rate

Modulation/Rate	WiFi standard (2.4 GHz)	Mean delay (ms)	Packet loss ratio (%)
DsssRate11Mbps	80211b	225	46
DsssRate2Mbps	80211b	435	35
DsssRate5_5Mbps	80211b	183	27
DsssRate11Mbps	80211n	59	11
ErpOfdmRate6Mbps	80211n	22	1.6
ErpOfdmRate12Mbps	80211n	10	3.7

The result of our experiments listed above parameters have a great impact on the network performance, even separately and often depend from each other. For example, one value of the packet size has a different impact on the network perfomance with various topologies. Tables 10 and 11 shows the affect of topology change on network performance. Table 11 represents the simulation results for the case when the network nodes are aligned at an equal distance from each other and data is transferred from one end to the other.

Table 10. Random waypoint mobility model

Packet size (KB)	Mean delay (ms)	Packet loss ratio (%)
64	1250	21
256	680	7
1024	775	44
2000	845	61

Table 11. Static line topology model

Packet size (KB)	Mean delay (ms)	Packet loss ratio (%)
64	1075	84
256	754	82,5
1024	1256	71
2000	1311	59

Thus, there are many parameters that can be considered in the cross-layer metrics development. Based on the conducted experiments, we evaluated the ranges of parameters influence on delay and packet loss ratio, as shown in Table 12.

Table 12. Summary table

Parameter name	Parameter value (set)	Mean delay (ms)		Packet loss ratio (%)	
		Min	Max	Min	Max
Transmitter power (dBm)	(7.5, 10, 15, 20)	0.4	412	0	63
WiFi standard	(b, g, n)	58	227	11	47
Packet size (KB)	(64, 128, 256, 512)	88	910	46	76
Source data rate (Kb/s)	(256, 521, 1000, 2000)	56	676	36	72
Nodes velocity (m/s)	(2.7, 5, 7.5, 10)	95	499	10	67
Number of nodes	(15, 25 40, 60)	16	398	4	64

Further, we have generated a list of parameters to consider in developing cross-layer metric, mainly for network MANET. In addition to parameters described in Table 12 we plan to consider: battery power, nodes' load, interference.

4 Conclusions and Future Work

It is impossible to develop routing metric, excluding the effect of another network MAC and PHY layers, since it is necessary to consider features of a wireless transmission. Depending on various network functioning conditions, such as landscape features, nodes mobility, traffic pattern type, the impact of parameters on the network performance may be different. One of the possible ways of regulating the degree of parameters importance of cross-layer metric is the use of weighting factors for each parameter.

In this survey, we presented an analysis of the various cross-layer routing metrics that have been proposed for routing protocols in wireless multihop networks. After that, we have formed the list of different layers' parameters used in such metrics. Then we have provided a number of experiments to estimate the influence of different parameters on network perfomance in the same initial conditions.

Future work goals are to develop cross-layer routing metric that take into account the features of wireless multihop networks based on proposed parameters and simulate this metric with the most widely used protocols, as AODV, OLSR, etc. and to analyze their performance over recently proposed routing metrics.

References

1. Ataee Bojd, E., Moghim, N.: A new connectionless routing algorithm using cross-layer design approach in MANETs. Automatika **57**(2), 514–524 (2016)
2. Gu, C., Zhu, Q.: A cross-layer routing protocol for mobile ad hoc networks based on minimum interference duration. In: Proceedings of the 2nd International Conference on Computer Science and Electronics Engineering, pp. 2070–2073 (2013)
3. Saheb, S.M., Dharmasa, B., Bhattacharjee, A.K.: Multipath routing protocol using cross-layer based QoS metrics for IEEE 802.11e WLAN. Int. J. Comput. Appl. **50**(10), 8–12 (2012)
4. Yadav, A., Sharma, T.: Cross-Layer Approach for Communication in Manet. Int. J. Comput. Sci. Mob. Comput. **4**(3), 285–292 (2015)
5. Sarfaraz Ahmed, A., Senthil Kumaran, T., Syed Abdul Syed, S., Subburam, S.: Cross-layer design approach for power control in mobile ad hoc networks. Egypt. Inform. J. **16**, 1–7 (2015)
6. Fatima, M., Gupta, R., Bandhopadhyay, T.K.: Route discovery by cross layer approach for MANET. Int. J. Comput. Appl. **37**(7), 15–24 (2012)
7. Natarajan, E., Devi, L.: Cross layer based energy aware routing and congestion control algorithm in MANET. IJCSMC **3**(10), 700–709 (2014)
8. Romdhani, L., Bonnet, C.: A cross-layer on-demand routing protocol for delay-sensitive applications. In: 16th IEEE International Symposium on Personal Indoor and Mobile Radio Communications (2005)
9. Yadav, A., Singh, Y.N., Singh, R.: Cross layer design for power control and link availability in mobile adhoc networks. Int. J. Comput. Netw. Commun. (IJCNC) **7**(3), 127–143 (2015)
10. Remya, K., Sangeetha, C.P., Suriyakala, C.D.: QoS Improvement in mobile ad hoc networks using cross layer optimization. In: National Conference on Recent Advances in Electrical and Electronics Engineering (NCREEE 2015), vol. 4(1), pp. 272–278 (2015)
11. Sanguankotchakorn, T., Dahal, R.: The cross-layer design for QoS routing in MANET AODV based on BER (2011)

12. Al-Khwildi, A.N., Khan, S., Loo, K.K., Al-Raweshidy, H.S.: Adaptive link-weight routing protocol using cross-layer communication for MANET. WSEAS Trans. Commun. **11**(6), 833–839 (2007)
13. Winter, R., Schiller, J.: A cross-layer mobility adaptation framework for ad hoc networks. In: Workshop on Applications and Services in Wireless Networks (ASWN), Berlin, Germany (2006)

An Improved Active Queue Management Algorithm for Time Fairness in Multirate 802.11 WLAN

Jianjun Lei(✉), Yingwei Wu, and Xu Zhang

School of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, China
{leijj,zhangx}@cqupt.edu.cn, infiniteone@foxmail.com

Abstract. In multirate 802.11 wireless local area network (WLAN), time unfairness is an inherent problem that slow stations occupy more time to transfer data and leave less time for fast stations, which is so called performance anomaly. The paper proposes an improved active queue management (IAQM) algorithm for fairly sharing network resources among all contending stations. Meanwhile, by setting different queue length and drop rate for each data flow with different destinations going through the access point (AP) according to their transmission rate, so that each station guarantees equal channel usage time. Therefore, the time fairness can be achieved and aggregate throughput can be improved. Both analytical and simulation results are provided to validate the effectiveness of the proposed IAQM algorithm, which can achieve good time fairness and a 30% improvement in aggregate throughput.

Keywords: Multirate WLAN · Performance anomaly · Time fairness · Queue length · Drop rate

1 Introduction

Many of the existing companies, organizations and communities provide wireless hotspots by operating over 802.11 WLANs that allows for the simultaneous usage of multiple data rates on a single wireless network, which leads to unfair bandwidth allocation, and stations competing for channel access according to IEEE 802.11 distributed coordination function (DCF) have comparable opportunity of channel access, regardless of transmission rate. The wireless station with lower data rate dominates shared channel usage time so that the throughput experienced by other stations transmitting at higher data rate will be drastically reduced. To some extent, it suppresses the network access to achieve higher aggregate throughput, and results in performance anomaly. To mitigate this problem, a lot of improvement mechanisms based on media access control (MAC) layer have been proposed [1, 2]. While these improvements are only suitable for wireless stations sending user datagram protocol (UDP) flow without considering the characteristics of transmission control protocol (TCP) flow. Each wireless station does not always participate in the competition of sending data in MAC layer for TCP flows under the TCP protocol congestion control.

R. Silhavy et al. (eds.), *Cybernetics and Mathematics Applications in Intelligent Systems*, Advances in Intelligent Systems and Computing 574, DOI 10.1007/978-3-319-57264-2_16

In this paper, we propose an IAQM algorithm to address the time fairness problem among stations. Different from the previous algorithms, which change the Contention Window (CW) value in station or aggregate the transport packet and frame to achieve time fairness, the IAQM algorithm sets different queue length and drop rate for each flows, and drops packets actively which are in sending queue in AP to restrain the sending of low rate stations indirectly. The main contributions are as follows: (1) the proposal of the sending queue model, which performs that the queue length of each virtual queue is stabilized at the respective ideal object value so as to ensure that time occupied by each flow to be equal; (2) the derivation of an analytical model for the packet drop rate calculation, which characterizes the data rate of the data flow; (3) the design detail of aforementioned mechanisms as well as a simulation evaluation of IAQM algorithm, which is an important contribution of our work.

The remainder of this paper is organized as follows. Section 2 briefly introduces the related work for the channel fairness research. Section 3 analyzes the performance of DCF and addresses the proposed algorithm in detail. Then, the simulation and analysis are demonstrated in Sect. 4. Finally, we draw a conclusion in Sect. 5.

2 Related Work

In this section, we summarize the related work under the MAC algorithms and internet protocol (IP) layer traffic control. Many international scholars have put forward the improvement measures to solve the problem of performance anomaly.

In [3, 4], under the certain assumptions, the throughput of the stations in DCF mechanism are inversely proportional to their minimum Contention Window (CWmin) value by some analyses. Time fairness can be realized by setting the initial value of CWmin for that is inversely proportional to its original bit rate. In [5], an optimal CWmin selection scheme is proposed, and the way to use the Arbitration Inter frame Space (AIFS) defined in IEEE 802.11e is discussed. In [6], a distributed analysis model is presented, which can achieve fairness of network by estimating the CW value according to the rate of each station. However, when a new station joins or leaves the network, its CW value always needs to be recalculated in these algorithms. In [7], the frame aggregation scheme in which the packets from higher layers are fragmented based on the bit rate of the station is proposed to achieve time fairness. At each transmission opportunity, the stations with high rate utilize a frame size equal to the Maximum Transmission Unit (MTU) whereas ones with low rate fragment their packets to transmit smaller frames. In [8], a cross-layer scheme is presented. It can degrade the number of bytes per frame transmitted by the low rate stations while allowing the stations with high rate to send full size frames by using IP path MTU discovery. But it is difficult to be implemented for the cross-layer mechanism. A dynamic and distributed frame aggregation mechanism is proposed in [9]. The frame aggregation approach is presented in a transmission opportunity (TXOP) that is regarded as the maximum channel occupation time of the IEEE 802.11e standard. When the channel occupancy time does not exceed TXOP, clients' awaiting frames can be aggregated. And clients with better channel conditions are allowed to transmit multiple zframes at a transmission opportunity while clients with low rate are not

allowed to perform aggregation. Its disadvantage is that frame aggregation will degrade the performance in the term of delay, and all these methods have no role in TCP flow.

In addition to MAC improvement methods, IP layer traffic control can also improve time unfairness. The existing AP queue management algorithm for multirate WLANs of [10] and TTPDE algorithm of [11] both selectively discard the packets whose transmission delay is the longest when AP is in congestion, and when AP needs to send packets, they also can selectively send packets which have the minimum delay value. This method makes the flows with high rate occupy the channel for a long time and the slow flows may be unable to access the channel and get into the state of "starvation". Though the efficiency of the network can be improved, it is at the cost of the fairness among stations.

To effectively solve the problems, an optimized active queue management algorithm is proposed. In the case of contention window without modification in the station, this algorithm assigns the downlink data flow merely from AP into multiple virtual data flows according to their different destinations. Meanwhile, it drops packets actively in a low possibility of high data rate flow and a high possibility of low data rate flow, affecting the sending rate indirectly, and time fairness can be achieved for both TCP and UDP flow.

3 Design Considerations and IAQM Algorithm

3.1 Performance of 802.11 DCF

According to 802.11 standards which are widely used in wireless systems up till now, DCF employs the CSMA/CA in stations with binary exponential backoff mechanism to control medium access, which provides the same channel occupation opportunity to all stations. Therefore, the aggregate throughput of the network can be seriously degraded to match the throughput of the station with the low data rate.

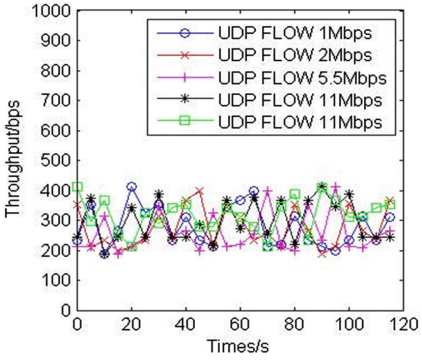


Fig. 1. Throughput of UDP flows

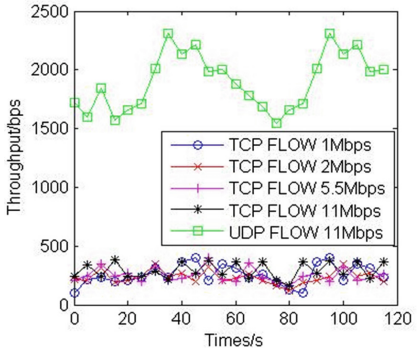


Fig. 2. Throughput of mixed UDP and TCP flows

The evaluation setting for traditional DCF backoff mechanism is described in Sect. 4. In Fig. 1, the throughput of all data flows is almost the same although at different rates, though with some oscillations. Figure 2 shows that the traditional DCF backoff mechanism does not suppress the throughput of TCP unfriendly flow such as the UDP data flow, so that the UDP data flow has more channel usage time and higher throughput, and the TCP data flow will not be guaranteed.

3.2 Sending Queue Model

To resolve the problem of performance anomaly, we divide the sending queue going through the AP into multiple virtual queues according to the destination of the data flow, so that to make sure each flow has the proportional data to send by taking into account the transmission rate of each station. Therefore, the object queue length of each data flow is also proportional to the data rate. To calculate the object queue length of each flow, the channel usage time of sending a packet should be calculated at first. If we neglect the propagation time, the channel usage time of sending a packet that the AP sends to a wireless station can be demonstrated as the following:

$$T_{total} = T_t + T_o \tag{1}$$

Form the Eq. (1), the overall channel usage time is composed of the transmission time T_t and the contention overhead delay T_o. The transmission time is given by $T_t = \frac{S_d}{r_i}$, where S_d is the packet size in bits and r_i is the rate of the data to the destination station i. The contention overhead delay is made up of the random backoff time and the transmission time of the control frame. Assume that there is no collision among data packets, the backoff time can be defined as half of the minimum CW value multiplied by time unit. So, the contention overhead delay can be chosen as a constant, which can be expressed by the following equation:

$$T_o = SIFS + DIFS + \frac{CW_{min}S}{2} + \frac{RTS + CTS + ACK + PHY_h}{r} \tag{2}$$

Where, short inter frame space (SIFS) is the duration time of a short inter frame space, the distributed inter frame space (DIFS) is the duration time of a distributed inter frame space, the unit time value can be defined as S, and the transmission rate can be expressed as r.

For the first come first served (FCFS) service discipline, each flow's queue length to different destination station should be inversely proportional to its channel usage time of sending a packet, so that the object queue length of flow i can be defined as:

$$l_i = L \times \frac{1}{\frac{S_d}{r_i} + T_o} \Bigg/ \sum\nolimits_{i=1}^{n} \frac{1}{\frac{S_d}{r_i} + T_o} \tag{3}$$

Where L is the aggregate flow's queue object length in AP. To avoid buffer overflow, we set L as $\frac{max_{th} + min_{th}}{2}$, where max_{th} is the max queue length and min_{th} is the min queue length.

3.3 Drop Rate

The drop rate is used to make the queue length be stabilized at the ideal objective value. The different drop rate is set for the different virtual queues, and the high data rate flow has a low drop rate and low data rate flow has a high drop rate. The data flow with high rate can consume more time to transmit data, and their data rate of destination stations can be adjusted indirectly according to the TCP congestion control mechanism.

We now explore the relationship between the throughput and the packet drop rate for TCP connection. Here, we make the assumption that the data rate and the packet size have been known.

There also are three preconditions for model inference: (1) The sender responds to a packet drop as a congestion indication by cutting the congestion window at least by half; (2) In the steady state, after a packet is dropped, the TCP sender increases its congestion window by at most one packet within each roundtrip time, until the congestion window again reaches its old value of packets; (3) The congestion window is set as maximum W when the TCP connection receives another packet drop.

By decreasing its window by at least half for each packet drop and increasing its window by at most one per roundtrip time afterwards, the TCP sender transmits $\frac{3}{8}W^2$ packets when a packet is dropped, which can be computed by Eq. (4).

$$\frac{W}{2} + \left(\frac{W}{2} + 1\right) + \ldots + W \approx \frac{3}{8}W^2 \tag{4}$$

The fraction p_{drop} of the sender's packets that are dropped is then bounded by the reciprocal of that value:

$$W \approx \sqrt{\frac{8}{3p_{drop}}} \tag{5}$$

When the congestion window reaches to W in the steady state, the average data rate is defined by:

$$T = W * M/R \tag{6}$$

Where M is the packet size and R is the roundtrip time.

After a packet is dropped in the steady state, it takes W/2 times M as the roundtrip time to turn back to the original value. And the average data rate between the drops of two packets is expressed as:

$$T = \left(\frac{W * M}{R} + \frac{\frac{W}{2} * M}{R} \right) / r_i \tag{7}$$

Through Eqs. (5), (6) and (7), the relationship between drop rate and data rate is governed by:

$$T = M/R\sqrt{\frac{2p_{drop}}{3}} \tag{8}$$

From Eq. (8), data rate changes when the drop rate changes with the decrease of the roundtrip time, which becomes obvious when the roundtrip time gets to 20 ms.

3.4 IAQM Algorithm Design

Our proposed IAQM algorithm creates different data flow queues for downlink flow towards different destination station through AP, and drops the transmitted packets in high rate with low probability and drops the transmitted packets in low rate with high probability. Also we set different queue lengths based on their data rate. Simultaneously, the data rate of the station can be impacted by setting the drop rate through TCP congestion control mechanism.

The pseudo code of IQAM algorithm is given as follows:

```
Input: packet p
1: I = classify(p)
2: l_i' = record _length(i)
3: avg=(1-w)*avg+w*∑l_i'
4: if (l_i' ≥ l_i) drop=1
5:     else if (avg≥ max_th) drop=1
6:         else if (avg<min_th) drop=0
7:             else calulate p_drop;
8:                 if (random[0,1]≤p_drop) drop=1
9:                 else drop=0
10:enqueue (p)
```

In IQAM algorithm, there is no real queue in each flow, and AP just records the number of packets going to different stations. It classifies packets at first based on the difference among different destination stations, then records the packet number of each flow, and drops the packet or enqueues.

4 Performance Evaluation

4.1 Experimental Design

In this section, we verify the performance of our algorithm via qualnet 6.1, and compare it with the traditional DCF backoff mechanism, the distributed algorithm (DA) [6], and the TTPDE algorithm [11].

The test scenarios mainly include five source stations 1–5 and five destination stations 6–10. The source stations send data to different destination nodes respectively through the AP, and all the stations in the network work on the same channel. The distances between the destination stations 6–10 and the AP are 100 m, 80 m, 60 m, 10 m and 10 m, so that the rates of the AP sending data to each destination station can be the data rate of 1 Mbit/s, 2 Mbit/s, 5.5 Mbit/s, 11 Mbit/s and 11 Mbit/s. The MAC layer of each wireless station adopts the IEEE802.11b protocol, and the basic parameters in the MAC layer and the physical layer are shown in Table 1.

Table 1. IEEE 802.11 system parameters

Parameter	Parameter value	Parameter	Parameter value
SIFS	10 us	ACK	112 bits
DIFS	50 us	r	1M bit/s
S	20 us	R	20 ms
RTS	160 bits	w	0.002
CTS	112 bits	Buffer size	100
max_{th}	60	PHY header	192 bits
min_{th}	40	CWmin	31

In this paper, two simulation experiments are carried out, namely UDP data flow transmission and the mixed UDP and TCP data flow transmission. First, the source stations 1 to 5 transmit the Constants Bit Rate (CBR) packets uniformly to the destination stations 6 to 10 through the AP. In the second experiment, the mixed UDP and TCP packets are sent. The sending source stations 1 to 4 send the FTP packets to the destination stations 6 to 9 respectively. The sending source station 5 sends the CBR packets to the destination station 10.

The performance evaluation metrics include fairness index and throughput. The fairness index of these experiments are mentioned in [12], which can be expressed as:

$$\mathrm{f} = \left(\sum_{i=1}^{n} T_i\right)^2 \Big/ n\left(\sum_{i=1}^{n} T_i^2\right) \tag{9}$$

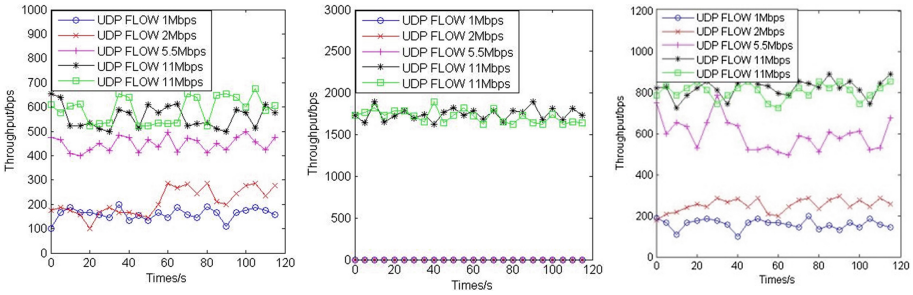
Where, T_i is the channel usage time of wireless station i. Additionally, the throughput is computed by the following equation:

$$t_i = S_d \Big/ \left(\frac{S_d}{r_i} + T_o\right) \tag{10}$$

Where, t_i is the throughput rate for sending one packet. The aggregate throughput of each flow should be in proportion to the data rate so that the network reaches time fairness.

4.2 UDP Data Flow

Figure 3 depicts the network throughput when stations transmit CBR packets. For DA, the throughput of each data flow is differentiated by layers, with a certain proportion of data rate, and so the time fairness can be reached, but the throughput is not in a high level, showed in Fig. 3(a). For TTPDE, the throughput of high data rate stations are guaranteed, but the throughput of low data rate stations is almost zero, as Fig. 3(b) shows. Figure 3(c) shows that the IAQM algorithm can also make the throughput of each data flow be with a certain proportion of data rate, and the throughput for all data rate stations keeps in a high level.



(a) DA (b) TTPDE (c) IAQM

Fig. 3. Throughput of DA, TTPDE and IAQM under UDP flows with different data rates

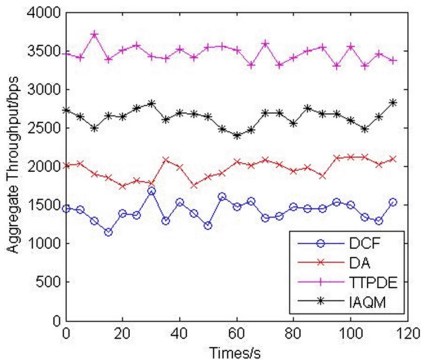


Fig. 4. Aggregate throughput of DCF, DA, TTPDE and IAQM under UDP flows

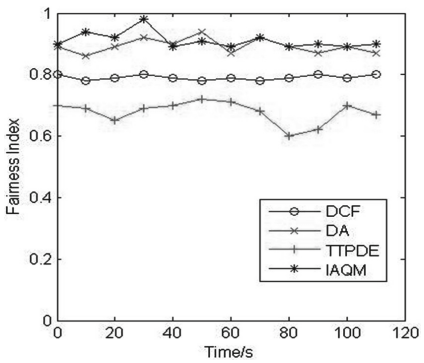


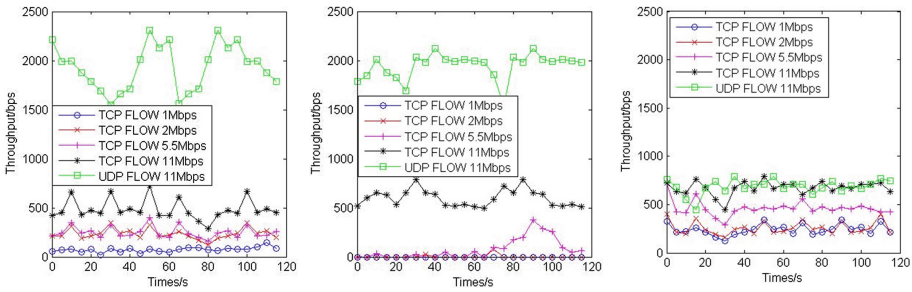
Fig. 5. Fairness index simulation for DCF, DA, TTPDE and IAQM algorithms

Figure 4 shows that the aggregated throughput of IAQM algorithm can achieve a 30% increase compared with the distributed algorithm. And the fairness indexes of four

algorithms have been demonstrated in Fig. 5. For IAQM, the value of indexes remains to be almost 1, which indicates that all data flows almost achieve the same channel usage time.

4.3 Mixed TCP and UDP Data Flow

Figure 6 shows the network throughput when stations transmit both FTP packets and CBR packets. As Fig. 6(a) shows, the distributed algorithm does also not protect the TCP flows, as a result, UDP flow has an extremely high throughput and the throughput of TCP data flow is very low. The TTPDE algorithm achieves the high throughput for high data rate flows, but a low throughput for the low data rate flows, showed in Fig. 6(b). Figure 6(c) shows that the IAQM algorithm restrains the UDP data flow effectively and improves the throughput of TCP data flows.



(a) DA (b) TTPDE (c) IAQM

Fig. 6. Throughput of DA, TTPDE and IAQM under mixed UDP and TCP flows with different data rates

In Fig. 7, the average throughputs are shown for the four algorithms DCF, DA and TTPDE algorithm do not achieve good throughput fairness and time fairness, while the

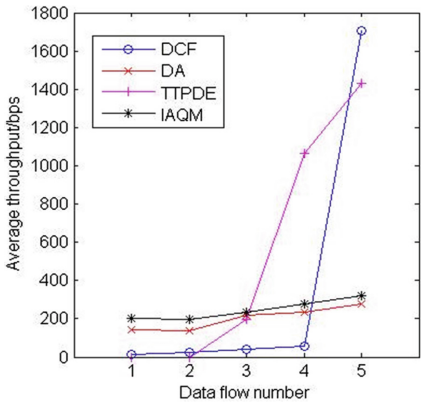


Fig. 7. Average throughput of DCF, DA, TTPDE and IAQM under mixed UDP and TCP flows

average throughput of IAQM changes with the increase of data rate, and it also performs well in terms of fairness.

5 Conclusion

This paper studies the issue of performance anomaly in multirate WLANs and proposes the IAQM algorithm. This algorithm sets different queue lengths for each data flow and drops actively packets which are in the sending queue in AP to suppress the sending of low rate stations indirectly. Meanwhile, we present the sending queue model and analyze the relationship between throughput and drop rate. Finally, we evaluate the IAQM algorithm under UDP, TCP and mixed UDP and TCP data flows respectively, and the simulation results show that our proposed algorithm can obtain good throughput and time fairness.

Acknowledgments. This research was supported by Program for Innovation Team Building at Institutions of Higher Education in Chongqing (CXTDX201601021), the National Science Foundation of Chongqing (cstc2014kjrc-qnrc40002) and Scientific and Technological Research Program of Chongqing Municipal Education Commission (KJ1500439).

References

1. Zhou, X., Zheng, C., He, X.: Adaptive contention window tuning for IEEE 802.11. In: 2015 22nd International Conference on Telecommunications (ICT), pp. 74–79. IEEE (2015)
2. Le, Y., Ma, L., Cheng, W., et al.: A time fairness-based MAC algorithm for throughput maximization in 802.11 networks. IEEE Trans. J. Comput. **64**, 19–31 (2015)
3. Li, B., Battiti, R.: Performance analysis of an enhanced IEEE 802.11 distributed coordination function supporting service differentiation. In: Proceedings of Quality for All, Cost 263 International Workshop on Quality of Future, Internet Services, Qofis 2003, Stockholm, Sweden, October 1–2, 2003, pp. 152–161 (2003)
4. Kim, H., Yun, S., Kang, I., et al.: Resolving 802.11 performance anomalies through QoS differentiation. J IEEE Communications Letters. **9**, 655–657 (2005)
5. Lin, P., Chou, W.I., Lin, T.: Achieving airtime fairness of delay-sensitive applications in multirate IEEE 802.11 wireless LANs. J IEEE Commun. Mag. **49**, 169–175 (2011)
6. Tarasyuk, O., Gorbenko, A., Kharchenko, V., et al.: Contention window adaptation to ensure airtime consumption fairness in multirate Wi-Fi networks. In: International Conference on Digital Technologies 2014, pp. 87–136 (2014)
7. Sdt, B.: 11b: Un Schema a Division de Temps Pour Eviter l'anomalie de la Couche MAC 802.11b. In: Cfip (2010)
8. Dunn, J., Neufeld, M., Sheth, A., et al.: A practical cross-layer mechanism for fairness in 802.11 networks. J Mob. Netw. Appl. **11**, 37–45 (2006)
9. Razafindralambo, T., Guérin Lassous, I., Iannone, L., et al.: Dynamic packet aggregation to solve performance anomaly in 802.11 wireless networks. In: ACM International Symposium on Modeling Analysis and Simulation of Wireless and Mobile Systems, pp. 247–254 (2006)

10. Huang, J.-W., Wang, J.-X.: Temporal fair active queue management algorithm for multi-rate 802.11 WLAN. J. Commun. **30**, 34–141 (2009)
11. Seok, Y.S., Park, J., Choi, Y.: Queue management algorithm for multi-rate wireless local area networks, pp. 231–234. IEEE Inc., Beijing (2003)
12. Chiu, D.M., Jain, R.: Analysis of the increase and decrease algorithms for congestion avoidance in computer networks. J. Comput. Netw. ISDN Syst. **17**, 1–14 (1989)

Control Theory Application to Complex Technical Objects Scheduling Problem Solving

Boris Sokolov[1], Inna Trofimova[2(✉)], Dmitry Ivanov[3], and Alekcey Krylov[4]

[1] St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences, University ITMO, St. Petersburg, Russia
sokol@iias.spb.su
[2] St. Petersburg State University, St. Petersburg, Russia
isolovyeva@mail.ru
[3] Berlin School of Economics and Law, Berlin, Germany
dmitry.ivanov@hwr-berlin.de
[4] St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences, St. Petersburg, Russia
kralex98@ya.ru

Abstract. We present a new model for optimal scheduling of complex technical objects (CTO). CTO is a networked controlled system that is described through differential equations based on a dynamic interpretation of the job execution. The problem is represented as a special case of the job shop scheduling problem with dynamically distributed jobs. The approach is based on a natural dynamic decomposition of the problem and its solution with the help of a modified form of continuous maximum principle blended with combinatorial optimization.

1 Model Formulation

We present the scheduling model where an CTO [1–4] is a networked controlled system that is described through differential equations based on a dynamic interpretation of the job execution. The job execution is characterized by (1) execution results (e.g., volume, time, etc.), (2) capacity consumption of the resources, and (3) CTO flows resulting from the delivery to the customer. We propose to use a two stage scheduling procedure in line with Chen and Pundoor [5]. A job control model (M1) is first used to assign jobs to suppliers, and then a flow control model (M2) is used to schedule the processing of assigned orders subject to capacity restrictions of the production and transportation resources. The basic interaction of these two models is that after the solving the job control model, the found control variables are used in the constraints of the flow control model. In additional models of resource and channel control, the material supply to resources and its consumption as well as setup times are represented.

1. A Dynamic Model of Job Control (model M1). We consider the mathematical model of job control. We denote the job state variable $x_{i\mu}^{(o)}$,

R. Silhavy et al. (eds.), *Cybernetics and Mathematics Applications in Intelligent Systems*, Advances in Intelligent Systems and Computing 574, DOI 10.1007/978-3-319-57264-2_17

where - indicates the relation to jobs (orders). The execution dynamics of the job $D_{\mu}^{(i)}$ can be expressed as (1).

$$\frac{dx_{i\mu}^{(o)}}{dt} = \dot{x}_{i\mu}^{(o)} = \sum_{j=1}^{n} \varepsilon_{ij}(t) u_{i\mu j}^{(o)}, \tag{1}$$

where $\varepsilon_{ij}(t)$ is an element of the preset matrix time function of time-spatial constraints, $u_{i\mu j}^{(o)}$ is a 0–1 assignment control variable.

Let us introduce Eq. (2) to assess the total resource availability time:

$$\dot{x}_{j}^{(o)} = \sum_{i=1}^{n} \sum_{\eta=1,\eta\neq i}^{n} \sum_{\mu=1}^{s_i} \sum_{\rho=1}^{p_i} u_{i\mu j}^{(o)}. \tag{2}$$

Equation (2) represents resource utilization in job execution dynamics. The variable $x_{j}^{(o)}$ characterizes the total employment time of the j-supplier. The control actions are constrained as follows:

$$\sum_{i=1}^{n} \sum_{\mu=1}^{s_i} u_{i\mu j}^{(o)}(t) \leq 1, \forall j; \quad \sum_{j=1}^{n} u_{i\mu j}^{(o)}(t) \leq 1, \forall i, \forall \mu; \tag{3}$$

$$\sum_{i=1}^{n} u_{i\mu j}^{(o)} [\sum_{\alpha \in \Gamma_{i\mu 1}^{-}} (a_{i\alpha}^{(o)} - x_{i\alpha}^{(o)}) + \prod_{\beta \in \Gamma_{i\mu 2}^{-}} (a_{i\beta}^{(o)} - x_{i\beta}^{(o)})] = 0; \tag{4}$$

$$u_{i\mu j}^{(o)}(t) \in \{0, 1\}; \tag{5}$$

where $\Gamma_{i\mu 1}^{-}, \Gamma_{i\mu 2}^{-}$, are the sets of job numbers which immediately precede the job $D_{\mu}^{(i)}$ subject to accomplishing of all the predecessor jobs or at least one of the jobs correspondingly, and $a_{i\alpha}^{(o)}, a_{i\beta}^{(o)}$ are the planned lot-sizes. Constraint (3) refers to the allocation problem constraint according to the problem statement (i.e., only a single order can be processed at any time by the manufacturer). Constraint (4) determines the precedence relations, more over this constraint implies the blocking of job $D_{\mu}^{(i)}$ until the previous jobs $D_{\alpha}^{(i)}, D_{\beta}^{(i)}$ have been executed. If $u_{i\mu j}^{(o)}(t) = 1$, all the predecessor jobs of the operation (job) $D_{\mu}^{(i)}$ have been executed. Note that these constraints are identical to those in MP models.

Corollary 1. The analysis of constraints (4) shows that control $\mathbf{u}(t)$ is switching on only when the necessary predecessor operations have been executed. $\sum_{i=1}^{n} u_{i\mu j}^{(o)} \sum_{\alpha \in \Gamma_{i\mu 1}^{-}} (a_{i\alpha}^{(o)} - x_{i\alpha}^{(o)}) = 0$ guarantees the total processing of all the predecessor operations, and $\sum_{i=1}^{n} u_{i\mu j}^{(o)} \prod_{\beta \in \Gamma_{i\mu 2}^{-}} (a_{i\beta}^{(o)} - x_{i\beta}^{(o)}) = 0$; of at least one of the predecessor operations. According to Eq. (5) controls contain the values of the

Boolean variables. In order to assess the results of job execution, we define the following initial and end conditions at the moments $t = T_0, t = T_f$:

$$x_{i\mu}^{(o)}(T_0) = 0; \quad x_{i\mu}^{(o)}(T_f) = a_{i\mu}^{(o)}. \tag{6}$$

Conditions (5) reflect the desired end state. The right parts of equations are predetermined at the planning stage subject to the lot-sizes of each job.

According to the problem statement, let us introduce the following performance indicators (7)–(9):

$$J_1^{(o)} = \frac{1}{2}\sum_{i=1}^{n}\sum_{\mu=1}^{s_i}(a_{i\mu}^{(o)} - x_{i\mu}^{(o)}(T_f))^2; \tag{7}$$

$$J_2^{(o)} = \sum_{i=1}^{\overline{n}}\sum_{\mu=1}^{s_i}\sum_{j=1}^{n}\int_{T_0}^{T_f}\alpha_{i\mu}^{(o)}(\tau)u_{i\mu j}^{(o)}(\tau)d\tau; \tag{8}$$

$$J_3^{(o)} = \frac{1}{2}\sum_{j=1}^{n}(T - x_j^{(o)}(T_f))^2. \tag{9}$$

The performance indicator (7) characterizes the accuracy of the end conditions' accomplishment, i.e. the service level of CTO. The goal function (8) refers to the estimation of an job's execution time with regard to the planned supply terms and reflects the delivery reliability, i.e., the accomplishing the delivery to the fixed due dates. The functions $\alpha_{i\mu}^{(o)}(\tau)$ is assumed to be known characterizes the fulfilment of time conditions for different jobs and time points of the penalties increase due to breaking supply terms respectively. The indicator (9) estimates the equal resource utilization in the CTO.

2. A Dynamic Model of Flow Control (Model M2). We consider the mathematical model of flow control in the form of Eq. (10):

$$\dot{x}_{i\mu j}^{(f)} = u_{i\mu j}^{(f)}, \quad \dot{x}_{ij\eta\rho}^{(f)} = u_{ij\eta\rho}^{(f)}. \tag{10}$$

We denote the flow state variable $x_{i\mu j}^{(f)}$, where indicates the relation of the variable x to flows. The control actions are constrained by maximal capacities and intensities as follows:

$$\sum_{i=1}^{\bar{n}}\sum_{\mu=1}^{s_i}u_{i\mu j}^{(f)}(t) \le \tilde{\tilde{R}}_{1j}^{(f)}, \ \sum_{\rho=1}^{p_i}u_{ij\eta\rho}^{(f)}(t) \le \tilde{\tilde{R}}_{1j\eta}^{(f)}, \tag{11}$$

$$0 \le u_{i\mu j}^{(f)}(t) \le c_{i\mu j}^{(f)}u_{i\mu j}^{(o)}, \ 0 \le u_{ij\eta\rho}^{(f)}(t) \le c_{ij\eta\rho}^{(f)}u_{ij\eta\rho}^{(o)}, \tag{12}$$

where $\tilde{\tilde{R}}_{1j}^{(f)}$ is the total potential intensity of the resource $C^{(j)}$, $\tilde{\tilde{R}}_{1j\eta}^{(f)}$ is the maximal potential channel intensity to deliver products to the customer $\overline{B}^{(\eta)}$ of results of CTO, $c_{i\mu j}^{(f)}$ is the maximal potential capacity of the resource $C^{(j)}$ for

the job $D_{\mu}^{(i)}$, and $c_{ij\eta\rho}^{(f)}$ is the total potential capacity of the channel delivering the product flow $P_{<s_i,\rho>}^{(j,\eta)}$ of the job $D_{\mu}^{(i)}$ to the customer $\overline{B}^{(\eta)}$ of results of CTO. The end conditions are similar to those in (6) and subject to the units of processing time. The goal functional of the flow control model are defined in the form of Eqs. (13) and (14):

$$J_1^{(f)} = \frac{1}{2}\sum_{i=1}^{n}\sum_{\mu=1}^{s_i}\sum_{j=1}^{n}[(a_{i\mu j}^{(f)} - x_{i\mu}^{(f)}(T_f))^2 + \sum_{\eta=1,\eta\neq i}^{n}\sum_{\rho=1}^{p_i}(a_{ij\rho\eta}^{(f)} - x_{ij\rho\eta}^{(f)}(T_f))^2]; \quad (13)$$

$$J_2^{(f)} = \frac{1}{2}\sum_{i=1}^{\overline{n}}\sum_{\mu=1}^{s_i}\sum_{j=1}^{n}\int_{T_0}^{T_f}\beta_{i\mu}^{(f)}(\tau)u_{i\mu j}^{(f)}(\tau)d\tau. \quad (14)$$

The economic meaning of these performance indicators correspond to Eqs. (7) and (8). With the help of the weighting performance indicators, a general performance vector can be denoted as (15):

$$\mathbf{J}(\mathbf{x}(t),\mathbf{u}(t)) = \| J_1^{(o)}, J_2^{(o)}, J_3^{(o)} J_1^{(f)}, J_2^{(f)} \| . \quad (15)$$

The partial indicators may be weighted depended on the planning goals and CTO strategies. Original methods (Gubarev et al. 1988) have been used to transform the vector $\mathbf{J}$ to a scalar form J_G.

The job shop scheduling problem can be formulated as the following problem of OPC: this is necessary to find an allowable control $\mathbf{u}(t)$ $t \in (T_0, T_f]$, that ensures for the model (1), (2), and (10) meeting the vector constraint functions $\mathbf{q}^{(1)}(\mathbf{x},\mathbf{u}) = 0$, $\mathbf{q}^{(2)}(\mathbf{x},\mathbf{u}) \leq 0$ (3), (5) and (10), (11), and guides the dynamic system (i.e., job shop schedule) $\dot{\mathbf{x}} = \varphi(t,\mathbf{x},\mathbf{u})$ from the initial state to the specified final state. If there are several allowable controls (schedules), then the best one (optimal) should be selected in order to maximize (minimize) J_G. In terms of optimal program control (OPC), the program control of job execution is at the same time the job shop schedule. We will refer to this problem of OPC as boundary problem (BP) [3,4].

The formulated model is a linear non-stationary finite-dimensional controlled differential system with the convex area of admissible control. Note that the BP is a standard OPC problem; see [6]. In fact, this model is linear in the state and control variables, and the objective is linear. The transfer of non-linearity to the constraint ensures convexity and allows to use interval constraints.

The representation of the CTO scheduling problem in terms of dynamic system (1)–(15) control problem lets apply for its analysis mathematical tools of the modern control theory [1–4]. For example, the qualitative analysis based on the control theory as applied to the dynamic system (1)–(15) provides the results listed in the Table 1. The table also presents possible directions of practical implementation (interpretation) for these results in real CTO scheduling.

One of the important problems in CTO control system (CTO CS) is the evaluation of goal abilities, i.e., potential of the system to perform its missions

in different situations. Thus, the preliminary analysis of information and technological and goal abilities (GA and ITA) of CTO CS is very important in practice and can be used to obtain reasonable means of the CTO exploitation under different conditions The problem of CTO CS GA and ITA evaluation and analysis can be solved on the basis of structure dynamics control models. These models have a form of nonstationary finite-dimensional differential dynamic systems (NFDDS) with reconfigurable structures. So the problem of GA and ITA evaluation can be regarded as a problem of NFDDS controllability analysis. The latter problem, in its turn, can be solved by the NFDDS attainability set $D(t, T_0, x(T_0))$ construction [7]. If the attainability set is obtained, the solvability of the previously stated boundary problems for structure-dynamics control (SDC) can be checked in accordance with the sets of initial X_0 and final X_f states $(x(T_0) \in X_0, x(T_f) \in X_f))$, with the considered period of time, with time-spatial, technical, and technological constraints.

Table 1. The main results of practical analysis of CTO control processes

Results of qualitative implementation	The main results of practical analysis of CTO control processes	The directions implementation of the results
1	Analysis of solution existence in the problems of CTO control	Adequacy analysis of the CTO control processes description in control models
2	Conditions of controllability and attainability in the problems of CTO control	Analysis CTO control technology realizability on the planning interval. Detection of main factors of CTO goal and information technology abilities
3	Uniqueness condition for optimal program controls in scheduling problems	Analysis of possibility of optimal schedules obtaining for CTO functioning
4	Necessary and sufficient conditions of optimality in CTO control problems	Preliminary analysis of optimal control structure, obtaining of main expressions for CTO scheduling algorithms
5	Conditions of reliability and sensitivity in CTO control problems	Evaluation of reliability and reliability and sensitivity of CTO control processes with respect to perturbation impacts and to alteration of input data contents and structure

Besides the general dynamic model of CTO CS SDC (the model M) its aggregated variants can be used for the attainability-set construction. Let us exemplify this approach via the models M_o, M_k. Interaction operations of the object B_j will be regarded as one aggregated operation, the channels $C_\lambda^{(j)}$ will be replaced by one general channel $C^{(j)}$. Besides, we prescribe $\theta_{i\kappa j\lambda} = 1\ \forall i, \kappa, j, \lambda$ and allow the interruptions of operations. So the aggregated models of object's IO and channels can be stated as follows [3,4]:

$$\dot{\tilde{x}}_i^{(o)} = \sum_{j=1}^{m} \varepsilon_{ij}(t)\tilde{u}_{ij}^{(o)}, \tag{16}$$

$$\dot{\tilde{x}}_{ij}^{(k)} = \sum_{l=1, l\neq i}^{m} \tilde{u}_{li}^{(k)} \frac{h_{li}^{(j)} - \widetilde{x}_{ij}^{(k)}}{\widetilde{x}_{lj}^{(k)}} \gamma_-(\widetilde{x}_{lj}^{(k)}), \tag{17}$$

where $\tilde{x}_i^{(o)} = \sum\limits_{\kappa=1}^{s_i} x_{i\kappa}^{(o)}$, $\tilde{u}_{ij}^{(o)} = \sum\limits_{\kappa=1}^{s_i} u_{i\kappa j}^{(o)}$ are the aggregating functions. The classes $\tilde{K}_\sigma^{(o)}, \tilde{K}_\sigma^{(k)}$ of allowable control inputs are defined as follows:

$$\tilde{K}_\sigma^{(o)} = \{\tilde{U}_\sigma^{(o)} = \|\tilde{u}_{ij}^{(o)}\| \,|\, \sum_{i=1}^{m} \tilde{u}_{ij}^{(o)} \le 1,\ \sum_{i=1}^{m} \tilde{u}_{ij}^{(o)} \le 1, \tilde{u}_{ij}^{(o)}\tilde{x}_{ij}^{(o)} = 0, \tilde{u}_{ij}^{(o)} \in \{0,1\}, \tilde{s}_\sigma^{(o)}\}$$

$$\tilde{K}_\sigma^{(k)} = \{\tilde{U}_\sigma^{(k)} = \|\tilde{u}_{ij}^{(k)}\| \| \sum_{i=1}^{m} \tilde{u}_{ij}^{(k)} \le 1, \tilde{u}_{ij}^{(k)} \in \{0,1\}, \tilde{s}_\sigma^{(k)}\},$$

where $\tilde{s}_\sigma^{(o)}, \tilde{s}_\sigma^{(k)}$ are function-theoretic constraints imposed on the classes of allowable controls. We assume that the control inputs are piecewise continuous functions. We introduce vector $\tilde{\mathbf{x}}^{(o)} = \|\tilde{x}_1^{(o)}, \ldots \tilde{x}_m^{(o)}\|^T$ and vector $\tilde{\mathbf{x}}^{(k)} = \|\tilde{x}_1^{(k)}, \ldots \tilde{x}_m^{(k)}\|^T$. Let $\tilde{\mathbf{x}}^{(o)}(t_0) = 0$, $\tilde{\mathbf{x}}^{(k)}(t_0) = \tilde{\mathbf{x}}_0^{(k)}$. Then the attainability set in the state space of the dynamic system (16), (17) can be obtained as follows:

$$\tilde{D}_{o,k} = \{\tilde{\mathbf{x}} | \tilde{x}_i^{(o)} = \int_{t_0}^{t_f} \sum_{j=1}^{m} \varepsilon_{ij}(\tau)\tilde{u}_{ij}^{(o)}(\tau)d\tau,\ \tilde{U}_\sigma^{(o)} \in \tilde{K}_\sigma^{(o)},$$

$$\tilde{x}_{ij}^{(k)} = \int_{t_0}^{t_f} \sum_{l=1}^{m} \tilde{q}_{lj}(\tau)\tilde{u}_{lj}^{(k)}(\tau)d\tau,\ \tilde{U}_\sigma^{(k)} \in \tilde{K}_\sigma^{(k)}\},$$

where $\tilde{\mathbf{x}} = \|(\tilde{x}^{(o)})^T, (\tilde{x}^{(k)})^T\|^T$, $\tilde{q}_{lj} = \frac{h_{lj}^{(j)} - \tilde{x}_{ij}^{(k)}}{\tilde{x}_{lj}^{(k)}} \gamma_-(\tilde{x}_{lj}^{(k)})$.

The following theorem [4] expresses characteristics of the attainability set.

Theorem. Let the functions $\varepsilon_{ij}(t)$ be nonnegative bounded functions having at most denumerable points of discontinuity, let the classes of allowable controls be defined by (16), (17), then the attainability set $\tilde{D}_{o,k}$ meets the following conditions:

(a) It is bounded, closed, and convex. It lies in the nonnegative or than of the space $X = \mathbf{R}^{m+mm}$;
(b) $\tilde{D}^{-}_{o,k} \subseteq \tilde{D}_{o,k} \subseteq \tilde{D}^{+}_{o,k}$, here

$$\tilde{D}^{-}_{(o,k)} = \{\tilde{\mathbf{x}} | 0 \le \tilde{x}^{(o)}_i \le \bar{\bar{\xi}}_i \tilde{\tilde{x}}^{(o)}_i, 0 \le \tilde{x}^{(k)}_{ij} \le \bar{\bar{\chi}}_i \varphi^{(k)}_{ij}, \bar{\bar{\xi}}_i \ge 0, \sum_{i=1}^{m} \bar{\bar{\xi}}_i = 1, 0 \le \bar{\bar{\chi}}_i \le 1\},$$

$$\tilde{D}^{+}_{(o,k)} = \{\tilde{\mathbf{x}} | 0 \le \tilde{x}^{(o)}_i \le \tilde{\tilde{x}}^{(o)}_i, 0 \le \tilde{x}^{(k)}_{ij} \le \bar{\bar{\chi}}_i \varphi^{(k)}_{ij}, 0 \le \bar{\bar{\chi}}_i \le 1\},$$

where $\tilde{\tilde{x}}^{(o)}_i = \int_{T_0}^{T_f} [\max_{j=1,\dots,m} \varepsilon_{ij}(\tau) d\tau]$ under the conditions $x^{(k)}_{ij} \equiv 0 \forall t, \forall i, \varphi^{(k)}_{ij} = \max_{j=1,\dots,m} \{h^j_{li}\} \forall j$.

The theorem is of high importance for the preliminary analysis of CTO CS control processes, as the calculation of the values $\tilde{x}^{(o)}_i, \varphi^{(k)}_{ij}$, is rather simple, while the sets $D^{-}_{(o,k)}, D^{+}_{(o,k)}$, let, in many cases, verify the end conditions and calculate the range of variation for the measures of CTO CS ITA.

2 Conclusions

Problems of CTO scheduling may be challenged by high complexity, combination of continuous and discrete processes, integrated production and transportation operations as well as dynamics and resulting requirements for adaptability. A possibility to address these issues opens the embedding of OPC into CTO scheduling and using its advantages in combination with advantages of mathematical programming (MP). Under the assumption that the introduction of the dynamic aspect of job arrival can have a significant impact on the solution procedure, this study presented a new original model for CTO scheduling as OPC of job execution dynamics blended with combinatorial optimization and based on a natural dynamic decomposition of the scheduling problem and its solution with maximum principle in combination with MP. The proposed substitution lets use fundamental scientific results of the OPC theory in CTO scheduling.

Acknowledgments. The research described in this paper is partially supported by the Russian Foundation for Basic Research (grants 15-07-08391, 15-08-08459, 16-07-00779, 16-08-00510, 16-08-01277, 16-29-09482-ifi-i, 16-07-00925, 17-08-00797, 17-06-00108, 17-01-00139, 17-20-01214), grant 074-U01 (ITMO University), project 6.1.1 (Peter the Great St. Petersburg Politechnic University) supported by Government of Russian Federation, Program STC of Union State "Monitoring-SG" (project 1.4.1-1), state order of the Ministry of Education and Science of the Russian Federation 2.3135.2017/K, state research 0073-2014-0009, 0073-2015-0007.

References

1. Ivanov, D.A., Sokolov, B.V.: Adaptive Supply Chain Management. Springer, Wiley and Sons, New York (2010)
2. Ivanov, D.A., Sokolov, B.V.: Dynamic supply chain scheduling. J. Schedul. **15**(2), 201–216 (2012)
3. Ohtilev, M.Y., Sokolov, B.V., Yusupov, R.M.: Intellectual Technologies for Monitoring and Control of Structure-Dynamics of Complex Technical Objects. Nauka, Moscow (2006)
4. Kalinin, V.N., Sokolov, B.V.: Optimal planning of the process of interaction of moving operating objects. Int. J. Diff. Eqn. **21**(5), 502–506 (1985)
5. Chen, Z.L., Pundoor, G.: Order assignment and scheduling in a supply chain. J. Oper. Res. **54**, 555–572 (2006)
6. Lee, E.B., Markus, L.: Foundations of Optimal Control Theory. Springer, Wiley and Sons, New York (1967)
7. Chernousko, F.L.: State Estimation of Dynamic Systems. SRC Press, Boca Raton (1994)

Protective Correction of the Flow in Mechanical Transport System

Stanislav Belyakov and Marina Savelyeva(✉)

Southern Federal University, Taganrog, Russia
beliacov@yandex.ru, marina.n.savelyeva@gmail.com

Abstract. In this paper, we will examine the problem of cost minimizing of the cargo moving in mechanical transport systems. Cost Index includes component, which reflects the loss on disaster recovery. We analyze ways to reduce the possibilities of accident initiation. We also consider the features of the adaptive routing. We define the conditions under which it can be used to manage by the intensity of the flow through separate network segments. We propose adaptive routing algorithm with protective correction of flow. The essence of the algorithm consists in installing high value of the cost of the transfer on separate segments. This value is fixed as a periodic event in the specified time window. We consider the factors that determine the value of protective correction of the parameters. We have identified the problem of finding the best values. We have proposed the mechanical transport system structure, which includes an intelligent module of instruction issue of protective correction. We consider the work principles of intelligent module based on case analysis with a hierarchical storage system of precedents.

Keywords: Mechanical transport system (MTS) · Dynamic routing · Adaptive routing · Protective correction

1 Introduction

Mechanical transport system (MTS) is a network built from elements of two types: conveyors and switches. Cargo is moved conveyors placed on belt. Switches act as network nodes, in these nodes the forwarding of cargo units is performed from one conveyor to another. An example of MTS is an automatic baggage handling system at the airports. The management system of conveyors and switches is implemented as a local area network of industrial controllers PLC (Programmable Logic Controllers) [1]. The controllers have the management task by switch of cargo direction and conveyor electric drive.

Each unit is provided with a cargo label. Addresses initial and final nodes are stored in this label. These data permit in the intermediate network nodes to determine the direction of the cargo unit transfer, solving the routing problem [2–6]. Due to the fact, how is constructed routing algorithm, depend such important factors as the cost of transportation, the risk of damage or loss of cargo, delivery efficiency [1, 7].

The problem of transportation management in the MTS can be formulated as follows:

R. Silhavy et al. (eds.), *Cybernetics and Mathematics Applications in Intelligent Systems*,
Advances in Intelligent Systems and Computing 574, DOI 10.1007/978-3-319-57264-2_18

$$\begin{cases} \sum_i l_i + w_{f_i} \to min, \\ t < t^*, \\ f_i \subseteq F, \end{cases} \tag{1}$$

where l_i is the transportation cost of each several cargo unit;
w_{f_i} are loss in the event of defect $f_i \subseteq F$;
F is set of possible defects that lead to emergency;
t^* is transportation time limit. The main means of solving the problem (1) is the routing. This is explained by the fact that:

- the transmission path of the cargo unit is not uniquely determined, and the route cost are different in MTS;
- the intensity of the cargo flow on an individual segment determines the possibility of defect occurrence.

Modern MTSs use dynamic routing [8, 9]. It helps to minimize the total cost of transportation. However, emergencies arising from overload segments are the result of an unacceptable increase flow intensity in some parts of the network. In this work, we propose a modification of the method of adaptive routing, allowing to solve this problem.

2 Routing Methods and Flow Control in MTS

The solution of the problem (1), using a fixed routing [10], is possible in the case of complete certainty the behavior of the MTS and outdoor environment. It means high reliability and stability of operational performance of MTS, strict conformance to the schedule of appearance and the completion of cargo flows, the stability properties of the cargo units. MTS in this case is described by a static model network [11], often used for calculations. The practical application of fixed routing is limited to the above conditions.

The fixed routing can be based on the dynamic network model [12]. Routing tables in the nodes are updated on a predetermined schedule. This approach is more adequate to the real situation, when the cargo is unstable and MTS parameters change over time. In this case, the complexity of the synthesis and analysis of dynamic models is much higher [5]. This creates difficulties in solving the problem (1) in real time scale.

Dynamic routing [13] solves problem (1) by adjusting of the routing tables in real time when changing the transport costs on individual segments. Cost is determined by the measured parameters network: direction selection speed, electric drive rod, elapsed time of device, etc. The disadvantage of dynamic routing is the inability to consider the transportation cost of the temporal parameters of traffic flows and properties of cargo units that are not available for the measurement. Because of that, it's possible overload. In addition, the significant role is playing by the lag of the mechanical part of the MTS. Routing table modification is completed much faster than the change in cargo flow value. And output stream may be unacceptably high in nodes, summing streams. The problem should be solved in advance decrease the flow value in dangerous situations.

The closest method of solution to the problem is to attract intelligent control mechanisms [14]. The incompleteness of the data, the time history of cargo flow sources make motivation for the use of intelligent observation over the MTS and predict of the cargo flow behavior. Attraction of expert-observer knowledge allows to generate dynamic routing strategy based on a holistic perception of the outdoor environment and MTS. The disadvantage of this approach is the lack of a protective mechanism of the flow correction to prevent accidents.

3 Protective Mechanism of the Flow Correction

Dynamic routing is implemented by changing the transfer value of individual segments of the MTS and warning about it neighboring nodes. Controller of each node corrects routing tables and send cargo units with the switch in the direction that minimizes the total cost of transportation. It will be observed, that dynamic routing does not control the flows, even though indirectly, affect their value. For instance, the low cost of transportation through the subnet stimulates the growth of the flow. Accordingly, a high value may lead to lower flow value. Since the danger an emergency is directly related to the flow value, we have an idea to use the cost of transportation as security facilities from accidents.

Protective flow correction is an artificial increase the transportation cost through the network segment to reduce the flow value. The parameters of the protective correction element are a pair (C_{s_i}, T_{s_i}), where C_{s_i} is cargo unit cost of transportation by the segment s_i; T_{s_i} is time window, during which the value of the cost is kept. Parameter determination of the protective correction is a non-trivial task, at the decision which found a few uncertainties:

- C_{s_i} should be chosen so as, don't to completely block flow through the segment and at the same time providing real decrease its rate. It requires an analysis of the number of cargo elements, that are in the MTS, the ways of their movement and changes in the rates of the input flows at the time of deciding upon the protective correction;
- T_{s_i} is determined in such a way so as don't to provoke an overload of other segments of the MTS. Figure 1 illustrates the general pattern of selection of value T_{s_i}. The greater its value is, the lower the probability P_{s_i} of occurrence of overload in the segment s_i. However, inevitably increases probability P_{N^-} of occurrence of overload on a subnet N^-, not consisting segment s_i. As follows from the qualitative analysis (Fig. 1), there is a compromise value $T^*_{s_i}$, deviation from which increases the probability of occurrence of accidents. As with the analysis of the parameters it need information about the network load, the response time to changes in the routing tables, variations of the rates of the input flows.

The need to consider the whole situation makes it unlikely the effective use of protective correction as the decentralized management tool of MTS. Means the following: the node controller of the MTS measures the flow rate v_{s_i}. If the threshold is

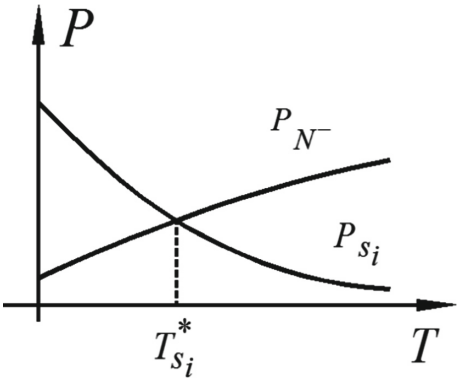


Fig. 1. Shows an illustration of changes in the probability of overload

exceeded a predetermined value $\overline{v}_{s_i}$, then the controller sets protective correction with given parameters (C_{s_i}, T_{s_i}). In this case, there is dangers:

– accidents within the segment s_i due to the fact that the input flow is not reduced immediately, and flow may increase for some time inertia system;
– occurrence of deadlock, when the segment s_i may remove the protective correction, if the input flow rate v_{s_i} decrease. But the input flow rate v_{s_i} cannot be reduced as long as it will not remove the protective correction and the segment does not restore previous capacity.

Thus, we can conclude about the necessity of central determining the parameters of protective correction. Management is implemented in a lack of information about the behavior of the outdoor environment and this MTS, that indicate necessity of application of intelligent principles of management system. Figure 2 shows the structure of the system. Intelligent control module (ICM) is included in the control network PLC. PLC is associated with control devices of MTS. PLC implement dynamic routing algorithm, and can perform the ICM commands on the flow protective correction. The command contains the following fields:

1. timestamp start of protective correction;
2. timestamp end of protective correction;
3. address of output of directional switch;
4. value of the transportation cost to the specified output.

The controller, receiving the command, implements the dynamic routing algorithm with fixed transmission cost by the output network segment.

ICM operates based on using the experience of observation of MTS [14]. Experience is represented by describing previously observed dangerous situations and decisions in these situations. The logical inference is based on the case-based reasoning [15]. Every new problem situation is compared with the known to find the nearest in meaning. The solution found situation is applied in the new situation. It should be highlighted that each of the known situations in the ICM knowledge base indicates a potential risk of accidents,

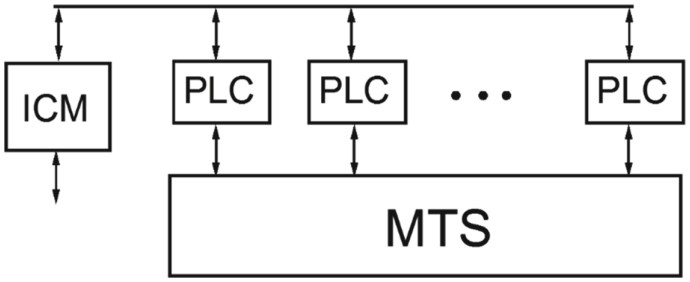


Fig. 2. Illustrates the structure of MTS with an intelligent control module.

i.e. it is forecast. The forecast is not absolutely reliable, so the "intelligence" of the system is manifested in the effort to prevent the occurrence of an abnormal situation and loss recovery.

4 Conclusion

The efficiency of flow protective correction is determined by the ratio of the loss on crash recovery and increased transportation costs. It follows from (1), protection from accidents makes sense if

$$\sum\nolimits_i l_i \ll \sum\nolimits_i w_{f_i} \tag{2}$$

Since protective correction leads to an increase transportation cost of the formula (2) can become the requirement to forecast reliability. Let P be the probability that the accident forecast is realized. Then the average loss from a wrong the forecast is

$$\overline{W} = (1 - P) \triangle L \tag{3}$$

where $\triangle L$ is the increase in the transportation cost during runtime of protective correction. Transform the formula (3) in (2) we find that

$$P \gg 1 - \frac{\sum_i w_{f_i}}{\triangle L} \tag{4}$$

The resulting expression reflects the requirement to knowledge of the intelligent system in the problem of flow protective correction.

Further research, in our view, should be in the direction of improving the presentation and use of knowledge just as inside one of MTS, so when moving knowledge between different systems.

Acknowledgments. This work has been supported by the Russian Foundation (under RF President), Project № MK-521.2017.8.

References

1. Yan J.,, Vyatkin V.: Distributed execution and cyber-physical design of baggage handling automation with IEC 61499. In: 9th International IEEE Conference on Industrial Informatics (INDIN 2011), Lisbon, pp. 573–578, July 2011
2. Bellman, R.: On a routing problem. Q. Appl. Math. **16**, 87–90 (1958)
3. Laporte, G.: The vehicle routing problem: an overview of exact and approximate algorithms. Eur. J. Oper. Res. **59**(3), 345–358 (1992)
4. Golden, B.L., Raghavan, S., Wasil, E.A. (eds.): The Vehicle Routing Problem: Latest Advances and New Challenges, vol. 43. Springer, New York (2008)
5. Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C.: Introduction to Algorithms, 3rd edn. 1312 p. MIT Press, Cambridge (2009)
6. Toth, P., Vigo, D. (eds.): Vehicle Routing: Problems, Methods, and Applications, vol. 18. SIAM, Philadelphia (2014)
7. Pang, C., Yan, J., Vyatkin, V., Jennings, S.: Distributed IEC 61499 material handling control based on time synchronization with IEEE 1588. In: IEEE International Symposium on Precision Clock Synchronization for Measurement, Control, and Communication, Munich, pp. 126–131, September 2011
8. Ash, G.R.: Dynamic Routing in Telecommunications Networks, 746 p. McGraw-Hill Professional, New York (1997)
9. Kodialam, M., Lakshman, T.V.: Dynamic routing of bandwidth guaranteed tunnels with restoration. In: Proceedings of the Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies, INFOCOM 2000, vol. 2, pp. 902–911. IEEE (2000)
10. Kleinrock, L., Gail, R.: Queueing Systems: Problems and Solutions, 240 p. Wiley-Interscience, New York (1996)
11. Guizani, M., Rayes, A., Khan, B., Al-Fuqaha, A.: Network Modeling and Simulation: A practical perspective, 304 p. Wiley, Hoboken (2010)
12. Mak, T., Cheung, P.Y.K., Lam, K.-P., Luk, W.: Adaptive routing in network-on-chips using a dynamic-programming network. IEEE Trans. Ind. Electron. **58**(8), 3701–3716 (2011)
13. Aksaray, D., Vasile, C.I., Belta, C.: Dynamic routing of energy-aware vehicles with temporal logic constraints. In: IEEE International Conference on Robotics and Automation (ICRA), pp. 3141–3146. IEEE (2016)
14. Belyakov, S., Savelyeva, M., Yan, J., Vyatkin, V.: Adaptation of material flows in mechanical transportation systems based on observation experience. In: Trustcom/BigDataSE/ISPA, vol. 3, pp. 269–274. IEEE, August 2015
15. Savelyeva, M.: The construction of the fuzzy route based on case-based reasoning. In: Belyakov, S., Rozenberg, I., Savelyeva, M. (eds.) Proceeding of the 19th International Conference on Soft Computing MENDEL, pp. 273–276 (2013)

Efficient MapReduce Matrix Multiplication with Optimized Mapper Set

Methaq Kadhum, Mais Haj Qasem(✉), Azzam Sleit, and Ahamd Sharieh

King Abdullah II School for Information Technology, Computer Science Department, University of Jordan, Amman, Jordan
methaq.kadhum@yahoo.com, mais_hajqasem@hotmail.com, {azzam.sleit,sharieh}@ju.edu.jo

Abstract. The efficiency of matrix multiplication is a popular research topic given that matrices compromise large data in computer applications and other fields of study. The proposed schemes utilize data blocks to balance processing overhead results from a small mapper set and I/O overhead results from a large mapper set. Balancing between the two processing steps, however, consumes time and resources. The proposed technique uses a single MapReduce job and pre-processing step. The pre-processing step reads an element from the first array and a block from the second array prior to merging both elements into one file. The map task performs the multiplication operations, whereas the reduce task performs the sum operations. Comparing the proposed and existing schemes reveals that the proposed schemes more efficiently consume time and memory.

Keywords: Hadoop · MapReduce · Matrix multiplication · Optimized mapper set

1 Introduction

Matrix multiplication is a fundamental operation in linear algebra with related real-life applications, such as matrix factorization, chemical system formulation, and graph analysis [14]. In addition to its naturally related applications, several problems are reducible by matrix multiplication. Thus, these problems should be investigated thoroughly to enhance the efficiency of implemented algorithms for matrix multiplication. Given the inputs of two matrices A and B, where the number of columns in A equals the number of rows in B, matrix multiplication produces matrix C with the number of rows equal to that in A and number of columns equal to that in B. The Brute-Force matrix multiplication algorithm for square matrices is given in Algorithm 1. The Brute-Force algorithm has a high processing complexity, of O(n3), but suffers from the massive memory lookup process required to locate each array element for multiplication. Over the years, several matrix multiplication algorithms have been proposed to reduce the cost and time of the matrix multiplication process [2, 15].

R. Silhavy et al. (eds.), *Cybernetics and Mathematics Applications in Intelligent Systems*, Advances in Intelligent Systems and Computing 574, DOI 10.1007/978-3-319-57264-2_19

Algorithm 1. Brute-Force Matrix Multiplication

```
Matrix Multiplication(A[1..n,1..n], B[1..n,1..n])
1:   FOR(i:= 1 to n)
2:     FOR(j:=1 to n)
3:       C[i,j]=0;
4:       FOR(k:=1 to n)
5:             C[i,j]= C[i,j]+A[i,k]*B[k,j]
6: EndFor
7:     EndFor
8:   EndFor
9:   return C
End
```

MapReduce is an algorithm design and processing paradigm that was proposed by Dean and Ghemawat in 2004 [4]. MapReduce enables efficient parallel and distributed computing and consists of two serial tasks, Map and Reduce. Each serial task is implemented with several parallel sub-tasks. Map task, the first task in MapReduce, accepts input for conversion into a different form as the output. In map task, both input and output data are formed using a series of elements with individual key-value pairs. The reduce task takes the map output and implements an aggregation process for pairs with identical keys. Although composing or implementing the algorithm in the map and reduce tasks for execution in MapReduce are nothing but trivial, the gain of such a decomposition process is massive. Thus, the program can be run over hundreds and thousands of parallel nodes over a cluster of machines [5].

Hadoop is a Java open-source platform used for developing MapReduce applications. This platform was developed by Google [6]. The Hadoop architecture is illustrated in Fig. 1.

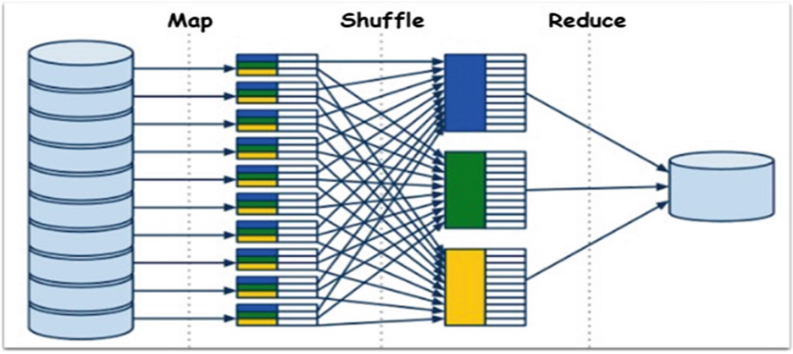


Fig. 1. Hadoop MapReduce architecture

As shown in Fig. 1, the Hadoop framework is responsible for distributing the input into the involved mappers. These mappers implement the map task, collect the results for sorting during the shuffle process, and feed and collect the output of the reducers.

The MapReduce paradigm has been used to decompose enormous tasks, such as data-mining algorithms. Specific MapReduce paradigms include: MapReduce with expectation maximization for text filtering [11], MapReduce with K-means for remote-sensing image clustering [12], and MapReduce with decision tree for classification [22]. Additionally, MapReduce has been used for job scheduling [23] and real-time systems [11].

Matrix multiplication that uses MapReduce has been proposed [9]. The earlier decomposition process of matrix multiplication involved two MapReduce tasks. However, problems arose with these decomposition processes: the processing overhead and file I/O overhead were obvious. Hence, it was necessary to re-decompose the matrix multiplication process in the MapReduce paradigm to enhance and decrease the computing overhead.

This paper proposes a technique for matrix multiplication. The technique uses a single MapReduce task with an optimized mapper set. The optimal number of mappers that formed the utilized mapper set is selected to balance the processing overhead results of a small mapper set and the I/O overhead results of a large mapper set. These two processes consume time and resources.

The rest of the paper is organized as follows: Sect. 2 reviews work that is closely related to the implemented MapReduce matrix multiplication task. Section 3 presents the proposed work for matrix multiplication and highlights the relation between proposed and previous techniques. Section 4 presents the experimental results. Finally, the conclusion is given in Sect. 5.

2 Related Work

The traditional sequential algorithms for matrix multiplication consume considerable space and time. To enhance the efficiency of matrix multiplication, the Fox [1], Cannon [14], and DNS [7] algorithms have been proposed to parallelize the matrix multiplication process. To maximize efficiency, these approaches balance inter-process communication, dependencies, and parallelism level. Parallel matrix multiplication depends on the independence of the multiplication process, which includes multiple independent element-to-element multiplications and multiple aggregations of independent multiplication results, as illustrated in Fig. 2.

Traditional parallel-based matrix multiplication was recently replaced with MapReduce, a parallel and distributed framework for large-scale data [3]. Typical MapReduce-based matrix multiplication requires two MapReduce jobs. The first job creates a pair of elements for multiplication by combining input arrays together during the map task. The reduce task of this job is inactive at this point. In the second job, the map task independently implements the multiplication operations on each pair of elements. The reduce job aggregates the results that correspond to each output element. The overall scheme of this technique is element-to-element, because each mapper implements

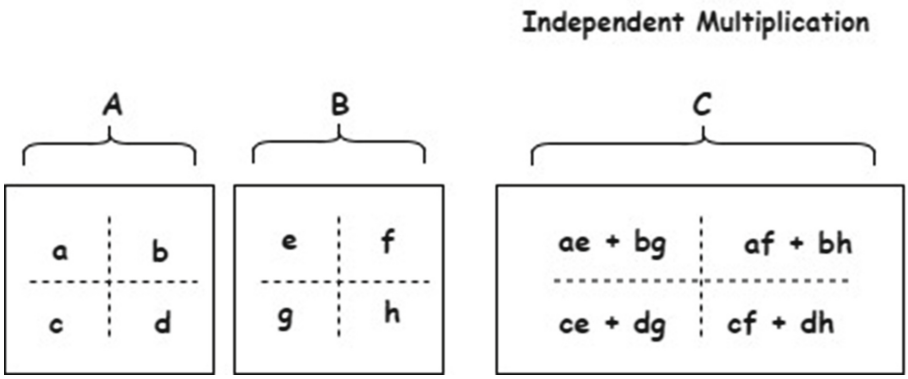


Fig. 2. Matrix multiplication independency process

Table 1. Element- by- element operations

Scheme		Input	Output
Element-by-element	Map	Files	$< a_{ij}, b_{kj} >$
	Reduce	–	
	Map	$< a_{ij}, b_{kj} >$	$<$ key, $a_{ij} * b_{kj} >$
	Reduce	$<$ key, $[a_{ij} * b_{kj} \ldots \ldots a_{ij} * b_{kj}] >$	$<$ key, $c_{ij} >$

multiplication element by element, as illustrated in Fig. 3. The operations implemented by the involved MapReduce jobs are presented in Table 1. Overall, one job is responsible for obtaining input elements from input arrays and the other is responsible for the actual matrix multiplication process. This approach is problematic given its requirement for high sorting and numerous map tasks. Note that the sorting process is implemented by the shuffle task in the MapReduce platform.

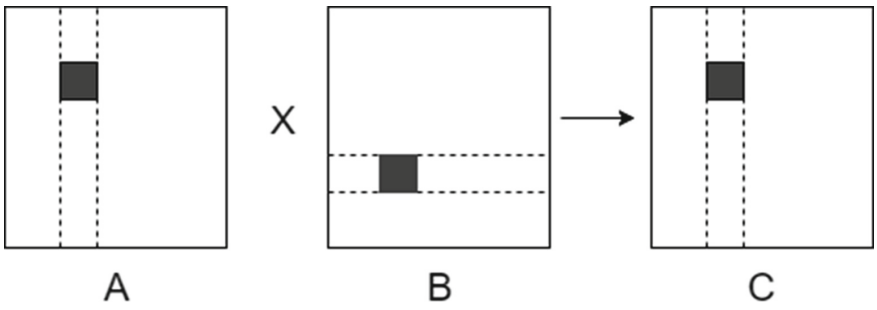


Fig. 3. Element- by- element matrix multiplication

A blocking scheme was proposed to overcome the disadvantages of the element-by-element scheme and to reduce overall computational cost. Sun et al. [24] proposed a MapReduce matrix factorization approach. Matrix multiplication, a significant part of

factorization, was carefully investigated to achieve efficiency. Two matrix multiplication jobs were used to accomplish the multiplication process. The process depended on the decomposition of the first matrix into row vectors and the second matrix into column vectors. Multiplication of the elements of these blocks was implemented on a single mapper. Thus, communication overhead and intermediate memory utilization were minimized. However, the number of computational process per-mappers increased.

Jianhua et al. [8] argued that such processes require time as the mapper computation costs are high and input writing consumes memory. Thus, two matrix multiplication jobs decomposed the first matrix into elements or columns vectors and the second matrix into rows instead of columns vectors. A single mapper multiplied the elements of these blocks. Then, aggregation was implemented over the producers. Therefore, the results of each mapper corresponded to multiple elements in the output array.

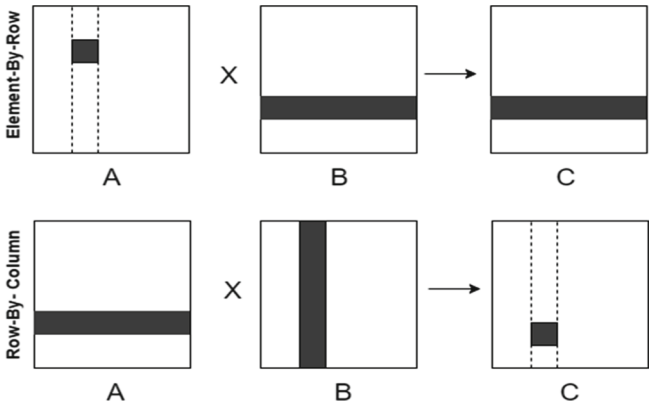


Fig. 4. Deng and Wu-schemes for matrix multiplication

Table 2. Proposed scheme operations

Scheme		Input	Output
Element-by-row-block	Pre-Process	Files	$< a_{ij}, b_{kj}, b_{kj} \ldots \ldots .. >$
	Map	$< a_{ij}, b_{kj}, b_{kj} \ldots \ldots .. >$	$< \text{key}, [a_{ij} * b_{kj}] \ldots \ldots \ldots [a_{ij} * b_{kj}] >$
	Reduce	$< \text{key}, c_{ij}, c_{ij} \ldots \ldots >$	$< \text{key}, c_{ij} + c_{ij} \ldots \ldots >$
Row-block-by-column-block	Pre-Process	Files	$< a_{ij}, a_{ij} \ldots \ldots b_{kj}, b_{kj} >$
	Map	$< a_{ij}, a_{ij} \ldots b_{kj}, b_{kj} >$	$< \text{key}, [a_{ij} * b_{kj}] \ldots \ldots \ldots [a_{ij} * b_{kj}] >$
	Reduce	$< \text{key}, c_{ij}, c_{ij} \ldots \ldots >$	$< \text{key}, c_{ij} + c_{ij} \ldots \ldots >$

Deng and Wu [16] presented the experimental results of the element-to-element scheme. Moreover, they presented block-based, element-to-column and row-to-column matrix multiplication, as illustrated in Fig. 4. The operations implemented by the involved MapReduce jobs are presented in Table 2. Their experiments showed that the element-to-row scheme ran faster than row-to-column. In turn, row-to-column was

faster than the element-to-element scheme. Moreover, the best scheme had medium input sizes and involved a medium number of mappers. Thus, their results suggested the need to balance input size with mapper number.

In addition to the blocking scheme, reducing the number of MapReduce jobs from two to one also reduced the overall computational cost for matrix multiplication. Therefore, inputs in the MapReduce Job should be as blocks. Each block should contain elements from both matrices to be multiplied. To reduce computational cost and memory consumption, Deng and Wu [4, 21] modified the way Hadoop read I/O files. In the HAMA project [21], a pre-processing stage was implemented for the same purpose.

3 Proposed Work

The goal of the proposed work is to enhance the efficiency of the matrix multiplication in MapReduce framework. This is implemented by balancing between the processing overhead results from using a small mapper set and the I/O overhead results from using large mapper set, which both leads to consume time and resources, based on our previous arguments in Sect. 2.

In the proposed technique, matrix multiplication is implemented as an element-to-block scheme, as illustrated in Fig. 5. In the first schema; first array is decomposed into individual elements, whereas the second array is decomposed into sub-row-based blocks, while the second schema; first array is decomposed into sub-row-based blocks, and the second array is decomposed into sub-column-based blocks. The number of mappers is determined by the size of the block that is generated for the second array and selected on the basis of the capability of the underlying mapper. Subsequently, a smaller block size increases the number of blocks, thus requiring more mappers and vice versa.

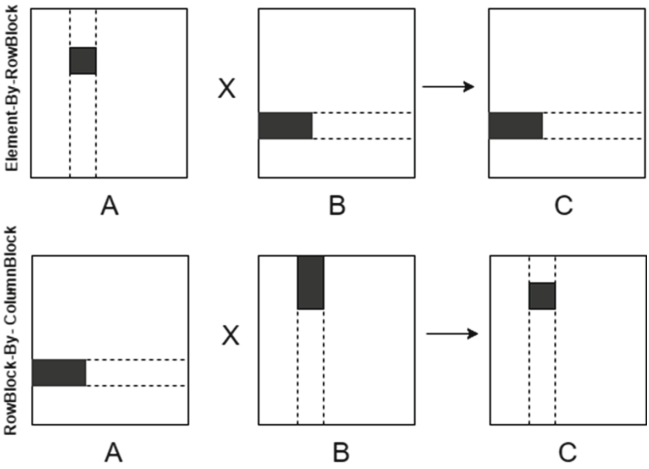


Fig. 5. Proposed schemes for matrix multiplication

This work uses a single MapReduce job. The map task, as listed in Table 2, is responsible for the multiplication operations, whereas the reduce task is responsible for the sum operations. The pre-processing step reads an element from the first array and a block from the second array, and then merges them into one file. Note that in matrix multiplication, the whole row in the first array has to be multiplied with the whole column in the second array to calculate the results of an element in the output. Thus, the results of each mapper in the proposed schemes are aggregated with other multiplication results in the reduce task.

Compared with existing schemes, the proposed work utilizes one MapReduce job instead of two. The number of multiplications handled by a mapper is dependent on the capability of the mapper, which is determined by block size. Previous work has investigated element-by-element (from the first and second arrays), element-by-column, and row-by-column multiplications. Varying the number of elements in rows and columns in different inputs revealed that the best result involves medium size input because the processing overhead at each mapper is ignored. Hence, to match the capabilities of that mapper, we proposed to vary the number of elements given to the mapper.

Unlike previous techniques, this work proposes to multiply an element by a block of elements. The block varies from a single element into a complete row. If the block size is equal to one, then the proposed work will be identical to an element-to-element scheme. However, if the block size is equal to the dimension of the input array, then the proposed work will be identical to the element-to-row/column scheme. Subsequently, the previous work is considered as a special case of our general proposed work. Table 3 compares the proposed and existing schemes.

Table 3. Proposed schemes vs. existing schemes

Existing schemes	Element- by-element	Element- by-row	Row- by-column
Process dependency and synchronization	$n^3 \rightarrow$ n	$n^2 \rightarrow$ n	$n^2 \rightarrow$ n
Number of mappers	n^3	n^2	n^2
Number of replicated elements among mappers	Minimum of: n	Minimum of: n	Minimum of: n
Shuffle traffic	n^3	n^2	n^2
Proposed Schemes	**Element- By-Column-Block**	**Column-Block By-Row-Block**	
Process dependency and synchronization	$n^2 * \#q \rightarrow$ n where q = size block	$n^2 * \#q \rightarrow$ n where q = size block	
Number of mappers	$n^2 * \#q$ where q = size block	$n^2 * \#q$ where q = size block	
Number of replicated elements among mappers	Minimum of: n	Minimum of: n	

4 Experiments and Result

The results of matrix multiplication using Hadoop for inputs with various size is presented. Sparse matrices of size n*n are randomly generated with numbers from 1–10. The experiments are conducted for various block size varied in the range [1−n]. In this work, we run a simple matrix multiplication process with size 100*100 on the platform with various block size varied in the range [1,10,15,20,25,30] in-order to determine the optimal length to be given to the mapper before running the actual job. The pre-experiments are shown in Table 4.

Table 4. Proposed scheme run time result

Block size	Run time (MS)
1	751235
10	543686
15	1688
20	53798
25	36314
30	237079

Based on experiments results with various block in the range [1,10,15,20,25,30], we determined that the optimal length of block size with a minimum run time is 15, So after that, we fixed the block sizes and run matrix multiplication process with various matrix sizes in existing schema and our proposed schema.

The results are reported as given in Table 5. As noted, the running time is cut down in the proposed scheme, especially, for element-by-column block scheme, in which the sorting

Table 5. Proposed scheme vs. existing schema run time result

Existing schema			
Matrix	Element-by-element	Element-by-row	Row-by-column
500 * 100	751235	105145	237079
500 * 500	1715043	120562	238688
1000 * 1000	2500256	195855	500256
2000 * 2000	2621523	543686	621523
4000 * 4000	2721523	534124	621523
Proposed schema			
Matrix	Element-by-columnblock	ColumnBlock by-rowblock	
500 * 100	1680	36314	
500 * 500	17716	36256	
1000 * 1000	42063	121542	
2000 * 2000	76685	121555	
4000 * 4000	78325	325478	

Table 6. Proposed scheme vs. existing schema memory consumption result

Existing schema			
Matrix	Element-by-element	Element- by-row	Row-by-column
500 * 100	36000	37004	36314
500 * 500	56310	55457	56214
1000 * 1000	110241	112400	114000
2000 * 2000	189254	189475	189124
4000 * 4000	212524	212142	212471
Proposed schema			
Matrix	Element-by-columnblock	Column-block by-roblock	
500 * 100	36766	36314	
500 * 500	55000	56310	
1000 * 1000	111454	111245	
2000 * 2000	190254	189456	
4000 * 4000	212441	212111	

process in the shuffle is reduced. As the matrix size growth, the stability of the proposed scheme in better compared to the existing schemes, which, seems to be almost linear.

The results for space consumption for the proposed and existing schemes are reported as given in Table 6. As noted, proposed and existing schemes almost identical but the proposed work takes slightly more spaces compared to others. Therefore, if the user cares about time our proposed schema is the best choice, but if he cares about the memory capacity he can choose from another algorithm.

Our algorithm is written in java and the experimental results are calculated for our Proposed Schemes and Existing Schemes on HP® core™ i7-5500U CPU @ 2.40 GHz /8 GB RAM.

5 Conclusion

A block-based matrix multiplication schemes were proposed in this paper. The proposed schemes balance between the processing overhead results from using a small mapper set and the I/O overhead results from using large mapper set, which both leads to consume time and resources. This balancing is optimizing by determining the optimal block size and number of involved mappers. The results show that the proposed schemes reduce both time and memory utilization.

6 Future Work

Our proposed schema is implemented on sparse algorithm, our future work will be on dense algorithm, in other hand, we can optimized reduce set.

References

1. Cannon, L.E.: A Cellular Computer to Implement the Kalman Filter Algorithm. No. 603-Tl-0769. Montana State Univ Bozeman Engineering Research Labs (1969)
2. Coppersmith, D., Winograd, S.: Matrix multiplication via arithmetic progressions. In: Proceedings of the Nineteenth Annual ACM Symposium on Theory of Computing, pp. 1–6. ACM (1987)
3. Catalyurek, U.V., Aykanat, C.: Hypergraph-partitioning-based decomposition for parallel sparse-matrix vector multiplication. IEEE Trans. Parallel Distrib. Syst. **10**(7), 673–693 (1999)
4. Dean, J., Ghemawat, S.: Mapreduce: simplified data processing on large clusters. In: OSDI, p. 10. USENIX (2004)
5. Dean, J., Ghemawat, S.: MapReduce: a flexible data processing tool. Commun. ACM **53**(1), 72–77 (2010)
6. Dean, J., Ghemawat, S.: MapReduce: Simplified data processing on large clusters. Commun. ACM **51**(1), 107–113 (2008)
7. Dekel, E., Nassimi, D., Sahni, S.: Parallel matrix and graph algorithms. SIAM J. Comput. **10**(4), 657–675 (1981)
8. Deng, S., Wenhua, W.: Efficient matrix multiplication in hadoop. Int. J. Comput. Sci. Appl. **13**(1), 93–104 (2016)
9. Fox, G.C., Otto, S.W., Hey, A.J.G.: Matrix algorithms on a hypercube I: Matrix multiplication. Parallel Comput. **4**(1), 17–31 (1987)
10. Lin, J., Dyer, C.: Data-intensive text processing with MapReduce. Synth. Lect. Hum. Lang. Technol. **3**(1), 1–177 (2010)
11. Liu, X., Iftikhar, N., Xie, X.: Survey of real-time processing systems for big data. In: Proceedings of the 18th International Database Engineering & Applications Symposium. ACM (2014)
12. Lv, Z., Hu, Y., Zhong, H., Wu, J., Li, B., Zhao, H.: Parallel K-means clustering of remote sensing images based on MapReduce. In: Wang, F.L., Gong, Z., Luo, X., Lei, J. (eds.) WISM 2010. LNCS, vol. 6318, pp. 162–170. Springer, Heidelberg (2010). doi: 10.1007/978-3-642-16515-3_21
13. Mahafzah, B.A., Sleit, A., Hamad, N.A., Ahmad, E.F., Abu-Kabeer, T.M.: The OTIS hyper hexa-cell optoelectronic architecture. Computing **94**(5), 411–432 (2012)
14. Norstad, J.: A mapreduce algorithm for matrix multiplication (2009). http://www.norstad.org/matrix-multiply/index.html. Accessed 19 Feb 2013
15. Thabet, K., Al-Ghuribi, S.: Matrix multiplication algorithms. Int. J. Comput. Sci. Netw. Secur. (IJCSNS) **12**(2), 74 (2012)
16. Seo, S., Yoon, E.J., Kim, J., Jin, S., Kim, J.S., Maeng, S.: Hama: An efficient matrix computation with the mapreduce framework. In: 2010 IEEE Second International Conference on Cloud Computing Technology and Science (CloudCom), pp. 721–726. IEEE, November 2010
17. Sleit, A., Al-Akhras, M., Juma, I., Alian, M.: Applying ordinal association rules for cleansing data with missing values. J. Am. Sci. **5**(3), 52–62 (2009)
18. Sleit, A., Dalhoum, A.L.A., Al-Dhamari, I., Awwad, A.: Efficient enhancement on cellular automata for data mining. In: Proceedings of the 13th WSEAS International Conference on Systems, pp. 616–620. World Scientific and Engineering Academy and Society (WSEAS), July 2009
19. Sleit, A., AlMobaideen, W., Baarah, A.H., Abusitta, A.H.: An efficient pattern matching algorithm. J. Appl. Sci. **7**(18), 269–2695 (2007)

20. Sleit, A., Saadeh, H., Al-Dhamari, I., Tareef, A.: An enhanced sub image matching algorithm for binary images. In: American Conference on Applied Mathematics, pp. 565–569, January 2010
21. Sun, Z., Li, T., Rishe, N.: Large-scale matrix factorization using mapreduce. In: 2010 IEEE International Conference on Data Mining Workshops. IEEE (2010)
22. Wu, G., et al.: MReC4.5: C4.5 ensemble classification with MapReduce. In: 2009 Fourth ChinaGrid Annual Conference. IEEE (2009)
23. Zaharia, M., et al.: Job scheduling for multi-user mapreduce clusters. EECS Department, University of California, Berkeley, Technical Report UCB/EECS-2009-55 (2009)
24. Zheng, J., Zhu, R., Shen, Y.: Sparse matrix multiplication algorithm based on MapReduce. J. Zhongkai Univ. Agric. Eng. **26**(3), 1–6 (2013)

Control of Time-Delay Systems with Parametric Uncertainty via Two Feedback Controllers

Radek Matušů(✉) and Roman Prokop

Faculty of Applied Informatics,
Centre for Security, Information and Advanced Technologies (CEBIA – Tech),
Tomas Bata University in Zlínnám,
nám. T. G. Masaryka 5555, 760 01 Zlín, Czech Republic
{rmatusu, prokop}@fai.utb.cz

Abstract. The main goal of this contribution is to present the application of polynomial approach-based design of two feedback controllers to time-delay plants with parametric uncertainty. Robust stability of designed control systems is analyzed via the families of their closed-loop characteristic quasi-polynomials, more specifically by means of the graphical method which combines the value set concept with the zero exclusion condition. A second order plus time delay plant with uncertain parameters is robustly stabilized (or intentionally got to the robust stability border) in the simulation example.

Keywords: Two feedback controllers · Time-delay systems · Parametric uncertainty · Robust stabilization · Polynomial control

1 Introduction

The time delays [1] and (parametric) uncertainties [2, 3] belong among the most studied phenomena in control theory since an appropriate control of the systems affected by them is motivated by many real applications. The control loop with two feedback controllers (TFC) represents a system in which the weight coefficients for two individual controllers can be chosen [4, 5]. It means that this structure offers a wide range of tuning options. Two extreme cases of possible choice correspond either to the standard one-degree-of-freedom (1DOF) control system, or (under some presumptions) to the two-degree-of-freedom (2DOF) configuration.

This paper is intended to present a possible approach to control of time-delay systems with parametric uncertainty by means of two feedback controllers. The control design is based on an approximation of time-delay term, the polynomial method [4, 5] and solution of Diophantine equations [6]. The subsequent robust stability analysis of the resulting closed-loop characteristic quasi-polynomials (including the original, non-approximated, time-delay term) uses the combination of the value set concept and the zero exclusion condition [2]. In the simulation example, a second order time-delay system with uncertain parameters is robustly stabilized by using TFC structure.

R. Silhavy et al. (eds.), *Cybernetics and Mathematics Applications in Intelligent Systems*,
Advances in Intelligent Systems and Computing 574, DOI 10.1007/978-3-319-57264-2_20

This contribution is a follow-up to the previous works [7–9] where time-delay free interval plants were robustly stabilized by the same approach.

2 Design of Control Systems with TFC

The structure of the continuous-time control system with TFC, namely $C_Q(s) = \tilde{q}(s)/\tilde{p}(s)$ and $C_R(s) = r(s)/\tilde{p}(s)$, and controlled plant $G(s) = b(s)/a(s)$ is shown in Fig. 1. It is adopted from [4, 5] with referred original inspiration in [10].

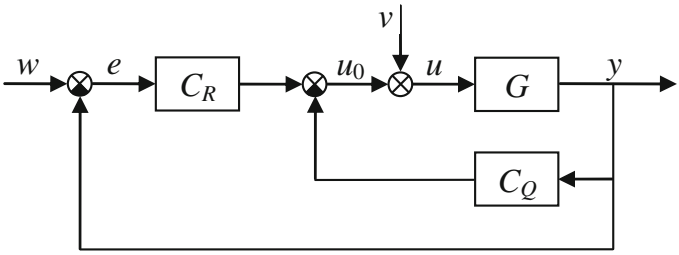


Fig. 1. Control loop with TFC

Note that the transfer function of the controlled plant is supposed in purely polynomial form, i.e. both numerator and denominator must be polynomials (and moreover $^\circ b(s) \leq {}^\circ a(s)$). However, if the time-delay plant $G_{TD}(s) = \tilde{b}(s)e^{-\Theta s}/\tilde{a}(s)$ is controlled, the time-delay term must be approximated in order to be suitable for polynomial control design. Among other possibilities, the application of the first order Padé approximation is convenient for this purpose:

$$e^{-\Theta s} \approx \frac{1 - \frac{\Theta}{2}s}{1 + \frac{\Theta}{2}s} \tag{1}$$

and thus:

$$G_{TD}(s) = \frac{\tilde{b}(s)}{\tilde{a}(s)}e^{-\Theta s} \approx \frac{\tilde{b}(s)\left(1 - \frac{\Theta}{2}s\right)}{\tilde{a}(s)\left(1 + \frac{\Theta}{2}s\right)} = \frac{b(s)}{a(s)} = G_A(s) \tag{2}$$

The controllers are designed by means of the polynomial technique [4, 5] which is convenient for ensuring of the main requirements such as stability and internal properness of the control system, asymptotic tracking of the reference signal and load disturbance rejection.

A basic Diophantine equation, crucial for control design, is:

$$a(s)\tilde{p}(s) + b(s)t(s) = d(s) \tag{3}$$

where $t(s) = r(s) + \tilde{q}(s)$. (Nominal) stability of control system from Fig. 1 is guaranteed for polynomials $\tilde{p}(s)$ and $t(s)$ obtained as a solution of the Eq. (3) with a stable right-hand side polynomial $d(s)$. In this contribution, both reference w and load disturbance v are assumed as the stepwise signals, i.e. $W(s) = w_0/s$, $V(s) = v_0/s$. Under this presumption, the asymptotic tracking and load disturbance rejection are fulfilled for the polynomials $\tilde{p}(s) = sp(s)$, $\tilde{q}(s) = sq(s)$.

The forms of polynomials $t(s)$, $r(s)$ and $q(s)$ are:

$$t(s) = \sum_{i=0}^{n} t_i s^i; \quad r(s) = \sum_{i=0}^{n} r_i s^i; \quad q(s) = \sum_{i=1}^{n} q_i s^{i-1} \tag{4}$$

with basic relations among their coefficients [5]:

$$\begin{aligned} r_0 &= t_0 \\ r_i + q_i &= t_i \quad \text{for } i = 1, \ldots, n \end{aligned} \tag{5}$$

The coefficients of the polynomials $r(s)$ and $q(s)$ can be gained on the basis of the calculated polynomial $t(s)$ and adjustable coefficients $\gamma_i \in \langle 0, 1 \rangle$ according to:

$$\begin{aligned} r_i &= \gamma_i t_i \quad &\text{for } i = 1, \ldots, n \\ q_i &= (1 - \gamma_i) t_i \quad &\text{for } i = 1, \ldots, n \end{aligned} \tag{6}$$

Obviously, the coefficients γ_i represent the weights for numerators of controllers' transfer functions. The unit coefficients γ_i for all i reduce the control system from Fig. 1 to the standard 1DOF configuration ($C_Q(s) = 0$). On the other hand, if $\gamma_i = 0$ for all i and moreover reference and load disturbance are stepwise signals, the control loop corresponds to the 2DOF structure [5].

Nevertheless, the control behavior can be particularly influenced by a choice of the right-hand side polynomial $d(s)$ in Diophantine Eq. (3). The simplest selection with multiple real roots will be used in this contribution.

More details of the method can be found in [4, 5, 7, 8].

3 Robust Stability of Time-Delay Systems with Parametric Uncertainty

Systems with parametric uncertainty are supposed to have known fixed structure (order) but on the other hand, their real physical parameters are known imprecisely. Transfer function which describes a time-delay plant with parametric uncertainty has a general form:

$$G_{TD}(s, q) = \frac{\tilde{b}(s, q)}{\tilde{a}(s, q)} e^{-\Theta(q)s} \tag{7}$$

where q is the vector of real uncertain parameters which is bounded by some uncertainty bounding set. Usually, the uncertain parameters are defined by intervals with minimal and maximal possible values (i.e. by using L_∞ norm).

A commonly used representative of (time-delay) systems with parametric uncertainty is the (time-delay) interval plant:

$$G_{TD}(s,\tilde{b},\tilde{a},\Theta) = \frac{\sum_{i=0}^{m}\left[\tilde{b}_i^-,\ \tilde{b}_i^+\right]s^i}{\sum_{i=0}^{n}\left[\tilde{a}_i^-,\ \tilde{a}_i^+\right]s^i}e^{-\left[\Theta^-,\Theta^+\right]s} \tag{8}$$

with mutually independent parameters bounded by means of their lower and upper limits.

The robust stability of the closed loop from Fig. 1 with the controlled plant (7) can be investigated through the robust stability of the family of closed-loop characteristic quasi-polynomials:

$$P_{CL} = \{p_{CL}(s,q) : q \in Q\} \tag{9}$$

where Q is the uncertainty bounding set and the structure of the uncertain quasi-polynomial is as follows:

$$p_{CL}(s,q) = \tilde{a}(s,q)\tilde{p}(s) + \tilde{b}(s,q)t(s)e^{-\Theta(q)s} \tag{10}$$

An elegant graphical method based on the combination of the value set concept with the zero exclusion condition [2] can be used for analyzing the robust stability of the family (9).

According to [2], the value set at given frequency $\omega \in \mathbb{R}$ is:

$$p_{CL}(j\omega,Q) = \{p_{CL}(j\omega,q) : q \in Q\} \tag{11}$$

The practical construction of the value sets can be done by substituting s for $j\omega$, fixing $\omega \in \mathbb{R}$ and letting q range over Q.

The zero exclusion condition for Hurwitz stability of the family of continuous-time (quasi-)polynomials (9) is defined [2]: Suppose invariant degree of (quasi-)polynomials in the family, pathwise connected uncertainty bounding set Q, continuous coefficient functions, and at least one stable member $p_{CL}(s,q^0)$. Then the family is robustly stable if and only if:

$$0 \notin p_{CL}(j\omega,Q) \quad \forall\omega \geq 0 \tag{12}$$

4 Simulation Example

Assume a second order time-delay plant with the uncertain parameters:

$$G_{TD}(s,\tilde{b},\tilde{a},\Theta) = \frac{\tilde{b}_0}{\tilde{a}_2 s^2 + \tilde{a}_1 s + \tilde{a}_0} e^{-\Theta s} = \frac{[0.7,\ 1.3]}{[0.7,\ 1.3]s^2 + [0.7,\ 1.3]s + [0.7,\ 1.3]} e^{-[0.7,\ 1.3]s} \tag{13}$$

i.e. each parameter can be perturbed up to ±30% of its unit mean value.

For the algebraic control design purpose, the transfer function (13) has to be approximated by the time-delay free nominal model. In the first step, the system with fixed mean values of the interval parameters is considered:

$$G_{TD-MEAN}(s) = \frac{1}{s^2 + s + 1} e^{-s} \tag{14}$$

The time-delay term in the model (14) is subsequently approximated according to (1), (2). It leads to the final nominal system:

$$G_A = \frac{b(s)}{a(s)} = \frac{-0.5s + 1}{0.5s^3 + 1.5s^2 + 1.5s + 1} \tag{15}$$

and thus the Diophantine Eq. (3) has the specific form:

$$\left(0.5s^3 + 1.5s^2 + 1.5s + 1\right)s\left(p_2 s^2 + p_1 s + p_0\right) + (-0.5s + 1)\left(t_3 s^3 + t_2 s^2 + t_1 s + t_0\right) = (s + m)^6 \tag{16}$$

The chosen right-hand side polynomial in (16) reveals that the case with multiple real roots is supposed. These multiple roots are selected as −1.5, i.e. $m = 1.5$. Furthermore, the weight coefficients from (6) are considered as $\gamma_1 = \gamma_2 = \gamma_3 = 0.3$(i.e. somewhere between 1DOF (30%) and 2DOF (70%) control configurations). These assumptions lead to the TFC:

$$C_Q(s) = \frac{5.7066s^2 + 15.5832s + 12.3238}{2s^2 + 12s + 33.6523}$$
$$C_R(s) = \frac{2.4457s^3 + 6.6785s^2 + 5.2816s + 11.3906}{2s^3 + 12s^2 + 33.6523s} \tag{17}$$

The corresponding family of closed-loop characteristic quasi-polynomials (9) with the uncertain parameters from (13) is:

$$
\begin{aligned}
p_{CL}(s,\tilde{b},\tilde{a},\Theta) = {} & 2\tilde{a}_2 s^4 + (12\tilde{a}_2 + 2\tilde{a}_1)s^3 + (33.6523\tilde{a}_2 + 12\tilde{a}_1 + 2\tilde{a}_0)s^2 + \cdots \\
& + (33.6523\tilde{a}_1 + 12\tilde{a}_0)s + 33.6523\tilde{a}_0 + \cdots \\
& + \tilde{b}_0 e^{-\Theta s}\left(8.1523s^3 + 22.2617s^2 + 17.6055s + 11.3906\right) \\
& \tilde{b}_0, \tilde{a}_2, \tilde{a}_1, \tilde{a}_0,\ \Theta \in [0.7,\ 1.3]
\end{aligned}
\tag{18}
$$

The value sets of the family (18) for the frequency range from 0 to 2.15 with the step 0.05 is shown in the Fig. 2. The uncertain parameters are sampled according to $\tilde{b}_0,\ \tilde{a}_2,\ \tilde{a}_1,\ \tilde{a}_0,\ \Theta = 0.7 : 0.1 : 1.3$. The zoomed version of the same plot, which provides the closer look to the neighborhood of the complex plane origin, is depicted in Fig. 3. It can be clearly seen that the origin of the complex plane (zero point) is excluded from the value sets. Moreover, the family contains a stable member. Consequently, the family of closed-loop characteristic quasi-polynomials (18) is robustly stable.

The obtained result of robust stability can be also visually confirmed by the control simulations from the Fig. 4. It depicts the output signals of 1024 "sampled plants" from the family (13) (four selected values (0.7, 0.9, 1.1, 1.3) for each of five uncertain parameters ($\tilde{b}_0,\ \tilde{a}_2,\ \tilde{a}_1,\ \tilde{a}_0,\ \Theta$)). Besides, the red curve represents the output signal of the nominal system (15). The simulation conditions were as follows: The stepwise reference signal changes from 1 to 2 in one third of the simulation time and the step load disturbance −0.5 affects the input to the controlled plant during the last third of simulation.

The selection of different weight coefficients would have no impact on the robust stability or instability of the control loop because the polynomial $t(s)$ would remain the same. It could influence "only" the control performance. On the other hand, the robust stability (as well as the control performance) can be influenced by a choice of the

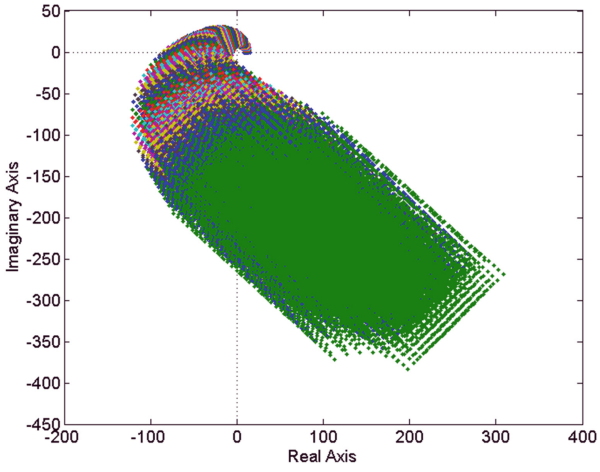


Fig. 2. Value sets for the family of closed-loop char. quasi-polynomials (18) – the full view

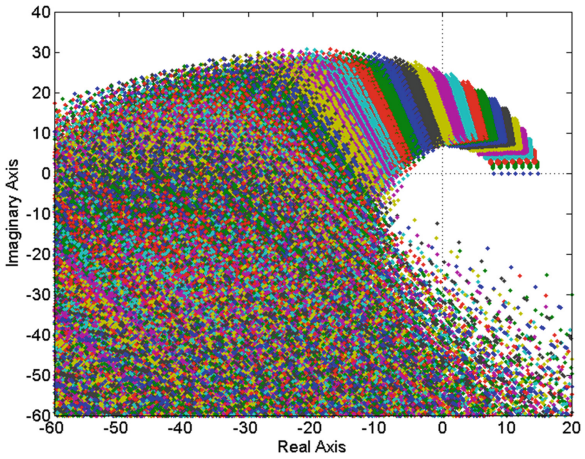


Fig. 3. Value sets for the family of closed-loop char. quasi-polynomials (18) – a zoomed view

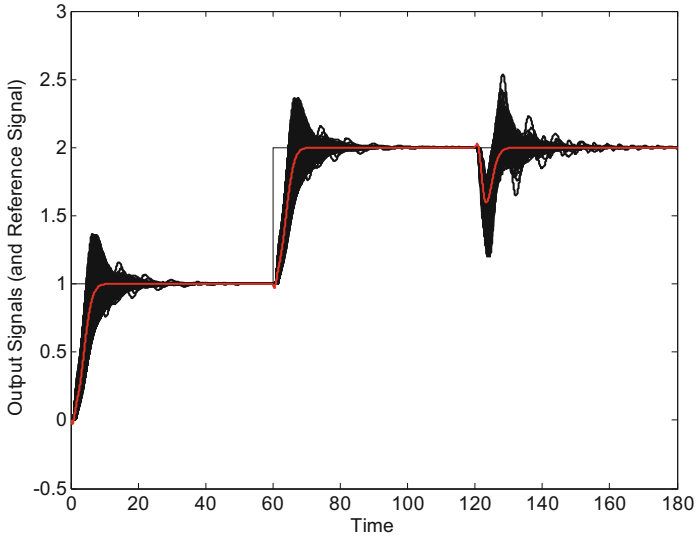


Fig. 4. Control of "sampled plants" from the plant family (13) by TFC (17)

multiple roots in (16). For example, the selection $m = 0.5865$ leads to the value sets which "touch" the complex plane origin (see Fig. 5) and consequently to the corresponding control simulations which are shown in Fig. 6. Obviously, this selection results in the control loop very close to the robust stability border. Similarly, another robust stability border control system could be obtained for a value near $m = 1.578$.

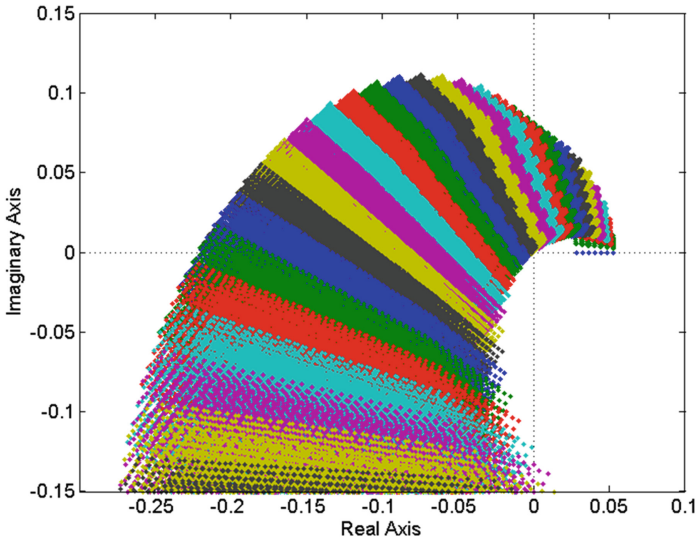


Fig. 5. Value sets for the family of closed-loop characteristic quasi-polynomials close to the robust stability border

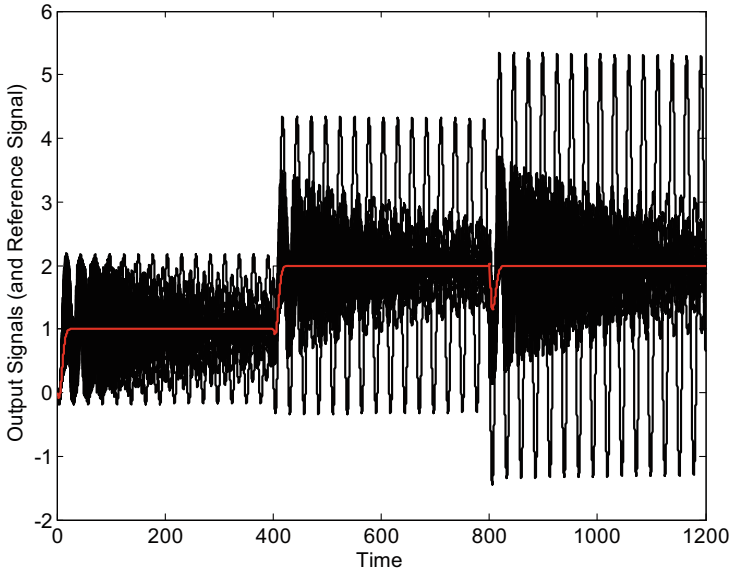


Fig. 6. Control of "sampled plants" from the plant family (13) – the robust stability border case

5 Conclusion

This contribution has been focused on robust stabilization of time-delay plants with parametric uncertainty by means of two feedback controllers designed on the basis of the polynomial approach. Robust stability of the control loops is tested using the graphical analysis of the families of closed-loop characteristic quasi-polynomials. The presented simulation example shows the robust stabilization of a second order plus time delay plant with the uncertain parameters. Moreover, the influence of the weight coefficients and the multiple pole placement has been discussed.

Acknowledgments. The work was supported by the Ministry of Education, Youth and Sports of the Czech Republic within the National Sustainability Programme project No. LO1303 (MSMT-7778/2014). This assistance is very gratefully acknowledged.

References

1. Richard, J.P.: Time-delay systems: an overview of some recent advances and open problems. Automatica **39**(10), 1667–1694 (2003)
2. Barmish, B.R.: New Tools for Robustness of Linear Systems. Macmillan, New York (1994)
3. Bhattacharyya, S.P., Chapellat, H., Keel, L.H.: Robust Control: The Parametric Approach. Prentice Hall, Englewood Cliffs (1995)
4. Dostál, P., Gazdoš, F., Bobál, V., Vojtěšek, J.: Adaptive control of a continuous stirred tank reactor by two feedback controllers. In: Proceedings of the 9th IFAC Workshop on Adaptation and Learning in Control and Signal Processing, Saint Petersburg, Russia (2007)
5. Dostál, P., Gazdoš, F., Bobál, V.: Design of controllers for time delay systems: integrating and unstable systems. In: Time-Delay Systems, pp. 113–126. InTech, Rijeka (2011)
6. Kučera, V.: Diophantine equations in control – a survey. Automatica **29**(6), 1361–1375 (1993)
7. Matušů, R., Prokop, R.: Robust stability of control systems with two feedback controllers and interval plants. In: Recent Advances in Systems – Proceedings of the 19th International Conference on Systems, Zakynthos, Greece, pp. 321–325 (2015)
8. Matušů, R.: Robust stabilization of interval plants by means of two feedback controllers. Inter. J. Circ. Syst. Sig. Process. **9**, 427–434 (2015)
9. Matušů, R.: Control of interval systems by using two feedback controllers. In: Proceedings of the 26th DAAAM International Symposium, Vienna, Austria, pp. 217–222 (2016). doi:10.2507/26th.daaam.proceedings.030
10. Ortega, R., Kelly, R.: PID self-tuners: some theoretical and practical aspects. IEEE Trans. Industr. Electron. **31**(4), 332–338 (1984)

Maze Navigation on Ball & Plate Model

Lubos Spacek(✉), Vladimir Bobal, and Jiri Vojtesek

Department of Process Control, Faculty of Applied Informatics,
Tomas Bata University in Zlin, Nad Stráněmi 4511, 760 05 Zlín, Czech Republic
{lspacek,bobal,vojtesek}@fai.utb.cz

Abstract. Today's CCD or CMOS image sensors are advanced enough to satisfy the need for accurate object detection and tracking. This leads to implementation of computer vision into industry, transportation, medicine, robotics and other sectors. The aim of this paper is to present steps needed to determine correct path through the maze constructed on a plate and navigate a ball along this path. Image processing techniques used here are simple enough to understand, so students can easily implement them to further extend educational capabilities of Ball & Plate model. The paper also shows the use of watershed transform, which can be extended for similar problems. The added maze thus provides excellent application for the model and simulates real-world issues in research and development.

Keywords: Maze navigation · Watershed transform · Ball & Plate model · Color segmentation · MATLAB

1 Introduction

The purpose of this paper is to navigate ball through maze by choosing the optimal control strategy. The polynomial approach to controller design with pole placement has a great advantage in comparison with other methods. It is very easy to make the whole design process automatic, which means if the plant or other parameters change, the controller can be quickly modified too. In addition, half of the poles placed are optimally calculated via minimization of linear quadratic (LQ) criterion [1,2].

It was possible to use many controller structures: 1DOF, 2DOF, cascade structure, ICM (Internal Model Control), controller with fuzzy supervision. The controller structure proposed here is two degree of freedom (2DOF) closed-loop controller structure, which provides separation of feed-back part (responsible for stabilization and disturbance rejection) and feed-forward part (responsible for reference tracking) [3]. This should provide better control over the model and its behavior.

The maze constructed on the plate needs to be transformed into digital data that can be precisely interpreted. Simple color segmentation proves to be a suitable tool and was chosen for its simplicity and speed. Another way to get the structure of maze walls would be edge detection and techniques such as Hough

R. Silhavy et al. (eds.), *Cybernetics and Mathematics Applications in Intelligent Systems*,
Advances in Intelligent Systems and Computing 574, DOI 10.1007/978-3-319-57264-2_21

transform [4]. For the next step, Watershed transform is used to obtain the path. It is also possible to randomly walk through the maze until correct path is found, but this is quite time consuming and redundant.

1.1 Ball & Plate Model

The Ball & Plate Apparatus is two dimensional model designed to control ball position and trajectory on the plate. The system is unstable with second order astatism and is suited for studying system dynamics, identification and design of control algorithms. Position of the ball is determined using camera located above the plate in specific distance. The plate is pivoted at its center and can rotate around two perpendicular axes using two stepper motors as seen in Fig. 1. In this arrangement, the model has two inputs (stepper motors voltages) and two outputs (2D coordinates of the ball). The hardware part of the model used here is marked CE151 and was built by Humusoft [5].

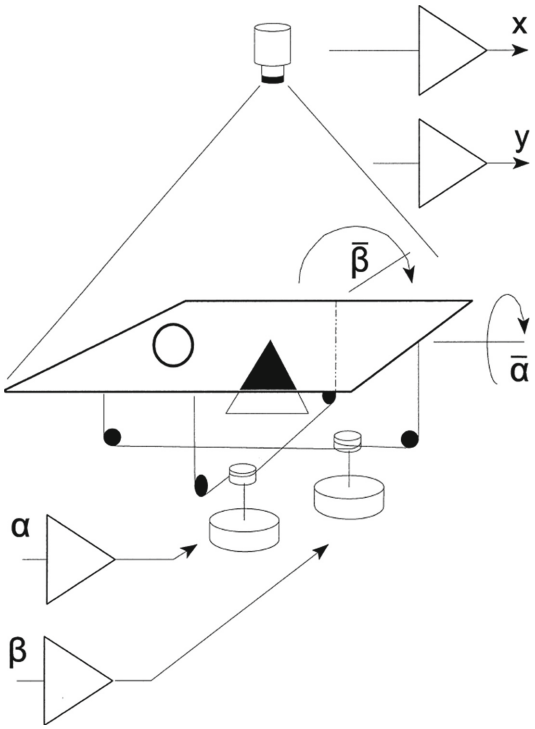


Fig. 1. CE151 Ball & Plate model diagram [5]

1.2 Maze Structure

A simple maze was constructed on the plate to further extend capabilities of the CE151 model. Maze walls were made from blue electrical tape only for the purpose of user and the camera. If its walls were higher, there would be no reason to implement more complex controller (ball would be controlled by walls themselves). The maze is thus only a 2D projection (for user and camera alike) of walls on the plate as seen in Fig. 2. Because the maze is constructed from the electrical tape, it is very easy to quickly change its appearance, which is also quite convenient.

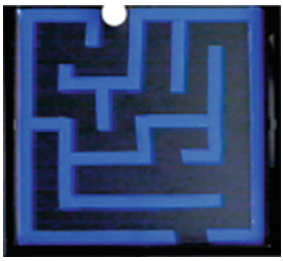


Fig. 2. Maze

For the sake of consistent results, the maze has to have one entrance, one exit and no loops. It is obvious that with more exits, the algorithm would have to pick one, which is fairly pointless and raises a question, whether it is still a maze. Loops are more logical within a maze, but again, one loop would create two new branches and the algorithm would have to pick one.

2 Methods

2.1 Color Segmentation and Pre-processing

Before the control process starts, a snapshot from the camera is taken in RGB color model. The ball can be on the plate, but should not cover blue walls (as seen in Fig. 2). Because maze walls are blue, the segmented color will be also blue. But taking only the blue RGB component is not enough, because presence of blue component does not necessarily mean the presence of blue color as perceived by humans, as shown in Table 1. Thus a simple formula for computing blueness of the image was used [6]:

$$b = B - max\,(R, G) \tag{1}$$

where R, G and B are components of RGB color model.

The blueness image can be used to create a binary mask by selecting appropriate threshold. To remove unwanted noise, only the largest continuous section (walls) should be selected. This guarantees that pixels which are not directly connected to walls (noisy pixels) will be removed. Resulting images are in Fig. 3.

Table 1. Color perception of RGB model and blueness calculation [6]

(R,G,B)	→	blueness
(255,0,0)	→	-255
(0,255,0)	→	-255
(0,0,255)	→	255
(127,127,255)	→	128
(255,0,255)	→	0
(0,255,255)	→	0

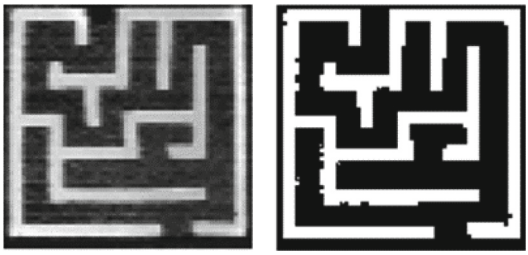


Fig. 3. Blueness image (left) and its binary mask (right)

2.2 Watershed Transform

The term watershed refers to a ridge that divides areas drained by different river systems as shown in Fig. 4. A catchment basin is the geographical area draining into a river or reservoir [7]. In image processing, it was introduced as a tool for

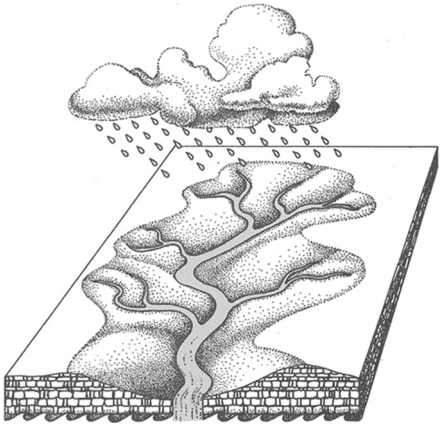


Fig. 4. Watershed drainage [10]

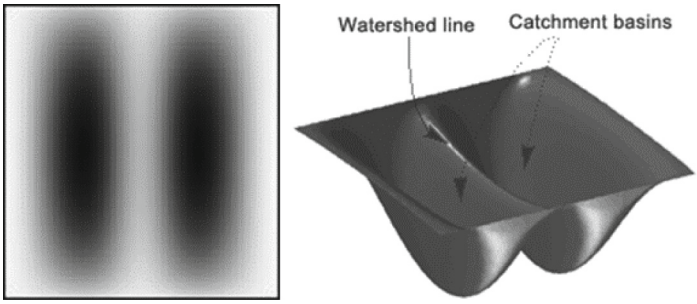


Fig. 5. Grayscale image and its topological relief [7]

Fig. 6. Watershed transform output for maze image

segmenting grayscale images by S. Beucher and C. Lantujoul in the late 70's [8]. It considers a grayscale image as a topographical relief (the grey level of a pixel represents the elevation of a point, where dark areas are low and bright areas are high). The example grayscale image and its 3D surface is in Fig. 5.

The `watershed` function in MATLAB [9] detects these watershed regions and outputs them in the matrix (Fig. 6) of the same size as the binary input image.

2.3 Post-processing and Trajectory Determination

The resulting binary matrix is again cleansed from noise after extraction of the path from output of the watershed transform. A sequence of reference values

Fig. 7. Post-processing steps

was thus obtained from the computed path in the form of binary matrix as seen in Fig. 7. For reference to be step-changing signal, only corner points of the path were chosen and to obtain a trajectory with directions, the algorithm walks through the binary matrix and saves row-column combinations as final reference value used in control scheme.

2.4 Controller Design

Plant identification is the first step in controller design and the Ball & Plate model was identified simply by positioning the ball to the center of the plate manually and step-changing plate's inclination. The step response of the system was thus obtained and properly identified. The model is considered to be symmetric, so only one plate angle and ball coordinate are taken into account. Small variations in the perpendicular direction were compensated by taking multiple measurements and uncompensated errors were taken as the part of the model to make identification more robust. Figure 8 shows plot of these measurements. Second plot shows their average and response of identified model (in both

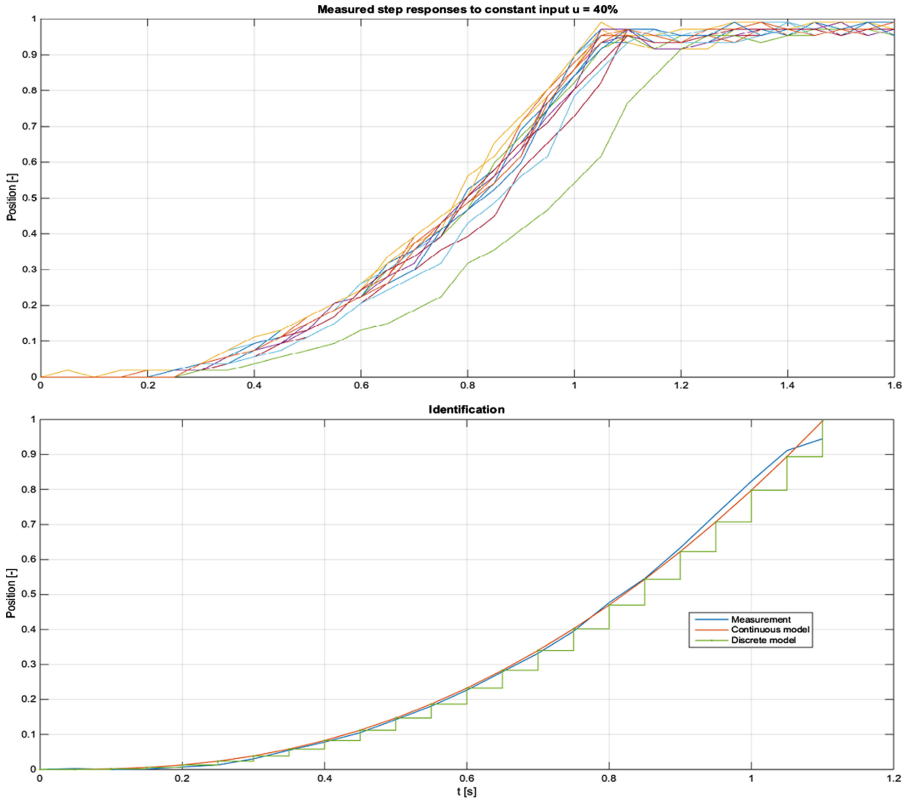


Fig. 8. Identification process

continuous and discrete form). The model was identified for general structure of continuous transfer function with double integrator and first order system dynamics as shown in following equation.

$$G(s) = \frac{C}{s^2\,(Ts+1)} = \frac{C}{Ts^3 + s^2} \tag{2}$$

where $G(s)$ is continuous transfer function with complex variable s and C, T are identified constant parameters. Discrete form of this transfer function can be written as follows:

$$G(z^{-1}) = \frac{B(z^{-1})}{A(z^{-1})} = \frac{b_1 z^{-1} + b_2 z^{-2} + b_3 z^{-3}}{1 + a_1 z^{-1} + a_2 z^{-2} + a_3 z^{-3}} \tag{3}$$

where $G(z^{-1})$ is discrete transfer function with complex variable z^{-1} and $B(z^{-1})$, $A(z^{-1})$ polynomials obtained from discretization. It is obvious that the identification of continuous transfer function is simpler, as it has only two unknown parameters and predetermined dynamics. The model represented in discrete transfer function has 6 unknown parameters.

The controller was designed for two degree of freedom (2DOF) closed-loop control system shown in Fig. 9, where $G(z^{-1})$ is the controlled plant, $C_b(z^{-1})$ is the feed-back part of the controller, $C_f(z^{-1})$ is the feed-forward part of the controller, $1/K(z^{-1}) = 1/(1 - z^{-1})$ is the summation part, $w(k)$ is reference signal, $n(k)$ is load disturbance and $v(k)$ is disturbance signal. For the sake of simplification, there will be assumed no disturbances acting on the system. The polynomial approach was chosen for the control system design, which is based on linear algebra. Coefficients of both feed-back and feed-forward parts of the controller were determined by minimization of quadratic criterion J shown in the following equation.

$$J = \sum_{k=0}^{\infty} \left\{ [e(k)]^2 + q_u [u(k)]^2 \right\} \tag{4}$$

where q_u is penalization constant, $e(k) = w(k) - y(k)$ is the error and $u(k)$ is the controller output. Closer lookup to polynomial method and minimization of

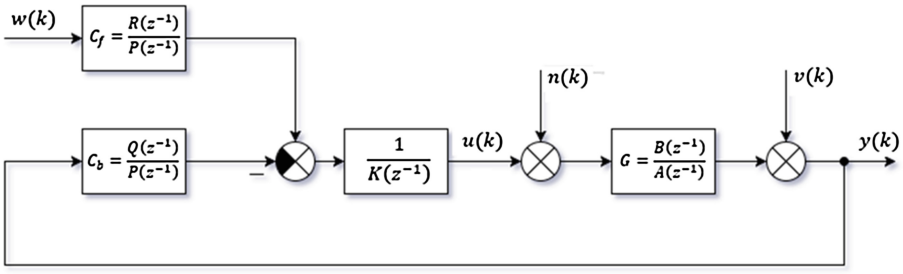


Fig. 9. Structure of 2DOF controller

LQ criterion is presented in [11]. This criterion was minimized using spectral factorization and with the help of Polynomial Toolbox in MATLAB [12].

3 Results

Described methods were implemented in an algorithm, which controls the process of path calculation. When the algorithm is complete the user is prompted to confirm the result (Fig. 10). This verification step by user is the last line of defense against reflections, bad lightning and unpredictable errors, because the algorithm detects incomplete routes and loops as seen in Fig. 11. The overall progress of the ball's trajectory and plate angles is shown in Fig. 12. The ball followed desired path with only minor setbacks caused by errors and model's unalterable limits.

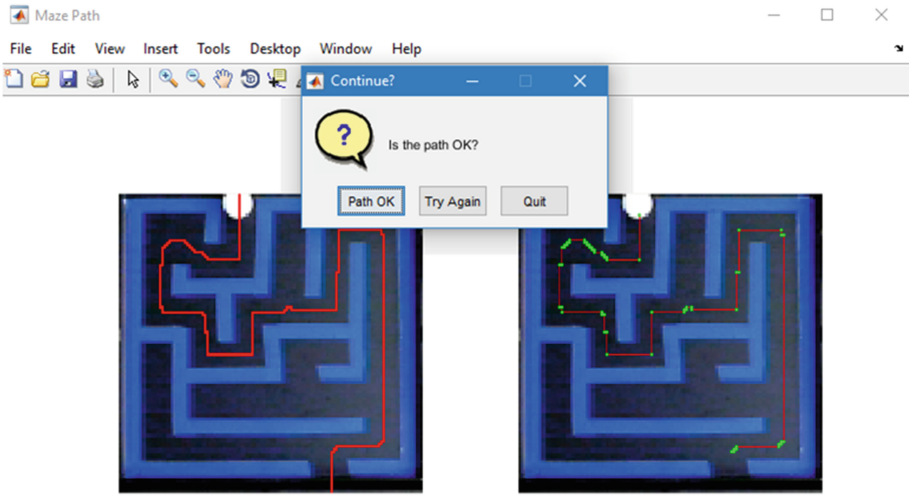


Fig. 10. Algorithm completion with user prompt

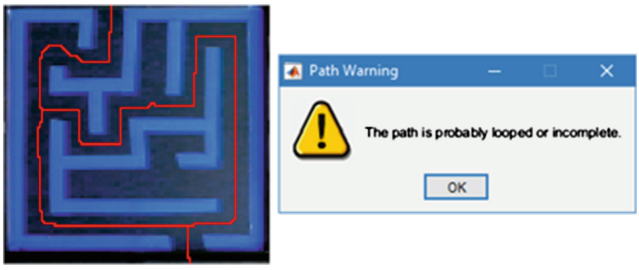


Fig. 11. Looped path with warning

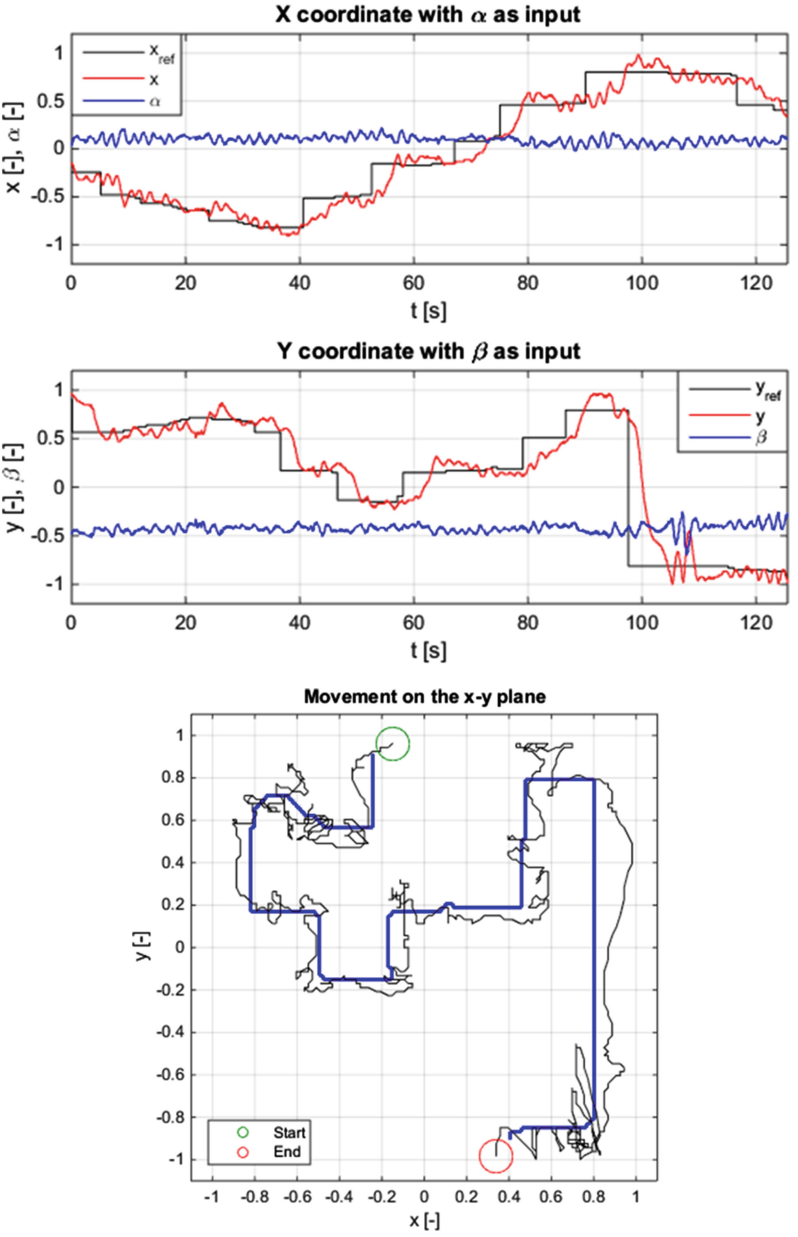


Fig. 12. Maze navigation

4 Discussions

This paper presented the navigation of ball through maze on Ball & Plate model. Although the model was already prepared from the hardware's side, all other steps needed to accomplish the task were taken and presented. The model was properly identified and the controller was designed based on this identification. Results of control are satisfying and one can conclude that used linear quadratic (LQ) controller is more than suitable for this kind of application.

The last step was to design an algorithm which would effectively solve constructed maze and navigate the ball through. With the use of image processing techniques and watershed transform, the solution was obtained and transformed into the set of reference values for designed controller. The ball followed the desired trajectory with success, which proves that methods presented in this paper were chosen appropriately. Additionally, watershed transform can be used for similar segmentation tasks in the same manner.

Acknowledgments. This article was created with support of the Ministry of Education of the Czech Republic under grant IGA reg. no. IGA/FAI/2017/009.

References

1. Bobál, V., Böhm, J., Fessl, J., Macháček, J.: Digital Self-tuning Controllers. Springer, London (2005)
2. Landau, I.D.: Digital Control Systems Design, Identification and Implementation. Springer, London (2007)
3. Matušů, R., Prokop, R.: Algebraic design of controllers for two-degree-of-freedom control structure. Int. J. Math. Models Methods Appl. Sci. **7**, 630–637 (2013)
4. Shapiro, L.G., Stockman, G.C.: Computer Vision. Prentice-Hall, New Jersey (2001)
5. Humusoft: CE 151 Ball & Plate Apparatus User's Manual, Prague (2006)
6. A simple image segmentation example in MATLAB. In: MATLABtricks. http://matlabtricks.com/post-35/a-simple-image-segmentation-example-in-matlab
7. The Watershed Transform: Strategies for Image Segmentation. In: Mathworks Technical Articles. http://www.mathworks.com/company/newsletters/articles/the-watershed-transform-strategies-for-image-segmentation.html
8. Couprie, M., Najman, L., Bertrand, G.: Algorithms for the topological watershed. In: Andres, E., Damiand, G., Lienhardt, P. (eds.) DGCI 2005. LNCS, vol. 3429, pp. 172–182. Springer, Heidelberg (2005). doi:10.1007/978-3-540-31965-8_17
9. Watershed transform. https://www.mathworks.com/help/images/ref/watershed.html
10. The Long Island Sound Watershed. http://soundbook.soundkeeper.org/chapter_ContentID_210_SectionID_6.htm
11. Bobál, V., Chalupa, P., Dostál, P., Kubalčík, M.: Digital control of unstable and integrating time-delay processes. Int. J. Circuits Syst. Sig. Process. **8**, 424–432 (2014)
12. PolyX. http://www.polyx.cz/

AEOC: A Novel Algorithm for Energy Optimization Clustering in Wireless Sensor Network

C. Parvathi[1(✉)] and Suresha[2]

[1] Department of Computer Science and Engineering, DBIT, Bengaluru, India
parvathi3125@gmail.com
[2] Department of Computer Science and Engineering, SVCE, Bengaluru, India
suresha_rec@rediffmail.com

Abstract. The area of Wireless Sensor Network (WSN) has witnessed various research contributions in the past decade for mitigating the issues of energy dissipation to ensure energy efficient routing and clustering. It was found that the existing technique doesn't have productive supportability towards addressing the energy efficiency as well as enhancing clustering performance in WSN. Hence, this paper presents a novel idea where the energy efficiency as well as clustering optimization was carried out by incorporating the selection mechanism of nodes. The paper also discusses the architecture and algorithm of the proposed technique with briefing of methodology that was adopted to accomplish the work. Finally, the paper exhibits the outcomes of the study very discretely and highlights the better clustering perspective of performing benchmarking the technique with existing standards.

Keywords: Wireless Sensor Network · Clustering techniques · Energy efficiency · Optimization · Network lifetime · Routing

1 Introduction

The rapid advancement into wireless network and communication technology, sensor dimension and its cost, sensor node design has laid a foundation of conceptualization of various applications in the different walks of the life using WSN [1]. A wireless sensor network comprises of three types of nodes e.g. sensor nodes, Cluster Heads (CH), and base stations [2]. At present, sophisticated applications of WSN is on rise and heavily demanded. Some of the significant applications of WSN are in (i) Defense application (ii) Environmental applications (iii) Healthcare applications (iv) home automation, etc. [3]. All such applications will require physical parameters to be measure and routed from its subnet to the sink and further sink to the collaborator where smart application and algorithm does its interpretations as per the requirement of the applications [4]. The important factors, which serve the guidelines to the design of a protocol or algorithm for WSN, are fault tolerance, sensor network topology, scalability, transmission media, hardware constraints and power consumption [5, 6]. The wireless sensor nodes are micro-electronic devices and it will be battery supported along with a little power is less than 0.5 Ah, 1.2 V [7, 8]. As most of the applications especially which are aimed to be

R. Silhavy et al. (eds.), *Cybernetics and Mathematics Applications in Intelligent Systems*,
Advances in Intelligent Systems and Computing 574, DOI 10.1007/978-3-319-57264-2_22

deployed into human inaccessible locations, feasibilities of replacement or harvesting or recharging these power sources are not feasible and yet it is far distant to get those days, thus there is power dissipation than a reversible consumption. A complete and systematic power analysis of a sensor node is important to identify power bottlenecks in the system, which can then be the target of aggressive optimization [9].

Hence, this paper presents a novel optimization technique towards leveraging the clustering process in wireless sensor network. Section 2 discusses about the existing research work followed by issues and challenges identification in Sect. 3. Section 4 presents the proposed methodology followed by elaborated discussion of algorithm implementation in Sect. 5. Comparative analysis of accomplished result is discussed under Sect. 6 followed by conclusion in Sect. 7.

2 Related Work

This section discusses about the existing research work being carried out toward techniques adopted for enhancing clustering performance. Our prior study [10] has discussed about existing routing techniques along with their clustering scheme. It was seen that the adoption of multivariate optimization problem was also found in various existing literatures. Basically, it was investigated in order to find the data reliability during data fusion, or routing, or aggregation in wireless sensor network. Rabbat and Nowak [11] have adopted multivariate optimization problem for enhancing the energy and communication behavior among sensors. Similar adoption of multivariate optimization was also seen in the work of He et al. [12], who have adopted cross-layer based approach in data aggregation using Lagrangian multiplier. Cao et al. [13] have adopted multivariate Bernoulli link framework for enhancing the essential routing principle among sensors. Usage of multivariate optimization was also witnessed in the work of Pushpalatha [14] where the authors have used iterative algorithm for addressing the localization issues in wireless sensor network. Further studies towards energy efficiency was carried out by Arumugam and Ponnuchamy [15], who have enhanced the conventional LEACH by applying optimization over cluster formations. Gholipour et al. [16] have focused on addressing the congestion issues by developing an artificial gradient field. Usage of gradient routing is also carried out by Kannan and Paramasivan [17] where the authors have used multiple hop based on-demand communication protocol to conserve energy. Javaid et al. [18] have developed a mechanism for performing distributed clustering by conserving loss of energy. Similar direction of work is also carried out by Jumira et al. [19] by introducing geographic routing technique. Amir et al. [20] have used evolutionary technique e.g. fuzzy logic and ant colony optimization to achieve energy conservation. A completely different form of technique called as ultrasonic frog was adopted as communication scheme by Xu et al. [21]. Xu et al. [22] have used orthogonal variable spreading factor for performing clustering and energy conservation in wireless sensor network. Hence, it can be seen that there are quite a good number of work being carried out for addressing the issues of energy efficient in wireless sensor network.

3 Problem Description

Although, various work has been carried out towards clustering as well as energy efficiencies in wireless sensor network, but it has been noticed that contribution of majority of the existing clustering limits to the selection of cluster-head only and completely ignores the selection process of the intermediate nodes. Although, there are also research work towards selection of energy efficient path but they doesn't again emphasize on clustering process with optimization. Moreover, majority of the optimization techniques implemented doesn't emphasize on communication overhead owing to its recursive nature of working principle. Hence, the problem statement is "Designing optimization of clustering is challenging aspect for balancing the problems of energy efficiencies and data delivery performance over constraint resources."

4 Proposed Methodology

An analytical research methodology is applied on proposed design methodology on the basis of 1st order radio-energy model [23]. The prime goal of proposed design is introducing a novel clustering algorithm for maintaining a proper balance between network lifetime and data delivery performance in wireless sensor network.

Figure 1 highlights the schematic architecture of the proposed AEOC utilizing both communication and energy modeling emphasizing selection of cluster head as well as selection of all the intermediate nodes (i.e. cluster head) in the course leading to destination node (i.e. base station). The sensor node in AEOC is deployed in highly distributed fashion within the simulation area. It was also ensured that the sensor node keep a well balance between high density and low sparsity formation when it comes to clustering using threshold-based approach. A very simple optimization algorithm is introduced that balances the allocation of lowered transmittance energy and higher data transmittance using multihop routing scheme in wireless sensor network. The next section discusses about algorithm design principle to further elaborate it.

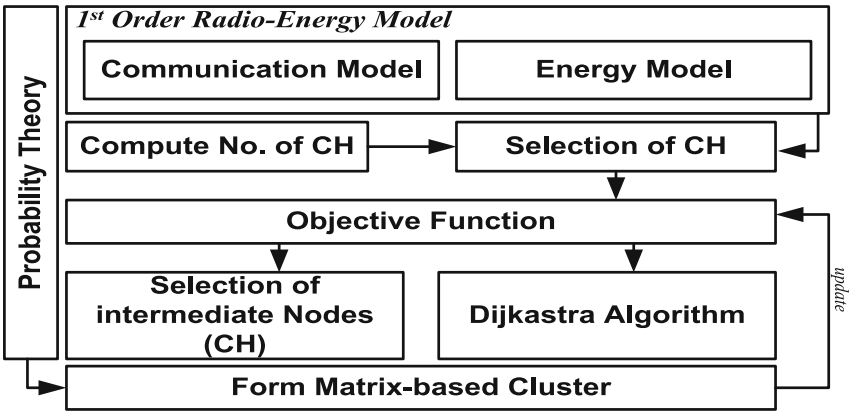


Fig. 1. Schematic architecture of AEOC

5 Algorithm Implementation

The prime purpose of the proposed technique is to implement a novel optimization technique of clustering that directly affects the energy dissipation system on the sensor node. The design and development of the proposed algorithm for energy modeling is totally based on first order radio-energy model and incorporates certain novel mechanism in a very simplified manner for (i) selection of cluster-head and (ii) energy-efficient routing. The formulation of the algorithm is carried out considering certain significant parameters e.g. (i) number of sensor, (ii) election probability, (iii) position of node, (iv) Energy for transmitting and receiving, and (v) control message. The steps of the proposed system AEOC is as shown below:

Algorithm for Energy Optimization Clustering in WSN

Input: N (number of sensor), p(election probability), n_{pos} (position of node), M (Matrix), E_{TX} / E_{RX} (Energy for transmitting and receiving), *msg* (control message).

Output: E_{TX} / E_{RX} (Optimized Energy for transmitting and receiving)

Start

1. N_{ch}→(i^2-N.p)
2. **for** n_{pos}→b
3. dump→M(N_{ch})
4. Apply objective function $\arg_{\min} \sum_{i \subset N} E_{TX}, \sum_{j \subset \phi} \alpha_j . f_j$
5. path_sel→vector(Dijkastra algo)
6. **if** (path <length(path_sel)-1)
7. E_{RX}→E_{RX}-E_{TX}*msg
8. **End of if**
9. **End of for**

End

The implementation of the above mentioned algorithm was carried out over specific set of *static sensor nodes* deployed in random manner over a *simulation area*. These two parameters are essential simulation parameters along with *test packets, initialized transmittance/receiving/amplification energy,* and *path loss exponent*. The operation of the algorithm is classified into mainly two steps: (i) selection of cluster head and (ii) establishing routing. Initially, a simulation area is designed that allows base station to be position in any position within it. All the sensors are randomly distributed within the simulation area. The algorithm initially computes the number of the possible cluster head just on the basis of total number of nodes N and probability of election p. The variable i in Line-1 will represent squared value of integers i = 1, 2 ….etc. This simple

equation will assist in converging the search of cluster head during the cluster head selection process, which is not found in existing clustering techniques (where the search for cluster head considers all the communicating nodes). Hence, greater deal of computational complexity is reduced in this step itself. Hence, a cluster is selected based on number of alive nodes and probability factor. Different from conventional circular clustering, we generate matrix-based clusters which are not only bigger in number but also can accommodate many sensors.

In order to formulate matrix-based clusters, the positions of the nodes are quite important. For all the eligible numbers which falls within simulation boundary *b* (Line-2), a new matrix is formulated i.e. M that stores all the number of cluster heads N_{ch} and hence it makes the accessing quite easier during the data aggregation process (Line-3). Logic is created to select the cluster heads from N_{ch} on the basis of higher value of residual energy with a condition that they must satisfy the objective function (Line-4). The objective function ensures that a cluster head must dissipate lower transmittance energy as possible as well as lowered communication overhead signified by $\alpha_{j.}$ The variable f_j is path loss exponent. The interesting point about the objective function is that how will the cluster head know to reduce the transmittance energy. The cluster head performed fusion of the data and now the fused data is forwarded either to base station directly (in case of single hop) or to another cluster head (in case of multi hop). Hence, the objective function is not really meant for single hop routing. However, in multihop, the intermediate nodes (which are other cluster heads) are selected on the basis of lowered transmittance energy to stable energy dissolution within the new routes leading towards another node (or cluster head) till it reaches base station. Not only this, a second layer of filtering is applied on selected path (based on lowered transmittance energy) using Dijkastra algorithm (Line-5). While using this shortest path algorithm, care was taken to incorporate standard first order radio-energy model where total energy is computed considering both transmittance and receiving energy, amplification energy, data fusion energy, and maximum possible iterations to obtain convergence. Hence, selection of the final energy-efficient shortest path is computed (Line-6) and receiving energy is also computed (line-7). The study thereby selects a path with lower travelling cost every time it (cluster head) will be required to select its neighbor cluster head for forwarding the message. In this process, the mechanism also computes receiving energy of the sensors. This scheme of clustering that emphasizes on intermediate sensors (i.e. another cluster head) and its selection technique is quite novel and is not found in literatures till date. Exploring the energy-efficient path further reduce the link cost for finally transmitting the data packet to the destination node.

Hence, the algorithm always ensures that there is reduced energy consumption by selecting appropriate intermediate cluster heads during routing process. Hence, the algorithm can be said to perform two logical steps (i) selection of cluster heads for all the matrix-based clusters and (ii) implements objective function in order to ensure that both routing overhead as well as energy dissipation. A novel clustering technique is incorporated which emphasize on the minimizing computational overhead while performing clustering operation. At the same time, the algorithm uses boundary-based concept in order to formulate a matrix-based clustering process to ensure higher number of participation of sensor in the process of data aggregation as well as clustering. The

information of the communicating nodes are stored and updated in a frequent interval of time due to matrix-based clusters. Hence, the process lowers down the response time and is readily applicable for the applications where response time as well as network lifetime requires proper balance. The effectiveness of the proposed study was assess using two prime performance parameter e.g. throughput and residual energy. The next section discusses about the results accomplished from the study.

6 Results Discussion

A simulation study was carried out using Matlab considering 500–1000 sensors bearing characteristics of MEMSIC nodes under simulation area of 1200 × 1500 m^2. As the study focuses on incorporating energy efficiency on its clustering technique; hence, it uses residual energy to check the extent of energy consumption. A hypothetical data of 2000–5000 bits were used over the experiments considering both constant and variable bit rate traffic with path loss exponent of 0.25. Throughput was also estimated for observing the behavior of the data delivery performance. For the purpose of benchmarking, the proposed system chooses to compare its outcome with LEACH algorithm [24]. Table 1 highlights the numerical outcome being accomplished while Fig. 2 highlights the graphical visualization of the numerical outcome.

Table 1. Numerical outcomes of proposed system

Simulation rounds	Residual energy (J)		Throughput (bps)	
	LEACH	AEOC	LEACH	AEOC
100	9.55	9.55	2755.02	2877.03
200	9.55	9.54	2862.77	3118.23
300	9.53	9.52	2879.58	3481.77
400	8.21	8.31	2891.10	3866.21
500	7.31	8.29	3018.87	3901.81
600	5.29	7.77	3353.46	4019.17
700	3.87	7.65	3561.17	4285.85
800	1.08	7.51	4014.03	4781.66
900	0	7.28	4521.11	4782.76
1000	0	6.01	4667.32	4987.64

A closer look into the outcome will show that AEOC offers better energy conservation scheme as compared to conventional hierarchical LEACH algorithm. The drainage rate of the energy for AEOC is quite slower as compared to LEACH algorithm for which reason the network lifetime is maximized. The prime reason behind this is computation of clustering takes place in the cluster set up process itself in a very converging manner that not only reduces the searching process of cluster head but also increases the throughput owing to its matrix-based clustering process. Different from conventional clustering process, the routing path selection is based on selection of intermediate cluster head with lower transmittance energy and further implying of shortest path technique to reach the destination node (i.e. another cluster head or base station).

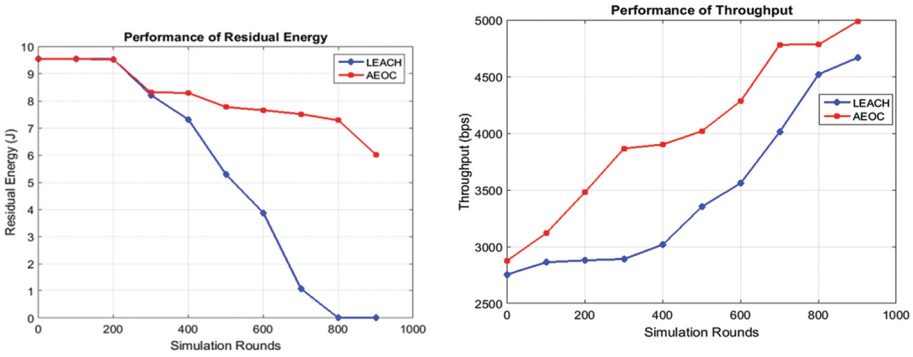


Fig. 2. Graphical outcome of comparative performance analysis

Along with this a first order radio-energy model ensures a standard uses of communication model to ensure better data delivery. The selections of the clusters are carried out considering their best positional information irrespective of any position of the base station. We have also checked for algorithm processing time to find that proposed AEOC consumes 0.2665 s while LEACH algorithm consumes 1.5226 s for 1000 simulation rounds. The numerical outcome also shows that sensors depletes its residual energy when it exceeds 800^{th} simulation rounds for LEACH while AEOC is found still with maximum proportion of residual energy even in 1000 rounds. The complete node death was witnessed on 2100 simulation rounds during our course of simulation trials. Hence, better lifetime is ensured by AEOC.

7 Conclusion and Future Work

The area of WSN is shrouded with various issues where energy issues are still found to be unsolved in research community. The main reason behind this is the frequent usage of similar clustering process ignoring the dynamicity in the possible networks. Hence, this paper has presented a novel thought for mitigating the issues of energy dissipation in WSN by incorporating the principles of optimization in a very simplified manner. The optimization will be carried out by balancing lowered transmittance energy and higher data delivery rate. A novel system of selection of intermediate route was also emphasized in this paper. The prime contribution of the paper is two folds, (i) selection of cluster head based on enhancing 1^{st} order radio-energy model and (ii) selection of energy efficient routing scheme by selecting the intermediate nodes efficiently followed by implying shortest path route. The study outcome was found to be highly in agreement with energy efficient clustering scheme as compared with hierarchical protocol with respect to residual energy and throughput. Our future work will be further to enhance the optimization scheme of clustering using novel bio inspired algorithm. Our focus will be to accomplish better convergence performance.

References

1. Kamila, N.K.: Handbook of Research on Wireless Sensor Network Trends, Technologies, and Applications. IGI Global, Hershey (2016)
2. Fahmy, H.M.A.: Wireless Sensor Networks: Concepts, Applications, Experimentation and Analysis. Signals and Communication Technology. Springer, Singapore (2016)
3. Gungor, V.C., Hancke, G.P.: Industrial Wireless Sensor Networks: Applications, Protocols, and Standards. CRC Press, New York (2013)
4. Gavrilovska, L., Krco, S., Milutinovi, V., Stojmenovic, I., Trobec, R.: Application and Multidisciplinary Aspects of Wireless Sensor Networks: Concepts, Integration, and Case Studies. Springer, New York (2010)
5. Akyildiz, I.F., Vuran, M.C.: Wireless Sensor Networks. John Wiley & Sons, Hoboken (2010)
6. Glisic, S.: Advanced Wireless Networks: Technology and Business Models. John Wiley & Sons, Hoboken (2016)
7. Garg, V.: Wireless Communications & Networking. Morgan Kaufmann, San Francisco (2010)
8. Xiao, Y., Chen, H., Li, F.H.: Handbook on Sensor Networks. World Scientific, Singapore (2010)
9. Issac, B., Israr, N.: Case Studies in Intelligent Computing: Achievements and Trends, Computers, 593 p. CRC Press, Boca Raton (2015)
10. Parvathy, C., Suresha: Existing routing protocols for wireless sensor network - a study. Int. J. Comput. Eng. Res. **4**(7), 8–27 (2014)
11. Rabbat, M., Nowak, R.: Distributed optimization in sensor networks. In: ACM - Proceedings of the 3rd International Symposium on Information Processing in Sensor Networks, pp. 20–27 (2004)
12. He, S., Chen, J., Yau, D.K.Y., Sun, Y.: Cross-layer optimization of correlated data gathering in wireless sensor networks. IEEE Trans. Mob. Comput. **11**(11), 1678–1691 (2012)
13. Cao, Z., He, Y., Liu, Y.: L^2: lazy forwarding in low duty cycle wireless sensor networks. IEEE Trans. Network. **23**(3), 922–930 (2015)
14. Pushpalatha, N., Anuradha, B.: Shortest path position estimation between source and destination nodes in wireless sensor networks with low cost. Int. J. Emerg. Technol. Adv. Eng. **2**(4), 6–12 (2012)
15. Arumugam, G.S., Ponnuchamy, T.: EE-LEACH: development of energy-efficient LEACH protocol for data gathering in WSN. EURASIP J. Wirel. Commun. **76**, 1–9 (2015). Springer
16. Gholipour, M., Haghighat, A.T., Meybodi, M.R.: Hop-by-hop traffic-aware routing to congestion control in wireless sensor networks. EURASIP J. Wirel. Commun. **15**, 1–13 (2015). Springer
17. Kannan, K.N., Paramasivan, B.: Development of energy-efficient routing protocol in wireless sensor networks using optimal gradient routing with on demand neighborhood information. Int. J. Distrib. Sens. Netw. **2014**, 1–7 (2014). Hindawi Publishing Corporation
18. Javaid, N., Rasheed, M.B., Imran, M., Guizani, M., Khan, Z.A., Alghamdi, T.A., Ilahi, M.: An energy-efficient distributed clustering algorithm for heterogeneous WSNs. EURASIP J. Wirel. Commun. **151**, 1–11 (2015). Springer
19. Jumira, O., Wolhuter, R., Zeadally, S.: Energy-efficient beaconless geographic routing in energy harvested wireless sensor networks. Concurrency and Computation: Practice and Experience. Wiley Online Library (2012)
20. Amiri, E., Keshavarz, H., Alizadeh, M., Zamani, M., Khodadadi, T.: Energy efficient routing in wireless sensor networks based on fuzzy ant colony optimization. Int. J. Distrib. Sens. Netw. **2014**, 1–17 (2014). Hindawi Publishing Corporation

21. Xu, M., Liu, G., Wu, H.: An energy-efficient routing algorithm for underwater wireless sensor networks inspired by ultrasonic frogs. Int. J. Distrib. Sens. Netw. **2014**, 1–12 (2014). Hindawi Publishing Corporation
22. Wu, X., Wang, Y., Liu, G., Li, J., Shu, L., Zhang, X., Chen, H., Lee, S.: Energy-efficient routing algorithms based on OVSF code and priority in clustered wireless sensor networks. Int. J. Distrib. Sens. Netw. p. 1 (2013). Hindawi Publishing Corporation
23. Gupta, S., Bhatia, V., Puri, V.: VPBC: a varying probability-based clustering for energy enhancement in WSN. In: Singh, R., Choudhury, S. (eds.) Proceeding of International Conference on Intelligent Communication, Control, and Devices. AISC, vol. 479, pp. 795–802. Springer, Singapore (2016)
24. Heinzelman, W.R., Chandrakasan, A., Balakrishnan, H.: Energy-efficient communication protocol for wireless microsensor networks. In: Proceedings of the 33rd Annual Hawaii International Conference on System Sciences, vol. 2, 4–7 January 2000, pp. 1–10 (2000). Li, L., Halpern, J.Y.: Minimum-energy mobile wireless networks revisited. In: IEEE International Conference on Communications, ICC 2001, Helsinki, Finland, June 2001

Large Networks of Diameter Two Based on Cayley Graphs

Marcel Abas(✉)

Faculty of Materials Science and Technology in Trnava,
Institute of Applied Informatics, Automation and Mechatronics,
Slovak University of Technology in Bratislava, Trnava, Slovak Republic
abas@stuba.sk

Abstract. In this contribution we present a construction of large networks of diameter two and of order $\frac{1}{2}d^2$ for every degree $d \geq 8$, based on Cayley graphs with surprisingly simple underlying groups. For several small degrees we construct Cayley graphs of diameter two and of order greater than $\frac{2}{3}$ of Moore bound and we show that Cayley graphs of degrees $d \in \{16, 17, 18, 23, 24, 31, \ldots, 35\}$ constructed in this paper are the largest currently known vertex-transitive graphs of diameter two.

Keywords: Degree · Diameter · Moore bound · Cayley graph · Networks

1 Introduction

Nowadays, large-scale networks (interconnection, optical, social, electrical, etc.) are a subject of very intensive study. Representing nodes of networks by vertices and communication lines by (directed) edges, networks can be modeled by (di)graphs. Below in Fig. 1 we can see a model of a simple computer network with computers $c_0, c_1, \ldots, c_7$.

Maximum communication delay and maximum communication lines connected to a node are the two main basic limitations on any network. These parameters correspond to the diameter and the maximum (out)degree, respectively, of the corresponding (di)graph. The next important property which a good model of a network might posses is simple and efficient routing algorithm. Since Cayley graphs are vertex-transitive, it is possible to implement the same routing and communication schemes at each node of the network they model [6].

The problem to find, for given diameter k and maximum degree d, the largest order $n(d, k)$ of a graph with given parameters, is in graph theory known as *degree-diameter* problem. There is a well known upper bound on the number $n(d, k)$ - *Moore bound*, which gives $n(d, k) \leq 1 + d + d(d-1) + d(d-1)^2 + \cdots + d(d-1)^{k-1}$ for all positive degrees d and diameters k.

The Moore bound for diameter $k = 2$ is $n(d, 2) \leq d^2 + 1$ and for degrees $d \geq 4$, $d \neq 7$ and $d \neq 57$ we have the bound $n(d, 2) \leq d^2 - 1$ [3]. The maximum order of a Cayley graph of diameter two and degree d is denoted by $C(d, 2)$ and

R. Silhavy et al. (eds.), *Cybernetics and Mathematics Applications in Intelligent Systems*,
Advances in Intelligent Systems and Computing 574, DOI 10.1007/978-3-319-57264-2_23

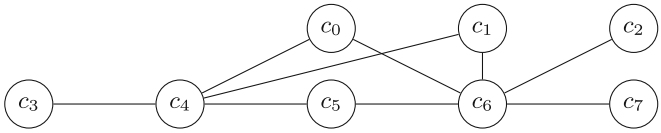


Fig. 1. A model of a simple computer network

for these graphs we have the following results. In [10] the authors constructed Cayley graphs of diameter two and of order $\frac{1}{2}(d+1)^2$ for all degrees $d = 2q-1$ where q is an odd prime power and the same authors gave a construction of Cayley graphs of diameter two and of order $d^2 - O(d^{\frac{3}{2}})$ for an infinite set of degrees d of a very special type [11]. It was shown in [1] that for all degrees $d \geq 4$ we have $C(d,2) \geq \frac{1}{2}d^2 - k$ for d even and $C(d,2) \geq \frac{1}{2}(d^2+d) - k$ for d odd, where $0 \leq k \leq 8$ is an integer depending on the congruence class of d modulo 8. In [4] the author has shown that lower as well as upper bounds on the number of vertices of Cayley graphs of diameter two and degree d for underlying dihedral groups are assymptotically $\frac{1}{2}d^2$. Finally, in [2] the author constructed for all degrees $d \geq 360756$ Cayley graphs of diameter two and of order greater than $0.684d^2$.

In this paper we give a construction of Cayley graphs of diameter two and of order $\frac{1}{2}d^2$ for every degree $d \geq 8$, with surprisingly simple underlying groups. For several small degrees we construct Cayley graphs of diameter two and of order greater than $\frac{2}{3}$ of Moore bound and we show that Cayley graphs of degrees $d \in \{16, 17, 18, 23, 24, 31, \ldots, 35\}$ constructed in this paper are the largest currently known vertex-transitive graphs of diameter two.

2 Preliminaries

For a given finite group Γ and a unit-free, inverse-closed generating set X of Γ, the Cayley graph $G = Cay(\Gamma, X)$ is a graph with vertex set $V(G) = \Gamma$ and with edge set $E(G) = \{\{g,h\} | g \in \Gamma, g^{-1}h \in X\}$. Since X is inverse-closed (that is $X = X^{-1}$), for every $g^{-1}h \in X$ we have $h^{-1}g \in X$. Therefore our Cayley graphs are undirected. It is well known that Cayley graphs are vertex transitive. The Cayley graph for underlying group $\Gamma = \mathcal{Z}_6$ and generating set $X = \{1, 3, 5\}$ is shown in Fig. 2 below. The edges corresponding to generators 1 and $(1)^{-1} = 5$ are drawn dashed.

Throughout this paper, an additive cyclic group of order n, with elements $\{0, 1, \ldots, n-1\}$ and identity element 0, will be denoted by $\mathcal{Z}_n$. Let $\Gamma_n = (\mathcal{Z}_n \times \mathcal{Z}_n) \rtimes \mathcal{Z}_2$ be a semidirect product of $\mathcal{Z}_2$ acting on $\mathcal{Z}_n^2 = \mathcal{Z}_n \times \mathcal{Z}_n$ such that the non-identity element of $\mathcal{Z}_2$ interchanges the coordinates of elements of $\mathcal{Z}_n^2$. That is $0 \in \mathcal{Z}_2 : (x,y) \to (x,y)$ and $1 \in \mathcal{Z}_2 : (x,y) \to (y,x)$. We will write the elements of Γ_n as triples (x,y,i) where $x, y \in \mathcal{Z}_n$ and $i \in \{0,1\}$. The inverse element to (x_0, x_1, i) is $(-x_i, -x_{i+1}, i)$ and for the product of two elements of Γ_n we have $(x_0, x_1, i) \cdot (y_0, y_1, j) = (x_0 + y_i, x_1 + y_{i+1}, i + j)$, where the indices are taken modulo 2.

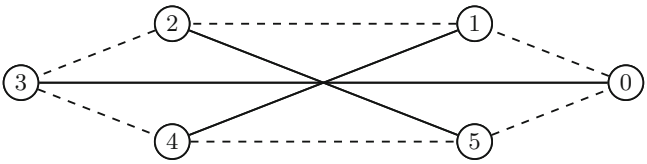


Fig. 2. Cayley graph for underlying group $\Gamma = \mathcal{Z}_6$ and generating set $X = \{1, 3, 5\}$

3 Large Cayley Graphs of Diameter Two

Theorem 1. *Let $r \geq 1$ be an integer, let $s, \epsilon \in \{0, 1\}$ and let $n = 4r + 2s + \epsilon$. Then there exists a Cayley graph of diameter two, degree $d = 2n - s + \epsilon$ and of order $\frac{1}{2}(d + s - \epsilon)^2$.*

Proof. We set $m = \lfloor \frac{n}{2} \rfloor = 2r + s$. Let the underlying group of the Cayley graph $G = Cay(\Gamma, X)$ be $\Gamma = \Gamma_n = (\mathcal{Z}_n \times \mathcal{Z}_n) \rtimes \mathcal{Z}_2$, defined in Sect. 2 and let the generating set X be the union $X = A \cup B \cup B^{-1} \cup C \cup C^{-1}$, where the sets A, B and C are defined as follows:

$A = \{a(i) | i \in \{0, 1, \ldots m + 2\epsilon - 2\}\}$, $a(i) = (i, -i, 1)$, $a(i)^{-1} = a(i)$
$B = \{b(i) | i \in \{1, 2, \ldots m\}\}$, $b(i) = (0, i, 1)$, $b(i)^{-1} = (-i, 0, 1)$
$C = \{c(i) | i \in \{0, 1, \ldots r\}\}$, $c(i) = (m - i, i, 0)$, $c(i)^{-1} = (-m + i, -i, 0)$

We can see that $|A| = m + 2\epsilon - 1$, $|B \cup B^{-1}| = n - \epsilon$ and $|C \cup C^{-1}| = 2r + \epsilon + 1$. Therefore the generating set X has order $(m + 2\epsilon - 1) + (n - \epsilon) + (2r + \epsilon + 1) = 2n - s + \epsilon$ and the Cayley graph G has order $|G| = |\Gamma| = 2n^2$. Since the degree of G is $d = |X| = 2n - s + \epsilon$, the graph G has order $\frac{1}{2}(d + s - \epsilon)^2$. To show that the Cayley graph has diameter two it is sufficient to show that every element of Γ_n is from X or it can be written as a product of two elements from X.

The rest of the proof is divided into two parts: in part (I) we generate elements of the form $(i, j, 0)$ and in the part (II) we show how to generate elements of the form $(i, j, 1)$. In the next, all calculations are performed modulo n. Note that if n is even then $-m = m$ and if n is odd then $-m = \lceil \frac{n}{2} \rceil = m + \epsilon$.

(I) Generating elements of the form $(i, j, 0)$.

(a) Elements of the form $(i, -i, 0)$.
(1) $0 \leq i \leq m - 2$
$(i, -i, 0) = a(m - 2) \cdot a(m - 2 - i) = (m - 2, 2 - m, 1)(m - 2 - i, -m + 2 + i, 1)$
(2a) n is odd
$(m - 1, -m + 1, 0) = a(m) \cdot a(1) = (m, -m, 1)(1, -1, 1)$
$(m, -m, 0) = a(m) \cdot a(0) = (m, -m, 1)(0, 0, 1)$
(2b) n is even
$(m-1, -m+1, 0) = c(r)^{-1} \cdot c(r-1+s)^{-1} = (m+r, -r, 0)(m+r-1+s, -r+1-s, 0)$
$(m, -m, 0) = b(m) \cdot b(m) = (0, m, 1)(0, m, 1)$ (if n is even then $m = -m$)
The other elements of the form $(i, j, 0)$ are inverses of the previous.

(b) Elements of the form $(i, j, 0)$, $j \neq -i$.
All the other elements in this part will be generated as products of generators from $A \cup B \cup B^{-1}$. It is easy to see that if $u, v \in \{(i', j', 1) | i', j' \in \mathcal{Z}_n\}$ and $u \cdot v = (i, j, 0)$, then $v \cdot u = (j, i, 0)$, $u^{-1} \cdot v^{-1} = (-j, -i, 0)$ and $v^{-1} \cdot u^{-1} = (-i, -j, 0)$. Therefore it is sufficient to show how to generate elements $(i, j, 0)$ for $0 \leq i \leq m$, $i \leq j < n - 1$.
(1a) $i = 0, 1 \leq j \leq m - 1$
$(0, j, 0) = b(m) \cdot b(m - j)^{-1} = (0, m, 1)(-m + j, 0, 1)$
(1b) $i = 0, j = m$
$(0, m, 0) = b(m) \cdot a(0) = (0, m, 1)(0, 0, 1)$
(1c) $i = 0, m + 1 \leq j \leq n - 1$
the elements $(0, j, 0)$ are inverses of those in (1a) and (1b)
(2) $1 \leq i \leq m, i \leq j \leq m$
$(i, j, 0) = b(j) \cdot b(i) = (0, j, 1)(0, i, 1)$
(3) $1 \leq i \leq m - 2 + \epsilon, m + 1 \leq j \leq n - 1 - i$
$(i, j, 0) = a(i) \cdot b(-i - j)^{-1} = (i, -i, 1)(i + j, 0, 1)$

(II) Generating elements of the form $(i, j, 1)$. There are exactly n congruence classes $C_0, C_1, \ldots, C_{n-1}$ of elements of the form $(i, j, 1)$ such that $(i, j, 1) \in C_k$ if and only if $i + j = k$, $k = 0, 1, \ldots, n - 1$. Since $(i, j, 1) \in C_k$ if and only if $(i, j, 1)^{-1} \in C_{-k}$, we will do the proof only for $k = 0, 1, \ldots, m$. For fixed k it is sufficient to show that either first or the second coordinate runs from 0 to $n - 1$.

(a) $1 \leq k \leq m - 1$
(1) The second coordinate is $0, 1, \ldots, r$:
$(k - j, j, 1) = c(j) \cdot b(m - k)^{-1} = (m - j, j, 0)(-m + k, 0, 1)$
(2) The second coordinate is $r + s, \ldots, 2r + s$:
$(m + \epsilon + k + j, m - j, 1) = b(m - k)^{-1} \cdot c(j) = (-(m - k), 0, 1)(m - j, j, 0)$
(3) The second coordinate is $2r + s + \epsilon, \ldots, 32r + s + \epsilon$:
$(m + k - j, m + \epsilon + j, 1) = b(m + \epsilon - k)^{-1} \cdot c(j)^{-1} = (-(m + \epsilon - k), 0, 1)(m + \epsilon + j, -j, 0)$
(4) The second coordinate is $3r + 2s + \epsilon, \ldots, n - 1$:
$(k + j, -j, 1) = c(j)^{-1} \cdot b(m + \epsilon - k)^{-1} = (m + \epsilon + j, -j, 0)(-(m + \epsilon - k), 0, 1)$

(b) $k = m$
(1) The first coordinate is $0, 1, \ldots, r$:
$(i, m - i, 1) = a(0) \cdot c(i) = (0, 0, 1)(m - i, i, 0)$
(2) The first coordinate is $r + s, \ldots, 2r + s$:
$(m - i, i, 1) = c(i) \cdot a(0) = (m - i, i, 0)(0, 0, 1)$
(3) The first coordinate is $2r + s + \epsilon, \ldots, 3r + s + \epsilon$:
$(m + \epsilon + i, -i, 1) = c(i)^{-1} \cdot a(0) = (m + \epsilon + i, -i, 0)(0, 0, 1)$
(4) The first coordinate is $3r + 2s + \epsilon, \ldots, n - 1$:
$(-i, m + \epsilon + i, 1) = a(0) \cdot c(i)^{-1} = (0, 0, 1)(m + \epsilon + i, -i, 0)$

(c) $k = 0$
(1a) n is even, the first coordinate is $0, 1, \ldots, r$:

$(i, -i, 1) = b(m) \cdot c(i) = (0, m, 1)(m - i, i, 0)$
(2a) n is even, the first coordinate is $r + s, \ldots, 2r + s$:
$(m - i, i, 1) = c(i) \cdot b(m) = (m - i, i, 0)(0, m, 1)$
(1,2b) n is odd, the first coordinate is $0, 1, \ldots, 2r + s$:
$(i, -i, 1) = a(i)$
(3) The first coordinate is $2r + s + \epsilon, \ldots, 3r + s + \epsilon$:
$(-m + i, m - i, 1) = b(m)^{-1} \cdot c(i) = (-m, 0, 1)(m - i, i, 0)$
(4) The first coordinate is $3r + 2s + \epsilon, \ldots, n - 1$:
$(-i, i, 1) = b(m) \cdot c(i)^{-1} = (0, m, 1)(m + \epsilon + i, -i, 0)$

4 New Record Cayley Graphs of Small Degrees

In this section we construct Cayley graphs of diameter two of large orders for several small degrees. Some of these graphs has order greater than 2/3 of Moore bound, and even the graph of degree 16 has order greater than 3/4 of Moore bound. These graphs were found using GAP (Groups, Algorithms, Programming - a System for Computational Discrete Algebra [5]).

Theorem 2. *There are the following lower bounds on the order of Cayley graphs of degrees* $d \in \{16, 21, 23, 28, 31, 37, 40, 46, 49, 54\}$ *and diameter 2:* $C(16, 2) \geq 200, C(21, 2) \geq 288, C(23, 2) \geq 392, C(28, 2) \geq 512, C(31, 2) \geq 648, C(37, 2) \geq 800, C(40, 2) \geq 968, C(46, 2) \geq 1152, C(49, 2) \geq 1352, C(54, 2) \geq 1568$.

Proof. We present a detailed proof for $d = 16$. The underlying group of the corresponding Cayley graph is the group $\Gamma = \Gamma_{10} = \mathcal{Z}_{10}^2 \rtimes \mathcal{Z}_2$ (as in Theorem 1, for n = 10) and the generating set $X = A \cup B \cup B^{-1} \cup C \cup C^{-1}$, where $A = \{(0, 0, 1)\}$, $B = \{(1, 0, 1), (1, 3, 1), (1, 7, 1), (5, 0, 1), (5, 2, 1)\}$ and $C = \{(5, 0, 0), (4, 1, 0), (3, 2, 0)\}$. We can see that the order of the Cayley graph $G = Cay(\Gamma, X)$ is $|G| = |\Gamma| = 200$ and its degree is $d = |X| = 16$. It can be verified by a straightforward calculation that the Cayley graph G has diameter 2. That is, $C(16, 2) \geq 200$. For the calculation one can use, for example, the following code (in GAP [5]):

```
n:=10;;
A:=[ [ 0, 0, 1 ] ];;
B:=[ [ 1, 0, 1 ], [1, 3, 1 ], [ 1, 7, 1], [5, 0, 1 ], [5, 2, 1] ];;
C:=[ [ 5, 0, 0 ], [ 4, 1, 0 ], [ 3, 2, 0 ] ];;
Xgen:=Union(A,B,C);;
for i in B do AddSet(Xgen,[(n-i[2]) mod n,(n-i[1]) mod n,1]); od;
for i in C do AddSet(Xgen,[(n-i[1]) mod n,(n-i[2]) mod n,0]); od;
times := function(x,y)
 if x[3]=0
  then return [(x[1]+y[1]) mod n, (x[2]+y[2]) mod n, (x[3]+y[3]) mod 2];
  else return [(x[1]+y[2]) mod n, (x[2]+y[1]) mod n, (x[3]+y[3]) mod 2];
 fi;
end;;
Generate:=[];;
```

```
for x in Xgen do for y in Xgen do AddSet(Generate,times(x,y)); od; od;
Print("Degree d =",Size(Xgen),"\n");
Print("Number of generated elements =",Size(Generate));
```

For degrees $d = 21, 23, 28, 31, 37, 40, 46, 49, 54,$ the underlying group of the corresponding Cayley graph is the group $\mathcal{Z}_n^2 \rtimes \mathcal{Z}_2$, with $n = 12, 14, 16, 18, 20, 22, 24, 26, 28,$ respectively, and the generating set is $X = A \cup B \cup B^{-1} \cup C \cup C^{-1}$, where $C = \{c(i) | i \in \{0, 1, \ldots r\}\}$, $c(i) = (m-i, i, 0)$, $m = \frac{n}{2}$, $r = \left\lfloor \frac{n}{4} \right\rfloor = \left\lfloor \frac{m}{2} \right\rfloor$. For the sets A and B we have:

$n = 12,\ m = 6,\ r = 3,\ \|G\| = 2 \cdot 12^2 = 288,\ d = 21$ $A = \{(0,0,1), (3,9,1)\}$ $B = \{(0,1,1), (0,2,1), (6,9,1), (4,0,1), (5,0,1), (1,5,1)\}$
$n = 14,\ m = 7,\ r = 3,\ \|G\| = 2 \cdot 14^2 = 392,\ d = 23$ $A = \{(0,0,1), (9,5,1)\}$ $B = \{(0,1,1), (0,2,1), (3,0,1), (12,6,1), (5,0,1), (7,13,1), (4,3,1)\}$
$n = 16,\ m = 8,\ r = 4,\ \|G\| = 2 \cdot 16^2 = 512,\ d = 28$ $A = \{(0,0,1), (1,15,1), (2,14,1)\}$ $B = \{(0,1,1), (0,2,1), (11,8,1), (6,14,1), (0,5,1), (9,13,1), (8,15,1), (4,4,1)\}$
$n = 18,\ m = 9,\ r = 4,\ \|G\| = 2 \cdot 18^2 = 648,\ d = 31$ $A = \{(0,0,1), (6,12,1), (7,11,1), (14,4,1)\}$ $B = \{(0,1,1), (2,0,1), (6,15,1), (1,3,1), (8,5,1), (17,7,1), (13,12,1), (11,15,1),$ $(6,3,1)\}$
$n = 20,\ m = 10,\ r = 5,\ \|G\| = 2 \cdot 20^2 = 800,\ d = 37$ $A = \{(0,0,1), (1,19,1), (2,18,1), (3,17,1), (10,10,1), (13,7,1)\}$ $B = \{(0,1,1), (5,17,1), (0,3,1), (9,15,1), (6,19,1), (17,9,1), (11,16,1), (3,5,1),$ $(0,9,1), (3,7,1)\}$
$n = 22,\ m = 11,\ r = 5,\ \|G\| = 2 \cdot 22^2 = 968,\ d = 40$ $A = \{(0,0,1), (1,21,1), (2,20,1), (3,19,1), (4,18,1), (5,17,1), (11,11,1)\}$ $B = \{(0,1,1), (9,15,1), (3,0,1), (11,15,1), (9,18,1), (2,4,1), (19,10,1), (3,5,1),$ $(14,17,1), (20,12,1), (11,0,1)\}$
$n = 24,\ m = 12,\ r = 6,\ \|G\| = 2 \cdot 24^2 = 1152,\ d = 46$ $A = \{(0,0,1), (1,23,1), (2,22,1), (3,21,1), (4,20,1), (5,19,1), (6,18,1), (7,17,1),$ $(12,12,1)\}$ $B = \{(0,1,1), (0,2,1), (0,3,1), (0,4,1), (0,5,1), (0,6,1), (0,7,1), (16,16,1),$ $(14,19,1), (1,9,1), (14,21,1), (12,0,1)\}$
$n = 26,\ m = 13,\ r = 6,\ \|G\| = 2 \cdot 26^2 = 1352,\ d = 49$ $A = \{(0,0,1), (1,25,1), (2,24,1), (3,23,1), (4,22,1), (5,21,1), (6,20,1), (7,19,1),$ $(8,18,1), (13,13,1)\}$ $B = \{(0,1,1), (0,2,1), (0,3,1), (0,4,1), (0,5,1), (0,6,1), (0,7,1), (0,8,1), (16,19,1),$ $(17,19,1), (3,8,1), (3,9,1), (13,0,1)\}$
$n = 28,\ m = 14,\ r = 7,\ \|G\| = 2 \cdot 28^2 = 1568,\ d = 54$ $A = \{(0,0,1), (1,27,1), (2,26,1), (3,25,1), (4,24,1), (5,23,1), (6,22,1), (7,21,1),$ $(8,20,1), (9,19,1), (14,14,1)\}$ $B = \{(0,1,1), (0,2,1), (0,3,1), (0,4,1), (0,5,1), (0,6,1), (0,7,1), (0,8,1), (0,9,1),$ $(17,21,1), (18,21,1), (3,9,1), (3,10,1), (14,0,1)\}$

which complete the proof of this theorem.

Table 1. Online record table of degree-diameter problem for Cayley graphs and our results [12]

Degree d	E. Loz	SS [10]	A [1]	NEW	Percentage
13	112				65.88
14	128				64.97
15	144				63.71
16	155			**200**	**77.82**
17	170			**200**	**68.96**
18	192			**200**	**61.53**
19	200				55.24
20	210				52.36
21		242		**288**	**65.15**
22		242		**288**	**59.38**
23			270	**392**	**73.96**
24			280	**392**	**67.93**
25		338		**392**	**62.61**
26		338		**392**	**57.90**
27			378	**392**	**53.69**
28			392	**512**	**65.22**
29			434	**512**	**60.80**
30			448	**512**	**56.82**
31		512		**648**	**67.35**
32		512		**648**	**63.21**
33		512		**648**	**59.44**
34		512		**648**	**56.00**
35			630	**648**	**52.85**
36			648		49.96
37		722		**800**	**58.39**
38		722		**800**	**55.36**
39			774	**800**	**52.56**
40			792	**968**	**60.46**
41			858	**968**	**57.55**
42			880	**968**	**54.84**
43			946	**968**	**52.32**
44			968		49.97
45		1058			52.22
46		1058		**1152**	**54.41**
47			1122	**1152**	**52.12**
48			1144	**1152**	**49.97**

Table 1. *(Continued)*

Degree d	E. Loz	SS [10]	A [1]	NEW	Percentage
49		1250		**1352**	**56.28**
50		1250		**1352**	**54.05**
51			1326	**1352**	**51.96**
52			1352		49.98
53		1458			51.88
54		1458		**1568**	**53.75**
55			1534	**1568**	**51.81**
56			1560	**1568**	**49.98**
57		1682			51.75

5 Conclusion and Remarks

In [7] the authors constructed an infinite family of vertex transitive non-Cayley graphs of degree d, diameter 2 and of order $\frac{8}{9}(d+\frac{1}{2})^2$. A simplified construction in terms of Abelian lifts of dipoles with loops and multiple edges was presented in [8]. It was shown in [9] that the maximum order of graphs of diameter 2 and degree d which arise as lifts of dipoles with loops and multiple edges, with voltage assignments in Abelian groups is $0.932d^2$. We can see that for degrees $d = 16, 21, 23, 28, 31$ the orders of Cayley graphs constructed in Theorem 2 are $\frac{8}{9}(d-c)^2$, where $c = 1, 3, 2, 4, 4$, respectively. For example for $d = 16$ we have order $200 = \frac{8}{9}(16-1)^2$. This observation suggest the following conjecture:

Conjecture 1. There is a positive constant c such that for any natural d there is a Cayley graph of diameter two degree d and of order at least $\frac{8}{9}(d-c)^2$.

Above, the reader can see a table of present record Cayley graphs of diameter two (retrieved from online table [12]), with our results added (bold font). It shows that, for example, for degrees $13 \leq d \leq 57$ our construction (plus Theorem 2) gives better results in 34 cases of total 45 degrees. The orders of Cayley graphs for degrees $3 \leq d \leq 12$ are not listed - these graphs were found, and shown to be optimal by Marston Conder [13]. Note that until now the largest known vertex-transitive graphs of diameter two have values $(d; 2) = (16; 162), (17; 170), (18; 192), (23, 24; 338), (31, \ldots, 35; 578)$ while our results give $(d; 2) = (16, 17, 18; 200), (23, 24; 392), (31, \ldots, 35; 648)$ (Table 1).

Acknowledgements. The research was supported by VEGA Research Grant No. 1/0811/14 and by the Operational Programme 'Research & Development' funded by the European Regional Development Fund through implementation of the project ITMS 26220220179.

References

1. Abas, M.: Cayley graphs of diameter two and any degree with order half of the Moore bound. Discrete Appl. Math. **173**, 1–7 (2014)
2. Abas, M.: Cayley graphs of diameter two with order greater than 0.684 of the Moore bound for any degree. Eur. J. Comb. **57**, 109–120 (2016)
3. Erdös, P., Fajtlowicz, S., Hoffman, A.J.: Maximum degree in graphs of diameter 2. Networks **10**, 87–90 (1980)
4. Erskine, G.: Diameter 2 Cayley graphs of dihedral groups. Discrete Math. **338**(6), 1022–1024 (2015)
5. GAP - Groups, Algorithms, Programming - a System for Computational Discrete Algebra. www.gap-system.org
6. Heydemann, M.C.: Cayley graphs and interconnections networks. In: Hahn, G., Sabidussi, G. (eds.) Graph Symmetry, Kluwer, pp. 167–224 (1997)
7. McKay, B.D., Miller, M., Širáň, J.: A note on large graphs of diameter two and given maximum degree. J. Combin. Theory Ser. B **74**, 110–118 (1998)
8. Šiagiová, J.: A note on the McKay-Miller-Širáň graphs. J. Comb. Theor. Ser. B **81**, 205–208 (2001)
9. Šiagiová, J.: A Moore-like bound for graphs of diameter 2 and given degree, obtained as abelian lifts of dipoles. Acta Mathematica Universitatis Comenianae **71**(2), 157–161 (2002)
10. Šiagiová, J., Širáň, J.: A note on large Cayley graphs of diameter two and given degree. Discrete Math. **305**(1–3), 379–382 (2005)
11. Šiagiová, J., Širáň, J.: Approaching the Moore bound for diameter two by Cayley graphs. J. Comb. Theor. Ser. B **102**(2), 470–473 (2012)
12. Degree-diameter problem: largest known Cayley graphs of diameter two. http://www.combinatoricswiki.org/wiki/The_Degree_Diameter_Problem_for_Cayley_Graphs#Cayley_Graphs_of_Diameter_Two
13. Degree-diameter problem: description of provably largest Cayley graphs. http://combinatoricswiki.org/wiki/Description_of_optimal_Cayley_graphs_found_by_Marston_Conder

Integrated S-AODV and DEL-CMAC Algorithm of Spatio Temporal Cross-Layer in Sensor Network

Shoba Chandra[1(✉)], Suresha Talanki[2], and Kiran Kumari Patil[3]

[1] Department of Information Science & Engineering, Sri Venkateshwara College of Engineering, Bengaluru, India
shoba19.svce@gmail.com

[2] Department of Computer Science & Engineering, Sri Venkateshwara College of Engineering, Bengaluru, India
suresha_rec@rediffmail.com

[3] School of Computing, REVA University, Bengaluru, India
kirankumari@reva.edu.in

Abstract. Cooperative Medium Access Protocol (CMAC) has been found to contribute towards energy efficiency among the sensor nodes; however, still there is less research work to prove their applicability on the sensor nodes when adhoc-based routing is adopted. The existing research work is reviewed towards utilizing cross-layer based approach for maximizing the layer interactivity and to enhance the computational efficiency in Wireless Sensor Network (WSN). Hence, the proposed system addresses increase in efficiency by incorporating a novel combinatorial policy of on demand adhoc routing with CMAC scheme over multihop network using cooperative transmission mechanism to bridge the communication gap between network layer and MAC layer. The outcome of the proposed system is found to excel better quality-of-service (QoS) performance in comparison to existing MAC protocol frequently used in WSN.

Keywords: Wireless Sensor Network · Cooperative Medium Access Protocol · Cross layer · Quality of service

1 Introduction

Wireless Sensor Network plays a potential role in upcoming technologies of pervasive computing e.g. Internet-of-Things [1, 2] where massive data collection takes place from multiple sources. Normally, the adoptions of sensors are always done for long-term usage and hence network lifetime becomes one selection factor for choosing the type of sensors for specific applications [18]. As the sensor nodes are quite smaller in size so it has limited computational capabilities and restricted resources [16]. There have been various studies in last decade, where cross layer approach has gained a pace in enhancing the communication performance of sensors [7, 19]. However, such studies have not yet achieved a benchmark and yet various open end problems still exist in present times [6]. Another most frequently adopted technique is Cooperative Medium Access Protocol or CMAC schemes that is particularly meant for addressing the problems of data delivery

R. Silhavy et al. (eds.), *Cybernetics and Mathematics Applications in Intelligent Systems*, Advances in Intelligent Systems and Computing 574, DOI 10.1007/978-3-319-57264-2_24

and energy consumption problems in Wireless Sensor Network [20]. Although, there has been couple of studies being done using CMAC scheme but it is still associated with limitations pertaining to Quality of Service (QoS). An efficient CMAC scheme should incorporate a better relay node selection process along with retention of higher degree of energy conservation. One of the biggest challenges in this regard is its usage of slots, which are not directly applicable for wireless adhoc based networks. It is highly essential that CMAC scheme should have significant compatibility as well as supportability towards communication standards supporting adhoc networks. Hence, this paper presents a technique where adhoc-based routing strategy is combined with CMAC scheme for leveraging the cross-layer performance in Sensor Network. Section 2 discusses about the existing research work followed by problem identification in Sect. 3. Section 4 discusses about proposed methodology followed by elaborated discussion of algorithm implementation in Sect. 5. Comparative analysis of accomplished result is discussed under Sect. 6 followed by conclusion in Sect. 7.

2 Related Work

This section discusses about the most recent implementation work being carried out towards cross layer approach in Wireless Sensor Network. Dobslaw et al. [8] have presented a cross layer approach using network layer and spanning MAC layer for accomplishing higher reliability factor with multiple base stations. Addressing the problems of energy consumption in Wireless Multimedia Sensor Network was seen in the work carried out by Kader et al. [4] using adaptive scheme of cross layer. The study also introduces a route scheduling technique for routing packets with multiple priorities. Su et al. [5] have presented a technique of cross layer for incorporating rate allocation policies. Data transmission using contention free approach was scheduled by the technique over single hop networks to show lower computational complexity in Underwater Acoustic Network. Considering correlation in the design process of cross layer based approach was found in the study of Das and Misra [8]. Gama et al. [9] have presented a Cooperative MAC scheme for conserving energy consumption in sensor network. The technique uses Markov-based approach for computing state of channel. The technique uses a novel selection strategy of relay nodes and adopts store and forward policy for cooperative transmittance. Kartsakli et al. [10] introduced a MAC scheme for bridging communication with cloud for data exchange with less error rate. Lee et al. [11] have introduced a novel optimization scheme where duty cycle is optimized over a new scheduling scheme for Energy Harvesting Sensor Network. Lin et al. [12] have presented a technique of distributed cross layered approach for catering up statistical QoS performance as well as energy efficiencies. Usman et al. [13] have used experimental approach for implementing cross layer approach using fuzzy logic for addressing energy problems in Smart Home Sensor Network. Xu et al. [14] have used optimization principle on cross layer approach using discrete temporal and stochastic approach. Nearly similar study towards energy efficiency has also been carried out by Yetgin et al. [17]. Liu et al. [21] and Shah et al. [22] have presented a cooperative MAC protocol using analytical modeling on cognitive radio. The next section discusses about the problems identified from existing research work.

3 Problem Description

From the prior section, various research work has already been carried out towards implementing cooperative MAC protocols in Sensor Network. However, the pitfalls are (i) very less work has maintained an equilibrium between communication and computation efficiencies using cross layer approach, (ii) less benchmarked studies, (iii) lesser work considering multiple input and single output with cooperative MAC scheme (as this is the best scheme to optimize the data delivery of the sensors), (iv) less emphasis on implementing adhoc on demand routing to minimize routing overhead using cross layer approach. It has been observed that usage of cooperative MAC protocols is less studied on Wireless Adhoc networking environment, where the challenges are multifold in presence of interference, noise, mobility, etc.

4 Proposed Methodology

The design and development of the proposed study was carried out considering analytical research methodology, where the prime contribution of the study is to obtain an enhanced communication performance in Wireless Sensor Network using cross layer approach.

Using S-AODV and DEL-MAC, the proposed system formulates its system model and communication model (Fig. 1). The system modeling consists of source/destination node, path loss exponent, fading and noise while the communication model consists of introducing novel packet structure, cooperative based transmission using symbol blocks, decoding, and followed by randomization of communication vector using spatial temporal coding approach. The proposed system also ensured that significant QoS enhancement being done using cross layer approach by faster interactivity between network and MAC layer. The next section elaborates about the algorithm implemented to accomplish this research goal.

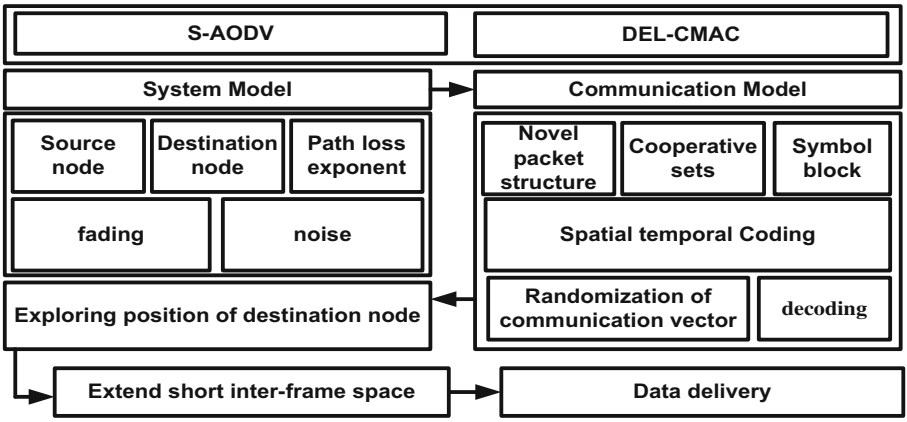


Fig. 1. Schematic architecture of proposed system

5 Algorithm Implementation

The proposed technique applies both S-AODV as well as DEL-CMAC protocol from our prior studies [6, 15] to further increase the interaction between network layer and MAC layer to enhance the power allocation policies among the sensors. The assumptions of the proposed algorithm are (i) that all the nodes can obtain their position information using a GPS (ii) the fading coefficient is highly independent over the communication channel as well as they are uniform over one frame. (iii) the noise over the receiver node has uniform power spectral density. The steps involved in the algorithm are shown below:

Algorithm for Integrated S-AODV and DEL-CMAC

Input: η (sensors), S_A (Simulation Area), σ_{sd} / $\sigma_{\phi d}$ (SNR), E_{TX} (Transmit energy), α_{sd} (channel gain), σ_o (decoding threshold), ϕ (Cooperative set), θ_ϕ (arbitrary vectors), β_{sd} (symbol blocks), δ (antenna)

Output: data transmission among the nodes

Start

1. $S_A \leftarrow \text{rand}(\eta)$
2. $\beta_{sd} = [\beta_o, \beta_1, \ldots. \beta_{\eta-1}]$, consider $\beta \sim \phi$ $[\phi \rightarrow |\phi|]$
3. In ϕ $[\beta \rightarrow f(\beta)]$, where $f(\beta) = \eta \times \delta$
4. $\theta = \{\theta_1, \theta_2, \ldots, \theta_\phi\} = \delta \times \phi$
5. $\alpha_{\phi d} = [\alpha_{1d}, \alpha_{2d}, \ldots. \alpha_{\phi d}]$
6. **For** i=1: η
7. Compute $\sigma_{Sd} \rightarrow E_{TX}|\alpha_{sd}|^2 / \eta_o$ and $\sigma_{\phi d} \rightarrow E_{TX}|\theta_\phi . \alpha_{\phi d}|^2 / \eta_o$ $[\sigma_{Sd} \geq \sigma_o$ and $\sigma_{\phi d} \geq \sigma_o]$
8. $\phi_k = \{d\} \,|\, \sigma_o \leq \sigma_{\phi_{k-1},d}, d \notin \bigcup_{s=1}^{k-1} \phi_s\}$
9. $s \xrightarrow{msg(pos\_req)} d$, $d \xrightarrow{rep(pos\_req)} s$
10. $s \xrightarrow{msg(RTS)} d, d \xrightarrow{rep(CTS)} s$
11. **End**

End

We amended the S-AODV implementation by modifying the control messages to facilitate better cross-layer implementation with spatio-temporal approach. At present, we use 5 types of control messages viz. (i) R_{REQ}, (ii) R_{REP}, (iii) R_{RTS}, (iv) R_{CTS}, (v) R_{ACK}. The R_{REQ} message consists of fields e.g. frame control, destination IP, source IP, position of source, hop count, sequence control, and frame check sequence. The R_{REP} message consists of same fields like R_{REQ} except it has additional field for list of candidate relays between source and destination. The R_{RTS} message consists of similar fields like R_{REP}

except it has additional field for duration. The fields of R_{CTS}, R_{ACK} is nearly the same as that of R_{RTS} with an exception that it doesn't have fields for candidate relays between source and destination. The sensor nodes (η) are randomly deployed in simulation area S_A (Line-1) followed by a communication model leading source *s* to transmit packet to destination *d* using multihop S-AODV protocol. A path loss exponent is considered along with computation of channel gain α_{sd} that is further subjected to standard fading process (Rayleigh). The next step of the algorithm considers symbol block forwarded from node *s* to *d* and empirically represented as $\beta_{sd} = [\beta_o, \beta_1, \ldots \beta_{\eta-1}]$ as shown in Line-2. Out of all the symbols, only the symbols that are completed of decoding (say β) is represented as cooperative set ϕ that will directly express the cardinality of sensors within this cooperative sets (Line-2). The algorithm then implements DEL-CMAC algorithm to ensure that all the sensors present in cooperative set ϕ should be able to integrally transmit the data to destination node *d*. Therefore, using spatial-temporal approach of coding, we can map the symbol block β into $f(\beta)$, where $f(\beta)$ directly represents product of sensors (η) and cardinality of antenna (δ) (Line-3).

For minimizing computational complexity associated with communication between network layer and MAC layer, it is essential that certain arbitrary communication vectors be formulated θ; hence, in such case, we can modify the mapping expression as $\beta \rightarrow f(\beta).\ \beta$. In this expression, β is a variable that corresponds to product of cardinality of antenna (δ) and cooperative set (ϕ) (Line-4). As the processing operations of the sensors are always carried out in local space so, β_k can be independently allocated for every sub-set *k* such that *k* belongs to main cooperative set ϕ. This property of cross-layer approach assist to implement in highly distributed fashion good enough for IoT based applications with sensors.

A communication vector $\alpha_{\phi d}$ is formulated in Line-5 that empirically represents coefficient of channel existing from sensors present in cooperative set ϕ and destination node *d*. Therefore, the signal that will be gathering at the destination node *d* can be expressed as channel matrix i.e. $y = f(\beta)\cdot\theta.\ \alpha_{\phi d} + \psi$, where ψ corresponds to noisy channel. Finally, we compute instantaneous signal-to-noise ratio in Line-7 for all the communication nodes with a condition that decoding is only possible if $\sigma_{Sd} \geq \sigma_o$ and $\sigma_{\phi d} \geq \sigma_o$. Finally, we amend the cooperative set (Line-8) followed by exploring the position of the destination node (Line-9) and transmission of the data (Line-10). The outcome of the algorithm is the delivery of the data in highly cost-effective manner and without losing many resources in the process of implementing cross-layer based approach. This results in faster data delivery process with lower energy consumption.

After receiving the data, the destination node *d* forwards and acknowledgement to its immediate source node to resist flooding. Owing to randomization of communication vector, the delay during the transmission between network layer and MAC layer is highly reduced. We also prolong the duration of short inter-frame space by adding maximum propagation delay to ensure that sensors present in cooperative set should obtain the data packet at approximately similar time duration. By doing this process, the system can ensure better synchronicity in the transmission process. Therefore, the proposed mechanism supports better quality of service in its cross-layer method based on spatial-temporal

approach over the wireless sensor network. The next section briefs about the outcome accomplished from the study.

6 Results Discussion

The proposed algorithm was simulated in NS-2 considering 1000 sensor nodes with slot time of 9 ms, data of 8000 bits, and data rate of 50 Mbps. The size of the MAC header is 270 bits. The R_{REQ} message consists of fields e.g. frame control (size: 16 bits), destination IP (size: 48 bits), source IP (size: 48 bits), position of source (size: 48 bits), hop count (size: 8 bits), sequence control (size: 16 bits), and frame check sequence (size: 32 bits). The study outcome is compared with conventional S-MAC [23] that is frequently used in wireless sensor nodes. The analysis of the outcome is carried out using the QoS parameters viz. delay, throughput, and energy consumption.

Simulated in Matlab, Fig. 2 shows that proposed system offer better delay and throughput performance as compared to S-MAC. A closer look into the study will show difference in energy performance for static (Fig. 2(d)) and mobility scenario (Fig. 2(c)). Energy consumption is more in static network as compared to mobility-based network. The reason for this outcome is that conventional S-MAC suffers from static duty cycle. Whereas the proposed system offers faster communication among the nodes using cross-layer approach that enables faster interaction and updating being done between network and MAC layer. The algorithm processing time of proposed system is found to be 0.2665 s while that of S-MAC is found to be 1.2744 s considering similar simulation environment. The cooperative process implemented using simple mathematical modeling exhibited in algorithm steps are the prime reason for enhanced QoS outcome as compared to frequently used S-MAC protocols.

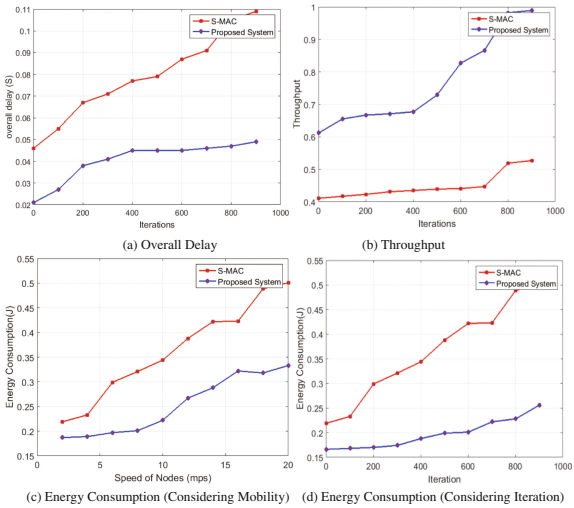


Fig. 2. Comparative analysis of proposed system

7 Conclusion and Future Enhancement

This paper presents a technique that has higher supportability of S-AODV as well as MAC protocol as a medium of implementing cross layer approach using spatial-temporal coding scheme. The purpose was mainly to perform more enhance and faster interconnectivity of network layer and MAC layer in protocol stack of Wireless Sensor Network. The prime contribution of this scheme is that it harnesses the potential characteristics of cooperative transmission that results in transmission of encoded data packet from cooperative group to destination node. This indirectly means that there is less likely to drop any packets be it prioritized or un-prioritized one in any form of traffic condition (both static and mobile). The study outcome was found to offer enhanced throughput, reduced delay, and lowered energy consumption with respect to S-MAC, which is widely implemented in maximum application of Wireless Sensor Network. The scheme is highly suitable for application that needs extensive resource on dynamic traffic condition. Our future work will be towards the direction of further optimizing the cross-layer based approach to ensure more improvement towards Quality-of-Service factors. A mathematical modeling-based method will be adopted with further improvement of interactivity between network and MAC layer. The resultant optimized cross-layer protocol can be used for specific IOT applications.

References

1. Ray, N.K., Turuk, A.K.: Handbook of Research on Advanced Wireless Sensor Network Applications, Protocols, and Architectures, IGI Global (2016)
2. Ranjan, R., Varma, S.: Challenges and implementation on cross layer design for wireless sensor networks. J. Wirel. Pers. Commun. **86**(2), 1037–1060 (2016). Springer
3. Dobslaw, F., Zhang, T., Gidlund, M.: QoS-aware cross-layer configuration for industrial wireless sensor networks. IEEE Trans. Ind. Inf. **12**(5), 1679–1691 (2016). doi:10.1109/TII.2016.2576964
4. Abd El Kader, M.E.E.D., Youssif, A.A.A., Ghalwash, A.Z.: Energy aware and adaptive cross-layer scheme for video transmission over wireless sensor networks. IEEE Sens. J. **16**(21), 7792–7802 (2016). doi:10.1109/JSEN.2016.2601258
5. Su, X., Chan, S., Bandai, M.: A cross-layer MAC protocol for underwater acoustic sensor networks. IEEE Sens. J. **16**(11), 4083–4091 (2016). doi:10.1109/JSEN.2015.2440301
6. Shoba, M., Meghana, D.S.: Enhancing network lifetime and energy efficiency using DEL-CMAC protocol for MANET's. Int. J. Inf. Technol. Comput. Eng. **3**(5) (2016). e-ISSN: 2455-5290
7. Khatri, U., Mahajan, S.: Cross-layer design for wireless sensor networks: a survey. In: IEEE International Conference on Computing for Sustainable Global Development, pp. 73–77 (2015)
8. Das, S.N., Misra, S.: Correlation-aware cross-layer design for network management of wireless sensor networks. IEEE IET Wirel. Sens. Syst. **5**(6), 263–270 (2015). doi:10.1049/iet-wss.2014.0110
9. Gama, S., Walingo, T., Takawira, F.: Energy analysis for the distributed receiver-based cooperative medium access control for wireless sensor networks. IET Wirel. Sens. Syst. **5**(4), 193–203 (2015). doi:10.1049/iet-wss.2013.0129

10. Kartsakli, E., et al.: Reliable MAC design for ambient assisted living: moving the coordination to the cloud. IEEE Commun. Mag. **53**(1), 78–86 (2015). doi:10.1109/MCOM.2015.7010519
11. Lee, S., Kwon, B., Lee, S., Bovik, A.C.: BUCKET: scheduling of solar-powered sensor networks via cross-layer optimization. IEEE Sens. J. **15**(3), 1489–1503 (2015). doi:10.1109/JSEN.2014.2363900
12. Lin, S.C., Akyildiz, I.F., Wang, P., Sun, Z.: Distributed cross-layer protocol design for magnetic induction communication in wireless underground sensor networks. IEEE Trans. Wirel. Commun. **14**(7), 4006–4019 (2015). doi:10.1109/TWC.2015.2415812
13. Usman, M., Muthukkumarasamy, V., Wu, X.W.: Mobile agent-based cross-layer anomaly detection in smart home sensor networks using fuzzy logic. IEEE Trans. Consum. Electron. **61**(2), 197–205 (2015). doi:10.1109/TCE.2015.7150594
14. Xu, W., Zhang, Y., Shi, Q., Wang, X.: Energy management and cross layer optimization for wireless sensor network powered by heterogeneous energy sources. IEEE Trans. Wirel. Commun. **14**(5), 2814–2826 (2015). doi:10.1109/TWC.2015.2394799
15. Shoba. M., Suresha: S-AODV: an adaptive method for improving AODV protocol for WSN. In: IEEE International Conference on Green Computing and Internet of Things, Noida, pp. 1016–1021 (2015)
16. Obaidat, M.S., Misra, S.: Principles of Wireless Sensor Networks. Cambridge University Press, Cambridge (2014)
17. Yetgin, H., Cheung, K.T.K., El-Hajjar, M., Hanzo, L.: Cross-layer network lifetime optimisation considering transmit and signal processing power in wireless sensor networks. IET Wirel. Sens. Syst. **4**(4), 176–182 (2014). doi:10.1049/iet-wss.2014.0049
18. Ragab, K., Abdullah, A.B.: Wireless sensor networks and energy efficiency: protocols, routing and management. Information Science Reference (2011)
19. Rashvand, H.F., Kavian, Y.S.: Using cross-layer techniques for communication systems, IGI Global (2012). doi:10.4018/978-1-4666-0960-0
20. Ju, P., Song, W., Zhou, D.: Survey on cooperative medium access control protocols. IEEE IET Commun. **7**(9), 893–902 (2013). doi:10.1049/iet-com.2012.0739
21. Liu, Y., Xie, S., Yu, R., Zhang, Y., Yuen, C.: An efficient MAC protocol with selective grouping and cooperative sensing in cognitive radio networks. IEEE Trans. Veh. Technol. **62**(8), 3928–3941 (2013). doi:10.1109/TVT.2013.2258952
22. Shah, G.A., Gungor, V.C., Akan, O.B.: A cross-layer QoS-aware communication framework in cognitive radio sensor networks for smart grid applications. IEEE Trans. Ind. Inf. **9**(3), 1477–1485 (2013). doi:10.1109/TII.2013.2242083
23. Ye, W., Heidemann, J., Estrin, D.: An energy-efficient MAC protocol for wireless sensor networks. In: Proceedings of the 21st International Annual Joint Conference of the IEEE Computer and Communications Societies (2002)

Robust Constrained Control: Optimization of 1 vs. 2 Closed-Loop Poles

Frantisek Gazdos(✉)

Faculty of Applied Informatics Nam, Tomas Bata University in Zlin,
Nam. T.G. Masaryka 5555, 760 01 Zlin, Czech Republic
gazdos@fai.utb.cz

Abstract. This paper presents optimization-based technique to design robust control system in case of control input limitations. The methodology uses the algebraic approach resulting in polynomial equations and a pole-placement problem to be solved. Closed-loop poles are optimized numerically with the help of the MATLAB computing system and its toolboxes for simulation and optimization. Suitable performance criteria and a procedure are suggested for this purpose. The case of 1 and 2 parameters optimization is illustrated on a nonlinear servo-system control design using both simulation and real-time experiments. Presented results prove the proposed methodology.

Keywords: Constrained control · Polynomial approach · Pole-placement problem · Robust control · Optimization · AMIRA servo-system

1 Introduction

When dealing with practical implementation of control strategies an engineer always has to face the problem of process variables limitations. The most crucial are the constraints on the control input signal – the controller's manipulated variable which is used to obtain required course of the controlled variable. This signal is always represented by a certain physical quantity, such as a flow rate, electric current or voltage etc. which obviously have some limitations. Besides amplitude limits of the control inputs there are very often limitations on the achievable speed of changes of the variables due to the used actuators, e.g. valves. These facts have to be carefully considered in the control system design procedure and simulation testing. Not respecting these limits can lead to serious consequences, especially when dealing with hardly controllable processes, e.g. unstable, with significant time-delay or with an inverse response [1]. In the literature there is a great number of classic methods dealing with this problem, often called anti-wind-up techniques applicable mainly to popular PI and PID controllers, e.g. [2, 3]. Among other modern control approaches the predictive control concept is also effective and popular in this field nowadays, e.g. [4, 5], although it is more computationally demanding.

Although there are many sources devoted to the robust control systems design and to the constrained control separately, simultaneous solutions of both these problems are still not so common; some representative solutions can be found in e.g. [6–10]. This

R. Silhavy et al. (eds.), *Cybernetics and Mathematics Applications in Intelligent Systems*,
Advances in Intelligent Systems and Computing 574, DOI 10.1007/978-3-319-57264-2_25

paper represents a contribution to this interesting and practically important topic. The methodology suggested in this paper is based on the usage of simulation and computing tools – the MATLAB environment and the systematic algebraic control approach transforming the control system design problem to the solution of polynomial equations, e.g. [11–13]. After formulation of basic control requirements the polynomial approach enables to find both suitable structure and parameters of controllers. Generally it can lead to more complicated structures of the resultant controllers than the classical PI or PIDs but this does not seem a serious problem nowadays when most of industrial controllers are implemented using PLCs. A natural part of the procedure for finding a suitable controller using the polynomial approach is the pole-placement problem solution, e.g. [14]. In this contribution this task is solved numerically using the standard MATLAB functions for nonlinear constrained optimization. The resultant poles (free tuning parameters) of the control loop are optimized with respect to both robustness and constraints on the controller's manipulated variable. For this purpose suitable control quality criteria and a corresponding procedure are suggested. The whole methodology is illustrated clearly on a representative example of control system design for the AMIRA DR300 servo-system. Different control approaches for this system can be found in e.g. works [15, 16].

The presented contribution is structured as follows: after this introductory section the paper starts by recalling basics from the polynomial control theory utilized in this work. Next part introduces the control quality criteria for subsequent optimization which is described in detail in the section later. Further parts present the illustrative example including brief description of the AMIRA DR300 servo-system, analyse the achieved results and suggest possible areas for future works. This paper extends previous authors' works [17–19] so that compares the results of optimization in the case of one versus two closed-loop poles (tuning parameters).

2 Theoretical Basics

This section recalls basics of the adopted polynomial approach for the usual control set-up of Fig. 1 and prepares the space for the methodology respecting both control input limitations and robustness of the resultant loop.

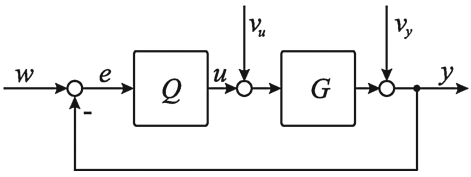


Fig. 1. Control set-up

Suppose the common control set-up of Fig. 1, where G denotes a plant to be controlled by a feedback controller Q utilizing information about the process controlled variable y and the reference (set-point) signal w via the control error e and generating

corresponding control input (manipulated variable) u. Signals v_u and v_y represent general disturbances.

Further assume that the plant and the controller can be approximated by transfer functions $G(s)$, $Q(s)$ with coprime polynomials $\{b(s), a(s)\}$ according to (1) while satisfying $\deg a(s) > \deg b(s)$ and $\deg p(s) \geq \deg q(s)$ (the argument s is the complex variable of the Laplace transform).

$$G(s) = \frac{b(s)}{a(s)}, Q(s) = \frac{q(s)}{p(s)} \tag{1}$$

Basic requirements for the control system introduced above are formulated in a common way as follows: *stability*, *asymptotic tracking of the reference signal*, *disturbances attenuation* and *inner properness*. Besides the above-mentioned general requirements, the control system should also be robust to cope with the real nonlinear plant (not only with an adopted simplified linear model) and possible disturbances. In addition the controller has to respect given physical limitations of the manipulated variable. All these tasks are discussed and solved further in this work.

It is straightforward that the stability condition will be fulfilled if the controller is given by a solution of the following polynomial equation with a stable polynomial $d(s)$ on the right side, e.g. [11–14]:

$$ap + bq = d. \tag{2}$$

This polynomial is called characteristic as it defines important properties, such as stability and periodic/aperiodic behavior. The Eq. (2), after a proper choice of the stable polynomial $d(s)$, is used to compute the unknown controller polynomials $q(s)$ and $p(s)$. Roots of the characteristic polynomial $d(s)$ are known as poles of the closed loop. Their proper placement influences not only stability of the loop but also the achieved control quality, i.e. a settling-time, overshoots, control input course etc. Therefore the so-called pole-placement problem is a natural part of the polynomial approach to control system design. In this work the poles are optimized numerically using the standard means of the MATLAB system to respect both limitations on the control input and robustness of the resultant loop.

Further assume that the reference and disturbances can be approximated by step-functions. Then it is also easy to show that to guarantee zero-control error in the steady state, the denominator controller polynomial $p(s)$ needs to be divisible by the s term, i.e. the controller has to include an integrator, which will be fulfilled for $p(s) = s\tilde{p}(s)$. Then the feedback controller $Q(s)$ in (1) will be

$$Q(s) = \frac{q(s)}{s\tilde{p}(s)}, \tag{3}$$

and the polynomial Eq. (2) defining stability reads: $as\tilde{p} + bq = d$.

The inner properness of the control system is satisfied if all its transfer functions are proper. With regard to the strictly proper plant transfer function and proper controller

(1), and taking into account solvability of (2), it is possible to derive the following formulas for degrees of the unknown polynomials q, $\tilde{p}$ and d:

$$\deg q(s) = \deg\ a(s), \quad \deg \tilde{p}(s) \geq \deg\ a(s) - 1, \quad \deg d(s) \geq 2\deg a(s). \tag{4}$$

For practical computation of the controller's polynomials $q(s)$ and $\tilde{p}(s)$ it is necessary to choose a suitable stable polynomial $d(s)$ appearing on the right side of the polynomial Eq. (2). This is the so called pole-placement problem mentioned earlier, e.g. [14]. Therefore we are seeking suitable poles p_i of the designed loop to fulfill given requirements. Hence $d(s)$ can be expressed as (5) for some poles (its roots) p_i. Then the control design procedure transforms to the optimization problem of finding the right poles providing the required control quality.

$$d(s) = \prod_{i=1}^{\deg\ d} \left(s - p_i\right) \tag{5}$$

In this work, it is suggested to choose the characteristic polynomial as:

$$d(s) = \prod_{i=1}^{\deg\ d} \left(s + \alpha_i\right) \tag{6}$$

for some reals $\alpha_i > 0$. This ensures stability of the loop (all poles are negative at positions $p_i = -\alpha_i$) as well as aperiodic behavior. Now the optimization task is to find optimal values of the parameters $\alpha_i > 0$, which is addressed in the next section.

3 Methodology

This part describes the used procedure for optimization of the closed-loop poles to meet the required control quality – loop robustness and limitation on the control input.

3.1 Control Quality Criteria

In this work the control quality is measured by two basic sub-criteria: one for assessing robustness of the designed loop – denoted as J_{rob} and the other for evaluating demands on the control input – indicated as J_u. As far as the loop robustness is concerned, a peak gain of the sensitivity function frequency response given by the infinity norm H_∞ is a good measure for this purpose [20]. Therefore to assess the robustness of the designed loop it is suggested to use the sensitivity function S and its infinity norm H_∞:

$$S = \frac{1}{1 + GQ} = \frac{ap}{ap + bq} = \frac{ap}{d}, \qquad J_{rob} = \|S\|_\infty = \sup_\omega |S(j\omega)|, \tag{7}$$

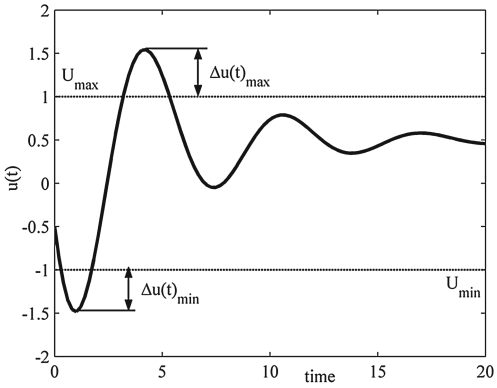


Fig. 2. Explanation of the sub-criterion J_u

where ω is the frequency. The 2nd sub-criterion J_u describing demands on the control input $u(t)$ is formed as follows. Let us define the achievable limits of $u(t)$ as U_{min} and U_{max}. Then, the control input has to stay within the following interval at all times: $u(t) \in < U_{min}; U_{max} > \forall t$. Further denote $\Delta u(t)_{\max}$ as the maximum overshoot of $u(t)$ above the given limit U_{max} and correspondingly $\Delta u(t)_{\min}$ the maximum undershoot under U_{min}. Then the sub-criterion J_u is computed simply as:

$$J_u = \Delta u(t)_{\max} + \Delta u(t)_{\min} \tag{8}$$

It is evident that the sub-criterion is equal to zero if the control input is within the desired limits and it is positive with higher values for $u(t)$ out of the required range. The situation is well illustrated in Fig. 2 with limits chosen as $U_{min} = -1$ and $U_{max} = 1$.

Having defined the criteria it is possible to formulate the optimization problem as:

$$\min_{\alpha} \; J_{rob}(\alpha) \text{ such that } \alpha_i \succ 0 \text{ and } J_u(\alpha) = 0, \tag{9}$$

where α is the vector of optimized parameters α_i. In other words, the goal is to find such stable aperiodic poles of the closed loop that provide most robust behavior and respect the given limits of the controller manipulated variable. The optimization procedure, described in the next section, is performed with the help of standard functions for optimization from the MATLAB computing system and its toolboxes.

3.2 Optimization Procedure

First, the closed-loop variables are scaled so that both the controlled output and also the reference are within the range from zero to one, i.e. $y(t), w(t) \in < 0;1 >$. Then the worst-case behaviour of the control system (regarding the changes in the reference) can be analysed by considering the reference change of magnitude one. Therefore the designed control system is analysed (simulated) facing this condition and the control quality

criteria J_{rob} and J_u from (9) are assessed for different values of the tuning parameters $\alpha_i > 0$. This is done with the help of MATLAB environment and its toolboxes for simulation and optimization. The procedure can be briefly described as:

- choose number of optimized parameters α_i;
- for every α_i choose an interval for optimization;
- find a solution of the problem specified in (9), i.e. find a minimum of the sub-criterion J_{rob} subject to the conditions such that $J_u = 0$ on the given region of α_i;
- collect the resultant parameters α_i and verify if they fulfill the given requirements.

If the algorithm for solution of the problem fails then there may not be such combination of α_i (under given conditions) to respect the limits of the control input $u(t)$. Then the designer has several basic possibilities: try to increase the number of optimized parameters α_i (if possible), use different control structures (e.g. with a pre-filter of the reference signal), or has finally no other way than enlarging the prescribed limits of the manipulated variable $u(t)$. The optimization algorithm uses a standard MATLAB function for nonlinear constrained minimization *fmincon*. It is a gradient-based method – the trust-region-reflective algorithm based on the interior-reflective Newton method, described in detail in, e.g. [21].

4 Case Study

The algorithm introduced in the previous sections is illustrated on the problem of designing a control system for the servo-system DR300 described further.

4.1 Servo AMIRA DR300

The DR300 servomechanism is a product of the AMIRA company, Duisburg, Germany. It consists of two identical motors connected by a mechanical clutch. The first one is used for control of the rotation speed or the shaft angle while the other (generator) can be used to simulate of load torque. The system is connected to a common PC by the Humusoft multifunction I/O card and controlled in the MATLAB environment using the Real Time Toolbox. The basic control task is to regulate the angular speed of the 1st motor despite possible variable load generated by the 2nd motor. The system is generally nonlinear with the static properties recorded in Fig. 3 (left) where the controlled variable (angular velocity) y is in revolutions per minute while the manipulated variable (control voltage) u is scaled to $\pm$ 1 machine unit.

As can be seen it has a dead zone and small hysteresis. Therefore the control is performed only in the most linear part aside from the saturation limits and the dead zone. A simplified mathematical model of this nonlinear system can be derived in the state-space form (10), e.g. [16], with $i(t)$ the motor current in [A], $\omega(t)$ the angular speed in [rad.s^{-1}], $u(t)$ the input voltage in [V] and $m_z(t)$ the load torque in [N.m].

$$\frac{d\,i(t)}{dt} = -\frac{R}{L}i(t) - \frac{k_e}{L}\omega(t) + \frac{1}{L}u(t), \qquad \frac{d\,\omega(t)}{dt} = -\frac{k_m}{J}i(t) - \frac{b}{J}\omega(t) - \frac{1}{J}m_z(t) \tag{10}$$

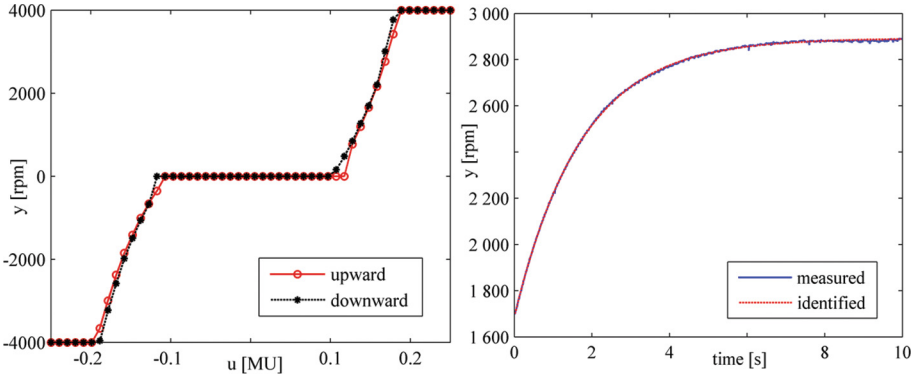


Fig. 3. Static characteristic (left) and measured/identified step-responses (right)

Constants R, L, k_e, k_m, b and J are parameters of the motor. The input voltage $u(t)$ is considered as the manipulated variable for the controlled variable – angular speed ω (t). The load torque $m_z(t)$ can be considered as a disturbance. An input-output model, where $U(s)$, Ω(s) are the Laplace transforms of the variables can be obtained as:

$$G(s) = \frac{\Omega(s)}{U(s)} = \frac{\frac{k_m}{LJ}}{s^2 + \left(\frac{R}{L} + \frac{b}{J}\right)s + \frac{Rb + k_e k_m}{LJ}}, \tag{11}$$

which shows that the model can be (in the linear region) identified as a 2nd order proportional system. Experimental identification provides the approximation (12), which shows that the system is stable aperiodic, relatively fast with time-constants $T_1 = 1.72$ $[s]$, $T_2 = 0.02$ $[s]$ and gain $k = 59766$ $[rpm/MU]$. The comparison of measured/identified step-responses in Fig. 3 (right) shows that the model approximates the system in the given region well.

$$G(s) = \frac{b(s)}{a(s)} = \frac{59766}{(1.72s + 1)(0.02s + 1)} = \frac{2073600}{(s + 0.58)(s + 59.52)} \tag{12}$$

Scaling of this model (so that the available range of $y(t)$ from 0 to 4000 $[rpm]$, and also of $u(t)$ from 0.1 to 0.2 $[MU]$, is transformed to 0–1) provides the following model:

$$G(s) = \frac{1.49}{(1.72s + 1)(0.02s + 1)} = \frac{51.84}{(s + 0.58)(s + 59.52)} = \frac{51.84}{s^2 + 60.10s + 34.70}, \tag{13}$$

which is further used for the controller design in the next section.

4.2 Control System Design and Experiments

Following the polynomial approach described briefly in the Sect. 2 a suitable feedback controller for this system is designed in the following general form:

$$Q(s) = \frac{q(s)}{p(s)} = \frac{q(s)}{s\tilde{p}(s)} = \frac{q_2 s^2 + q_1 s + q_0}{s\left(\tilde{p}_1 s + \tilde{p}_0\right)}. \tag{14}$$

Unknown coefficients of the controller are obtained by the solution of the polynomial Eq. (2) for a given stable polynomial $d(s)$. This polynomial, in the general form (5), (6), must be according to (4) of the 4th degree. First, let us try if its simplest form (15) with only one tuning parameter $\alpha > 0$ (consequently there are 4 identical closed-loop poles located at $p_{1,2,3,4} = -\alpha$) will enable to find a suitable robust controller respecting the given control input limits.

$$d(s) = (s + \alpha)^4 \tag{15}$$

The free tuning parameter α is optimized numerically via the procedure suggested in the Sect. 3 to respect both robustness of the loop and limits of the control input signal $u(t)$ which were in this case defined as:

$$u(t) \in \langle 0.1;0.2\rangle\ [MU] \quad \forall t, \text{ after scaling:} u(t) \in \langle 0;1\rangle\ [-] \quad \forall t \tag{16}$$

In this simple case of only 1 tuning parameter it is possible to obtain easily course of the sub-criterion J_u (8) assessing the control input signal with respect to the given limitations (16), depending on the parameter α, see Fig. 4 - left.

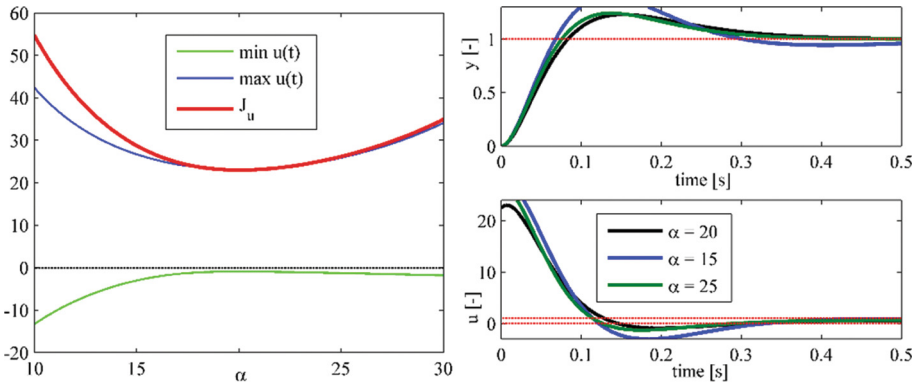


Fig. 4. Sub-criterion J_u with α (left) and worst-case simulation responses (right)

From the left graph, where the solid red curve represents the criterion assessing the control input J_u, it is evident that there is no such value of α for which the criterion is equal zero. In other words, for this simplest choice of the characteristic polynomial d with only one tuning parameter (15) there is no such value of α that provides the control

input in the required range (16). The lowest value of the criterion is Ju ≈ 23.0 for approximately α ≈ 20. Presented simulation responses of Fig. 4 (right) confirm this result and present also small variations of the tuning parameter (± 25%).

As suggested in the Sect. 3.2, due to the negative results, the number of optimized poles is enlarged so that the characteristic polynomial has now the form:

$$d(s) = \left(s + \alpha_1\right)^2 \left(s + \alpha_2\right)^2, \tag{17}$$

with 2 tuning parameters $\alpha_1, \alpha_2 > 0$ (consequently there are two double closed-loop poles located at $p_{1,2} = -\alpha_1$, $p_{3,4} = -\alpha_2$). The suggested optimization procedure stays the same. Complex MATLAB computations provided the graphs of Fig. 5 where the control input criterion J_u depends on α_1, α_2.

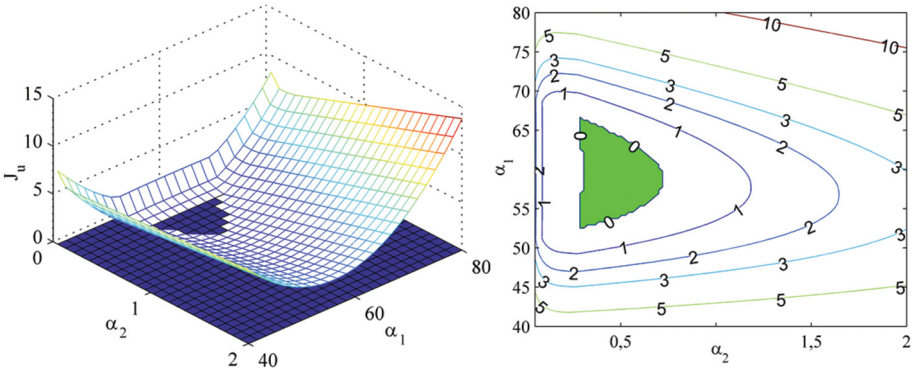


Fig. 5. Sub-criterion J_u with α_1, α_2 and its contour plot

As can be seen, there exists a region where $J_u = 0$. In other words there is a combination of parameters $\alpha_1, \alpha_2 > 0$ for which the control input stays in the prescribed limits. The detailed contour plot of the criterion shows the admissible area more clearly (green-filled area). Now the task is to find the most robust controller in this region, i.e. the minimum of the J_{rob} criterion (7) assessing the loop robustness. Further MATLAB computations provided Fig. 6 where the robustness criterion depends on the optimized parameters, with the green-filled admissible interval for $J_u = 0$. In the detailed contour plot the criterion in the admissible region for $J_u = 0$ (green area) does not change much and it is relatively small with is minimal values around $J_{rob} < 1.01$. Therefore it is suggested to pick the center of this area as the "optimal" setting to enable both – safe constrained control and loop robustness as well.

Consequently, the best suggested setting of the 2 tuning parameters under given conditions is for $\alpha_1 = 59$, $\alpha_2 = 0.5$. Closed-loop simulation responses for this setting are provided in Fig. 7 (left graph), together with their small variations ± 25%.

From the recorded simulation responses above it is clear that the suggested setting provides the control input in the desired range while both the variations ± 25% give much bigger control action (and finally converge asymptotically to the same value around time t = 10 [s] as the simulation model is linear). Real-time measurements on

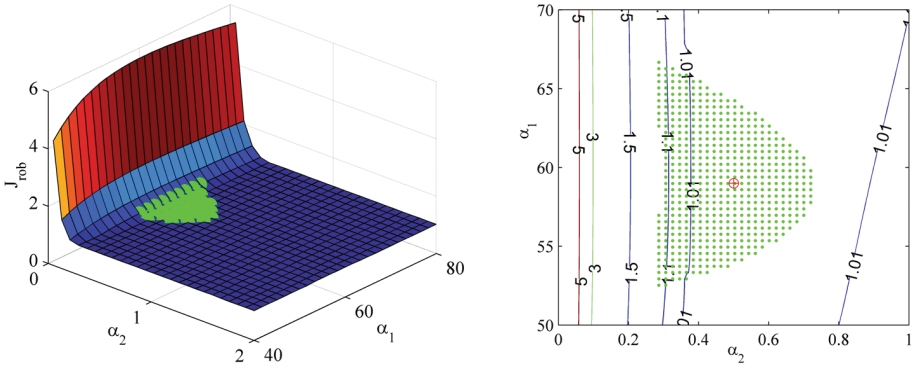


Fig. 6. Sub-criterion J_{rob} with α_1, α_2 and its contour plot

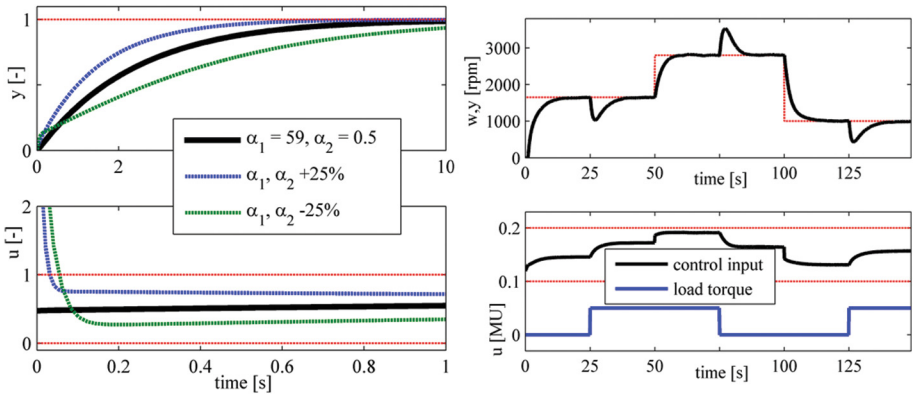


Fig. 7. Worst-case simulation responses (left) and real-time control for suggested setting (right)

the AMIRA DR300 servo-system are recorded in the right part of the figure, where both the reference tracking in different operating points and disturbance attenuation (represented by a variable load) are presented. From the recorded data it is obvious that the control system is stable, without overshoots and with good tracking not only in the identified operating point but also in different regions which shows robustness of the design to model mismatch. Disturbances represented by a variable load are also compensated well, and what is important the control input signal is within the required limits at all times.

4.3 Discussion of the Results

From the presented information and experiments it is possible to summarize:

- as generally expected, the simplest choice of the closed-loop characteristic polynomial with only 1 multiple pole limits achievable performance so that the required range of $u(t)$ or desired robustness cannot be fulfilled under the given set-up;

- more tuning parameters and consequently more different closed-loop poles provide better performance for both measures – control input demands and robustness (the robustness criterion in case of 1 tuning parameter is around 1.3 for "best" setting while for 2 tuning parameters is below 1.01 as presented in Fig. 6);
- it is not a coincidence that suggested "optimal setting" of tuning parameters (for $\alpha_1 = 59, \alpha_2 = 0.5$, resulting in closed-loop poles $p_{1,2} = -59, p_{3,4} = -0.5$) is very close to the poles of the identified model ($p_1 = -59.52, p_2 = -0.58$), i.e. when we prescribe closed-loop behaviour close to the dynamics of the original system, the system is more robust and also the control input is in a reasonable range;
- it is also expected that other control set-ups, e.g. with a pre-filter of the reference, can improve the results;
- optimization of all the closed-loop poles may also bring improvements, however, with limited possibilities for graphical interpretation and inspection.

5 Conclusions

This contribution presented one possible approach to robust constrained control – the approach is based on the direct optimization of closed-loop poles to meet both robustness and limitations on the control input. For this purpose suitable criteria have been suggested together with the optimization procedure which fruitfully exploits the capabilities of the MATLAB computing system. The presented results of constrained robust control of the AMIRA DR300 servo-system show applicability of the methodology under real conditions. Further works will compare different control configurations and investigate the achievable results for optimization of all the close-loop poles with different optimization methods.

Acknowledgments. This work, as a part of the project "Development and Applications of Advanced Process Control Methods", was supported by the excellence projects strategy of Tomas Bata University in Zlin. This support is greatly acknowledged.

References

1. Stein, G.: Respect the unstable. IEEE Control Syst. Mag. **23**(4), 12–25 (2003)
2. Saberi, A., Stoorvogel, A.A., Sannuti, P.: Control of Linear Systems with Regulation and Input Constraints. Springer, London (2000)
3. Glattfelder, A.H., Schaufelberger, W.: Control Systems with Input and Output Constraints. Springer, London (2003)
4. Camacho, E.F., Bordons, C.: Model Predictive Control. Springer, London (2004)
5. De Doná, J.A., Goodwin, G.C., Seron, M.M.: Anti-windup and model predictive control: reflections and connections. Eur. J. Control **6**(5), 467–477 (2000)
6. Campo, P.J., Morari, M.: Robust control of processes subject to saturation nonlinearities. Comput. Chem. Eng. **14**(4–5), 343–358 (1990)
7. Miyamoto, S., Vinnicombe, G.: Robust control of plants with saturation nonlinearity based on coprime factor representations. In: 35th IEEE Conference on Decision and Control, Japan, pp. 2838–2840 (1996)

8. Huba, M.: Robust constrained PID control. In: International Conference Cybernetics and Informatics, Vyšná boca, Slovak Republic, pp. 1–18 (2010)
9. Vozák, D., Veselý, V.: Stable predictive control with input constraints based on variable gain approach. Int. Rev. Autom. Control **7**(2), 131–139 (2014)
10. Torchani, B., Sellami, A., Garcia, G.: Robust sliding mode control of class of linear uncertain saturated systems. Int. Rev. Autom. Control **6**(2), 134–146 (2013)
11. Kučera, V.: Diophantine equations in control – a survey. Automatica **29**, 1361–1375 (1993)
12. Hunt, K.J.: Polynomial Methods in Optimal Control and Filtering. Peter Peregrinus Ltd., London (1993)
13. Anderson, B.D.O.: From Youla-Kucera to identification, adaptive and nonlinear control. Automatica **34**, 1485–1506 (1998)
14. Kučera, V.: The pole placement equation. A survey. Kybernetika **30**(6), 578–584 (1994)
15. Bobál, V., Kubalčík, M., Chalupa, P., Dostál, P.: Self-tuning control of nonlinear servo system: comparison of LQ and predictive approach. In: 17th Mediterranaen Conference on Control and Automation, Thessaloniki, Greece, pp. 240–245 (2009)
16. Roubal, J., Augusta, P., Havlena, V.: A brief introduction to control design demonstrated on laboratory model servo DR300 – AMIRA. Acta Electrotechnica et Informatica **5**(4), 1–6 (2005)
17. Gazdoš, F., Marholt, J.: Simulation approach to robust constrained control. Int. Rev. Autom. Control **7**(5), 578–584 (2014)
18. Gazdoš, F., Marholt, J.: Optimization of closed-loop poles for robust constrained control. In: 20th International Conference on Process Control, PC 2015, Strbske Pleso, Slovakia, pp. 158–163 (2015)
19. Gazdoš, F.: Optimization of closed-loop poles for limited control action and robustness. In: 2nd International Afro-European Conference for Industrial Advancement, AECIA 2015, Paris, France, pp. 385–396 (2015)
20. Skogestad, S., Postlethwaite, I.: Multivariable Feedback Control: Analysis and Design. Wiley, Chichester (2005)
21. Coleman, T.F., Li, Y.: An interior, trust region approach for nonlinear minimization subject to bounds. SIAM J. Optim. **6**, 418–445 (1996)

Machine Learning Approaches to Electricity Consumption Forecasting in Automated Metering Infrastructure (AMI) Systems: An Empirical Study

A. Jayanth Balaji(✉), D.S. Harish Ram, and Binoy B. Nair

Department of Electronics and Communication Engineering,
Amrita School of Engineering, Coimbatore Amrita Vishwa Vidyapeetham,
Amrita University, Coimbatore 641112, India
{a_jayanthbalaji,ds_harishram,b_binoy}@cb.amrita.edu

Abstract. In a Smart grid, implementation of value-added services such as distribution automation (DA) and Demand Response (DR) [1] rely heavily on the availability of accurate electricity consumption forecasts. Machine learning based forecasting systems, due to their ability to handle nonlinear patterns, appear promising for the purpose. An empirical evaluation of eight machine learning based systems for electricity consumption forecasting, based on Extreme Learning machines (ELM), Ensemble Regression Trees (ERT), Artificial Neural Network (ANNs) and regression is presented in this study. Forecasting systems thus designed, are validated on consumption data collected from 5275 users. Result indicate that ELM based electricity consumption forecasting systems are not only more accurate than other systems considered, they are considerably faster as well.

Keywords: Machine learning · Extreme Learning machines (ELM) · Ensemble Regression Trees (ERT) · Artificial Neural Network (ANNs) · Hodrick-Prescott (HP)

1 Introduction

Automated Metering Infrastructure (AMI) is an integral part of a smart grid. AMI not only allows the consumers facilities such as real time consumption monitoring and automated billing etc., it also allows the power utilities to implement value-added services such as DA and DR. A typical AMI system is represented in Fig. 1 [2]. One of the major aspects of AMI is its ability to learn customer electricity consumption behavior and accordingly optimize the quality of smart grid services delivered to the user. A detailed analysis of the issues related to load forecasting and DR has been presented in [3,4]. AMI applications to scheduling

The CER Electricity Dataset is sourced from Irish Social Science Data Archive (ISSDA).

R. Silhavy et al. (eds.), *Cybernetics and Mathematics Applications in Intelligent Systems*, Advances in Intelligent Systems and Computing 574, DOI 10.1007/978-3-319-57264-2_26

of residential microgrids [5], weather based load normalization and increasing the forecast accuracy for hourly forecast consumption is done in [6] and the user segmentation based on users power consumption profile forecasting for hourly power consumption data and thereby maximizing the benefits of smart grid services especially on DR is done in [7].

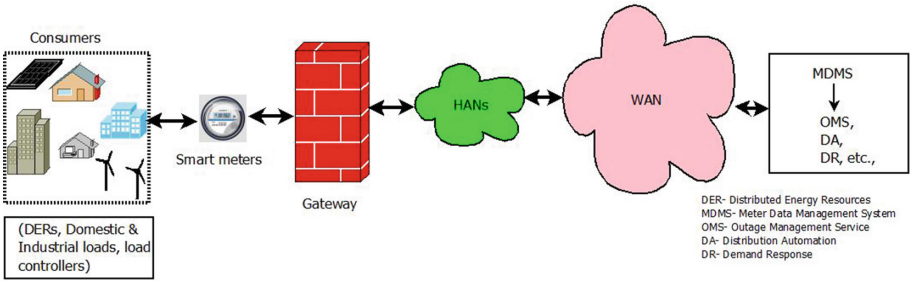


Fig. 1. AMI block diagram [2]

However, it is observed that there are few studies on identification of suitable techniques for electricity consumption forecasting. It is also observed that the consumption patterns of users change according to the time of the day and according to the seasonal variations. This aspect has also not received much attention. Consumption behavior is also seen to be very different for domestic consumers when compared to industrial users. The earlier work by the authors on the topic [2] attempted to empirically address these aspects of consumption forecasting by analyzing forecast performance of machine learning based systems for 485 Small and Medium Enterprise consumers. This paper presents a much more comprehensive study with a total of 5275 users and four different forecasting techniques.

Rest of the paper is organised as follows: Sect. 2 presents the description of the proposed system and the methodology, Sect. 3 details the experimental results and conclusions are presented in Sect. 4.

2 System Description and Methodology

The block diagram of the proposed system is presented in Fig. 2.

2.1 Dataset and Preprocessing

All the models were developed for CER electricity dataset acquired from Irish Social Science Data Archive (ISSDA) [8] comprising half hourly interval electricity consumption data for 6445 users, subcategorized as (a) Small and Medium Enterprise (SME) consisting of 485 users (b) Residential (RES) consisting of

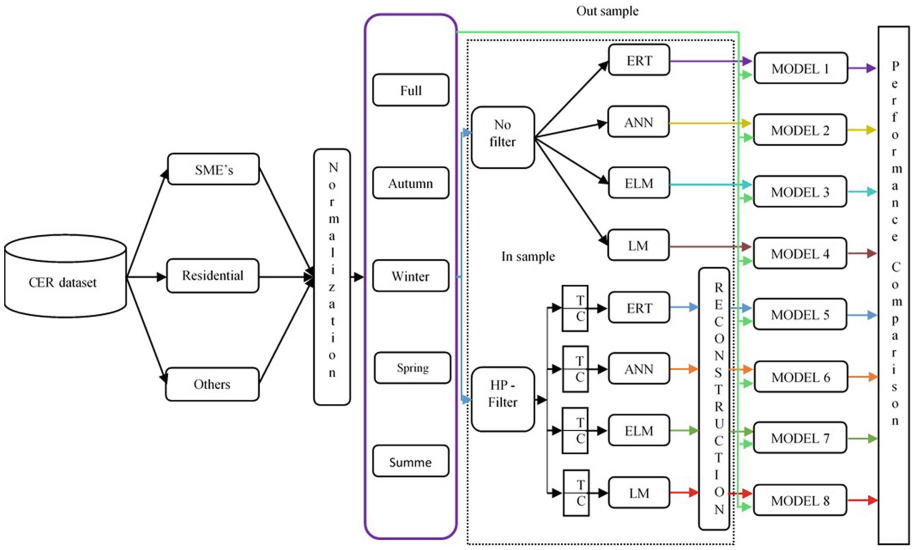


Fig. 2. Proposed system block diagram

4225 users and (c) 'Others' (OTH) consisting of 1735 users. The data has been recorded over the period 00:00 h, 14/07/2009 to 23:59 h, 29/12/2010. It was observed on inspection that in the OTH category, the consumption data for 1170 users is not available for the entire duration listed above and was not considered for further analysis. The total number of users considered, then, is 5275.

As the first preprocessing step, the consumption data for each user is split into four subsets: summer (May to July), winter (November to January), autumn (August to October) and spring (February to April). Different forecasting models are trained for each season separately. For comparison purposes, the models were also trained on the data without seasonal splits.

As the next step, after normalization of the data, the trend and cyclic components of the data are separated using Hodrick-Prescott (HP) filter [9]. HP filter has been successfully used in [10–13] as a preprocessing technique for financial forecasting systems. It has also been reported to offer an enhancement in performance for electricity consumption forecasting systems in [2]. Machine learning based forecasting systems described later in the paper are separately trained to forecast trend and cyclic components. Finally both the forecasts are added to yield the estimated consumption for that instant. Systems that forecast the consumption based directly on the normalized data alone (i.e. not employing the HP filter) are also evaluated.

Next step in the process is to train a machine learning system to estimate the electricity consumption. Three different machine learning techniques as well as linear regression (LR) models are investigated for the purpose. Each of these techniques is described in the next sub-section.

2.2 Machine Learning Techniques Considered

A total of four techniques, namely, ELM, ANN, ERT and LR were evaluated in the present study. Hence a total of eight different forecasting models were evaluated (see Fig. 2). Each of the techniques used and the system parameters are described below.

2.2.1 Extreme Learning Machines (ELMs)

ELMs are basically single hidden layer feedforward neural network (SLFNN), however in ELMs the weights connecting input and hidden layer are randomly fixed and not updated. The hidden to output layer weights are learnt in a single step with the help of Pseudoinverse technique which makes the architecture much faster and generalized when compared to any other neural network, with minimal training errors and smaller norm of weights [14,15]. Kernel based ELMs have also been proposed [16]. The results of employing kernel ELMs have been encouraging, as seen in [16,17]. A detailed treatment on kernel ELMs can be found in [15]. In the proposed work, the ELM Models are designed using linear kernel with regularization coefficient as '10'. The kernel function and the regularization coefficient value were arrived at, using trial-and-error. The estimated

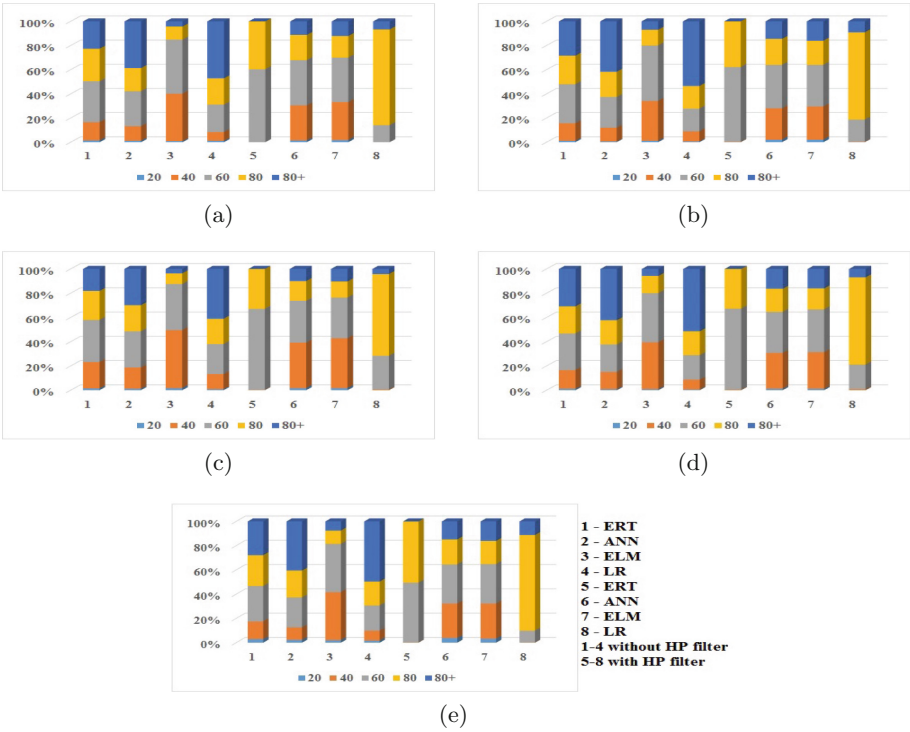


Fig. 3. MdAPE plots for all models with RES category users (a) all seasons together (b) Autumn (c) Winter (d) Spring and (e) Summer.

output from the ELM, $\widehat{Y}$ can be expressed as follows:

$$\widehat{Y} = W_{H-O}\sigma_A(W_{I-H}X) \tag{1}$$

where: W_{I-H} = input - hidden layer weight vector, W_{H-O} = hidden - output layer weight vector and σ_A = activation function.

Algorithm:

Step 1: Initialize W_{I-H} with random weights.

Step 2: Compute W_{H-O} using pseudoinverse:

$$W_{H-O} = \sigma_A(W_{I-H}X)^{+}Y \tag{2}$$

2.2.2 Artificial Neural Networks (ANNs)

ANNs have been very widely used for forecasting purposes [18]. In the present study single hidden layer feedforward ANNs trained using Levenberg-Marquardt (LM) learning algorithm and 5 hidden neurons (number of hidden neurons is arrived at, using trial and error) are employed.

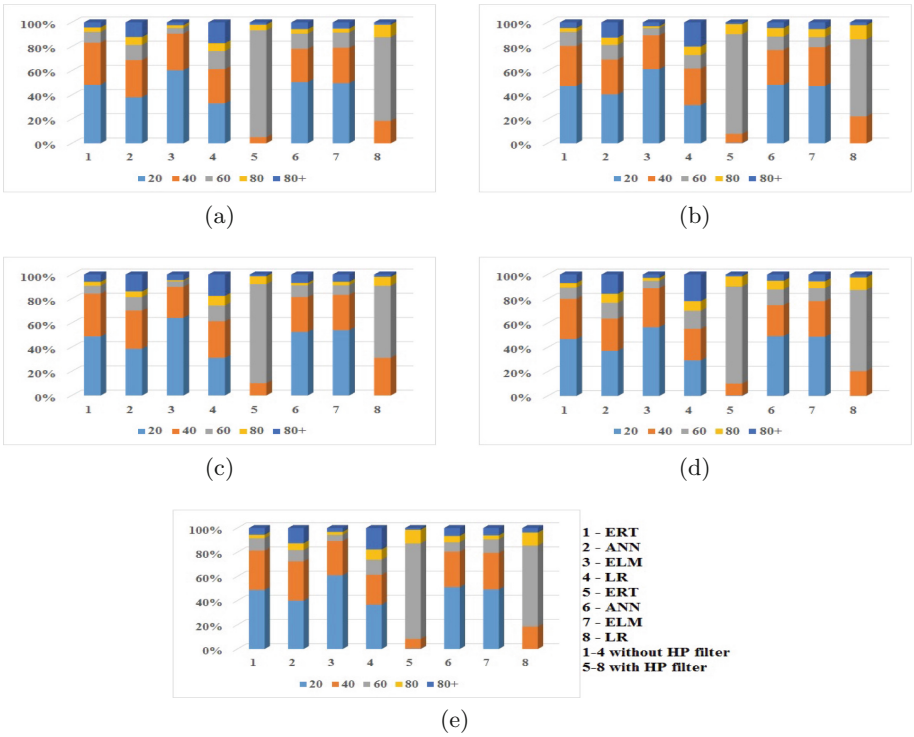


Fig. 4. MdAPE plots for all models with SME users (a) all seasons together (b) Autumn (c) Winter (d) Spring and (e) Summer.

2.2.3 Ensemble Regressing Trees (ERTs)

ERTs are forecast models which are generally weighted combinations of several regression trees which contributes to increase in predictive performance of regression tree. ERT models in the present study are developed using bagging [19] and number of trees is 5 (selected using trial and error).

2.2.4 Linear Regression (LR)

This are the simplest curve fitting algorithms available. The coefficients for the LR models considered in the present study were identified using simple Ordinary Least Squares technique.

3 Results and Analysis

Due to the large amount of data to be processed, a representative subset from the full set of users was selected for each of the three categories SME, RES and OTH. All the models were first validated on these subsets and only the best

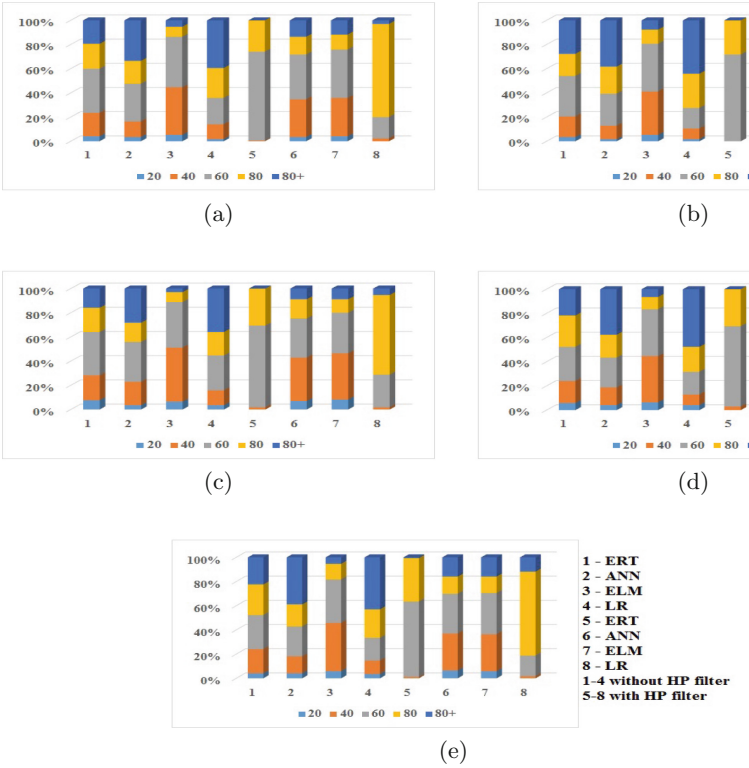


Fig. 5. MdAPE plots for all models with OTH users - (a) all seasons together (b) Autumn (c) Winter (d) Spring and (e) Summer.

performing models were considered for evaluation on the complete datasets. The subset selection criterion, as proposed in [20] is given below:

$$\text{sample size} = \frac{(Z^2(p)(1-p))}{C^2} \tag{3}$$

where: $z = 1.96$ for 95% confidence level, $p = 0.5$ (variability, maximum is 0.5) and $c = 0.05$ (i.e. 5%) is the confidence interval.

$$\text{For Finite population sample size} = \frac{samplesize}{1 + \frac{samplesize-1}{TotalPopulation}} \tag{4}$$

The nominal sample size required is found to be 215 for 'Small and Medium Enterprise users', 352 for 'residential' users and 315 for 'others' with a confidence level as 95% and confidence interval as '5'. Performance metric considered for the present study is Median Absolute Percentage Error (MdAPE) [21]. MdAPE has also been used as the performance measure in [2].

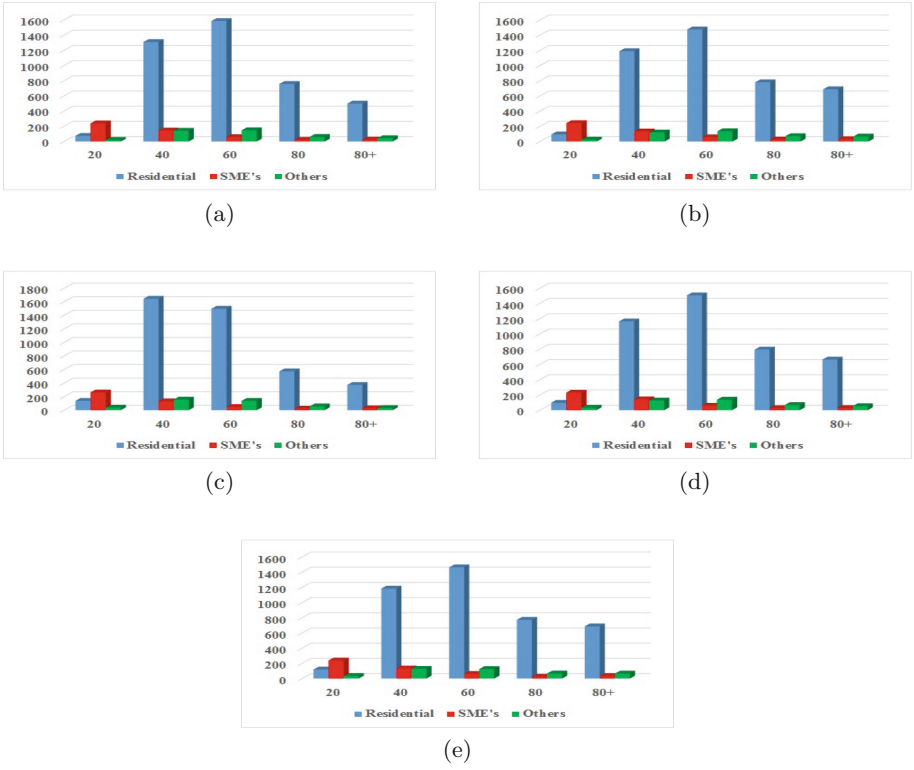


Fig. 6. MdAPE Histogram for RES, SME and OTH consumers using HP-ELM 2-hour ahead forecasting system for (a) all seasons together (b) Autumn (c) Winter (d) Spring and (e) Summer.

All the systems under consideration are validated for their 2, 4, 6 and 8 - hour ahead forecast capabilities. Since data is available at half hourly intervals, this translates to 4, 8, 12 and 16 step-ahead forecasts. Sliding window technique was used to re-train the forecasting systems with window length of 10. i.e.,

$$\widehat{y}(t+4) = f(y(t), y(t-1), ..., y(t-9)) \tag{5}$$

where: $\widehat{y}(t+4)$ = estimated two hour ahead consumption, $f(.)$ = machine learning based forecasting technique and $y(t)$ = consumption at time instant 't'.

Forecasting results for all the eight models (1–8 in Figs. 3, 4 and 5) are presented in the form of stacked column charts in figures below. The five colors in the charts represent the MdAPE ranges 0–20%, between 20 and 40%, 40 to 60%, 60 to 80% and >80% (legends 20, 40, 60, 80 and 80+ in Figs. 3, 4 and 5).

It can be observed from the results presented in Figs. 3, 4 and 5 that ELM based models tend to outperform all the other models considered. Hence, the

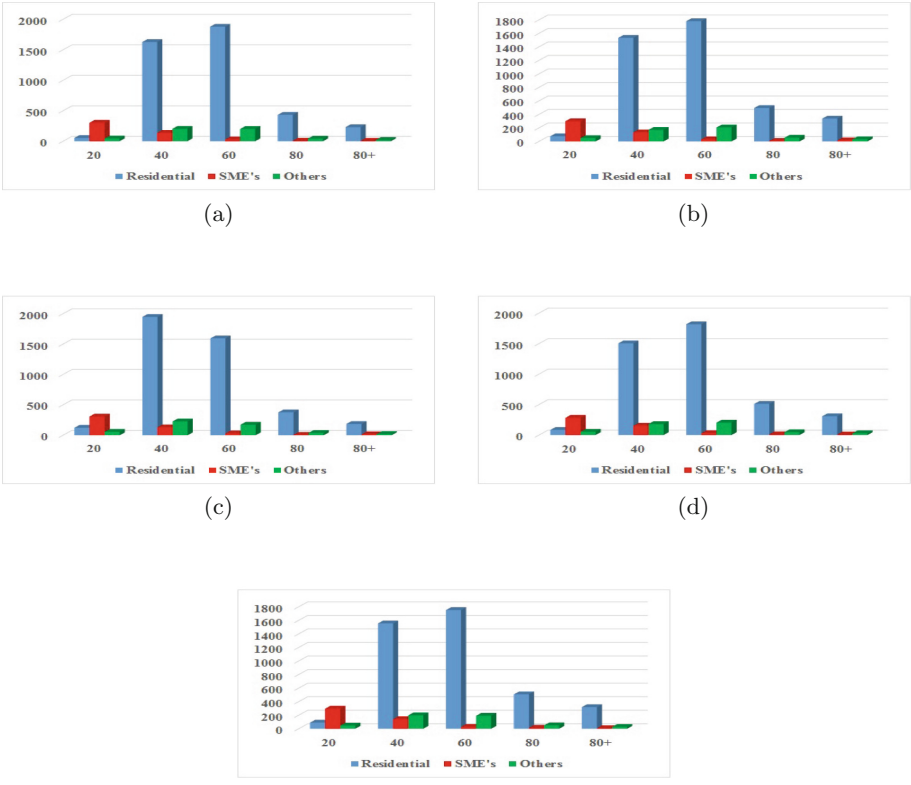
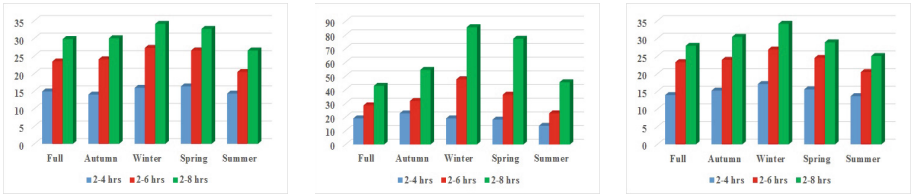


Fig. 7. MdAPE Histogram for RES, SME and OTH category consumers using ELM (without using HP filter) 2-hour ahead forecasting system for (a) all seasons together (b) Autumn (c) Winter (d) Spring and (e) Summer.



(a) (b) (c)

Fig. 8. Percentage increase in MdAPE between different forecast periods (a–c) forecasting system trained using ELMs without splitting the time series using HP filter for Residential, SMEs and Others category users respectively

two ELM based models (HP-ELM and ELM without HP preprocessing) were evaluated on the entire set of 5275 users. The results are presented in Figs. 6 and 7. It can be observed from the Fig. 8 that the 2-hour ahead forecasts were consistently better than the longer horizon forecasts of 6 and 8 h-ahead. It was also observed that forecasting system trained with ELMs and without using HP filter produce better result with consistency than the rest of systems considered. Employing HP filter for separation of trend and cyclic components and trained using ELMs and ANN also generated better forecasts, albeit slightly worse than the without HP-ELM system discussed earlier. It was also evident that linear regression models were consistently worse off compared to the other three techniques considered.

4 Conclusion

Based on the results presented above, it can be concluded that shorter term horizon forecasts tend to be more accurate while employing the proposed forecasting systems. It is also observed that ELM based forecasting systems are able to generate better forecasts when compared to ANN or ERT based systems. Due to the forecasting performance exhibited by ELM based systems in the present study, ELMs can be considered to be good candidates for generating electricity consumption forecasts as a part of a larger AMI system.

References

1. Siano, P.: Demand response and smart gridsA survey. Renew. Sustain. Energy Rev. **30**, 461–478 (2014)
2. Jayanth Balaji, A., Harish Ram, D.S., Nair, B.B.: Modeling of consumption data for forecasting in automated metering infrastructure (AMI) systems. In: Silhavy, R., Senkerik, R., Oplatkova, Z.K., Silhavy, P., Prokopova, Z. (eds.) Automation Control Theory Perspectives in Intelligent Systems. AISC, vol. 466, pp. 165–173. Springer, Cham (2016). doi:10.1007/978-3-319-33389-2_16
3. Chan, S., et al.: Load/Price forecasting and managing demand response for smart grids: methodologies and challenges. IEEE Sig. Process. Mag. **29**(5), 68–85 (2012)

4. Zhao, H., Tang, Z.: The review of demand side management and load forecasting in smart grid. In: 2016 12th World Congress on Intelligent Control and Automation (WCICA) (2016)
5. Tasdighi, M., et al.: Residential microgrid scheduling based on smart meters data and temperature dependent thermal load modeling. IEEE Trans. Smart Grid **5**(1), 349–357 (2014)
6. Hong, T., et al.: Long term probabilistic load forecasting and normalization with hourly information. IEEE Trans. Smart Grid **5**(1), 456–462 (2014)
7. Kwac, J., et al.: Household energy consumption segmentation using hourly data. IEEE Trans. Smart Grid **5**(1), 420–430 (2014)
8. ISSDA. http://www.ucd.ie/issda/data/commissionforenergyregulationcer/
9. Hodrick, R., Prescott, E.: Postwar U.S. business cycles. In: Real Business Cycles A Reader, pp. 593–608 (1998)
10. Nair, B.B., et al.: Clustering stock price time series data to generate stock trading recommendations: an empirical study. Expert Syst. Appl. **70**, 20–36 (2017)
11. Nair, B.B., Mohandas, V.: An intelligent recommender system for stock trading. Intell. Decis. Technol. IDT **9**(3), 243–269 (2015)
12. Nair, B.B., Mohandas, V.: Artificial intelligence applications in financial forecasting-a survey and some empirical results. Intell. Decis. Technol. IDT **9**(2), 99–140 (2015)
13. Nair, B.B., et al.: A stock trading recommender system based on temporal association rule mining. SAGE Open **5**, 2 (2015)
14. Huang, G.-B., et al.: Extreme learning machine: a new learning scheme of feedforward neural networks. In: 2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No. 04CH37541)
15. Huang, G.-B., et al.: Extreme learning machine: theory and applications. Neurocomputing **70**(1–3), 489–501 (2006)
16. Huang, G.-B., et al.: Extreme learning machine for regression and multiclass classification. IEEE Trans. Syst. Man Cybern. Part B (Cybern.) **42**(2), 513–529 (2012)
17. Huang, G.-B.: An insight into extreme learning machines: random neurons, random features and Kernels. Cogn. Comput. **6**(3), 376–390 (2014)
18. Specht, D.F.: The general regression neural network-Rediscovered. Neural Netw. **6**(7), 1033–1034 (1993)
19. Breiman, L.: Bagging predictors. Mach. Learn. **24**(2), 123–140 (1996)
20. Israel, G.D.: Determining sample size. University of Florida Cooperative Extension Service, Institute of Food and Agriculture Sciences, EDIS (1992)
21. Gooijer, J.G.D., Hyndman, R.J.: 25 years of time series forecasting. Int. J. Forecast. **22**(3), 443–473 (2006)

Simulation of a Single-Component System Using the Trajectories Method Taking into Account the Scheduling Preventive Maintenance

Mikhail V. Zamoryonov, Vadim Ya. Kopp, Olga V. Chengar(✉), and Yuri L. Rapatskiy

The Federal State Autonomous Educational Institution of Higher Education, Sevastopol State University, Sevastopol, Russian Federation
zamoryonoff@gmail.com, {v_kopp,u.l.Rapatskiy}@mail.ru, olga.chengar@gmail.com

Abstract. This article describes the functioning of a single-component system based on scheduling preventive maintenance. Its semi-Markov model using the trajectories method is constructed considering the following assumption: the time of preventive maintenance is much less than the time between failures and the time of system recovery. In addition, as a rule, the preventive maintenance is conducted out of working time fund, which justifies the assumption of its momentary execution. The employed method of trajectories is applicable only for the discrete systems. Thus, while modeling the current system, which is the system with a continuous phase space of states, the algorithm of phase consolidation for the transition to a system with a discrete phase space of states was applied. This method allows to obtain the exact solution of the task of finding the distribution function in the Laplace images, as opposed to the known methods. In order to validate the obtained results there has been a comparison of mathematical expectations of time staying in subset of operational conditions obtained due to the distribution function detected and based on the theorem of the average stationary stay time of the semi-Markov process in the subset of states.

Keywords: Semi-Markov system · Stationary distribution · Method of trajectories · Repeated entering's · Hidden failures · Means of control

1 Introduction

In solving the problem of reliability and efficiency improvement of technical systems there is the task to develop a sound operating strategy. The operation strategy (maintenance rules) is built on the basis of: objective data on the technical system (safety performance and maintainability); the specific characteristics of the system (the system structure, the failure indication characteristics, the presence of a built-in performance monitoring); the data on the operation conditions. The operation strategy should have the property of optimality by some measure that indicates the system performance and operation. The selection of the optimal maintenance strategy allows you to achieve the best results due to the reorganization of the

R. Silhavy et al. (eds.), *Cybernetics and Mathematics Applications in Intelligent Systems*, Advances in Intelligent Systems and Computing 574, DOI 10.1007/978-3-319-57264-2_27

operational rules with no additional forces and resources [1–3]. We would also note that the optimization conduct requires the construction of correct mathematical models of the studied systems functioning.

The problem of maintenance both in our country and abroad received much attention as the rational approach to this issue can lead to a significant economic benefit. There is a large number of publications [4–19] dedicated to the issues of theoretical. However, let us note that mostly these tasks contain steady-state characteristics, and the distribution function (DF) of time to failure of the system in view of the preventive maintenance is not considered. Though, in the simulation of complex systems based on the preventive maintenance of individual components it is advisable to use a hierarchical approach to the construction of the general model system, for which it is necessary to have information on the DF of operation time of individual elements.

The purpose of this article is to build a semi-Markov (SM) model of a single-component system based on the scheduling preventive maintenance using the method of trajectories.

2 Problem Formulation

Let us consider the operation of such a system. The time of its fail-safe operation – RV α_1 with DF $F_1(t) = P(\alpha_1 \le t)$, system recovery time – RV β_1 with DF $G_1(t) = P(\beta_1 \le t)$. At random times (at intervals α_2 with DF $F_2(t) = P(\alpha_2 \le t)$ the preventive maintenance is held, the duration of preventive maintenance – RV β_2 with DF $G_2(t) = P(\beta_2 \le t)$. Preventive maintenance is carried out, if the start of maintenance fell into the system work period. After the preventive maintenance is done the system operation starts from the beginning (the operating properties of the system completely update). RV α_1, α_2, β_1, β_2 are assumed to be independent, having finite mathematic expectations and variances; DFs $F_1(t)$, $G_1(t)$, $G_2(t)$ have the densities $f_1(t)$, $g_1(t)$, $g_2(t)$.

To describe the operation of the system we use the PMR $\{\xi_n, \theta_n; n \ge 0\}$ and the corresponding SMP $\xi(t)$. The physical states of the system: 1 - the system is operational, 0 - the system is recovering, 2 - the system is in preventive maintenance. Let us extend PSS entering the semi-Markov states:

$$E = \{221, 210, 102x, 112x\},$$

where 221 – is the preventive maintenance carrying out (instantaneous state); 210 – the preventive maintenance is completed, the system began to operate from the start; 102x – the system has completed its work, the recovery has begun; the time remaining until the preventive maintenance $x > 0$; 112x - the system recovery has completed; the time remaining prior to the preventive maintenance $x > 0$.

The graph of such a system is shown in Fig. 1.

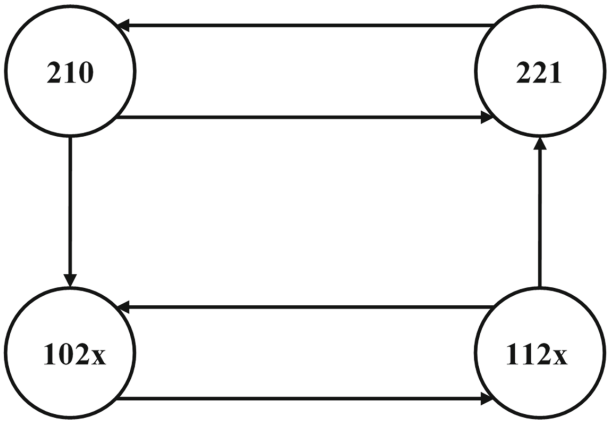


Fig. 1. The graph of a single-component system based on scheduling preventive maintenance.

3 Problem Solution

Considering the fact that the time of preventive maintenance is much less than time to failure and the recovery time after a failure, and the fact that preventive maintenance can be carried out outside of the working fund (for example, during the two-shift operation – in the third shift) when modeling the system we neglect the time for preventive maintenance. Taking into account the above, we have the subsets of efficient $E_+ = \{221, 210, 112x\}$ and inefficient state $E_- = \{102x\}$

The probabilities of transition of the SM system, the stationary distribution of the embedded Markov chain (EMC) and times of stay in the states are defined in [7]

$$P_{210}^{221} = P\{\alpha_1 > \alpha_2\} = \int_0^\infty \bar{F}_1(y) dF_2(y); P_{221}^{210} = 1;$$

$$p_{210}^{102x} = P\{\alpha_2 - \alpha_1 \in dx\} = \int_0^\infty f_1(y) d_y F_2(x+y);$$

$$p_{112x}^{102y} = P\{x - \alpha_1 \in y\} = f_1(x-y),\ x \geq y;$$

$$P_{102x}^{112x} = 1; P_{112x}^{221} = P\{\alpha_1 > x\} = \bar{F}_1(x).$$

$$\begin{cases} \rho_1 = \rho_0 \int_0^\infty \bar{F}_1(y) dF_2(y) + \int_0^\infty \bar{F}_1(y) \rho_{112}(y) dy; \rho_0 = \rho_1; \rho_{112}(x) = \rho_{102}(x); \\ \rho_{102}(x) = \rho_0 \int_0^\infty f_1(y) d_y F_2(x+y) + \int_x^\infty f_1(y-x) \rho_{112}(y) dy, x > 0; \\ \rho_0 + \rho_1 + \int_0^\infty \rho_{102}(x) dx + \int_0^\infty \rho_{112}(x) dx = 1. \end{cases}$$

$$\rho_{112}(x) = \rho_0 \int_0^\infty h_1(y) d_y F_2(x+y),$$

where $h_1(y) = \sum_{n=1}^{\infty} f_1^{*(n)}(y)$ – is the density of the recovery function.

$\theta_{221} = \beta_2$, $\theta_{210} = \alpha_1 \wedge \alpha_2$, $\theta_{112x} = \alpha_1 \wedge x$ and $\theta_{102x} = \beta_1 \wedge x$, where $\wedge$ - is sign of minimum.

To determine the DF of time to failure of the system we use the trajectories method based on the theorem about the DF times of stay in a subsets states in view of the repeated enterings [20]. The method consists of several steps.

In the first step we make the transition from a system with discrete-continuous phase space of states to the system with discrete states, using the algorithm of phase consolidation [20, 21].

To do this we need to find the DF of stay times in the states 102x and 112x and the transition probabilities. First of all we determine the stationary distribution of CMV for these states:

$$\hat{\rho}_{102} = \hat{\rho}_{112x} = \rho_0 \int_0^\infty h_1(y) \bar{F}_2(y) dy;$$

$$\rho_0 = \frac{1}{2 + 2\int_0^\infty h_1(y) \bar{F}_2(y) dy};$$

$$\rho_{210} = \rho_{221} = \rho_0.$$

Next, we determine the transition probabilities:

$$\hat{P}_{112}^{221} = \frac{\int_0^\infty \int_0^\infty h_1(y) f_2(x+y) \bar{F}_1(x) dx dy}{\int_0^\infty h_1(y) \bar{F}_2(y) dy};$$

$$\hat{P}_{112}^{102} = \frac{\int_0^\infty \int_0^\infty h_1(y) f_2(x+y) F_1(x) dx dy}{\int_0^\infty h_1(y) \bar{F}_2(y) dy};$$

$$\hat{P}_{210}^{221} = \int_0^\infty \bar{F}_1(y) f_2(y) dy;$$

$$\hat{P}_{210}^{102} = \int_0^\infty F_1(y) f_2(y) dy,$$

and the DF of time stay of the system in these states:

$$\hat{F}_{102}(t) = G_1(t) + \bar{G}_1(t) \cdot \left(1 - \frac{\int_0^\infty h_1(y) \cdot \bar{F}_2(t+y)dy}{\int_0^\infty h_1(y) \cdot \bar{F}_2(y)dy} \right);$$

$$\hat{F}_{112}(t) = F_1(t) + \bar{F}_1(t) \cdot \left(1 - \frac{\int_0^\infty h_1(y) \cdot \bar{F}_2(t+y)dy}{\int_0^\infty h_1(y) \cdot \bar{F}_2(y)dy} \right).$$

The second step is to extract all the possible transition trajectories from the subset E_+ into the subset E_-:

$$T_1 = \{112x\};$$
$$T_2 = \{221, 210, 112x\}.$$

In the third step we determine that in each state of the trajectory the system falls at least once during the stay in this trajectory with the probability of 1.

In the fourth step according to the theorem on the DF of the system stay times in the states considering the repeated enterings [20] into the field of Laplace images we determine by the Laplace density and residence time distribution function of the system in the states of trajectories based on repeated returns according to the following formula:

$$f_i^\theta(s) = \frac{f_i(s)}{c_i - (c_i - 1)f_i(s)}.$$

The fifth step. We determine the probability of each trajectory:

$$P_1^T = \hat{P}_{112}^{102};$$
$$P_2^T = \hat{P}_{112}^{221}.$$

The sixth step. In accordance with the result of the Theorem 2 above, we find that the DF of time spent by the system in each trajectory.

$$F_1^T(t) = \hat{F}_{112}(t);$$
$$F_2^T(t) = \hat{F}_{112}(t) * F_{210}^\theta(t),$$

where $*$ - is a sign of the convolution operation, and $F_{221}^\theta(t)$ - is not taken into account, as 221 – is and instantaneous state.

In the seventh step, we define the required $F_{rez}(t)$ DF of time spent by the system in the subset E_+:

$$F_{rez}(t) = F_1^T(t) \cdot P_1^T + F_2^T(t) \cdot P_2^T$$

Let us consider the example of the simulation of such system with known parameters of distribution of random variables.

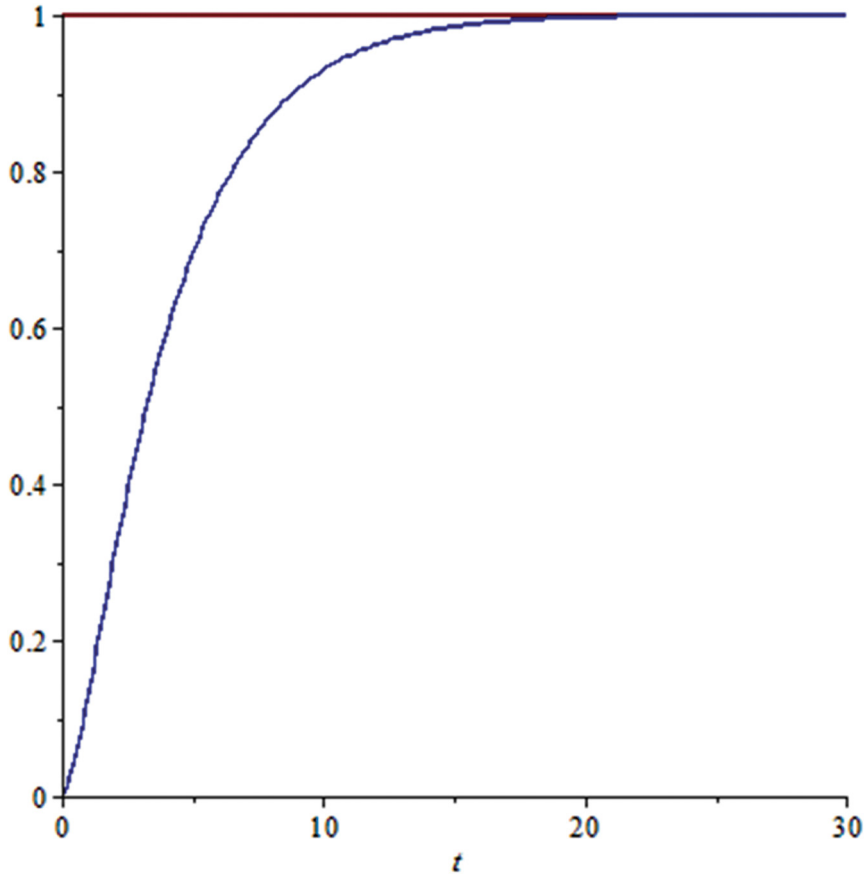


Fig. 2. The view of the DF stay time of the system in subset E_+

The initial data for modeling are the distribution functions $F_1(t)$, $F_2(t)$, $G_1(t)$ and $G_2(t)$; they are distributed by the generalized Erlang law of the second order with the parameters λ_1, λ_2; μ_1, μ_2; υ_1, υ_2; γ_1, γ_2 accordingly; and

$$f_1(t) = \frac{\lambda_1 \lambda_2 (e^{-\lambda_1 t} - e^{-\lambda_2 t})}{\lambda_2 - \lambda_1},$$

where λ1 = 0,3333 ч–1, λ2 = 1,000 ч–1;

$$f_2(t) = \frac{\mu_1\mu_2(e^{-\mu_1 t} - e^{-\mu_2 t})}{\mu_2 - \mu_1},$$

where μ1 = 0,0833 ч–1, μ2 = 0,250 ч–1;

$$g_1(t) = \frac{\upsilon_1\upsilon_2(e^{-\upsilon_1 t} - e^{-\upsilon_2 t})}{\upsilon_2 - \upsilon_1},$$

where υ1 = 6,667 ч–1, υ2 = 20,0 ч–1;

$$g_2(t) = \frac{\gamma_1\gamma_2(e^{-\gamma_1 t} - e^{-\gamma_2 t})}{\gamma_2 - \gamma_1},$$

where γ1 = 2,667 ч–1, γ2 = 8,00 ч–1.

Figure 2 shows the required $F_{rez}(t)$ distribution function obtained in this study.

Let us compare the expected value of the function we received and the expected value defined by the expression [22]:

$$T_+ = \frac{\sum\limits_{i \in M_+} m_i \rho_i}{\sum\limits_{e \in E \subset M_+} \sum\limits_{j \in M_-} P_{ej} \rho_i} \quad (1)$$

The expectation of the distribution function we have received is 4.194805194805194805 h, whereas the one defined by the expression (1) - 4.194805194805194805 h.

It is clear that the expected values are the same.

4 Conclusions

The comparison of the expected values of the stay times in the subset of the operating conditions obtained on the base of the DF detected in this study and on the base of the theorem of the mean stationary stay time of the semi-Markov process in the subset of states, has proved the correctness of the obtained results.

We plan to apply this method to simulate more complex technical maintenance in automated production in the further researches.

Acknowledgments. The study was completed as the base part of the state order of the Ministry of Education and Science of the Russian Federation № 2014/702 (project № 4000) and as the support of the grant RFBR (project № 15-01-05840).

References

1. Barzilovich, E.J.: Models maintenance of complex systems. In: Proceedings of Manual for Schools, 231 p. Higher, SK (1982)
2. Barlow, R., Belyaev, J.K., Bogatyrev, V.A., et al.: Reliability of technical systems: manual. In: Ushakov, I.A. (ed.) Radio and Communications, 606 p. (1985)

3. Bayhelt, F., Frank, P.: Reliability and maintenance. Mathematical bypass. In: Radio and Communications, 392 p. (1988)
4. Balan, S.: The organization of the repair and maintenance of engineering equipment. In: Proceedings of Allowance for Students. Mechnikov National Universities, Astroprint, Odessa, 165 p. (1997)
5. Borisov, Y.S.: Organization of repair and maintenance of equipment. In: Mechanical Engineering, 360 p. (1978)
6. Gertsbakh, I.B.: Prevention models (Theoretical Foundations of preventive maintenance planning). Sov. radio, 216 p. (1969)
7. Glech, S.G.: Determining the optimal points for the prevention of cell technology. Vestn. SevGTU: Coll. scientific. tr., Sevastopol, Vyp. 36, pp. 169–176 (2002)
8. Glech, S.G.: Determination of the optimal technological aspects of prevention cell to instantly replenished reserve time. Collection of scientific works SIYaEiP, Issue 5, pp. 187–193. SIYaEiP, Sevastopol (2001)
9. Druzhinin, G.V.: The processes of maintenance of automated systems. Energy, 273 p. (1973)
10. Chestnut, V.A.: Optimal maintenance tasks. Knowledge, 121 p. (1981)
11. Chestnut, V.A.: Semi-Markov model maintenance process. Knowledge, 91 p. (1987)
12. Korlat, A.N., Kuznetsov, V.N., Novikov, M.M., Turbin, A.F.: Semi-Markov model of renewable systems and queuing systems. Kishinev Shtiintsa, 276 p. (1991)
13. Manshin, G.G.: Preventive management regimes of complex systems. In: Science and Technology, Minsk, 255 p. (1976)
14. Manshin, G.G., Barzilovich, E.Y., Voskoboev, V.F.: Methods of preventive maintenance ergonomics systems. In: Science and Technology, Minsk, 222 p. (1983)
15. Manshin, G.G., Ignatov, V.A.: Elements of optimal maintenance products theory. In: Konovalova, E.G. (ed.) Science and Technology, Minsk, 191 p. (1974)
16. Manshin, G.G., Brick, S.V.: Quality assurance of functioning of automated systems. In: Science and Technology, Minsk, 221 p. (1986)
17. Obzherin, Y.E., Glech, S.G.: Analysis of the technological reliability of the cell, taking into account the instantaneous prevention. Matamet simulates th in education, science and industry: Coll. scientific. tr. AS, St. Petersburg Department IHEAS, Saint-Petersburg, pp. 123–127 (2000)
18. Obzherin, Y.E., Glech, S.G.: The impact of prevention on the reliability of technological cell. In: Applied Problems of Gas and Fluid Mechanics: Articles IX International Scientific and Engineering. Conference Scientists of Ukraine, Russia, Belarus, 25–29 September 2000. Izd SevGTU, Sevastopol, pp. 51–56 (2000)
19. Obzherin, Y.E., Glech, S.G.: The impact of prevention on the reliability of technological cell instantly replenished with time reserve. In: Applied Problems of Gas and Fluid Mechanics: Proceedings of the X International Scientific and Engineering Conference Scientists of Ukraine, Russia, Belarus, 25–29 September 2001, pp. 135–141. Izd SevGTU, Sevastopol (2001)
20. Zamorënov, M.V., Koop, V.Y., Obzherin, Y.E., Zamorënova, D.V.: Testing paths method on the example of the production element simulation operation process devalues failures. News TSU Technical Science, 2 pm, Part 1 of Tula State University, vol. 8, pp. 57–71 (2015)
21. Koroljuk, V.S., Turbin, A.F.: Mathematical Foundations of phase merging complex systems, 217 p. Naukova Dumka, Kiev (1978)
22. Koroljuk, V.S.: Stochastic models of systems. In: Turbin, A.F. (ed.) Sciences, 208 p. Dumka, Kiev (1989)
23. Koroljuk, V.S., Turbin, A.F.: Markov renewal processes in problems of system reliability. In: Sciences, 236 p. Dumka, Kiev (1982)

Analysis of the IoT WiFi Mesh Network

Piotr Lech(✉) and Przemysław Włodarski

Department of Signal Processing and Multimedia Engineering,
Faculty of Electrical Engineering, West Pomeranian University of Technology,
Szczecin, Sikorskiego 37, 70-313 Szczecin, Poland
{piotr.lech,przemyslaw.wlodarski}@zut.edu.pl

Abstract. This paper presents a conception of designing wireless sensor networks in mesh topology that perform their IoT tasks applying popular WiFi standards. Cheap IoT modules involve compromise between reliability and the price. Phenomena that occur in real wireless sensor network depends on many factors that are sometimes not well defined. Statistical analysis of the packet delays and failure rates for different scenario paths in our experimental network helps to identify anomaly nodes.

Keywords: IoT · WiFi · ESP8266 · NodeMCU · Mesh network

1 Introduction

Nowadays, Internet of Things (IoT) more often becomes a better replacement for home automation. A lot of applications in the commercial solutions leads to a significant drop in prices for IoT devices as well as project investments [2]. Migration from licensed and closed system, which is dedicated for particular application, to IoT technology causes greater flexibility in implementations and a much wider range of applications through an integration of the popular network services. Of particular interest is the utilization of WiFi technology due to the possibility of easy expansion and integration with the existing network. On the one hand, we can maintain operating strategy in a closed environment, isolating IP network designed for IoT applications from global network, or we can focus on an open systems integrated with public Internet network. Due to the commonness of wireless networks, we can not exclude the simultaneous coexistence and diffusion of networks for the both aforementioned cases. On the contrary, it is advisable to share the same infrastructure due to reduction in the costs of construction and operation [3]. Certain habits in the design process related to the industry communication standards for sensor networks [1], such as mesh networks, suffers from some difficulties in implementation.

In IP networks implemented using WiFi standard, there are two fundamental models of connections that are implemented in the following topologies:

– Ad-Hoc - devices within the range can communicate with each other,
– AP - devices communicate with each other via an access point.

R. Silhavy et al. (eds.), *Cybernetics and Mathematics Applications in Intelligent Systems*,
Advances in Intelligent Systems and Computing 574, DOI 10.1007/978-3-319-57264-2_28

Implementation of the mesh topology encounters problems with the need for additional procedures related to routing. There are many protocols that support the process of mesh network service within the IP network, for example: B.A.T.M.A.N. (Better Approach To Mobile Adhoc Networking), Babel (a distance-vector routing protocol for IPv6 and IPv4 with fast convergence properties), HWMP (Hybrid Wireless Mesh Protocol). The use of these protocols requires the full implementation of TCP/IP stack and significant computing power which limits their implementations. However, it should be noted that not all equipment (for WiFi communication) support particular protocol as in the case of HWMP protocol implementation. The use of advanced microcontroller greatly increases the building cost of network, and many times we need to get information about the slow process through periodic measurement, using free microcontroller resources to propagate data among network nodes. A wide range of simple WiFi communication modules made as System on Chip (SoC), that in addition to handling communication standards enable data acquisition by embedded GPIO ports. It allows for the construction of low cost sensor networks based on widely used wireless network standard - WiFi. However, when too little computing power does not allow for the implementation of advanced algorithms that support routing in IP networks with mesh topology, it is possible to create a simplified network while maintaining features of the mesh network. It is assumed that the network replicates messages and routing path is fixed and static. Multiplied packets are removed at sink - the element of sensor network where all messages are sent. We show how the processes in our experimental IoT mesh network influence network performance during normal operation.

2 Testing Environment and Research Strategy

The main purpose for testing environment is to implement WiFi communication in mesh IP topology (Fig. 1) with some limitations related to the construction of particular devices. To accomplish this task, the developer version (NodeMCU) of communication module based on SoC chip ESP8266 is used. This module has 10 GPIO ports (each of which can handle PWM, I^2C and 1-wire) and embedded 10-bits ADC and USB-to-UART converter. NodeMCU also has the installed firmware with TCP/IP stack that supports programming using the *Lua* scripting language. An interesting feature on the ESP8266 is the simultaneous operation in AP (access point) and STA (station) modes. Unfortunately, according to the documentation, it is possible to handle up to 4 connections to the STA-type clients. The basis of the operation of the single node in our experimental mesh network is the communication in AP+STA mode. The communication strategy among nodes is based on the transmission of the single message to the nodes with numbers that are higher than the number of the receiving station. All nodes have permanently assigned numbers that increases from source (RPi 1) in the direction of the sink node (RPi 2) [5]. According to Fig. 1 one can select the following routing paths: 1-2-5, 1-2-3-5, 1-2-4-5, 1-3-5 and 1-4-5. The node numbers are closely related to their IP addresses. The messages are sent over the UDP

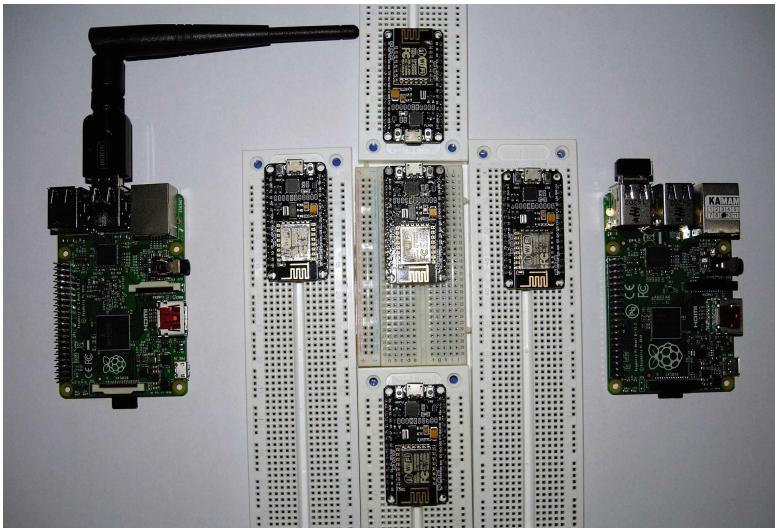


Fig. 1. Set of test devices

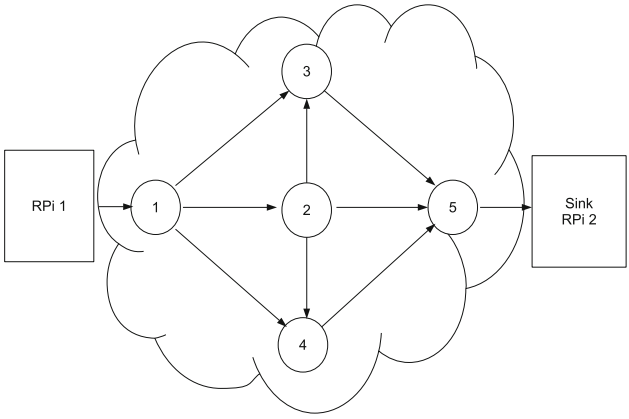


Fig. 2. Topology

protocol, multiplied messages are removed in the node with higher number or in the sink node. Because the connectionless UDP protocol is used, it is possible that message can be dropped causing non-zero loss rate. The source of messages is the microcomputer Raspberry Pi v.2 (RPi 1) which sends packets to the node 1 (see Fig. 2). The NodeMCU modules duplicate and send packets according to the aforementioned strategy. All traffic flush at the second Raspberry Pi (RPi 2) - the sink node connected via point-to-point to the node 5. At the same time, RPi 2 serves as a recorder for received messages. All nodes and microcomputers have synchronized clocks over the network. Each node records the reception

time of packet, writes it into the appropriate field in the message and forward the message to the next node. Nodes that perform message duplication, sends them in one loop - from lowest to highest number associated with node. Time recording for received packets at particular nodes make it possible to observe phenomena occurring during data transfer and can be useful for communication modeling [5].

3 Data Analysis

The total measurement time took 20 min 2 s and 198 ms. During this time a lot of time stamps in milliseconds represented by integers were recorded. Each scenario consists of 1000 series of time stamps registered at the time of packet arrival in each node. First time stamp in each series is always related to rpi1 device, next time stamps were captured at different nodes depending on scenario, and the last time stamp is always fifth node. In case of failure, when packet was dropped on any node, a 0 value was recorded and this kind of data is analyzed in Subsect. 3.2. All data for the whole measurement were stored in one CSV-formatted file. All fields as well as example values are presented in Table 1.

The analysis is divided into two parts: delay and failure, based on data that are split into five time series, according to the scenario path. The relationships among scenarios as well as data structure is investigated.

3.1 Delay Analysis

In order to analyze delays, time series that represent different scenarios path were extracted from the data file. In order to skip failures, arrays filled with zeroes $\{0, 0, 0, 0\}$ for 4-nodes paths and $\{0, 0, 0, 0, 0\}$ for 5-nodes paths were filtered from previously obtained sequences. Then all samples were differentiated in each random sequence to obtain interarrival times of incoming packets.

Because of the connection speed differences among nodes that resulted in different levels of intervals, as well as for comparison purposes, all random sequences were shifted to 0–4 scale by changing mean values. As a result, it was possible to compare differences in distributions for different scenario paths using quantile plots (Figs. 3 and 4). Additionally, all values were randomized for better visualization. Dots that are outside the red boxes mean the level of inconsistency with the distribution for the base path 1-2-5. Since there are many combinations for this type of comparison, one decided to choose only two series, for first and fifth node, related to 1-2-5 path. One can see that for the first node (Fig. 3) distributions are more consistent than for the fifth node (Fig. 4). This may be due to the distance associated with the number of nodes that every single packet had to pass.

Further studies on the interval distributions were focused on the relationship between histograms and probability density function. Figure 5 shows example histograms for different paths and nodes and their mean values. As it can be seen all histograms have equal number of packets for the corresponding intervals.

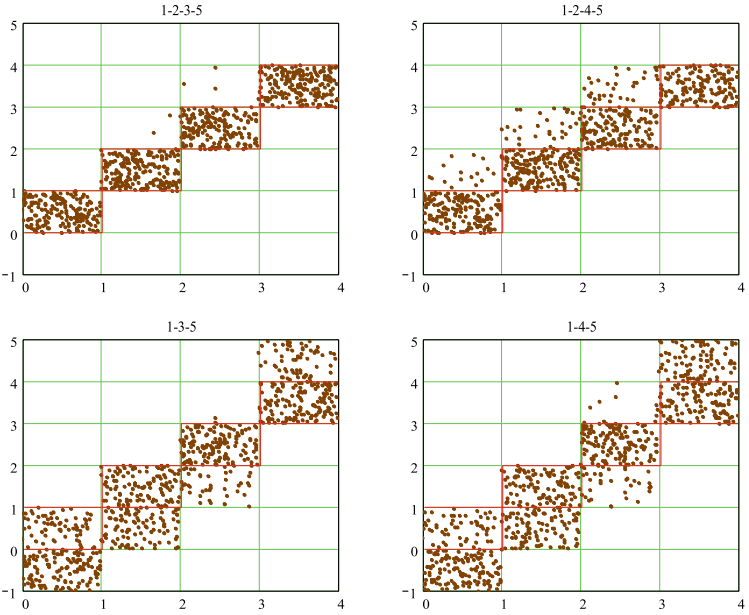


Fig. 3. Randomized quantile plot of packet reception intervals for the first node

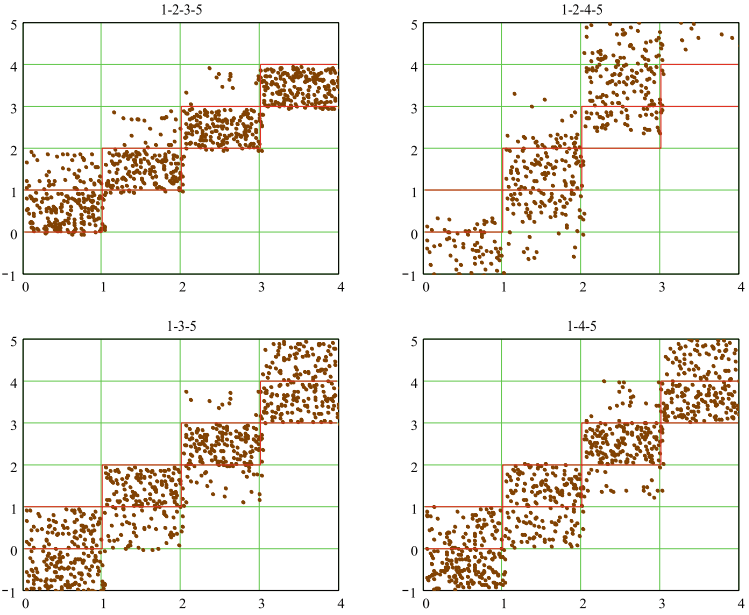


Fig. 4. Randomized quantile plot of packet reception intervals for the fifth node

Table 1. Recorded data structure - fields and an example

Field name	Field type	Example
series number	*integer*	957
word *rpi*1	*string*	*rpi*1
time stamp [ms] at *rpi*1	*integer*	1481737848652
scenario path #1	*string*	$1- > 2- > 5- > rpi2$
time stamps [ms] for scenario #1	*array*	{1481737848656, 1481737848782, 1481737848984, 1481737849010}
...	...	...
scenario path #5	*string*	$1- > 4- > 5- > rpi2$
time stamps [ms] for scenario #5	*array*	{1481737848656, 1481737848885, 1481737849138, 1481737849164}

Table 2. Deviation of mean interval values for different paths, in [%]

Scenario path	Node 1	Node 2	Node 3	Node 4	Node 5
1-2-5	0.019	0.01	-	-	0.115
1-2-3-5	0.015	0.009	0.066	-	0.288
1-2-4-5	0.014	0.008	-	0.822	0.827
1-3-5	0.004	-	0.025	-	0.198
1-4-5	0.007	-	-	0.006	0.041

Thus, it is supposed that all random sequences have uniform distribution. Table 2 shows the absolute deviation of mean interval values related to the theoretical

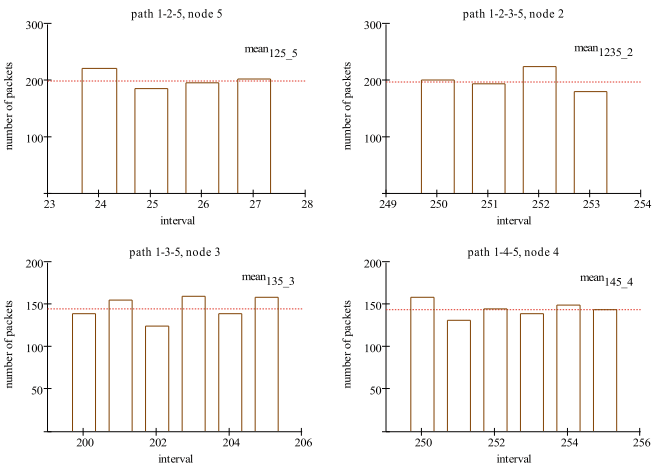


Fig. 5. Histograms and mean values for different nodes and paths

Table 3. Relative variances of intervals for different paths

Scenario path	Node 1	Node 2	Node 3	Node 4	Node 5
1-2-5	0.252	2.058	-	-	4.215
1-2-3-5	0.271	3.752	3.468	-	3.743
1-2-4-5	1.637	1.231	-	**61.275**	**68.098**
1-3-5	6.186	-	1.479	-	1.504
1-4-5	0.465	-	-	2.602	0.233

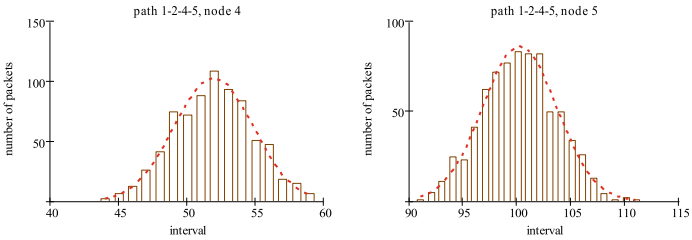


Fig. 6. Two histograms for the selected random variables from Table 3

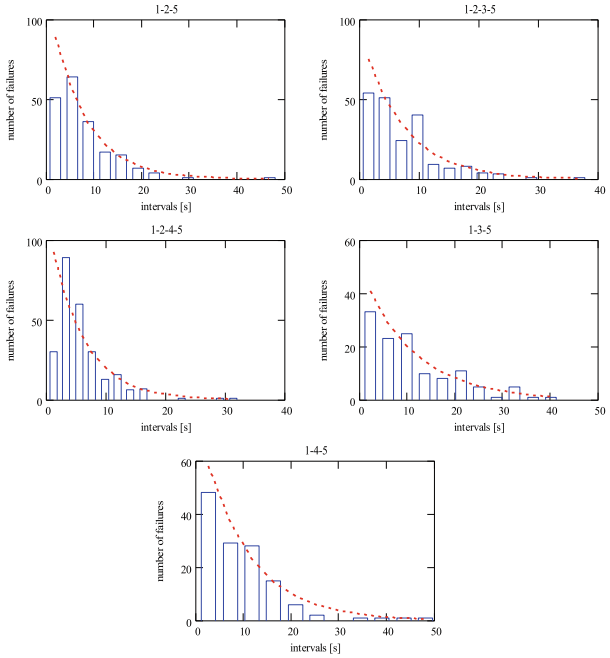


Fig. 7. Histograms and theoretical curves for inter-failure times

value for the uniform distribution, that is $abs(\tilde{x}/\bar{x} - 1) \cdot 100$ [%], where $\tilde{x}$ is the estimated mean for random sequence and $\bar{x}$ is the theoretical mean for uniform distribution which equals to the mean of minimum and maximum interval value. Table 3 shows the absolute variances related to the theoretical variance for the uniform distribution, that is $abs(V(\tilde{x})/V(\bar{x}) - 1) \cdot 100$ [%], where $V(\tilde{x})$ is the estimated variance for random sequence and $V(\bar{x})$ is the theoretical variance for the uniform distribution which equals to $((max(x) - min(x) + 1)^2 - 1)/12$ (for discrete random variable). All values in Table 2 are below 1%, relative variances in Table 3 are also not too high except two values for nodes 4 and 5 in the 1-2-4-5 scenario path (over 60%). The high values indicates that these random sequences have different distribution, in this case normal - see Fig. 6.

3.2 Failure Analysis

Failure analysis is based on the assumption that all failures were independent and not correlated with the operation of the devices. The natural distribution for such phenomena is the exponential distribution. Because failures was strictly related to the whole scenario paths, and not to the particular nodes, there was exactly five random sequences analyzed. All of them had similar properties related to the exponential distribution, especially as far as the first two statistical moments are concerned. Histograms presented in Fig. 7 have $K = floor(\sqrt{N})$ bins. Estimation of λ parameter for the probability density function $f_{exp}(y) = \lambda e^{-\lambda y}$ as well as

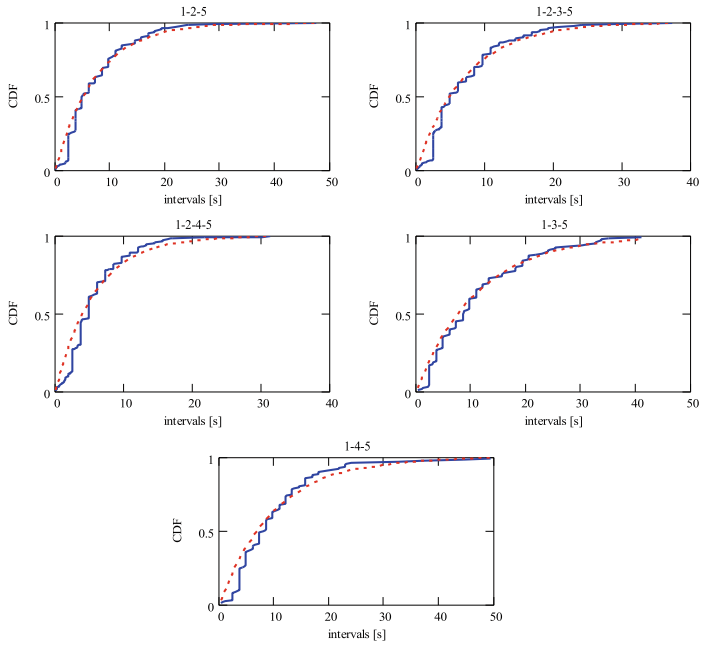


Fig. 8. Empirical and theoretical cumulative distributions for different scenarios

for the cumulative distribution function $F_{exp}(y) = 1 - e^{-\lambda y}$ is based on the mean value of y: $\tilde{\lambda} = 1/\bar{y}$ [6]. These $\tilde{\lambda}$ values are used to draw theoretical probability density function in Fig. 7 and cumulative distribution function in Fig. 8 (dashed red lines).

4 Conclusions

IoT technologies often apply the ideas used in sensor networks. However, in the case of IoT projects and mesh topology it requires a full TCP/IP stack and an additional routing algorithm that involves high level of computational complexity. As a result, the implementation of the aforementioned elements requires more expensive microcomputer systems with integrated operating system (embedded). With the compromise related to the simplification of routing algorithms, additional resources are released for any tasks, i.e. measurement, peripheral communication, etc. Simultaneously, the network becomes stable and reliable, which was confirmed by the test carried out. Statistical analysis of failures shows no surprises, and their correspondence to the exponential distribution is common in the environment where there are no factors influencing the operation. However, delay analysis shows that the distribution is closely related to uniform, except two cases when normal distribution should be applied. It was probably an anomaly that occurred between 4th and 5th node in the 1-2-4-5 scenario path. An application of several statistical methods shows how the communication process can be monitored and detect possible anomalies. This problem could be the starting point for future research, regarding the network structure as well as communication in noisy and congested environment.

References

1. Chong, C.-Y., Kumar, S.P.: Sensor networks: evolution, opportunities, and challenges. Proc. IEEE **91**(8), 1247–1256 (2003)
2. Guinard, D., Vlad, T.: Building the Web of Things. Manning Publications (2015)
3. Choubey, P.K., Pateria, S., Saxena, A., Vaisakh Punnekkattu Chirayil, S.B., Jha, K.K., Sharana Basaiah, P.M.: Power efficient, bandwidth optimized and fault tolerant sensor management for IOT in smart home. In: Proceedings of IEEE International Advance Computing Conference (IACC), pp. 366–370 (2015)
4. Maksimovic, M., Vujovic, V., Davidovic, N., Milosevic, V., Perisic, B.: Raspberry Pi as Internet of Things hardware performances and constraints. In: Proceedings of IcETRAN: First International Conference on Electrical, Electronic and Computing Engineering, pp. 1–6 (2014)
5. Hussain, M.I., Dutta, S.K., Ahmed, N., Hussain, I.: A WiFi-based reliable network architecture for rural regions. ADBU J. Eng. Technol. **3** (2016)
6. Forbes, C., Evans, M., Hastings, N., Peacock, B.: Statistical Distributions, 4th edn (2011). ISBN 978-0-470-39063-4

The Experience of Building Cognitive User Interfaces of Multidomain Information Systems Based on the Mental Model of Users

M.G. Shishaev[1,2], V.V. Dikovitsky[1,2(✉)], and L.V. Lapochkina[3]

[1] Murmansk Arctic State University, Egorova st. 15, Murmansk 183038, Russia
shishaev@arcticsu.ru, dikovitsky@gmail.com
[2] Institute for Informatics and Mathematical Modelling of Technological Processes of the Kola Science Centre RAS, 24A, Fersman st., Apatity 184209, Russia
[3] Federal State Autonomous Educational Institution of Higher Education, «Northern (Arctic) Federal University named after M.V. Lomonosov», Severnaya Dvina Emb. 17, Arkhangelsk 163002, Russia
l.lapochkina@narfu.ru

Abstract. The article describes the methodological basis of the synthesis of cognitive interfaces for multidomain information systems. A definition of the cognitive user interface as well as approaches to its formal assessment are given. Special attention is paid to the semantic and perceptual aspects of cognitive interfaces, the concept of relevance, pertinence, user stereotypes. Application of the user experience in the task of information retrieval is described. One of the possible ways of obtaining and record-keeping of user preferences is constructing a model of user interests in the form of a formalized mental model. An approach which makes it possible to increase the relevance of the search results is presented.

Keywords: Cognitive user interface · Multidomain IS · Semantic model · User interface · Cognitive · Relevance · Pertinence · Perceptual stereotypes

1 Introduction

Over the past decades, an information system (IS) has undergone an impressive evolution. Modern IS are extremely diverse: in the scale – from a microchip to global systems; in the functionality – from the trivial data storage to the storage of artificial intelligence. The volume of stored information has increased. According to experts, the total amount of data stored on the Internet doubles every two years. The modern person receives and processes as much information within a month as the person of the XVII century did during a lifetime. Growth of volumes of IS data and increase in their functional capabilities have led to widespread occurrence of large information systems targeted at different categories of users. By "different categories of users" we understand persons belonging to different age groups, different social strata, with different cultural backgrounds and areas of professional interests, etc. We call such systems knowledge-based multidomain information systems (KBMDIS). Such systems can be contrasted with

R. Silhavy et al. (eds.), *Cybernetics and Mathematics Applications in Intelligent Systems*,
Advances in Intelligent Systems and Computing 574, DOI 10.1007/978-3-319-57264-2_29

specialized or problem-oriented IS which aim at solving a limited spectrum of related tasks or providing information support for a single community of users. The examples of multidomain information systems are news sites, resources providing background information on a certain territory or other objects of interest to users of different categories, all sorts of internet portals, and others.

Multidomain systems have specific requirements to the quality of the user interface (UI). They must provide a convenient and user-friendly mechanism of accessing and interpreting information for different categories of users. The set of interface properties which providing its user-friendly character and promoting effective understanding of the information transmitted within the interface is what we call "cognitive properties", the interface which has these properties is called "*cognitive interface*". For different categories of users with differences in their representations of the world (mental models) it is difficult to provide the cognitive properties within a static user interface. The solution is to create a dynamic interface adapted for the current user. Nowadays there already exist some information technologies that solve this problem – Cascading Style Sheets, tag clouds, etc. However, the mentioned technologies solve only the technical problem of generating user interfaces and thus answer the question "How to form the interface?", leaving the question "What should the interface for this user be like?" unanswered. To find a correct answer to this question we need clear criteria of quality (cognitive properties) of the interface and appropriate tools for identification of user mental stereotypes. Cognitive quality of the interface depends on the degree of correspondence between interface and mental stereotypes of users. For a formal definition of this correspondence it is necessary to have a formal representation of stored information. On the other hand, it is also necessary to comprehend and formally represent the regularities of human perception and interpretation of information. Next there will be considered the main concepts associated with the efficiency of information transfer between humans and machines – cognition, relevance, pertinence, and others, on the basis of which it is possible to determine the mentioned appropriate measures. Special techniques of information retrieval on the basis of the user mental model are described in the third part of the article. In the final part of the paper the results of the implementation of the described techniques are given.

2 Definition of Cognitive User Interface

Cognitive interface (CI) can be represented as an interface that provides the correct formation of concepts in the human-computer interaction. In this case "correctness" means the ability to effectively handle these concepts (including the deduction of new knowledge) in order to achieve the user's goals. One of the first models describing the process of human cognition from the standpoint of cognitive processes is a model of early selection by Broadbent [1]. According to this model, information signals come from the outside and undergo filtering in order to be memorized. Initially Broadbent suggested that filtering is carried out on the basis of physical parameters of signals only (volume, tone, colour, etc.) But later J. Gray and A. Wedderburn proved that the channel selection is carried by taking into account the semantics of the incoming information. [2] Thus, the cognitive process involves one

more component which provides semantic analysis of the incoming information. We call this component "semantic analyzer" or "mentality" (Fig. 1).

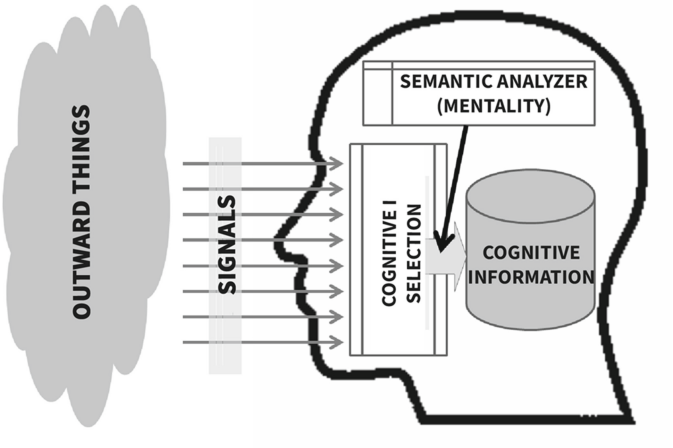


Fig. 1. The model of the cognitive process.

The information system plays the role of the outside world, forming some images to be read by a human within the user interface during the human-computer interaction. These images are formed intentionally as a part of a controlled process. It means that the information system can and should be active and have intrinsic intentionality. We assume that if the signals generated by the system are consistent with the mentality of a human user, then the cognitive process will be more effective – a person can quickly and accurately build correct mental images that constitute the cognitive information. It means that a part of selective functions will be delegated from the human to the information system (Fig. 2). Information system should have a "mental model of user stereotypes" that contains a representation of the user's expectations.

CI - is the interface that implements the functions of the human cognition and provides correct and efficient (in terms of speed) formation of concepts on the basis of the perceived signals. The theoretical limit of this proportion depends on the accuracy of the model of user mentality and the accuracy of the semantic data model of the information system. The correctness of concept interpretation is provided by the correctness of semantics of the interface. The interface should contain relevant components only. The secondary components should be discarded. This, in its turn, is provided by the correctness of the model of mental stereotypes and the correctness of selection procedures. The method of imaging selected information is also important: if the perception stereotypes are taken into account in visualization, then the cognitive process can be significantly accelerated. Thus, two phases of the cognitive interface formation can be identified, and correspondingly two aspects of the UI cognition – perceptual and semantic ones (Fig. 3).

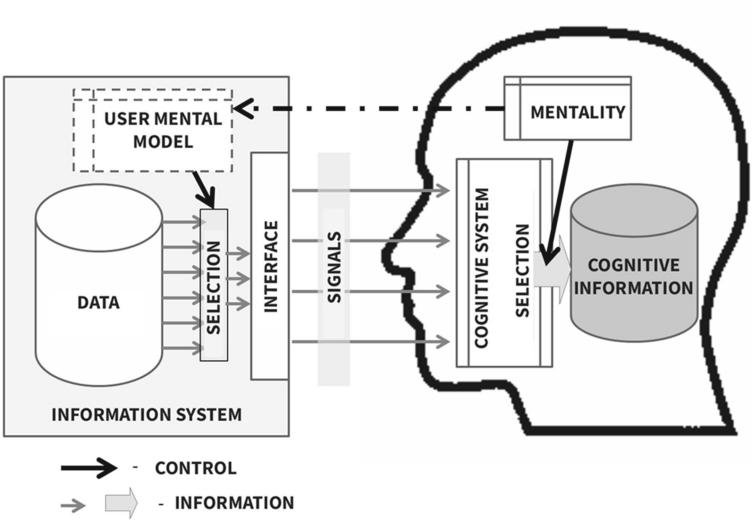


Fig. 2. The model of the cognitive process.

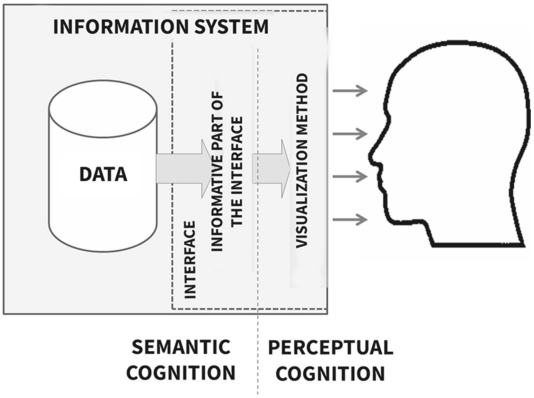


Fig. 3. Two components of the cognitive interface.

The information system can and should form a set of information flows transmitted within the human-machine interface for facilitating the cognitive process, i.e. the identification of significant factors and their combinations for subsequent decision-making. The main question in the synthesis of the cognitive interface is defining the most significant factors for the user. In the case of specialized (problem-oriented system) it is easy to give an answer to this question – the factors influencing the solution to the problem are known beforehand, as well as their usual interpretation and the way of visual presentation and structuring. However, things get complicated if we need to present information from one and the same set to heterogeneous groups of users who solve different tasks and are characterized by different stereotypes of information perception and interpretation [3]. In this case, the system should first identify the user, and then create a

representation of the information that corresponds to the specific character of the task and the subjective characteristics of the user. Experience [4] shows that a high degree of such correspondence ensures efficiency of the interface in terms of speed and accuracy of the information perception.

Users do not always have a complete and clear view of the problem and information required to solve it. The user is able to formulate a complete/accurate search query if their subjective representation is complete and clear. If the user has little idea about what he needs, then he should have an opportunity of choosing from the available information. These are the causes of there being two approaches to the organization of effective man-machine communication – search approach (effective in the first case) and navigation approach (effective in the second case).

To characterize the correspondence of the interface to the users' mental stereotypes we use well-known concepts used when considering information search tasks – relevance and pertinence. Relevance is a measure of correspondence of search results to the task pointed out in the request. We distinguish between substantive and formal relevance. [5] However, is necessary to admit that the query wording itself is a subjective language expression of the user's expectations. Pertinence is a measure of correspondence of the received information to the informational needs of the user [6]. Figure 4 illustrates the difference between these concepts.

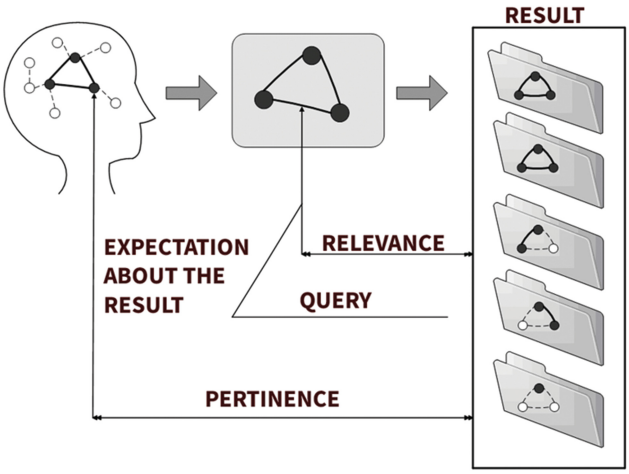


Fig. 4. Relation between the concepts.

It is obvious that getting an objective evaluation of pertinence as a result of computations is impossible due to the impossibility of completely accurate modeling of the user's expectations. Some authors even speak in favor of futility of any attempts to evaluate the quality of search results based on formal evaluations of subjective users' expectations. [7] Thus, one of the key problems is the dynamics of users' expectations. However, we believe it is possible to come close to this estimation. The problem of dynamics of users' expectations can be solved by ensuring permanent monitoring of user activity, and correcting on its basis the model of the user mental stereotypes.

One technique for such monitoring is put forward in [8]. This paper suggests building a model of mental stereotypes of the user in the form of a semantic network which is constantly updated according to user queries and semantic structure of documents entering the focus of the user's attention. Being in the focus of the user's attention is identified by the period of time spent on working with documents which users select from the search results.

The search query is not formulated clearly when the navigation interface is used. However, while surfing the user always has some expectations about the contents of the information system. This allows us to use the same concept of pertinence to evaluate the cognitive properties of the navigation interface. In this case, the result of the search should be understood as a way of structuring information databases presented to the user. And the pertinence can be formally defined as a discrepancy between the structure of information databases and the way of structuring knowledge typical of a particular person. A method of numerical evaluation of such discrepancy was proposed in [3].

3 Application of Mental Models of Users Formed in the Automated Mode to Improve Search Pertinence

The role of the user's individuality both in estimating results and choosing a search pattern is noted in many works [9, 12, 13]. In [9] it is pointed out that users prefer to access information by information retrieval systems (IRS) rather than through the direct navigation. In [10] the correlation between information needs, effectiveness of IRS, experience and characteristics of the user is explained. In [11, 12] the user's involvement in the search process is considered, the concept of «human-computer information retrieval» (HCIR) including various aspects of the human-computer interaction in information retrieval is suggested. In [13] it is shown that taking into account the implicit feedback in the form of the user's behavior when ranking results can increase the search efficiency by 21%. One of the possible ways of obtaining and record-keeping of users' preference is constructing a model of users' interests. For example, in [14] it is proposed to use the user model based on the user questioning to optimize information retrieval of multimedia files. In [15] it is suggested to use an associative network of lexical relations between words for modeling cognitive processes of the user and query optimization. In [4] we proposed a method for automatically receiving users' preferences in the form of a mental model of the user based on considering statistics of the user's interaction with the information system.

As far as information retrieval is concerned, the mental model allows specifying the query context and restricting the search area. The user mental model (UMM) is an associative semantic network, where the set of nodes denotes domain concepts used by the user, and the set of edges denotes a variety of weighted relations between concepts. UMM is generated automatically on the basis of the user's queries and statistics of their work with the information system. The weight coefficient of the relation between concepts increases if related concepts are used together, for example in one query, and it decreases in other cases. Interaction of the user with the information system can be represented by the following algorithm (Fig. 5):

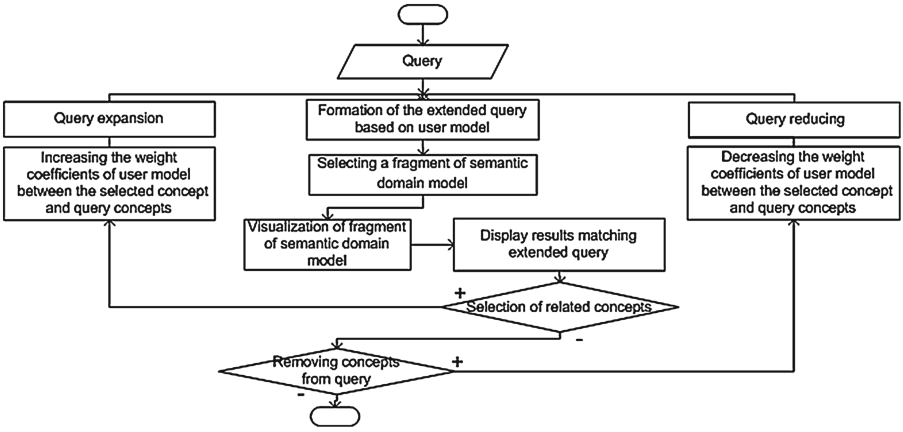


Fig. 5. Algorithm of the user interaction.

The interaction consists in iterative extending the user's queries on the basis of UMM, providing a possibility to correct the query, as well as to take into account the user's preferences by adjusting the weighting coefficients between the concepts of UMM. The weighting coefficients of relations between concepts increase in case of their joint use. The prevalence of one of the concepts of UMM over another is defined by the weight of edges. The peculiarity of this interaction is a possibility to set negative weighting coefficients (subtractive relations), indicating the absence of importance of concepts of this type for this particular user.

The document search on a pre-indexed collection considering weighting coefficients of relations in the mental model of the user includes the following steps:

Query formation in terms of the semantic domain model (SDM). Semantic domain model is a semantic net formed by analyzing and integrating semantic structures of the indexed documents [4]:

Extended query formation. The extended query contains relations and concepts from SDM (Fig. 6).

$$EQ = f_q(Q, KB) = \{C^Q, L^Q | (Eq(c_i^Q, c_j^{KB}) > 1 - \varepsilon)\}, \\ C^Q \subset C, L^Q \subset L\, i = \overline{1, N_Q}, j = \overline{1, N_{KB}} \quad (1)$$

where KB – a semantic domain model, C^q – a set of SDM concepts from the query, L^Q – relations over C^q, $f_q()$ - function that assigns an SDM fragment to the query, $Eq()$ – names similarity evaluation function, ε – concept similarity estimation error.

(a) Query extension based on the weights of relations and limiting contexts of query based on subtractive relations:

$$EQ = \{C^Q, L^Q\} \cup \{C', L' | l{:}c_i \in C^Q, c_j \in C', |\overline{w_k}| > x\}, \\ C' \subset C, L' \subset L, l \in L' \quad (2)$$

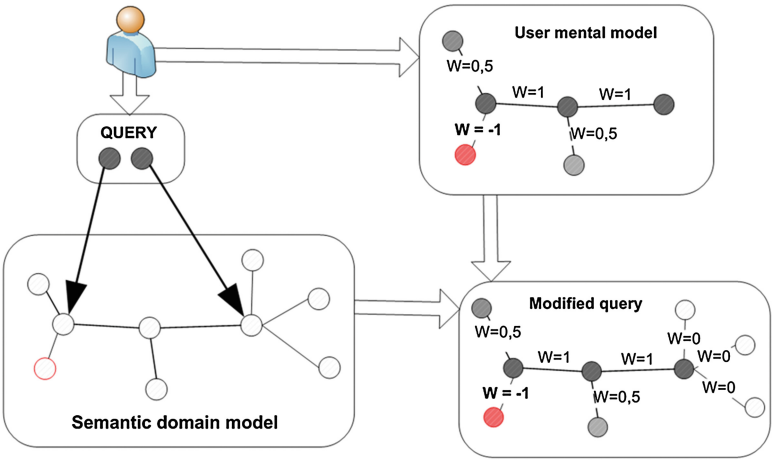


Fig. 6. Extended query formation.

where C' – a set of SDM concepts related to C^Q by L' relations, x – a coefficient of the inclusion of SDM relations in an extended query.

2. Receiving the documents corresponding to the extended query:

$$D = \{d_i | C^{d_i} \cap C^Q \neq \emptyset\}; i = \overline{1,n} \tag{3}$$

where C^{d_i} – a set of SDM concepts from a document d_i, C^Q – a set of SDM concepts from EQ

3. Ranking of the documents:

$$\begin{aligned} R(d_k) &= \sum_{L_{d_k}} (f_u(\bar{w}_k, r) - \sum_{L'_{d_k}} (f_u(\bar{w}_k, r), \\ L_{d_k} &= \{l^d | (c_i, c_j \in d_k) \wedge (tp \in \{synonymOf,\ HyponymOf,\ associateWith\})\} \\ L'_{d_k} &= \{l^d | (c_i, c_j \in d_k) \wedge (tp \in \{subStract\})\},\ i,j = \overline{1,n}, k = \overline{1,m}, tp \in Tp \end{aligned} \tag{4}$$

where $f_u(\bar{w}_k, r)$ – the function returns the weights of relations from the set L_{d_k} for the r^{th} category of users, Tp – a set of relation types.

Thus, if a UMM has subtractive relations between concepts, then documents containing these concepts will have a lower priority after ranking.

The search result is a set of documents ranked in order of descending estimation $R()$. Thus, the use of formal mental models in information retrieval allows taking into account users' preferences formalized in the form of weight coefficients relations over concepts, as well as carrying out automatic ranking of results by taking into account mental models of the user.

4 Evaluation

The search method was evaluated by searching in a pre-indexed collection of 14 thousand documents from problem-oriented IS which was used by different groups of user. The experts estimated search results by ten queries. The criteria for estimation are the search speed and the time taken to satisfy the information needs expressed by one request; accuracy - matching results to the query; and results completeness - coverage of mentioned objects in the search results.

$$Precision = \frac{|D_{rel} \cap D_{retr}|}{|D_{retr}|}, \text{ Recal} = \frac{|D_{rel} \cap D_{retr}|}{|D_{rel}|} \tag{5}$$

where D_{rel} - relevant documents, D_{retr} - results.

The experts used linguistic scale for evaluating the alternatives. Evaluation of *i-th* alternative was made by *j-th* expert by the formula:

$$v_{ij} = 1 - \frac{(l-1)}{k}, \tag{6}$$

where l – linguistic school index values; k – scale values number.

The experts used this formula for alternatives evaluations:

$$s_i = \sum_{j=1}^{n} v_{ij} \tag{7}$$

The evaluation results are shown in Table 1.

Table 1. Evaluations results.

Estimated characteristics	Search method	
	Input line interface	Method described
Search speed estimation	0,4	0,7
Results accuracy	0,9	0,9
Results completeness	0,7	0,9
Ratings average value	0,7	0,9

Experimental results demonstrate that user mental model can be used to implement the information retrieval method.

5 Conclusion

The advanced intelligent technologies can significantly improve the efficiency of information systems. The article suggests a definition of the cognitive user interface, mental model of the user and application of them for the query formation and contextual search.

As a result it can be concluded that the pertinence is more important than relevance. The difference between these concepts is increasing in the case of multidomain information systems that are aimed at meeting the heterogeneous user requirements. Also the problem gets more complicated with the growth of subject area dynamics. Applying the mental model of the user generated with the help of the proposed scheme (based on feedback) improves the search results in terms of pertinence.

References

1. Broadbent, D.E.: Perception and Communication. Elmsford, New York (1958)
2. Gray, J.A., Wedderburn, A.A.I.: Grouping strategies with simultaneous stimuli. Q. J. Exp. Psychol. **12**, 180–184 (1960)
3. Shishaev, M.G., Lomov, P.A., Dikovitsky, V.V.: Formalization problem of constructing cognitive user interfaces for multidomain information resources. Proc. Kola Sci. Centre RAS **4/2013**(17), 90–97 (2013)
4. Dikovitsky, V.V., Shishaev, M.G.: Technology of formation of adaptive user interfaces for multidomain information systems of industrial enterprises. Inf. Resour. Russia **1**, 23–26 (2014)
5. Financial Dictionary. http://dic.academic.ru/contents.nsf/fin_enc/
6. Zhdanova, G.S.: Glossary of terms in computer science at the Russian and English, Moscow (1971)
7. Avetisyan, R.D.: Theoretical Foundations of Computer, Moscow, 168 p (1997)
8. Dikovitsky, V.V.: Methods of intellectual data processing and presentation in multi-subject information systems of industrial enterprises. SPIIRAS Proc. **42**, 56–76 (2015)
9. Liawa, S.: Information retrieval from the World Wide Web: a user-focused approach based on individual experience with search engines. Comput. Hum. Behav. **22**, 501–517 (2006)
10. Al-Maskari, A., Sanderson, M.: A review of factors influencing user satisfaction in information retrieval. J. Am. Soc. Inf. Sci. Technol. **61**(5), 859–868 (2010)
11. Kelly, D.: Methods for evaluating interactive information retrieval systems with users. Found. Trends Inf. Retrieval **3**(1–2), 1–224 (2009)
12. Marchionini, G.: Toward human-computer information retrieval Bulletin. In: Bulletin of the American Society for Information Science, June/July 2006. http://www.asis.org/Bulletin/Jun-06/marchionini.html
13. Agichtein, E., Brill, E., Dumais, S.: Improving web search ranking by incorporating user behavior information. In: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR 2006, New York, NY, USA, pp. 19–26 (2006)
14. Chai, W., Vercoe, B.: Using User models in music information retrieval systems. In: Proceedings of ISMIR (2000). http://ciir.cs.umass.edu/music2000/posters/chai.pdf
15. Wettler, M., Glockner-Rist, A.: Cognitive processes in information retrieval: production rules and lexical nets. In: Mental Models and Human-Computer Interaction, pp. 243–255 (1991)

Implementation of Synthetic Aperture Radar and Geoinformation Technologies in the Complex Monitoring and Managing of the Mining Industry Objects

Maria R. Ponomarenko[1] and Ilya Yu. Pimanov[2(✉)]

[1] St. Petersburg Mining University, Saint Petersburg, Russia
pnmry@yandex.ru
[2] St. Petersburg Institute of Informatics and Automation, Russian Academy of Science, Saint Petersburg, Russia
pimen@list.ru

Abstract. Design, planning and management of opencast and underground mining require safety control of mining operations. Geodynamic monitoring of mining areas is necessary for operational forecasting and prevention of dangerous deformation processes. The identification of geodynamic active zones and forecasting of geodynamic risks are based on systematic observations of the surface and mining facilities. A promising method of obtaining timely spatial information to solve the problems mentioned is the satellite radar imagery. The integration of radar products and intelligent information systems improves the efficiency and accuracy of data analysis. The paper presents the methods of radar image processing in order to conduct comprehensive monitoring of the Earth's surface and infrastructure in mining enterprises. For efficient use of thematic processing products the results were placed on the web server in the information-analytical system "RegionView" providing distributed access to spatial data through the web interface and standard protocols.

Keywords: Synthetic aperture radar · Earth surface monitoring · Interferometry · Geoinformation systems · Temporal data model · Web cartography

1 Introduction

The effective functioning of the mining enterprise requires proactive managing of the mining process. Management of mining operations is performed on the basis of operational information about the state of the objects included in the production. Real-time data acquisition and analysis are necessary for the correct decision-making.

One of the key tasks of management is safety control. Mining processes always lead to large-scale technogenic change of the environment connected with the extraction of minerals, construction of mining facilities, and support of mining operations. Open-pit and underground mining may cause earth surface displacements and deformations of mining objects that pose hazard to life of miners and reduce the economic efficiency of the mining process. In this connection, systematic observations of the surface and facilities located in mining areas are necessary to ensure the safety of mining operations. The

R. Silhavy et al. (eds.), *Cybernetics and Mathematics Applications in Intelligent Systems*, Advances in Intelligent Systems and Computing 574, DOI 10.1007/978-3-319-57264-2_30

organization of these observations refers to the management objectives and protection of natural and engineering structures from the harmful effects of mining. Operational forecasting of dangerous geodynamic processes, timely warning and prevention of deformations and constant control of surface stability reduce damage caused by dangerous effects of deformational processes.

Nowadays, mining companies use a wide range of methods and techniques for geodynamic monitoring: ground-based instrumental observations, aerial survey, laser scanning, radar imagery [1–3]. Traditionally used methods of levelling are labor-intensive and provide a limited number of observed points. Levelling as well as aerial survey and laser scanning also depend on weather conditions. Aerial survey and laser scanning provide high information content of observations and detailed surface models but these methods are characterized by a high cost of the work [4]. The Global Navigation Satellite System (GNSS) technologies allow obtaining precise data regardless of the season and time of day [5, 6]. However, this method is pointwise: measurements are taken directly in the areas of displacements and therefore it is impossible to obtain data in dangerous zones [7].

The problem of surface deformation monitoring is increasingly being solved now using satellite radar imagery. The synthetic aperture radar (SAR) technology has significant advantages in comparison with the above-mentioned methods. SAR operates independently of natural light and cloud cover receiving high spatial resolution data in near real-time and in a wide swath [8]. Above all, the radar can obtain areal data for multiple points located throughout the observed surface including hazardous and remote locations.

All these capabilities make SAR an effective means of complex geodynamic monitoring of the surface [9]. SAR provides a spatial basis for the identification of geodynamic active zones. The method of radar interferometry is used to generate digital elevation models (DEM) [10, 11]. Constructed DEMs are further included in the displacement detection. The monitoring of surface subsidence and deformations of technological objects is performed using displacement maps built on the results of interferometric processing [12–15]. SAR images are also a source of data for large-scale mapping of the mining infrastructure with wide opportunities of change detection of the objects geometry and state [4, 16]. Despite the fact that spaceborne radar technologies are actively used to observe mining areas, the question of analyzing and presenting the results of the SAR data processing and their further implementation in the management process remains open.

High-performance analysis of monitoring data requires use of intelligent information technologies. An effective mean of integration, storage and analysis of the data from different and usually distant sources is the analytical information system [17–19]. In mining information and management systems are now applied mostly for the management of human resources and process operations of plants.

In this article, we introduce an approach to the processing, analysis and presentation of the SAR data for monitoring the Earth's surface and infrastructure in mining enterprises using information analytical system "RegionView".

2 Methods

Geodynamic and deformation monitoring includes observations of surface subsidence in the undermined territories, deformations of buildings and structures, displacements of rock dumps, stability control of ledges and pit walls. These problems can be solved through the integrated application of SAR and GIS technologies.

SAR-based all-weather systematic observations will regularly provide researchers with high precision and operational spatial data of the surface dynamics over a large area of coverage. In addition, archive radar data are used to analyze the rock mass geomechanical state and its temporal variability.

In general, the monitoring system includes complementary ground and space segments. The space segment is based on radar data and includes the following tasks:

- monitoring of stable processes;
- identification of potentially dangerous geodynamic zones;
- supplement and refinement of ground-based observations;
- source data provisioning for the organization of ground-based observations (identifying areas where it is necessary to conduct ground-based observations).

Remote sensing data, in their turn, are supplemented by the results of ground measurements (refined with control points). Subsequently, dangerous areas identified by the radar are monitored with high accuracy and frequency using the methods of field observations. The organization of satellite radar monitoring involves creating a system for:

- acquiring and processing radar data;
- analyzing the processed images;
- presentation of surface dynamics.

Reaching these aims requires development of data-processing techniques. Displacement monitoring requires a set of input data including high-precision spatial basis, terrain model, and measured values of deformations. These data can be obtained by the results of radar data processing.

Digital elevation models (DEM) and surface displacement maps are generated using radar interferometry. Radar interferometry is a direct method of surface deformation determination. The accuracy of the displacement estimation using persistent scatterers interferometry reaches first millimeters in height.

In addition to the monitoring of the Earth's surface, it is necessary to control the state of infrastructure and transport network, as well as the boundaries of the mining allotment in general. The construction and regular updating of large-scale maps and plans are carried out by the results of radar image interpretation. Area measurement focuses on the analysis of amplitude layer in contrast to the displacement monitoring based on the usage of phase component images. The results are used to control the area of mining lease boundaries, infrastructure and transport network, as well as for the calculation of the volume of laid out rocks. This information is useful for the analysis of the dynamics of the opencast mining implementation.

The storage and visualization of measurement results play an important role in the analysis of geodynamic processes. The correct presentation of monitoring data ensures

their efficient use while the ability to view "historical" data allows tracing the change dynamics of the object state.

Continuous updates of data required for monitoring are challenging as large amounts of heterogeneous spatial data need to be stored and visualized. This problem is solved using the temporal data model (TDM) and geoinformation technologies.

3 Approbation

3.1 Study Area

The study area is the central part of the Kola Peninsula, located in the Arctic Circle – the city of Kirovsk, Murmansk region, where apatite-nepheline ore is mined using open-pit and underground technologies. Conducting ground-based measurements in this territory is very difficult due to a small amount of light and cloudless days and the long period of snow cover. The spatial coverage of performed measurements is limited due to the large number of hazardous areas. For these reasons, there is no current spatial data for a large part of the mining lease.

3.2 Data Sources

As the input data we used Sentinel-1 data (Table 1). Sentinel-1 carries C-SAR instrument – radar antenna operating in C-band with a wavelength of ~5.5465763 cm. C-band imagery is less sensitive to the effects of vegetation heterogeneity in comparison with the X-band sensing. Interferometric Mode provides a moderate geometric resolution (5 m by 20 m) and a large swath width (250 km).

Table 1. Parameters of the input Sentinel-1A images.

Parameters	Sentinel-1A
Band	C-band
Wavelength	5.5465763 cm
Mode	Interferometric Mode (IW)
Spatial resolution	5 × 20 m
Swath	250 km
Polarization	VV, VH

Sentinel-1 images were obtained using Sentinels Scientific Data Hub that provides free and open access to all Sentinel missions. Sentinel-1 images were selected based on the data requirements for the interferometric processing (orbital parameters, temporal and spatial baselines, etc.). In addition, the period of sensing was also taken into account: summer season was selected because the territory is covered with snow most of the year.

Images in the visible range, archive topographic maps, and results of field observations were used as additional sources of data to control and refine the processing results.

3.3 Processing Chain

Measurement of the surface displacements is performed on the basis of SAR products. Integration, storage and visualization are based on the information analytical system "RegionView". The processing chain included the following basic steps:

- preparation of images (data selection and download, orbit calibration, etc.);
- interferometric and thematic processing for DEM generation, displacement estimation and obtaining spatial data;
- post-processing: vectorization, generalization, geometry simplification, smoothing, mapmaking;
- storing in the database, visualization of results.

Data processing was carried out in open-source software. Interferometric and thematic processing were performed in the Sentinel 1 Toolbox, phase unwrapping – in Snaphu. QGIS was used for the post-processing. Interferometric processing consisted of the following steps:

1. Coregistration of the images.
2. Range and azimuth filtering.
3. Interferogram formation.
4. Coherence estimation.
5. Topographic phase correction (for the displacements).
6. Interferogram filtering.
7. Phase unwrapping.
8. DEM generation.
9. Displacement estimation.
10. Terrain correction.

Terrain data obtained at the first stage of processing was used for the displacement estimation. With a use of GIS the results of the interferometric processing were transformed into DEM and displacement map. The processing of the amplitude images was performed with the methods of texture analysis, unsupervised classification, composite images, manual and semi-automatic interpretation and included the following steps:

1. Generating RGB composites.
2. Terrain correction.
3. Image unsupervised classification.
4. Object interpretation.

The results of this part of the work is a set of GIS layers for the main classes of terrain objects such as industrial and urban areas, road network, hydrography, and vegetation.

The storage of the resulting data was organized using the TDM. TDM is used to store both input data (RGB representation of satellite images) and the results of thematic processing of satellite images. The key feature of applied TDM is bi-temporality. This means fixing the time of data relevance and the transaction time (the time of data recording into the storage) which greatly facilitates the search of the results related in time as it allows storing the information about the data lifecycle.

Publication of the results was performed on the basis of service "RegionView" developed in SPIIRAS (Fig. 1). RegionView is a modular distributed system comprising server applications, GIS servers and database servers. All components are designed with the use of open-source software and do not require the purchase of paid licenses. The modular principle of construction and the use of standard protocols of data exchange provide flexible placement of the system components. RegionView is a tool with professional features that does not require professional knowledge in geoinformatics, computer and information technology. The system allows data viewing using WMS (Web Map Service) and WFS (Web Feature Service) standard protocols (without transmission and download), co-editing and access control [20, 21].

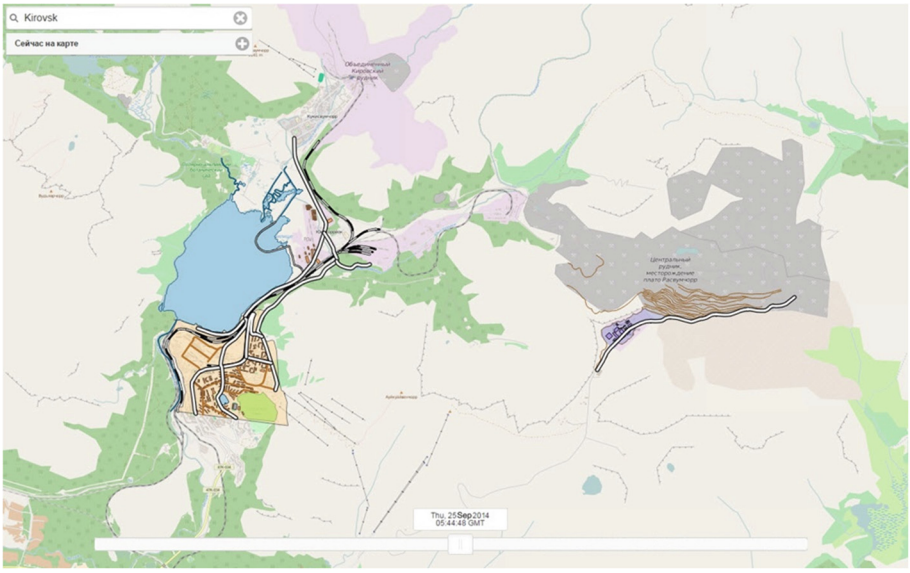


Fig. 1. "Region-View" system interface with loaded results of thematic processing.

3.4 Results

Surface terrain and displacement data were used to study geodynamic processes, identify active areas, monitor the deformations of mining facilities, refine the scheme of geodynamic zoning, assess the geodynamic activity, and predict geodynamic risks in the area of mining operations. The performance assessment requires usage of extended interferometric techniques with a larger amount of test areas.

The results of image composite classification and interpretation provided the spatial information on the mining lease objects that can further be used for the updating the topographic maps. Methods used during thematic processing of amplitude layers improved the accuracy of the interpretation and the reliability of object recognition.

4 Discussions

With this paper, we presented an approach of surface deformation monitoring that integrates SAR data and information analytical technologies. Processing chain was described that allows integration of SAR products into information analytical system. Methods were presented that improve the accuracy of object interpretation.

The results of the survey lead to the conclusion that radar imagery data can be successfully used for the geodynamic monitoring and operational forecasting of dangerous geodynamic phenomena caused by the development of deposits also in semi-automatic mode. The study confirms the feasibility of the development of radar data processing and interpretation using GIS and information analysis systems.

Commonly used methods of deformation monitoring of the study area include ground measurements with furtherer visualization through the sets of computer-aided design (CAD) layers. A first performance assessment indicates that use of information systems can significantly expand the capabilities of monitoring by:

- rapid update of the data by using SAR data;
- historical data acquisition;
- flexible visualization and joint analysis of available spatial data.

Today, SAR products are being actively integrated into operational monitoring and decision support systems to control natural objects and processes, forecast and perform damage control of natural disasters connected with surface displacements such as volcanic eruption, landslides, flood, etc. [22]. Deformations analyzed in this study are technogenic and have different form of appearance. The monitoring of mining objects puts its own requirements to the frequency, spatial coverage, accuracy and visualization of the measurements.

Comparing with latest researches in SAR-based monitoring of mining areas our approach provides extended methods of storing, visualization and analysis of the processing results.

Further development aims at including heterogeneous spatial data as information systems combining various monitoring results (geodetic measurements, remote sensing data) and geoinformation (GIS) technologies in mining provide production safety [23–25].

Acknowledgments. The research described in this paper is partially supported by the Russian Foundation for Basic Research (grants 15-08-08459,16-07-000925, 16-08-00510, 17-08-00797, 17-06-00108, 17-01-00139), supported by Government of Russian Federation, Program STC of Union State "Monitoring-SG" (project 1.4.1-1), state order of the Ministry of Education and Science of the Russian Federation №2.3135.2017/K, state research 0073–2014–0009, 0073–2015–0007.

References

1. Odijk, D., Kenselaar, F., Hanssen, R.: Integration of leveling and InSAR data for land subsidence monitoring. In: Proceedings 11th FIG Symposium, Santorini, Greece (2003)
2. Ogundare, J.O.: Precision Surveying: The Principles and Geomatics Practice. John Wiley & Sons, Inc., Hoboken (2015)
3. Kashnikov, Y.A., Musikhin, V.V., Lyskov, I.A.: Radar interferometry-based determination of ground surface subsidence under mineral mining. J. Min. Sci. **48**, 649–655 (2012)
4. Paradella, W.R., Ferretti, A., Mura, J.C., Colombo, D., Gama, F.F., Tamburini, A., Santos, A.R., Novali, F., Galo, M., Camargo, P.O., Silva, A.Q., Silva, G.G., Silva, A., Gomes, L.L.: Mapping surface deformation in open pit iron mines of Carajás Province (Amazon Region) using an integrated SAR analysis. Eng. Geol. **193**, 61–78 (2015)
5. Chen, G., Cheng, X., Chen, W., Li, X., Chen, L.: GPS-based slope monitoring systems and their applications in transition mining from open-pit to underground. Int. J. Min. Min. Eng. **5**(2), 152–163 (2014)
6. Panzhin, A.A., Panzhina, N.A.: Satellite geodesy-aided geodynamic monitoring in mineral mining in the Urals. J. Min. Sci. **48**, 982–989 (2012)
7. Mazzanti, P.: Remote monitoring of deformation. an overview of the seven methods described in previous GINs. Geotech. Instrum. **30**, 24–29 (2012)
8. Hanssen, R.F.: Radar Interferometry: Data Interpretation and Error Analysis, 308 p. Kluwer Academic Publishers, Dordrecht (2001)
9. Rheault, M., Bouroubi, Y., Sarago, V., Nguyen-Xuan, P.T., Bugnet, P., Gosselin, C., Benoit, M.: Integrated SAR technologies for monitoring the stability of mine sites: application using TerraSAR-X and RADARSAT-2 Images. Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci. **XL-7/W3**, 1057–1062 (2015)
10. Crosetto, M.: Calibration and validation of SAR interferometry for DEM generation. ISPRS J. Photogramm. Remote Sens. **57**(3), 213–227 (2002)
11. Mirzaee, S., Motagh, M., Arefi, H.: Assessment of referance height models on quality of TanDEM-X DEM. ISPRS Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci. **XL-1-W5**, 463–466 (2015)
12. Wegmuller, U., Walter, D., Spreckels, V., Werner, C.L.: Nonuniform ground motion monitoring with TerraSAR-X persistent scatterer interferometry. IEEE Trans. Geosci. Remote Sens. **48**(2), 895–904 (2010)
13. Zhang, H., Wang, C., Tang, Y.: Subsidence monitoring in coal area using timeseries InSAR combining persistent scatterers and distributed scatterers. Int. J. Appl. Earth Obs. Geoinf. **39**, 49–55 (2015)
14. Jiang, L., Lin, H., Ma, J., Kong, B., Wang, Y.: Potential of small-baseline SAR interferometry for monitoring land subsidence related to underground coal fires: Wuda (Northern China) case study. Remote Sens. Environ. **115**, 257–268 (2012)
15. Du, Z., Ge, L., Li, X., Ng, A.H.-M.: Subsidence monitoring over the Southern Coalfield, Australia using both L-Band and C-Band SAR time series analysis. Remote Sens. **8**, 543 (2016)
16. Ponomarenko, M.R., Pimanov, I.Yu.: Processing of SAR amplitude images with posting the results on web server. J. Sib. Fed. Univ. Eng. Technol. **9**(7), 994–1000 (2016). doi: 10.17516/1999-494X-2016-9-7-994-1000
17. Merkuryeva, G., Merkuryev, Y., Sokolov, B.V., Potryasaev, S., Zelentsov, V.A., Lektauers, A.: Advanced river flood monitoring, modelling and forecasting. J. Comput. Sci. **10**, 77–85 (2015). doi:10.1016/j.jocs.2014.10.004

18. Sokolov, B.V., Zelentsov, V.A., Brovkina, O., et al.: Intelligent integrated decision support systems for territory management. Artif. Intell. Perspect. Appl. **347**, 321–331 (2015). doi: 10.1007/978-3-319-18476-0_32
19. Zelentsov, V.A., Krylenko, I.N., Pimanov, I.Yu., Potryasaev, S.A., Sokolov, B.V., Akhtman, Y.: Bases of construction of the remote sensing data processing, storage and visualization system based on service-oriented architecture. Izv. vuzov Pribiristroenie. **58**(3), 241–243 (2015) doi:10.17586/0021-3454-2015-58-3-241-243
20. Mochalov, V.F., Markov, A.V., Grigorieva, O.V., Zhukov, D.V., Brovkina, O.V., Pimanov, I.Yu.: Remote sensing for environmental monitoring. complex modeling. In: Silhavy, R., Senkerik, R., Oplatkova, Z.K., Silhavy, P., Prokopova, Z. (eds.) Automation Control Theory Perspectives in Intelligent Systems. AISC, vol. 466, pp. 497–506. Springer, Cham (2016). doi:10.1007/978-3-319-33389-2_47
21. Zelentsov, V.A., Potryasaev, S.A., Pimanov, I.J., Nemykin, S.A.: Creation of intelligent information flood forecasting systems based on service oriented architecture. In: Silhavy, R., Senkerik, R., Oplatkova, Z.K., Silhavy, P., Prokopova, Z. (eds.) Automation Control Theory Perspectives in Intelligent Systems. AISC, vol. 466, pp. 371–381. Springer, Cham (2016). doi:10.1007/978-3-319-33389-2_35
22. Meyer, F.J., McAlpin, D.B., Gong, W., Ajadi, O., Arko, S., Webley, P.W., Dehn, J.: Integrating SAR and derived products into operational volcano monitoring and decision support systems. ISPRS J. Photogramm. Remote Sens. **100**, 106–117 (2015)
23. Hannemann, W., Brock, T., Busch, W.: GIS for combined storage and analysis of data from terrestrial and synthetic aperture radar remote sensing deformation measurements in hard coal mining. Int. J. Coal Geol. **86**, 54–57 (2011)
24. Shanjun, M., Qiaoxi, L., Mei, L.: Design and development of safety production management information system based on a digital coalmine. Procedia Earth Planet. **1**(1), 1121–1127 (2009)
25. Blachowski, J.B., Milczarek, W., Stefaniak, P.: Deformation information system for facilitating studies of mining-ground deformations, development, and applications. Nat. Hazards Earth Syst. Sci. **14**, 1677–1689 (2014)

Lightning Impulse Voltage Evaluation

Nopphadon Khodpun and Krisada Vilailak(✉)

Electrical Engineering Branch, Faculty of Engineering, Vongchavalitkul University, Nakhonratchasima, Thailand
{noppadon_kho,krisada_vil}@vu.ac.th

Abstract. This research studied and developed software for lightning impulse voltage parameters evaluation. Full lightning impulse voltage of 9 cases in TDG program had been used as references for software tested. This software created 2 types of voltage waveform which are mean curve and approximate real curve. Kalman Filter had been used for mean curve. QR algorithm had been used for approximate real curve. The experimental results show that the purported software can evaluate all lightning impulse parameters as IEC 61083-2 standards in every case.

Keywords: Lightning impulse voltage evaluation · Kalman filter · QR algorithm

1 Introduction

Testing electrical devices, which must be set up in high voltage system outdoor, needs to be evaluated with different types of electrical voltages in order to meet the IEC 60060-1 standard (IEC 1989), IEC 60060-2 standard (IEC 1994). Types and sizes of voltage used for testing insulator level or highest voltage of equipment were determined. For impulse voltage was determined as

- Lightning Impulse Voltage
- Switching Impulse Voltage

Parameter value was focused, as it is a feature of impulse voltage, such as peak value, front time, time to half, etc. Up to now, impulse voltage evaluation has been improved during disturbances, and a computer has been used for calculating parameter value as it is convenient, fast and accurate compared to other evaluations at that time. Wave features were imitated via a computer, providing different features in accordance with different math equations. This study developed a program evaluating parameter of lightning voltage impulse, applying Kalman filter and QR algorithm which is an approach for analyzing Eigen problem, then comparing accuracy. Lightning voltage impulse features referred to TDG program as IEC 61083-2 standard (IEC 1996).

2 Literature Review

2.1 Analyzing Lightning Voltage Impulse Without Disturbances

Lightning Voltage Impulse can be evaluated as for parameter value of wave feature as the following (Fig. 1)

R. Silhavy et al. (eds.), *Cybernetics and Mathematics Applications in Intelligent Systems*, Advances in Intelligent Systems and Computing 574, DOI 10.1007/978-3-319-57264-2_31

– Peak value is an actual value of voltage waveform
– Front time, $T_1 = 1.67*T$ as T is a difference of time that voltage is 30% of the peak; time at the voltage is 90% of the peak; and the point that a line dragged through 2 lines to cross was called virtual origin, '0'
– Time to half value, T_2 is time from is the time from virtual origin through the peak, and the voltage is 50% of the peak.

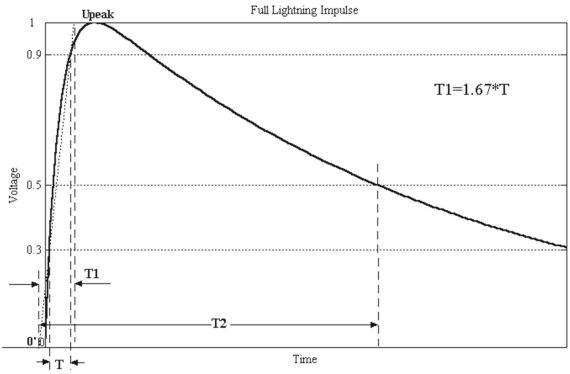


Fig. 1. Parameter of lightning voltage impulse according to IEC 60060-1

2.2 Analyzing Lightning Voltage Impulse with Disturbances

According to IEC 61083-2, with TDG, disturbance criteria is as the following

– Noise is less than 1% of the peak of voltage impulse
– Main frequency of vibrate on lightning voltage impulse is more than 500 kHz
– The total of vibrate and overshoot are less than 5% of the waveform peak
– The time of overshoot is less than 1 μs.

However, IEC 61083-2 determined parameter evaluation of lightning voltage impulse that in case the voltage meets the standard, calculate parameter from average voltage (case (a), (b)) and if the voltage does not meet the standard, calculate parameter from an actual voltage (case (c), (d)) (Fig. 2).

2.3 Test Data Generator Program (TDG)

TDG or Test Data Generator is a program with IEC 61083-2 (IEC 1996), generating feature in order to test with the program developed in some labs. 15 cases of voltage feature (IEC 61083-2 determining a correct parameter value for each case) consisted of lightning voltage impulse (1.2/50 μs) as of 9 cases, lightning voltage impulse ($T_1 \approx$ 0.5 μs) as of 3 cases, lightning voltage impulse (250/2500 μs) as of 2 cases, and lightning voltage impulse (8/20 μs) as of 1 case. This research studied only on 9 cases of lightning voltage impulse such as feature in case 1, 3, 4, 6, 8, 9, 11, 13 and 14, respectively.

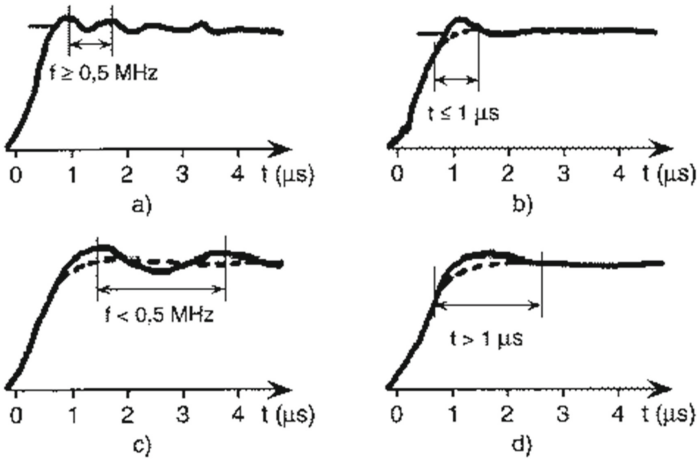


Fig. 2. Lightning voltage impulse with disturbances parameters

This research selected sampling rate as 100 MHz (duration time between data point is as of 10 ns) according to IEC 61083-1 (IEC 1991), and 10,000 points of data points (the number of points before actual data is about 10%) in order to illustrate back wave up to 50% of the peak (reference waveform of case 11 shows the time of half peak ≈ 90 μs).

2.4 Literature Review

IEC 60060-1 determined standard of lightning voltage impulse as 1.2/50 μs from

$$M(t) = A\left(e^{\alpha(t-t_0)} - e^{\beta(t-t_0)}\right) \tag{1}$$

Previous studies determined standard of lightning voltage impulse, which parameter in the Eq. (1) is A=1.03*U_{peak}, $\alpha = -1/(68.5\ \mu s)$ and $\beta = -1/(0.405\ \mu s)$ for evaluating lightning voltage impulse ((Perez et al. (1995); Wong et al. (1999); Brede et al. (1999) and Boaventura (2003)).

2.4.1 Kalman Filter

Kalman filter is a digital filter used for evaluating state vector from status in both continuous time and discrete time. Moreover, it is a linear approach from status equation of discrete time.

Perez et al. (1995) (1996) suggested Kalman filter for evaluating lightning voltage impulse by using model equation of the impulse as

$$M(t) = A_0\left(e^{-a(t+t_0)} - Be^{-b(t+t_0)^d}\right) \tag{2}$$

This method was tested with all TDG, and it found that the value met the standard except waveform in case 14, which got the peak and time of front wave that did not meet the standard.

3 Methodology

3.1 Initial Value Calculated for Average Waveform Simulation

From the Eq. (1) considering back wave, when $\left(-\mathrm{Ae}^{\beta(t-t_0)}\right)$ approach 0, it will show the equation back wave as

$$M(t) = \mathrm{Ae}^{\alpha(t-t_0)} \tag{3}$$

And $A_{initial}$ can be calculated from the equation of A_0 at the time $t = t_0$

$$\therefore A_{\text{initial}} = A_0 e^{\alpha_{\text{initial}}(t_0 - T_0)} \tag{4}$$

After getting $A_{initial}$ and $\alpha_{initial}$, it can calculate $\beta_{initial}$ from the equation

$$\therefore \beta_{\text{initial}} = \frac{-1}{n} \sum_{i=1}^{n} \text{abs}(\beta_i) \tag{5}$$

Therefore, it will have $A_{initial}$, $\alpha_{initial}$, $\beta_{initial}$ and t_0 as default for evaluating parameter of lightning voltage impulse, generating average waveform with Kalman filter.

3.2 Average Waveform Simulation

This research used model equation of lightning voltage impulse as

$$M(t) = \mathrm{Ae}^{\alpha(t-t_{01})} - \mathrm{Be}^{\beta(t-t_{02})} \tag{6}$$

The method of generating average waveform started with back wave first, then the waveform giving different result from the actual waveform, and back wave will be put together with front wave. Then, put the two parts together as a full model of lightning voltage impulse. Generating average waveform used Kalman filter, supposing that fixed value in the Eq. (6) such as A, B, α, β, t_{01} and t_{02} are random variable x(t), determining the two equations as the following

$$x_{k+1} = \Phi x_k \text{ and } z_k = M(x_k, t_k) + N_k$$

make improved equations

$$P_k = (I - K_k H_k) P_k^- \tag{7}$$

$$H_k = \left[\frac{\partial h}{\partial x}\right]_{x_k = x_k} \tag{8}$$

Then, adjust x_k and P_k to have value for further round, and make variance (Matrix P) reduce in each round of calculation until it was less than the criteria. Thus, it will get proper function value (Fig. 3).

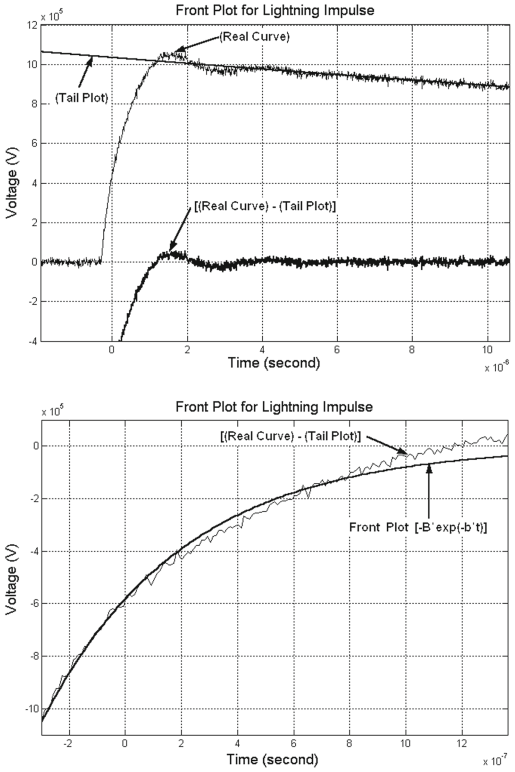


Fig. 3. Lightning impulse waveform, back wave and front wave (reference waveform case 8)

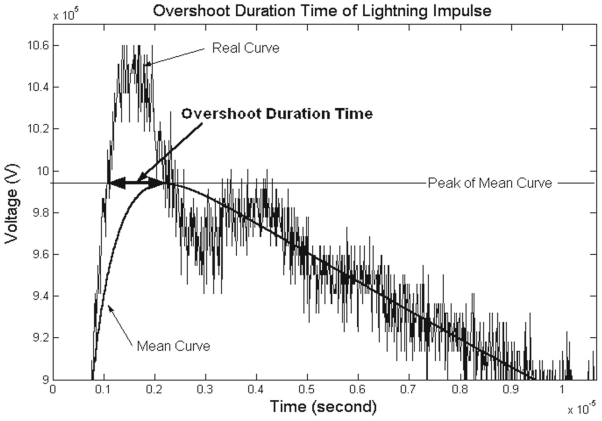


Fig. 4. Evaluating time of overshoot (reference waveform, cases 8 peak)

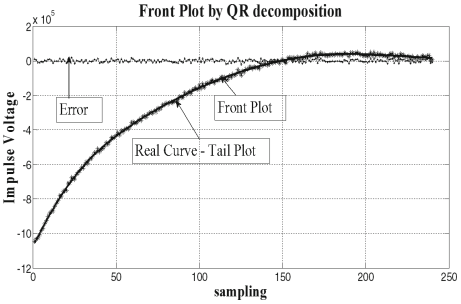


Fig. 5. Simulated waveform for actual waveform using QR

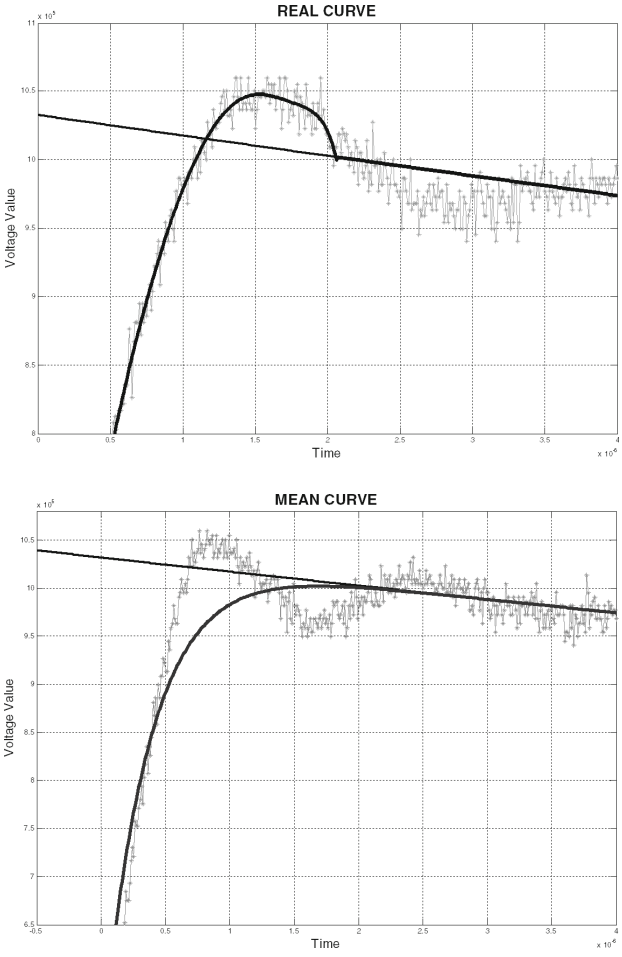


Fig. 6. The result of program testing and reference waveform of case 8 and 9 (peak)

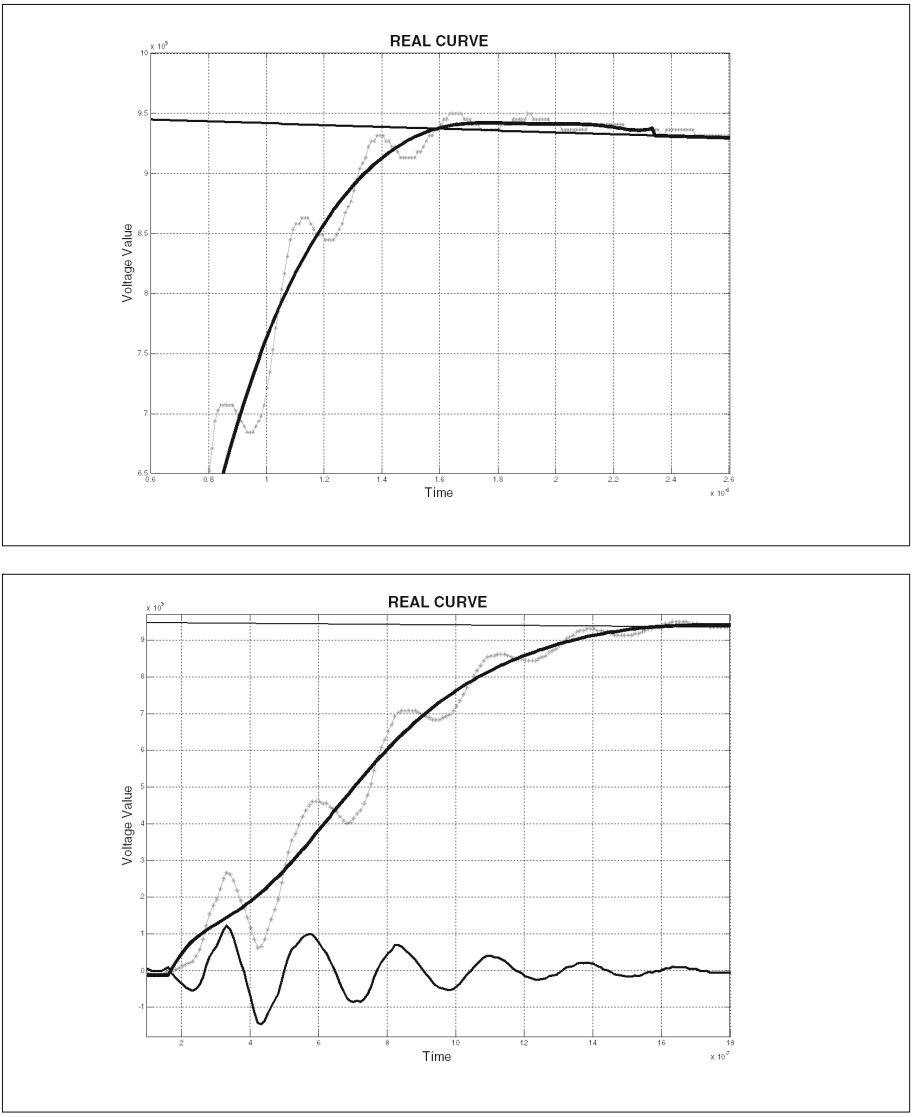


Fig. 7. The result of program testing and reference waveform of case 11 (Peak and front wave)

3.3 Analyzing Residue Waveform

Residue waveform is a difference of waveform between actual waveform and simulating average waveform. Analyzing residue waveform for evaluating parameter of disturbance, it found that residue waveform compose of noise, vibrate and overshoot.

3.3.1 Evaluating Frequency of Vibrate

Analyzing residue for evaluating frequency of vibrate started with creating back wave of lightning voltage impulse, and got residue waveform which was a difference of actual waveform and simulated waveform as Fig. 8. Then, selected the starting point of data from voltage of residue waveform as 0 on the right-hand side for 512 points and evaluated the size of frequency occurred. The biggest frequency was considered as 390 kHz

3.3.2 Evaluating Overshoot

Ganacho et al. (1997) suggested the method of evaluating time that has overshoot. overshoot is the time that voltage of actual waveform has more voltage than peak of average waveform as Fig. 4.

3.4 Simulating Actual Waveform

Simulating actual waveform for evaluating parameter of this study used QR (QR algorithm) to apply for calculating linear equation

$$\begin{aligned} & A = \mathrm{QR} \\ \text{when } & Q = \text{ orthogonal matrix } (Q^T Q = I) \\ & R = \text{ upper matrix and the result in diagonal of eigenvalue of A} \end{aligned} \tag{9}$$

So, the result of equation is as

$$x = R^{-1} Q^T b \tag{10}$$

When putting coefficient of polynomial, it will get waveform for actual waveform as Fig. 5.

4 Results and Discussions

4.1 Results

The program evaluating lightning voltage impulse in this study did accuracy test with 9 cases of waveform of lightning voltage impulse such as 1, 3, 4, 6, 8, 9, 11, 13 and 14. The result is shown in Table 1.

4.2 Result Graphs from Impulse Voltage Waveform Program

See Figs. 6 and 7.

4.3 Discussion

Comparing the result with 9 cases of reference waveform found that parameter value is in the scope of standard, which means that the developed program is accurate. However,

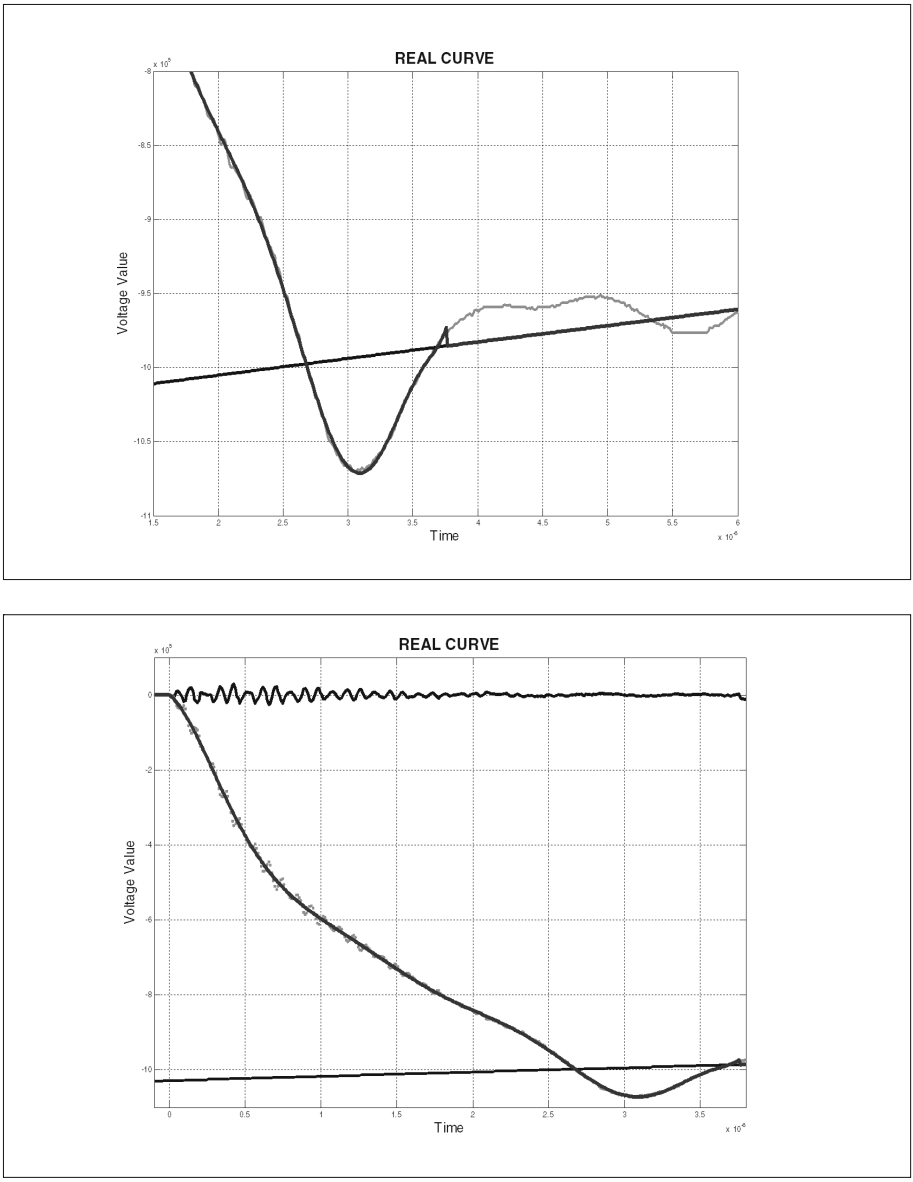


Fig. 8. The result of program testing and reference waveform of case 13 (Peak and front wave)

for case 13, the program evaluated frequency of vibrate first and the frequency did not meet the standard. So, the program evaluated parameter of waveform from actual waveform immediately, in which the result was not different.

Table 1. Showing parameter evaluated from the program, comparing to standard in each case

Reference wave form	Boundary standard parameters					Result parameters				
	Peak (MV)	T_1 (μs)	T_2 (μs)	F_O (kHz)	T_O (μs)	Peak (MV)	T_1 (μs)	T_2 (μs)	F_O (kHz)	T_O (μs)
1	1.04–1.06	0.81–0.87	57.5–62.5	–	–	1.051	0.84	60.3		
3	1.04–1.06	1.6–1.7	45–49	<500	–	1.052	1.66	47.9	390.6	
4	0.96–0.99	1.0–1.1	48–52	>500	–	0.976	1.05	49.7	585.9	0.64
6	1.04–1.06	0.81–0.87	57.5–62.5	–	–	1.050	0.85	60.1		
8	1.04–1.06	1.6–1.7	45–49	<500	–	1.048	1.69	46.4	390.6	
9	0.96–0.99	1.0–1.1	48–52	>500	–	0.976	1.094	49.6	585.9	
11	0.94–0.96	1.07–1.19	82–91	–	>1	0.953	1.128	87.1	–	1.05
13	−1.08−1.06	3.40–3.76	56–62	–	>1	−1.072	3.57	58.7	390.6	–
14	−0.97−0.95	1.85–2.05	43–47	–	<1	−0.965	1.889	45.2	585.9	0.91

5 Conclusion

From the study, equation model can create voltage impulse well from the basis of double exponential function. However, setting default for calculating lightning voltage impulse standard 1.2/50 μs ($A=1.03*U_{peak}$, $\alpha = -1/(68.5\ \mu s)$ and $\beta = 1/(0.405\ \mu s)$) caused error since standard scope of parameter of reference waveform from TDG in each case contains a lot of difference.

The result of the program evaluating lightning voltage impulse and reference waveform from TDG in 9 cases of lightning voltage impulse shows that the result of the program met the standard. So, it means that algorithm and the method using in the research is accurate and appropriate for lightning voltage impulse evaluation.

Acknowledgement. Without the research presentation fund of Vongchavalitkul University, this research would not have been possible.

References

Boaventura, W.C.: Modelling impulse voltage test waveforms using time-domain fitting based on Prony's method. Int. Symp. High Voltage Eng., 253–258 (2003)

Brede, A.P., Werle, P., Gockenbach, E., Borsi, H.: A new method of determining the mean curve of lightning impulses according to IEC 60060-1. Int. Symp. High Voltage Eng. **467**, 74–77 (1999)

Ganacho, F., et al.: Evaluation procedures for lightning impulse parameter in case of waveforms with oscillations and/or overshoot. IEEE Trans. Power Delivery **12**(2), 640–649 (1997)

International Electrotechnical Commission: High – Voltage Test Techniques Part 1: General Definitions and Test Requirements, 2nd edn. IEC std. 60060-1, 11 1989

International Electrotechnical Commission: High – Voltage Test Techniques Part 2: Measuring System, 2nd edn. IEC std. 60060-2, 11 1994

International Electrotechnical Commission: Evaluation of software used for the determination of the parameters of impulse waveforms: Digital recorders for measurements in high - voltage impulse tests – Part 2. IEC std. 61083-2 (1996)

Perez, J., Martinez, J.: Kalman filter algorithm for digitally recorded lightning impulse parameter evaluation. IEEE Trans. Power Delivery **10**(4), 1713–1719 (1995)
Perez, J., Martinez, J.: Digitally recorded lightning impulse with overshoot parameter evaluation by using Kalman filtering method. IEEE Trans. Power Delivery **11**(4), 1005–1014 (1996)
Wong, K.C.P., et al.: Digital measurement of lightning impulse parameters using curving fitting algorithm. Int. Symp. High Voltage Eng. **467**, 193–196 (1999)

Pattern Recognition for Predictive Analysis in Automotive Industry

Veronika Simoncicova, Lukas Hrcka(✉), Lukas Spendla, Pavol Tanuska, and Pavel Vazan

Faculty of Materials Science and Technology in Trnava, Institute of Applied Informatics, Automation and Mechatronics, Slovak University of Technology, Trnava, Slovak Republic
{veronika.simoncicova,lukas.hrcka,lukas.spendla, pavol.tanuska,pavel.vazan}@stuba.sk

Abstract. Predictive maintenance (PdM) techniques are designed to help identify the condition of devices in order to predict when maintenance should be performed. The ultimate goal of PdM is to perform maintenance at a scheduled point in time when the maintenance activity is most cost-effective and before the equipment loses performance within a threshold. Currently, reducing service costs and losses due to downtime is one of the ways to increase your profits and success in the market. We tried to identify problem messages and failures from the manufacturing data example set from car body work. Two different data sets were joined and we designed a process to identify message and failure alerts preceding errors.

Keywords: Predictive maintenance · Failures · Analysis · Error

1 Introduction

In today's modern, fast changing and competitive business environment, companies often face problems of fast and right decision-making, which would improve their competition on the global market and would lead to increased production. Each industrial company is provided with equipment, robots and devices used to achieve the goal. Keeping continuity of the process without long and unexpected failures caused by unprofitable devices is a huge problem in a great industry company possibly resulting in a financial loss of the company. The solution to the problems can be achieved if failures and machine outages are minimized on the behalf of planned maintenance.

Poorly maintained equipment and devices may lead to more frequent equipment and device failures, poor utilization of equipment and devices and delayed production schedules. In industrial environment, maintenance activities are typically intended to reduce failures of machinery creating condition for increasing device availability and consequently increasing productivity. Production data collected in real-time contains valuable information and knowledge that could be integrated within prediction systems to improve decision making and enhance productivity [1].

Every day, maintenance planning faces the challenge of how to ensure the maximum availability of machines while simultaneously minimizing material consumption for

R. Silhavy et al. (eds.), *Cybernetics and Mathematics Applications in Intelligent Systems*, Advances in Intelligent Systems and Computing 574, DOI 10.1007/978-3-319-57264-2_32

maintenance and repairs – all with an eye to guaranteeing product quality. The traditional maintenance strategy in the company is based on operative maintenance of manufacturing devices after problems arise or after a device stops working properly resulting in losing time and money. New technology, sophisticated tools and model influence the efficiency of maintenance means and techniques enabling failure prediction before their inception. Therefore, preventive maintenance method was introduced utilising new software and technology. On the basis of empiric experience and knowledge, the devices were repaired in appropriately designed intervals having impact on the maintenance method efficiency and achieving decreased counts of device failures.

Nowadays advanced technology and analytical tools were needed proposing more sophisticated maintenance methods to identify and evaluate several measured values and consequently, to predict, when the maintenance of devices is needed. Predictive maintenance (PdM) techniques offer a viable solution to this dilemma.

Many authors have described different strategies and methods for maintenance management. Batenam [2] described three basic types of maintenance programs, including reactive, preventive and predictive maintenance. Preventive and predictive maintenance represent two proactive strategies utilised by companies to avoid equipment errors.

Since Barlow and Hunter [3] proposed the minimal repair model in 1960, a lot of optimal maintenance strategies have been developed and implemented for improving system reliability, preventing system failures and reducing maintenance costs.

In many cases, the preventive maintenance philosophy, and at times, even a less sophisticated predictive maintenance program is adopted for the equipment. These essential machines do not need to have the same monitoring instrumentation requirements as critical machines [4].

Big manufacturing companies currently collect, store and process large quantities of data not only from control level of manufacturing process but also from higher levels of hierarchical control. These data represent ideal baseline for implementation of predictive maintenance due to the fact that data from lower level of hierarchical control is of greatest importance for predictive maintenance. This trend closely related to integration of Industry 4.0 concept, including Internet of Things technologies, into manufacturing, which serves as ideal basis for data analysis of any kind, including predictive maintenance [5].

Predictive maintenance consists. of several steps [6]:

- Trend analysis: Reviews the data to find if the asset being monitored is on an obvious and immediate downward slide toward failure. Typically, a minimum of three monitoring points are recommended for arriving at a trend accurately as well as a reliable measure is necessary to find out if the condition is deprecating linearly.
- Pattern recognition: Decodes the causal relations between certain type of events and machine failures. For example, after being used for a certain product run, one of the components used in the asset fails due to stresses unique to that run.
- Critical range and limits: Tests to verify if a set of data is within a critical range limit (set by professional experience). However, machine learning schemes can be adopted to eliminate user intuition for setting these limits.

- Statistical process analysis: Existing failure record data (retrieved from warranty claims, data archives and case-study histories) is driven through analytical procedures to find an accurate model for the failure curves and the new data is compared against those models to identify any potential failures.

In our paper, we examine trend analysis pattern recognition using Rapid miner, one of the most used open source predictive analytics platform for data analysis. It is accessible as a stand-alone application for information investigation and as a data mining engine for the integration into own products. Rapid miner provides an integrated environment for data mining and machine learning procedures, including [7]:

- extracting the data from different source systems; transforming the data and loading into a data warehouse (DW) or data repository of other applications,
- data pre-processing and visualization,
- predictive analytics and statistical modelling, evaluation, and deployment.

Providing learning schemes, models and algorithms from WEKA and R scripts makes it even more powerful [7].

Rapid miner provides a graphical user interface (GUI) to design and execute analytical workflows. Those workflows are called Processes and consist of multiple Operators. A graphical user Interface (GUI) allows connecting the operators with each other in the process view. Each independent operator carries out one task within the process and forms an input for another operator in a workflow. Analysing data retrieved at the beginning of the process is the major function of a process [7].

Rapid miner offers large amount of different operators, which can be easily extend with existing extensions. There are packages for text processing, web mining, weka extensions, R scripting, series extension, Python scripting, anomaly detection and more [7].

2 Chapter Materials and Methodology of Experiment

The main goal of this paper is to propose a methodology of joining data, analysing and acquiring information to be used for predictive maintenance in an industrial company. The first step is export from databases SAP PM and SEGA. System SAP PM contains data from car body work written in by employees. Whenever failure occurs in the production and the failure affects the production process, the employee writes the failures in to system SAP PM.

Employees record three kinds of failures:

- significant failures, which stop the manufacturing,
- minor failures, which do not stop the manufacturing
- and preventive maintenance.

The SEGA system contains records about incorrect statuses of robot, which are loading automatically from the robots and technologies. The messages contain information regarding the settings of processes. Two groups of data are loaded into system SEGA:

- messages logs not influencing the operations of robots and
- significant and minor error logs.

Minor error is causing reset of the system and manufacture is not influented. Significant error can caused stop the manufacture. The gained data was then loaded to the software Rapidminer, which is a data science software platform developed by the company of the same name that provides an integrated environment for machine learning, deep learning, text mining, and predictive analytics.

The next important part is pre-processing data, where data are modificated using necessary operators in Rapidminer. The exact description of the modification data is explained in my previous publication "Data Pre-processing from Production Processes for Analysis in Automotive Industry" [8].

Following the initial analysis of data obtained from SAP PM, we identified the most common errors reported by robots in 2015. We selected a specific failure and based on an expert advice of maintenance. The most problematic part consists of robots that are responsible for tightening the screws so-called FDS heads. FDS head is a device which is fixed to the robot arm (Fig. 1). This device automatically sends errors logs and message logs.

ExampleSet (546 examples, 0 special attributes, 3 regular attributes)

Row No.	Názov vybavenia	Názov technického miesta	count(Názov vybave... ↓
28	125110R10 FDS skrut. hlava G	VW-BA-H04-H04A-A1A2	48
373	216640R06 FDS skrut. hlava G	VW-BA-H04-H04A-A5A4	39
49	125310R04 FDS skrut. hlava G	VW-BA-H04-H04A-A2A2	34
530	ST 225340	VW-BA-H04-H04A-A2A2	34
488	ST 135420	VW-BA-H04-H04A-A3A1	26

Fig. 1. The list of failures

Robots entered a huge amount of logs from which I can not accurately identify significant errors. We eliminated this problem by combining with a set of SAP PM. Employees write errors with a certain time delay while the robots entered errors accurate to the second, so it was necessary to create an algorithm written in Python, which calls haystacksearch, where needles are represented by failures reported and repaired by the employees and haystack is mirroring the errors from the event logs. The processing script based on the pandas and numpy python libraries supported by RapidMiner as well, allowed us to test and to use parts of the code in the analysis pipeline.

At first the reported failures had to be processed and grouped together, since we needed multiple occurrences of the same failure to find correlations in the errors from the event logs. Not all reported errors could have been directly matched, as they were written into the report by the employees. Therefore, we had to use regular expressions and partial searches, to identify the occurrences of the same errors serving as the basis for the logged error search.

Between the reported failures and logged errors no connections existed, except for the time of the occurrence. Since there are great disproportions between reported and

real failure times, we could not pair these times directly. Therefore, we had to search the reported error times in backward direction. We also used the error location and failed component if available to search only the relevant errors in the error log. Part of the algorithm is depicted Fig. 2. Using this algorithm, we clearly identified the times when a given failure occurred and thus, we were able to link both datasets.

```
dfNeedles["Anlage"] = dfNeedles.apply(lambda row: TechPoint2Anlage(row["Názov technického miesta"]), axis=1)

reEquipNamePatter = re.compile("(1|2|3|4|5)\d{5}R\d{2}|ST\s\d{6}|ARG\d")
def EquipName2Equip(sEquipName):   #Parses Device out of Alart Text column
    mgEquip = reEquipNamePatter.search(sEquipName)
    if mgEquip is None:
        return numpy.NaN
    sEquipDesc = mgEquip.group(0)
    if len(sEquipDesc) < 4:
        return numpy.NaN
    return sEquipDesc

dfNeedles["Equip"] = dfNeedles.apply(lambda row: EquipName2Equip(row["Názov vybavenia"]), axis=1)
```

Fig. 2. The part of the algorithm code

After joining the two data sets, the process was modelled to determine the specific error messages. Consequently, by linking the two data sets, we detected serious mistakes in the data exported from the SEGA system if compared to the SAP PM data. Part of the process you can see in Fig. 3.

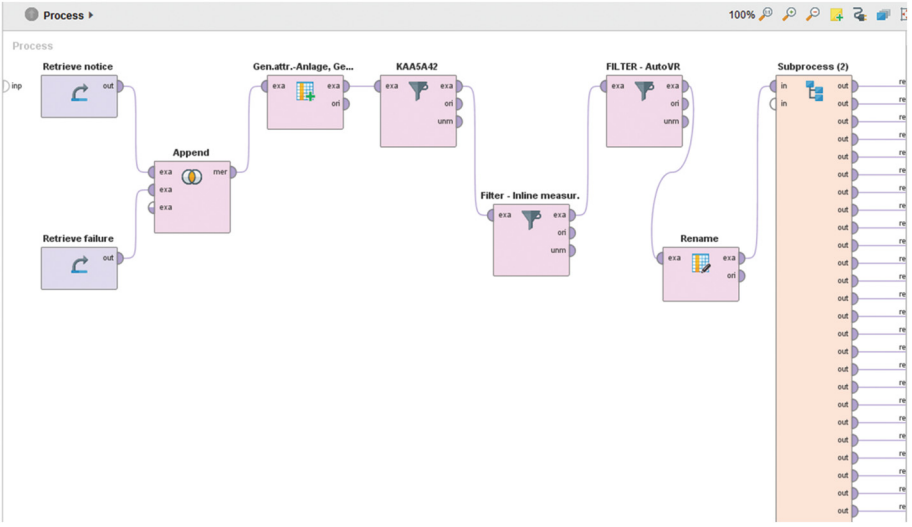


Fig. 3. The part of process

Then we created a process where we downloaded the data of robot - significant error logs, minor error logs and message logs into the software RapidMiner by the operator "Retrieve data". Their mutual connection was subsequently achieved using the operator Append. Data were modified and we filtered unnecessary logs. The total number of error states of the FDS head was 54, due to the limited time horizon of the data samples only

46 errors were identified and assigned. The algorithm provided assigned to each error the exact time of occurrence, space and area of distribution. Combining these data results in their penetration representing a set of equipment with accurate positioning and timing. The identified errors were selected, and with them all the other error logs and message logs that occurred exactly two hours before significant errors that we identified using an algorithm.

3 Chapter Reached Results

The resulting data represents error messages occurring before serious production disruptions. The number of errors and messages in time is shown in Fig. 4, where we can see the timeline on the x axis and the number of errors on the y axis. From the output, we have time set, when the equipment most commonly transmitted errors and messages.

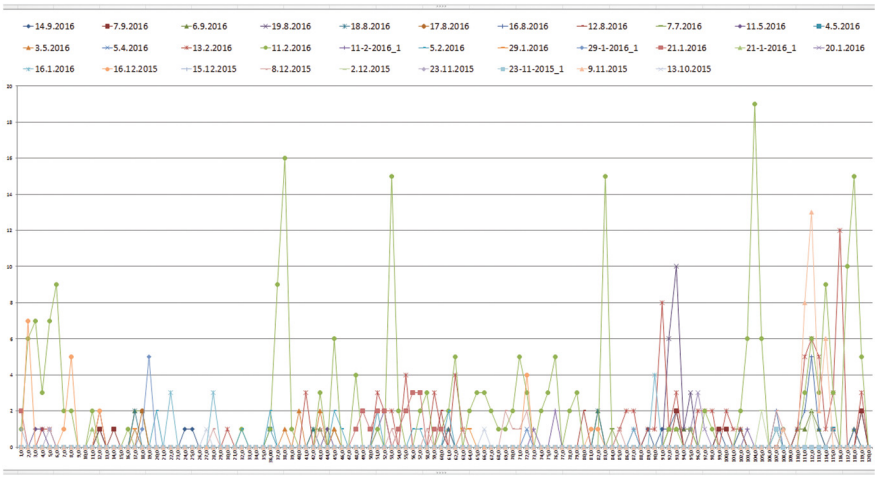


Fig. 4. The graph of errors and messages in time line

Furthermore, we have specified the same errors in terms of the position in which they are incurred. Figure 5 shows the number of errors to specific robots before each significant error.

The alarm text is the last refinement. We also compared the number of errors on the basis of selectivity, we were able to create an accurate list of errors that occur before a failure and we also exactly determined the time point of occurrence, the place of the error origin as well as the particular robot responsible for the error alert.

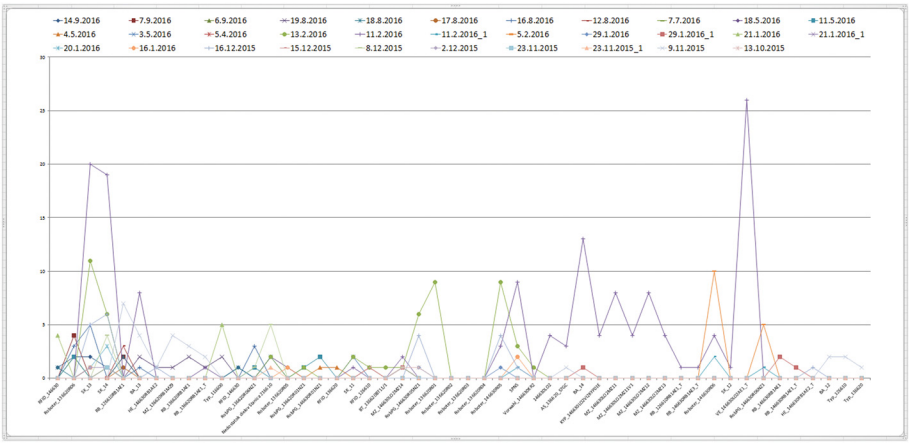


Fig. 5. The graph of errors and messages in the division

ExampleSet (27 examples, 0 special attributes, 4 regular attributes)

Row No.	Text alarmu	Gerate	Rozdiel_cas...	count(error...
1	AS_136620_S7GC nájdené dva alebo viac krokov	AS_136620_S7GC	112	1
2	BA_12 Jednotlivý pohyb aktívny	BA_12	111	1
3	BA_12 Jednotlivý pohyb aktívny	BA_12	112	2
4	BA_13 Jednotlivý pohyb aktívny	BA_13	108	1
5	BA_13 Jednotlivý pohyb aktívny	BA_13	111	1
6	BA_13 Jednotlivý pohyb aktívny	BA_13	113	1
7	BA_13 KWE7 aktívne	BA_13	113	1
8	HE_146630RB1AE2 Porucha blokovania	HE_146630RB1AE2	114	1
9	MZ_136620RB1MZ9 Porucha valec a	MZ_136620RB1MZ9	108	1
10	MZ_136620RB1MZ9 Porucha valec b	MZ_136620RB1MZ9	108	1
11	MZ_136620RB1MZ9 Časová kontrola pohybu späť	MZ_136620RB1MZ9	108	1
12	MZ_136620RB1MZ9 Časová kontrola pohybu vpred	MZ_136620RB1MZ9	108	1
13	RB_126610RB1AE1 BE3 Kontrola zastavenia motora snímaèo	RB_126610RB1AE1	111	3
14	RB_126610RB1AE1 BE3 Kontrola zastavenia motora snímaèo	RB_126610RB1AE1	112	2
15	RB_126610RB1AE1 BE3 Kontrola zastavenia motora snímaèo	RB_126610RB1AE1	114	1
16	RB_126610RB1AE1 BE4 Kontrola pozície vpredu snímaèom	RB_126610RB1AE1	111	3
17	RB_126610RB1AE1 BE4 Kontrola pozície vpredu snímaèom	RB_126610RB1AE1	112	2
18	RB_126610RB1AE1 BE4 Kontrola pozície vpredu snímaèom	RB_126610RB1AE1	114	1
19	RB_136620RB1AE1 BE1 Kontrola zadnej polohy snímaèom	RB_136620RB1AE1	114	1
20	RB_136620RB1AE1 Kontrola taktu nie je zrušená	RB_136620RB1AE1	112	1

Fig. 6. The list of errors

4 Conclusion

Our main goal was to find messages and failures alerting of identified serious failure. We disposed of a list of particular failure (Fig. 6) messages and a list of failures causing the incorrect status of robots and devices.

Our main objective was searching for the messages and errors causing the identified significant errors. We obtained a list of specific errors (see Fig. 6), to be further tested and linked to messages and errors caused by this particular outage.

Within the ongoing research, subsequent findings and further data we receive from the technology will be added to the currently presented outcome in order to find out more specific and detailed results and to capture entirely the possible effects of the respective error. All results will be used to further research in predictive maintenance.

References

1. Wang, X.Z., McGreavy, C.: Automatic classification for mining process operational data. Ind. Eng. Chem. Res. **37**, 2215–2222 (1998)
2. Bateman, J.: Preventive maintenance: standalone manufacturing compared with cellular manufacturing. Ind. Manag. **37**, 19–21 (1995)
3. Barlow, R.E., Hunter, L.C.: Optimum preventive maintenance policies. Oper. Res. **1960**, 90–100 (2006)
4. Scheffer, C., Girdhar, P.: Machinery vibration analysis & predictive maintenance, vol. 6 (2004)
5. Kagermann, H., Wahlster, W., Helbig, J.: Recommendations for implementing the strategic initiative INDUSTRIE 4.0 Frankfurt(2013)
6. Predictive maintenance, Internet. http://www.simafore.com/blog/bid/180786/4-ways-predictive-analytics-can-improve-equipment-maintenance
7. Today´s autmotive markets must move beyond traditional strategies Internet. https://rapidminer.com/industry/automotive/
8. Simoncicova, V., Hrcka, L., Tadanai, O., Tanuska, P., Vazan, P.: Data pre-processing from production processes for analysis in automotive, industry [Internet] (2016). http://www.ceciis.foi.hr/app/public/conferences/1/ceciis2016/papers/DKB-3.pdf

Methodology and Structure Adaptation Algorithm for Complex Technical Objects Reconfiguration Models

Anton Pashchenko[1(✉)], Pavel Okhtilev[1], Semen Potrysaev[1], Yury Ipatov[2], and Boris Sokolov[1,3]

[1] Saint Petersburg Institute of Informatics and Automation, Russian Academy of Sciences (SPIIRAS), Saint Petersburg, Russia
{pashchenkoae,pavel.oxt,semp}@mail.ru, sokol@iias.spb.su
[2] Volga State University of Technology, Yoshkar-Ola, Russia
ipatov_ya@list.ru
[3] Saint Petersburg National Research University of Information Technologies, Mechanics and Optics (ITMO), Saint Petersburg, Russia

Abstract. Complex-technical object (CTO) is the main object of investigation. In the paper are shown how the problem of CTO functional reconfiguration can be sovled in the terms of proposed CTO structural dynamics control theory. General formal description of CTO structure-dynamics control (SDC) including its functional reconfiguration is suggested. New approach to structure adaptation of CTO functional reconfiguration models is developted. This approach is based on concept of integrated modeling and simulation.

Keywords: Complex-technical objects · Models of CTO reconfiguration · Structure adaptation algorithm

1 Introduction

Complex technical object (CTO) is the main object of our investigation in the paper. Classic examples of complex objects are: control systems for various classes of moving objects such as surface and air transport, ships, space and launch vehicles, geographically distributed heterogeneous networks, flexible computerized manufacturing and ets [1–3]. The preliminary investigations confirm that the most convenient concept for the formalization of CTO control processes is the concept of an active mobile object (AMO). In general case, it is an artificial object (a complex of devices) moving in space and interacting (by means of information, energy, or material flows) with other AMO and objects-in service (OS).

There are different external and internal, objective and subjective perturbation impacts altering operation conditions of CTO [1, 2]. The perturbation impacts initiate CTO structure dynamics. Control inputs are being produced to compensate the influence of perturbation factors. In this case, CTO structure dynamics control (SDC) should be organized [1]. By structure dynamics control we mean a process of control inputs producing and implementation for the CTO transition from the current macro-state to a given one.

R. Silhavy et al. (eds.), *Cybernetics and Mathematics Applications in Intelligent Systems*, Advances in Intelligent Systems and Computing 574, DOI 10.1007/978-3-319-57264-2_33

Figure 1 shows the structural model of CTO operability evolution in a graphic form. The nodes of the graph represent the CTO macro-states: intact, failure, operability, non-operability, correct functioning, and in-correct functioning.

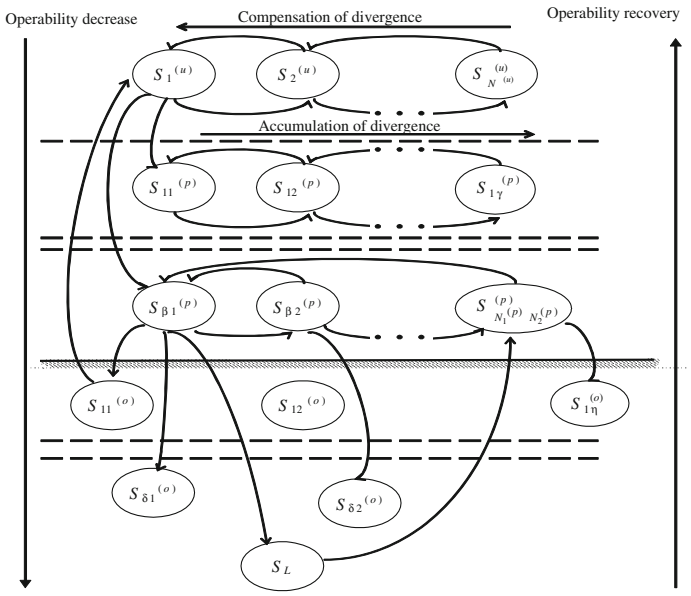


Fig. 1. Structural model of CTO operability evolution

Figure 1 shows three main classes of the CTO states: *the classes of complete operability and intactness; the class of partial operability; the class of non-operability*. The last macro-state S_L at the bottom of Fig. 1 represents non-operability of all CTO elements and maximal losses. Each arc of the graph in Fig. 1 denotes CTO transition from some macro-state to another one.

The arcs of the graph denote transitions of CTO from one macro-state to another macro-state. Vertical arcs denote operability decrease or recovery. Horizontal arcs denote accumulation or compensation of parameters mismatch. Distinct types of arcs are used for the above types of transitions in Fig. 1. The arcs drawn downwards represent D-transitions, the arcs drawn from left to right represent A-transitions, the arcs drawn upwards are used for R-transitions, the arcs drawn right to left denote C-transitions.

The dashed liner in Fig. 1 separates operability states and failure states. In accordance with the above definitions CTO operability dynamics is a combination of the following simultaneous processes bounded up with CTO technical state alteration: degradation process (D–processes); recovery processes (R-processes); processes of parameter divergence accumulation (A-processes); processes of parameter divergence compensation (C-processes).

Therefore, one of the goals of CTO structure dynamics control is permanent maximization of operability level for CTO and CTO elements. For these purposes, CTO reconfiguration should be used. Reconfiguration is a process of the CTO structure

alteration with a view to increase, to keep, or to restore the level of CTO operability, or with a view to compensate the loss of CTO efficiency as caused by the degradation of its functions [1, 4, 5].

Figure 2 shows classification of CTO typical reconfiguration tasks. In practice, for decision CTO typical reconfiguration tasks simulation models (first of all multi-agents models) are widely used [1–6]. However, analysis shows that at the present stage of the development of the theory and technology of simulation of CTO reconfiguration (CTO structure dynamic control in general case) we can no longer continue our research without in any way addressing issues of its interaction with other modelling theories and technologies.

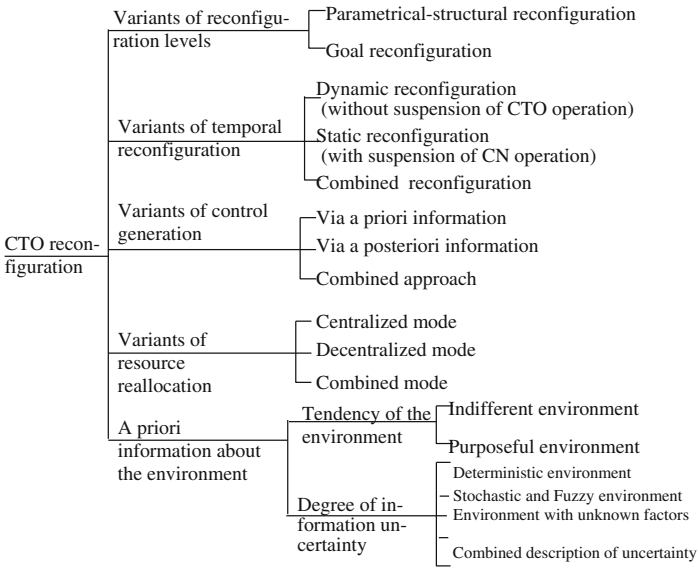


Fig. 2. Classification of CTO typical reconfiguration tasks

In this case methodology and technology of integrated modeling should be used [1, 2, 7–15]. By integrated modeling and simulation (IMS) of any CTO we mean the methodology and technology of multiple-model description of CTOs and the combined use of methods, algorithms and multi-criteria techniques for analysis, synthesis and selection of the most preferred management decisions related to the creation, use and development of CTOs in different, dynamically changing, internal and external environments. [1, 10].

The main advantage of the IMS is that the combined use of alternative models, methods, and algorithms compensates for their shortcomings and limitations and enhances their strengths. Our investigation have shown that existence various alternative descriptions for CTO elements and control subsystems gives an opportunity of adaptive models selection (synthesis) for program control under changing environment.

The analysis of known investigations on the subject [1, 3] confirms that the traditional tasks of CTO control should be supplemented with procedures of structural and parametric adaptation of models, algorithms and conforming use special control software

(SCS). In the paper, we propose new approach to structure adaptation of CTO reconfiguration models, which is based on concept of IMS.

2 Approach

Let us introduce some notation for problem definition. Let $A = \{A_i,\ i \in N = \{1, \ldots, n\}\}$ be a set of CTO business processes (BP) (and corresponding control functions) to be implemented at some node of CTO at a given time interval $T = [t_0, t_f]$. To achieve the CTO goals during the interval T, the BP have to be fulfilled. We distinguish between the functions of goal definition, planning (long term and operational planning), real-time control, CTO states analysis, external situation analysis and coordination. The set $A = \{A_i,\ i \in N = \{1, \ldots, n\}\}$ is related to sets of informational-technological operations $D^{(i)} = \{D_\zeta^{(i)},\ \zeta \in K = \{1, \ldots, s_i\}\}$, that are necessary for implementation of BP A_i, $i = 1, \ldots, n$. Let $B = \{B_j,\ j \in M = \{1, \ldots, m\}\}$ be a set CTO main elements and subsystems. Each element B_j can include technical facilities $C^{(j)} = \{C_\lambda^{(j)},\ \lambda \in L = \{1, \ldots, l\}\}$ with appropriate computer equipment and software. Technical facilities are used for implementation of control functions [1, 3, 15].

Let $E(t) = \left\| e_{ij}(t) \right\|$ be a known matrix function, with $e_{ij}(t) = 1$ in case of the subsystem B_j is carrying out the function A_i at time t in accordance with time-spatial, technical and technological constraints, $e_{ij}(t) = 0$ otherwise.

In our paper models (analytical, simulation and combined (integrated) models) describing structural states $S = \{S_\delta\} = \{S_1, S_2, \ldots, S_{K_\Delta}\}$ are used for optimal distribution of BP and control functions among subsystems of CTO, for technological operations planning and for evaluation of CTO efficiency (in other words this process we name as process of CTO functional reconfiguration). The following characteristics of CTO efficiency can be used: the total number of functions implemented in subsystems during the interval T, the total number of BP in given macro-states, the total number of technological operations executed over the time interval T, the total time of operations over the time period T. The above-mentioned characteristics can have stochastic or fuzzy interpretation if uncertainty factors are present [1, 2, 16].

The following dynamic model of functions distribution can be used for evaluation of CTO efficiency during its functional reconfiguration [1].

$$\dot{x}_i^{(\phi)} = \sum_{j=1}^{m} \varepsilon_{ij}(t) u_{ij}^{(\phi)};\ \dot{x}_{i\zeta j}^{(0)} = \sum_{\lambda=1}^{l} b_{i\zeta j\lambda} u_{i\zeta j\lambda}^{(0)};\qquad \dot{y}_{ij}^{(\phi)} = v_{ij}^{(\phi)}; \tag{1}$$

$$\sum_{j=1}^{m} u_{ij}^{(\phi)} \left[\sum_{\alpha \in \Gamma_{i1}} \left(a_\alpha^{(\phi)} - x_\alpha^{(\phi)} \right) + \prod_{\beta \in \Gamma_{i2}} \left(a_\gamma^{(\phi)} - x_\gamma^{(\phi)} \right) \right] = 0; \tag{2}$$

$$\sum_{\lambda=1}^{l} u_{i\zeta j\lambda}^{(0)} \left[\sum_{\nu \in \Gamma_{i\zeta 1}} \left(a_{i\nu j}^{(0)} - x_{i\nu j}^{(0)} \right) + \prod_{\mu \in \Gamma_{i\zeta 2}} \left(a_{i\mu j}^{(0)} - x_{i\mu j}^{(0)} \right) \right] = 0; \tag{3}$$

$$\sum_{i=1}^{n} u_{ij}^{(\phi)}(t) \le 1;\ \forall j;\quad \sum_{j=1}^{m} u_{ij}^{(\phi)}(t) \le 1;\ \forall i;\quad u_{ij}^{(\phi)}(t) \in \{0,1\}; \tag{4}$$

$$\sum_{j=1}^{m}\sum_{\lambda=1}^{l} u_{i\zeta j\lambda}^{(0)}(t) \le 1,\ \forall i, \forall \zeta;\quad \sum_{i=1}^{n}\sum_{\zeta=1}^{s} u_{i\zeta j\lambda}^{(0)}(t) \le 1, \tag{5}$$

$$\forall i, \forall \zeta;\ u_{i\zeta j\lambda}^{(0)}(t) \in \{0, u_{ij}^{(\phi)}\}; \tag{6}$$

$$v_{ij}^{(\phi)}\left(a_{isj}^{(0)} - x_{isj}^{(0)}\right) = 0;\ v_{ij}^{(\phi)}(t) \in \{0,1\}; \tag{7}$$

$$x_i^{(\phi)}(t_0) = x_{i\zeta j}^{(0)}(t_0) = y_{ij}^{(\phi)}(t_0) = 0; \tag{8}$$

$$x_i^{(\phi)}(t_f) = a_i^{(\phi)};\ \left(a_{i\zeta j}^{(0)} - x_{i\zeta j}(t_f)\right) y_{ij}^{(\phi)}(t_f) = 0; \tag{9}$$

$$x_i^{(\phi)}(t_f) = a_i^{(\phi)};\ \left(a_{i\zeta j}^{(0)} - x_{i\zeta j}(t_f)\right) y_{ij}^{(\phi)}(t_f) = 0; \tag{10}$$

$$J_0 = \sum_{i=1}^{n}\sum_{j=1}^{m} v_{ij}^{(\phi)}(t_f);\ J_1^{(j)} = \sum_{i=1}^{n} v_{ni}^{(\phi)}(t_f);\ J_2 = T - \sum_{i=1}^{m} y_{nj}^{(\phi)}; \tag{11}$$

where $x_i^{(\phi)}(t)$ is equal to total duration of the business process A_i fulfillment in subsystem B_j as $u_{ij}^{(\phi)}(t) = 1$; the variable $x_{i\zeta j}^{(0)}$ express the current state of the technological operation $D_\zeta^{(i)}$; $y_{ij}^{(\phi)}$ is equal to the time passed after A_i completion in B_j until the time $t = t_f$; $a_\alpha^{\phi}, a_\alpha^{(0)}, a_\gamma^{(0)}, a_{i\nu j}^{(0)}, a_{i\mu j}^{(0)}$ are given values setting end conditions for $x_i^{(\phi)}(t)$, $x_\alpha^{(\phi)}(t)$, $x_\gamma^{(\phi)}(t)$, $x_{i\nu j}^{(0)}(t), x_{i\mu j}^{(0)}(t)$ at time $t = t_f$; $u_{ij}^{(\phi)}, u_{i\zeta j\lambda}^{(0)}, v_{ij}^{(\phi)}$ are control inputs. Here $u_{ij}^{(\phi)}(t) = 1$ if BP A_i is being executed in the subsystem B_j at time t, $u_{ij}^{(\phi)}(t) = 0$ otherwise; $u_{i\zeta j\lambda}^{(0)}(t) = 1$ if the technological operation $D_\zeta^{(i)}$ is executed in the technical facility $C_\lambda^{(j)}$, $u_{i\zeta j\lambda}^{(0)}(t) = 0$ otherwise; $v_{ij}^{(\phi)} = 1$ if BP A_i was implemented in the subsystem B_j, $v_{ij}^{(\phi)} = 0$ otherwise. Here the sets Γ_{i1}, Γ_{i2} include the numbers of functions that are direct predecessors of the control function A_i. The set Γ_{i1} indicates predecessors connected by logical "and", the set Γ_{i2} indicates predecessors connected by logical "or". The sets $\Gamma_{i\zeta 1}, \Gamma_{i\zeta 2}$ include the numbers of technological operations $D_\nu^{(i)}$ and $D_\mu^{(i)}$ that are direct predecessors of the operation $D_\zeta^{(i)}$. The subscripts 1 and 2 express the type of logical connection as stated above. Therefore, constraints (2) and (3) define allowable sequences of control functions and technological operations. Constraints (4) and (5) specify that each BP at each time can be carried out only in one subsystem B_j ($i = 1, \ldots, n,\ j = 1, \ldots, m$) and conversely, each subsystem B_j can carry out only one BP A_i at the same time. Similar constraints are used for technological operations $D_\zeta^{(i)}$ that are executed at the technical facility $C_\lambda^{(j)}$. Constraints (6) permits to interrelate CTO BP and control functions models of optimal

distribution and models for technological operations planning and for evaluation of CTO efficiency. Expression (7) states switching-on conditions for the auxiliary control input $v_{ij}^{(\phi)}(t)$. Expressions (9), (10) and (11) specify end conditions for the state variables at the time $t = t_0$, $t = t_f$, R^1 is a set of positive real numbers. The functionals (indexes) J_0, J_1, J_2 are quality measures for distribution of BP in CTO. Here J_0 is equal to total number of functions by the time $t = t_f$, J_1 is equal to the number of subsystems the function A_i is implemented by the time $t = t_f$, J_2 expresses the elapsed time for implementation of all necessary functions.

Now the verbal description of a CTO functions-distribution problem (in other words CTO functional reconfiguration problem) can be presented as follows. It is necessary to select the best variants of functions distribution among the nodes of CTO for each structural state $S = \{S_\delta\} = \{S_1, S_2, \ldots, S_{K_\Delta}\}$ of CTO (under known time spatial, technical and technological constraints) and to find the best variants of functions implementation. The structural states of CTO should be sorted according to their preference. The preference relation can be expressed through quality functions characterizing efficiency of CTO and its structural and technologic characteristics.

The described problem belongs to the class of multi-criteria choice problems with finite sets of alternatives (structural states of CTO). Different methods and algorithms for decision this problem are proposed [1–3, 10, 14, 15]. Moreover, our investigation have shown the proposed logical-dynamic multiple-model description of CTO SDC permits to organize procedure of parametric and structural adaptation for corresponding logical-dynamic models. In the paper [2], we presented procedures of parametric and structural adaptation for CTO SDC logical-dynamic models, which were based on evolutionary modeling. Here, we propose algorithm of structural adaptation for CTO SDC logical-dynamic models, which is based on fuzzy clusterization.

3 Results

Let us consider algorithm for structural adaptation of CTO SDC models. This algorithm is based on fuzzy clusterization. The clusterization is performed in advance, before CTO reconfiguration execution. The set of CTO macro-states is being divided into equivalence classes: $S = \{S_\delta\} = \{S_1, S_2, \ldots, S_{K_\Delta}\}$ such that each class corresponds to a certain structure θ of the multiple-model complex (1)–(11). Moreover, proposed models permit to generate (synthesis) different variants of concrete logical-dynamic model and its macro-states due to logical constrains (2). The structural adaptation is being performed during the CTO operation phase. It is necessary to recognize the current multi-structural macro-state and its membership of certain equivalence class σ_θ ($\theta = 1, \ldots, \Theta$). The task of fuzzy clusterization can be described as follows: the set of multi-structural macro-state $S = \{S_1, S_2, \ldots, S_{K_\Delta}\}$ is given; each multi-structural macro-state have a finite set) $\vec{J}_\delta$, ($\delta = 1, \ldots, K_\Delta$) of parameters (indices) characterizing different aspects of CTO functioning. It is necessary to find an optimal partition (or coverage) of the set S. The partition is a finite set $\sigma = \{\sigma_1, \ldots, \sigma_\Theta\}$ of fuzzy clusters (equivalence classes), each cluster being described by a certain model complex. The problem can be written as follows:

$$\sum_{\theta=1}^{\Theta}\sum_{\delta}^{K_\Delta}\mu_{\theta\delta}^2||\vec{J}_\delta-\vec{I}_\theta||\rightarrow\min_{\{\mu_{\theta\delta}\}\{\vec{I}_\theta\}}, \tag{12}$$

$$\sum_{\theta=1}^{\Theta}\mu_{\theta\delta}=1;\ \delta=1,\ldots,K_\Delta;\ \mu_{\theta\delta}\geq 0,\ \forall\theta,\ \forall\delta, \tag{13}$$

$$\vec{I}_\theta=\frac{1}{\sum\limits_{\delta}^{K_\Delta}\mu_{\theta\delta}^2}\sum_{\delta}^{K_\Delta}\mu_{\theta\delta}^2\vec{J}_\delta, \tag{14}$$

where $\vec{J}_\delta$ are given vectors of parameters (indices) characterizing the multi-structural macro-state S_δ; $\vec{I}_\theta$ are required centers of clusters (point of m'-dimensional space), each vector $\vec{I}_\theta$ represents some poly-model complex; variables $\mu_{\Theta\delta}$ are the grades of S_δ membership of the class σ_θ; $\left\|\vec{J}_\delta-\vec{J}_\theta\right\|=\left\{\sum\limits_{K'=1}^{m'}\left[\left(J_{\delta K'}-I_{\theta K'}\right)^2\right]^{1/2}\right\}$ is a norm of the vector $\vec{J}_\delta-\vec{J}_\theta$ (the distance in Euclidean space R^m `). The functional (12) is equal to a weighted variance characterizing spread of points $\{\vec{J}_\delta\}$ with respect to centers $\vec{I}_\theta$ [1, 2, 16]. The weights belong to the set $\{\mu_{\theta\delta}^2\}$. The objective of the adaptation task is to find the centers and the grades of membership $\mu_{\theta\delta}$ (and the weights $\mu_{\theta\delta}^2$) such that the constraints (13), (14) are satisfied and the variance is minimal [1, 16].

The problem is multi-extremal and rather complex, the alternatives depend on two groups of parameters: $\{\mu_{\theta\delta}\}$, $\{\vec{J}_\delta\}$. Here can be used an iterative procedure with alternative receiving of $\mu_{\theta\delta}$ and $J_{\delta K'}$, $k'=1,\ldots,m'$. The procedure can start when the initial values are assigned to parameters of one group. Let us consider the method of initial approximation of the vector $\vec{\mu}_\delta^{(o)}=||\vec{\mu}_{1\delta}^{(o)},\ldots,\vec{\mu}_{\theta\delta}^{(o)}||$. Experts use the method of pairwise comparison [17] to put in order the clusters σ_θ (poly-model complexes) for each multi-structural macro-state S_δ. They compare the pairs of S_δ inclusions in each σ_θ ($\theta=1,\ldots,\Theta$). This results in the matrix of pairwise comparison:

$$\mathbf{c}^{(\delta)}=\left\|\mathbf{c}_{\theta,\theta'}^{(\delta)}\right\|;\theta,\theta'=1,\ldots,\Theta. \tag{15}$$

where $\mathbf{c}_{\theta,\theta'}^{(\delta)}$ is a value characterizing the degree of preference of S_δ inclusion in the class σ_θ with respect to S_δ inclusion in the class $\sigma_{\theta'}$. Then the task of search for $\vec{\mu}_\delta^{(o)}$ is reduced to solving of the algebraic equations (the matrix from is used):

$$\left(\mathbf{c}^{(\boldsymbol{\delta})}-\rho_{\max}^{\delta}\overline{E}\right)\vec{\mu}_\delta^{(o)}=0;\ \theta,\theta'=1,\ldots,\Theta. \tag{16}$$

where $\overline{E}$ is a unitary matrix of dimension $\Theta\times\Theta$; $\rho_{\max}^{(\delta)}$ is a maximal proper number of the matrix $\mathbf{c}$; $\vec{\mu}_\delta^{(o)}$ is the initial vector of S_δ membership grades of each class σ_θ. In a general case we receive an intersection of sets $\sigma_\theta\cap\sigma_{\theta'}\neq\varnothing$, $\theta,\theta'\in\{1,\ldots,\Theta\}$ rather than a

partition. To form a partition (a single-valued correspondence between sets of multi-structural states and poly-model complex) the threshold l_θ should be given for each σ_θ. In this case S_δ belongs to σ_θ if and only if $\mu_{\theta\delta} \geq l_\theta$. Thus, the general procedure of (12)–(14) solving includes the following main steps.

Step 1. The vector $\vec{I}_\theta^{(r)}$ ($r = 0, 1, 2 \ldots$ is a an iteration number) is received from (14) for given values $\vec{\mu}_\delta^{(o)}$ ($\delta = 1, \ldots, K_\Delta$); $\vec{\mu}_\delta^{(o)} r = r + 1$.

Step 2. The next approximation of the vector $\vec{\mu}_\delta^{(r)}$ for each S_δ is received from the formula:

$$\mu_{\theta\delta}^{(r)} = \frac{1}{L_{\theta\delta}^{(r)}}, \quad L_{\theta\delta}^{(r)} = \sum_{\theta'=1}^{\Theta} \left(\frac{||\vec{J}_\delta - \vec{I}_\theta^{(r)}||}{||\vec{J}_\delta - \vec{I}_{\theta'}^{(r)}||} \right)^2. \tag{17}$$

Development of the formula is given in [1, 16]. It can be seen that the S_δ membership grade of the cluster σ_θ (of the poly-model complex M_θ) is inversely proportional to the sum of squares of the rations of the distance between the point $\vec{J}_\delta$ and the center of the cluster $\vec{I}_\theta^{(r)}$ to the distances between this point and the centers of the other clusters $\vec{I}_{\theta'}^{(r)}$.

Step 3. If the condition $\eta(\mu_{\theta\delta}^{(r)}, \mu_{\theta\delta}^{(r-1)}) \leq \varepsilon'$ is satisfied for all $\mu_{\theta\delta}^{(r)}$, then the iterations are terminated, otherwise we return to **step 1**. Here $\eta(\,,\,)$ is some measure of deviation, for example, $\eta(\,,\,) = \left| \mu_{\theta\delta}^{(r)} - \mu_{\theta\delta}^{(r-1)} \right|$; ε' is a given threshold.

At the CTO operation phase the following activity is needed to select the model structure M_θ ($\theta = 1, \ldots, \Theta$) adequate to the situation. A monitoring of CTO reconfiguration dynamics is performed periodically with a given interval. During the monitoing the current multi-structural macro-state S_δ ($\delta = 1, \ldots, K_\Delta$) of CTO is determined. Then $\mu_{\theta\delta}$ is received from formula (17) and is compared with a given threshold l_θ. If $\mu_{\theta\delta} \geq l_\theta$ then it is advisable to use the multiple-model complex M_θ for a description of S_δ and corresponding process of CTO reconfiguration. Thus, subject to the situation, an adaptive choice of a multiple-model structure is performed to receive the best [in the sense of formula (9)] description of CTO reconfiguration.

4 Summary

The processes of CTO operation are non-stationary and nonlinear. It is difficult to formalize various aspects of CTO. There are no strict criteria of decision making for CTO control and no a priori information about many CTO parameters. Besides, the CTO operation is accompanied by external and internal, objective and subjective perturbation impacts. The perturbation impacts initiate the CTO structure-dynamics control (SDC) including reconfiguration and predetermine a sequence of control inputs compensating the perturbation. The above-mentioned CTO peculiarities do not let produce an adequate description of control processes in existing and designed CTO on a basis of single-class models. That is why the multiple-model description of CTO structure-dynamics control processes were proposed. Classification and features analysis of CTO reconfiguration

were carried out. We proposed general formal description of CTO structure-dynamics control (SDC) including reconfiguration in the paper. Original algorithm of structural adaptation for CTS SDC models was suggested. The algorithm was based on the methods of fuzzy clusterization. Now this algorithms is implemented in intelligent information systems for operational river-flood forecasting [18].

Acknowledgements. The research described in this paper is partially supported by the Russian Foundation for Basic Research (grants 15-07-08391, 15-08-08459, 16-07-00779, 16-08-00510, 16-08-01277, 16-29-09482-ofi-i, 17-08-00797, 17-06-00108, 17-01-00139, 17-20-01214), grant 074-U01 (ITMO University), project 6.1.1 (Peter the Great St.Petersburg Politechnic University) supported by Government of Russian Federation, Program STC of Union State "Monitoring-SG" (project 1.4.1-1), state order of the Ministry of Education and Science of the Russian Federation №2.3135.2017/K, state research 0073–2014–0009, 0073–2015–0007, International project ERASMUS +, Capacity building in higher education, № 73751-EPP-1-2016-1-DE-EPPKA2-CBHE-JP, Innovative teaching and learning strategies in open modelling and simulation environment for student-centered engineering education.

References

1. Okhtilev, M.Y., Sokolov, B.V., Yusupov, R.M.: Intellectual Technologies of Monitoring and Controlling the Dynamics of Complex Technical Objects. Nauka, Moskva, 409 p. (2006)
2. Silhavy, R., Senkerik, R., Oplatkova, Z.K., Prokopova, Z., Silhavy, P. (eds.): Intelligent Systems in Cybernetics and Automation Theory. AISC, vol. 348. Springer, Cham (2015). doi: 10.1007/978-3-319-18503-3
3. Skurikhin, V.I., Zabrodskii, V.A., Kopeichenko, Y.: Adaptive Control Systems for Manufacturing. Mashinostroenie, Moscow (1989)
4. Newman, M.E.J.: The structure and function of complex networks. SIAM Rev., 45, 167–256 (2003). Peregudov, F.I., Tarrasenko, F.P.: Introduction to Systems Analysis. Vysshaya Shkola, Moscow (1989)
5. Van der Velde, W.E. Control System Reconfiguration. In: Proceedings of American Control Conference 1984, vol. 3, pp. 1741–1745 (1984)
6. Arnott, D.: Decision support systems evolution: framework, case study and research agenda. Eur. J. Inf. Syst. **13**(4), 247–259 (2004)
7. http://www.liophant.org/scsc
8. http://www.scs.org
9. http://www.wintersim.org
10. http://www.simulation.su
11. Michalevich, Z.: Genetic Algorithms + Data Structures = Evolution Programs, 387 p. Springer, Heidelberg (1996)
12. Glover, F., Kochenberger, G.: Handbook of Metaheuristics, 570 p. Kluwer Academic Publishers, Dordrecht (2003)
13. FIPA Agent Management System. http://www.fipa.org/specs/fipa00023/XC00023H.html
14. Wooldridge, M.: An Introduction to Multi-agent Systems. Wiley, New York (2002)
15. Ivanov, D., Sokolov, B., Pavlov, A.: Dual problem formulation and its application to optimal redesign of an integrated production–distribution network with structure dynamics and ripple effect considerations. Int. J. Prod. Res. **51**(18), 5386–5403 (2013)
16. Zaden, L.A.: Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic. Fuzzy Sets Syst. **90**, 111–127 (1997)

17. Saaty, T.L.: Theory and application of the analytical network process. RWS, Pittsburg (2005)
18. Alabyan, A.M., Krylenko, I.N., Potryasaev, S.A., Sokolov, B.V., Yusupov, R.M., Zelentsov, V.A.: Development of intelligent information systems for operational river-flood forecasting. Herald Russ. Acad. Sci. **86**(1), 24–33 (2016)

Characterization of the Current Conditions of the ITSA Data Centers According to Standards of the Green Data Centers Friendly to the Environment

Leonel Hernandez[1(✉)] and Genett Jimenez[2]

[1] Department of Telematic Engineering, Engineering Faculty, Institución Universitaria ITSA, Barranquilla, Colombia
lhernandezc@itsa.edu.co

[2] Department of Industrial Process Engineering, Engineering Faculty, Institución Universitaria ITSA, Barranquilla, Colombia
gjimenez@itsa.edu.co

Abstract. Data Center is a specially conditioned space to house all equipment and systems. When it points especially conditioning it means that a data center is a place that you have the following installed: air conditioning, stabilized power supply, uninterrupted power supply, structured wiring, fire prevention systems, access control systems surveillance cameras, alarms fire contraindications, temperature and humidity control. The purpose of this document is to explain to the reader in detail how to improve the condition and use of all devices must necessarily contain the data center, to make it more friendly to the environment and at the same time it means a reduction in the cost operation for the company. The results of measurements of data centers of the University Institution ITSA will be shown as an example.

Keywords: Energy · Environment · Devices · Router · Data Center

1 Introduction

Since 2006, there are many articles in the network on green technology that have defined a Green Data Center as the one in whose operation, maintenance and planning is efficient in the expenditure of energy. Although the important thing in the background is whether companies and customers of the information technologies make it by the savings that they had an impact on the electricity bill or really because they are aware of the importance of the rational and efficient management of the power demand of these facilities. The question then would be like knowing that my data center or CPD (tally center) is efficient enough to be cataloged as green data center.

When we are faced with this question, we must not only think about reducing the consumption of electrical energy, but that we require that the company will provide clean energies and management systems that allow you to make sustainable and economically profitable its operation.

R. Silhavy et al. (eds.), *Cybernetics and Mathematics Applications in Intelligent Systems*, Advances in Intelligent Systems and Computing 574, DOI 10.1007/978-3-319-57264-2_34

The problem stems from the fact that the demand for green energy has only been a reality in the last five years, which means that most of the data center can take years to operate with low efficiencies and even more with designs and technologies that were not within its conception an aggregate value of energy efficiency that can categorize as friendly with the environment. The above on the basis that the data center were designed, built and until recently operated under the precepts of continuity in the service and robustness, this last translated as high levels of redundancy.

As Mohd Nor says [1], our world is depending more and more on technology, especially IT. Almost the majority of the main activities that we execute today, communications, financials, online business, video, training, entertainment, education, applications are hosted in data centers that need to be available 24 × 7. Due to this, energy consumption is very high in data centers, which contributes to the detriment of the environment. This study wants to make an analysis of the state of the data centers of the University and serve as a starting point for a future design of an environmentally friendly green data center.

2 Design of Green Data Center: A Literature Review

There are important studies and research about green data center, it is found in the scientific literature. In this literature review, some of these investigations are mentioned.

To verify if data center is green, a several metrics can be taken into account, these performance metrics can be seen in a taxonomy study done by Wang [2]. For energy efficiency aspects, several studies have been developed [3–6], in which it explains how to build green data centers taking into account the guidelines that contribute to improve the efficiency of the use of energy, without affecting the information processing, and that have a great impact on the design of the data center.

One of the topics that has contributed to the efficient use of energy in the green data centers is virtualization, consisting in that in a single server can be configured and managed multiple machines [7, 8].

It can be found some examples about how to make a green data center, and build it [9]. For example, Google [10] spent a total of $25,000 to optimize this room's airflow and reduce air conditioner use. A $25,000 investment in plastic curtains, air return extensions, and a new air conditioner controller returned a savings of $67,000/year.

3 Data Center

A Data Processing Center (CPD) is a large space where it is located the electronic equipment, which stores all of the information in an organization. According to Agustí-Torra et al. [11], "a data center is an infrastructure built to provide IT services such as massive data storage, content delivery network (CDN), web, e-mail, server hosting, enterprise-class applications, or on-demand computing", among others.

The electric energy is the key to all the operations that are performed in the data centers. Without a reliable power supply, this type of infrastructure could not function. However, it is not enough to be connected to the mains traditional, but it is necessary to

count with support equipment able to provide sufficient energy for the proper functioning of the CPD in the event of a blackout. Therefore, the electrical generators (energy storage systems) are a fundamental part to ensure the service in case of power outages.

The data centers are connected to the internet via redundant Gigabit Ethernet connections, so that in the event of a fall a line, the service will continue to operate without problems. Due to the large amount of valuable information that is stored on servers hosted in the CPD, security is paramount to prevent any type of information theft or another series of problems. Video surveillance services and presence of personnel 24 h a day are some measures, which all data centers should implement to ensure the security of its customer data.

An optimum temperature is essential to get the maximum performance from the machines installed there that is why the data centers use air-conditioning systems that maintain the temperature of the chambers in a strip of between 15° and 25°, avoiding overheating of the servers.

3.1 Current Trends in the Data Center

At present there are different ideas for the design of a Data Center robust, reliable and high productivity, for this have been developed most optimal trends, which comply with the objectives of an efficient design. Today the convergence of multiple applications and systems is deployed, such as the voice (VoIP), data (10 GbE), Video IP, distribution AV, wireless, security, PoE and applications.

3.2 High Availability in the Data Center

To determine the criticality of the electrical systems and their redundancy, Zuñiga argues that it is essential to know the levels of availability required by the business to determine the total power, density per rack, available spaces and make a projection of the estimated growth [12].

Standards play a key role here. Should be considered several, within which they emphasize TIA 942, promoted by BICSI, the agency that studies and empowers designers and installers. It is also important to apply the concepts of Uptime Institute, who certifies the levels of availability TIER, administration and operation of Data Centers.

3.3 Avoiding Risks and Failures in the Operation

Experts in the design and implementation of data centers argue that one of the main risks that must be monitored is to ensure that the systems of redundancy and operational procedures work to perfection.

The main risks that could affect the operation of a Data Center are unanticipated power outages, hikes and low voltage electrical and pollution - noise and harmonics - which affect the availability of the users and generate millions of dollars in losses to the companies. All of them are possible to avoid or at least to control.

According to the opinion of Leonardo Covalschi, CEO of Synapsis, each Data Center must meet minimum standards that guarantee the client tranquility and support with regard to what you can get [12].

4 Green Data Center

A Green Data Center is defined as one in which the mechanical systems, electrical, lighting and computation are designed for maximum efficiency and environmental impact [13].

4.1 Energy Problem in Data Center

The core of the Internet are the data centers: brains that process most of the data that run through this. The amount of data increases as the society grows and increases their dependence on technology so that continually undertakes to data centers to improve their storage capacities, processing and transmission. For example, the Facebook data center stores more than 240 billion photos, and this figure continues to increase at a rate of 350 million per day [14], so that your infrastructure is forced to increase steadily. Similarly, a data center in a small company can have a very rapid growth.

Moore's Law of world energy consumption of data centers indicates that this doubles every 5 years. The consumption of a data center is not small, may require 100 to 200 times per square meter, the energy consumed by a common office [15]. For this reason, data centers have become the target of new government regulations. Faced with this situation, has shown great concern, and so has improved the efficiency of the devices and developed software and metrics for such purposes. In spite of the fact that these modifications are good, are not sufficient: comprehensive strategies are needed to optimize efficiency in data centers.

4.2 Energy Efficiency in the Data Center

Energy efficiency in data centers is related to a decrease of the total energy consumption and an increase in the efficiency in the use of this. A few years ago, the concept of energy efficiency in data centers was very subjective, since that often was not clear how to measure the energy, where you should measure, or what units to use. Before this inconvenience, data centers needed a measure as the liters of petrol per kilometer traveled to measure efficiency. For this reason, was developed the metric of efficiency called PUE (Power Usage Effectiveness), which is the standard of energy efficiency in data centers more internationally accepted. In spite of the fact that the issue of energy efficiency goes beyond the PUE, this is a good indicator of how it is handling the energy in a data center.

4.3 The METRIC PUE, the Standard of Efficiency More Used

The PUE (Power Usage Effectiveness) is a metric created by the Organization the Green Grid, which measures the effectiveness in the use of the energy of a data center, and is calculated by the following formula:

$$PUE = \frac{Total\ power\ of\ data\ center\ equipment}{Total\ power\ of\ IT\ equipment} \quad (watts) \tag{1}$$

The Green Grid sets the following levels of efficiency, shown in Table 1 [16]:

Table 1. Levels of efficiency according to the Green Grid

PUE	Levels of efficiency
3	Very inefficient
2.5	Inefficient
2	Average
1.5	Efficient
1.2	Very efficient

4.4 Best Practice in Energy Efficiency

With the purpose of reducing the PUE and achieve efficiency, below is a basic guide to best practice in energy efficiency, using as a point of departure the cascade effect.

To achieve high-energy efficiency, it is essential to use its equipment efficient, since they generate an effect of saving energy in cascade, along the trajectories of energy on the data center. The energy consumed in a data center is distributed in the following 3 paths: Toward the power equipment and you, toward the cooling equipment and toward the "Other equipment".

If the IT equipment consumes less, you can generate a cascade effect [17, 18] for energy savings through the first 2 trajectories. This is because when it reduces the consumption of IT equipment, these powers and generate less heat, then the cooling

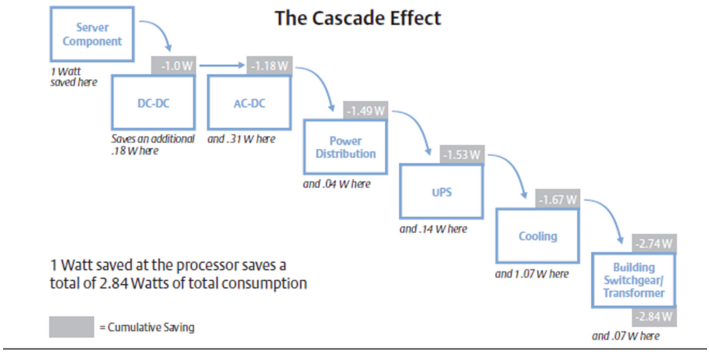


Fig. 1. Cascading effect.

equipment consume less energy, and thus the power equipment generate even less heat. Figure 1 shows an example of the cascade effect [17, 18], and how if it reduces 1 W of consumption in the IT equipment, you can obtain a total saving of 2.7 W of power consumption in the data center.

4.5 IT Equipment

IT equipment are the most important in terms of efficiency because they are the source of the cascade. The following are recommended practices and technologies:

- Virtualize servers on server's blade-type.
- Disconnect or establish timetables off to the IT equipment with little use.
- Activate the economic modes and use the software for monitoring of energy that many of these teams bring.
- Get energy-efficient servers that meet the standards of quality more important.

4.6 Server Virtualization on Servers' Blade-Type

The servers are of vital importance because they carry out the data processing in a data center. Due to that consume a large part of the energy of the data center, it is essential to reduce their consumption and get more out of their computing resources. Server virtualization on server's blade-type is a good way to achieve these objectives. Server virtualization is based on the concept of virtual machine [19]. A virtual machine is a software that can emulate the hardware of a computer, such as a server, a storage device, or even a network and run programs as if it were a real computer. Server virtualization makes use of the virtual machines to put multiple servers, called virtual servers, to run on a physical server, as shown in Fig. 2:

Fig. 2. Illustration of the concept of virtual server

Server virtualization on server's blade-type is a Synergy interesting as regards energy saving, as these two technologies together; you get a better use of the computing resources of the servers, mixed with a reduction for space used by these. By deploying them, you must have certain considerations regarding the cooling equipment. In addition, after deploying must not be surprised if the PUE data center increases, because it reduces the energy consumption of IT equipment, also reducing the consumption of power and cooling equipment, but not on a proportional basis. Figure 3 illustrates a real

example of this [19], in which it notes that virtualization on servers blade-type causes a reduction in the total power consumed, 1000 kW to 672 kW, and a decrease in the efficiency of 37% to 50%.

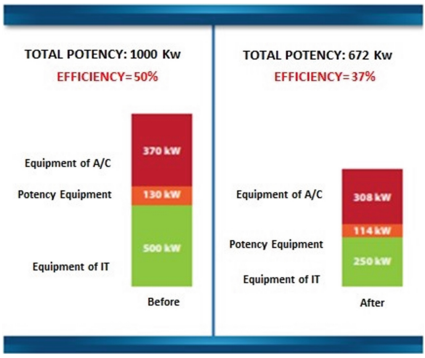


Fig. 3. Virtualization on servers, blade type effect on the efficiency

4.7 IT Equipment with Little Use

It is necessary to switch off the computer when it is not expected to be used, this by 2 important reasons. The first is that in spite of the fact that there are appliances that consume less power, such as a kitchen or clothes dryer, are lit during intervals of up to several months, so that through the life, energy costs and cooling can overcome the computer [20]. The second reason is that the difference in energy consumption between the IT equipment on, while in use or without being on it, is very little. For example, a server that works to 100% of its capacity, can consume 300 W, while a 2%, 200 W (Fig. 4). This is because, regardless of the percentage of CPU utilization, the internal electronics remains energized but not execute instructions. To avoid this unnecessary consumption, you must remove the data center equipment that is not being used and

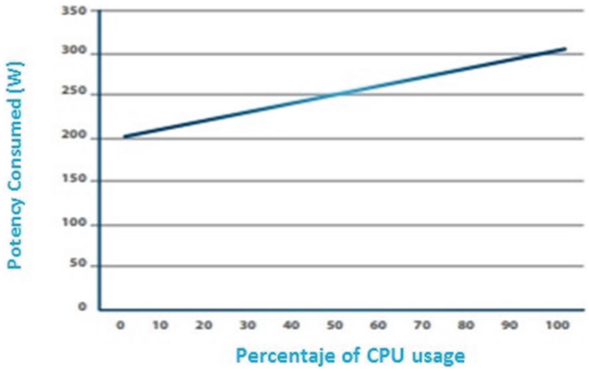


Fig. 4. Power consumed by a server common vs. percentage of use of your CPU

establish timetables to turn off the rest of the devices during extended periods will not use [21].

4.8 Power Equipment

The power equipment carries the energy to the IT equipment. The power flows from the onslaught of the building or from the electric generator to the IT equipment, through a chain of transmission, distribution and conversion of energy, mainly composed of:

- Electric Generator: Generates electrical energy from fossil or other sources of energy.
- Transfer: Choose between taking the power from the mains or the electric generator, depending on whether the network is enabled or not, respectively.
- Switchgears: Protects, controls and allow the distribution of the energy.
- UPS (Uninterruptible Power Supply): Supports energy to the most critical equipment, such as IT.
- Distribution Panels: Distribute the energy in the various electrical circuits.
- Wires: Carry energy

4.9 Refrigeration and Air Flow

The 99% of the electrical energy consumed in a data center is transferred in the form of heat into space. Unless this heat is removed, the temperature of the data center is incremented until the point at which the IT equipment overheats and fails. The IT equipment is designed to work in certain temperature ranges so that there must be a cooling system to maintain an optimum temperature by removing the heat generated.

The following are recommended technologies of refrigeration equipment: precision air conditioning, type air units closely coupled cooling and compressor technologies like digital scroll compressors.

These technologies should be preferably accompanied by the following best practices: measure the cooling system taking into account present and future loads; ensure adequate coordination of air conditioning units, so that there are no cases in which, for example, one humidified while another dehumidified. The control systems can be ordered to monitor and prevent such conflicts; adjust the temperature of the data center at a level not too cold [22].

4.10 Accommodation of Computers

The accommodation of equipment is important, otherwise, you can submit distributions of temperatures little homogeneous [23]. Figure 5 shows an arrangement of equipment that is very bad, in which the IT equipment as usual, takes the cool air from the front and expels it as hot air through the rear, however, all computers will be accommodated in the same direction, so it happens a heating effect in string, which causes the temperature distribution in the data center is not very homogeneous. To avoid these problems, you must employ techniques of accommodation equipment. A very easy to implement

is to split the corridors in hot and cold so that the air conditioners deposited cold air in the cold aisles and collect the hot air from the hot. Figure 6 illustrates this.

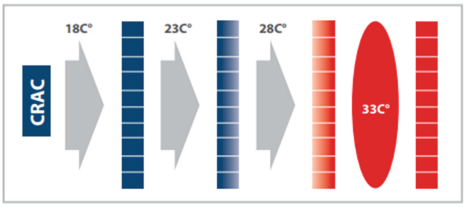


Fig. 5. Effect of orienting the IT equipment in the same address

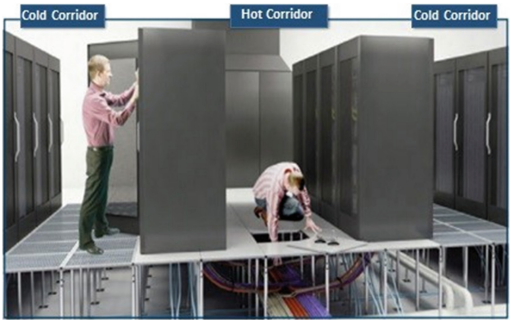


Fig. 6. Technique of divided into hot and cold aisles

To prevent the hot air from mixing with the cold, you can use cold aisle containments (cold aisle containers), as shown in Fig. 7.

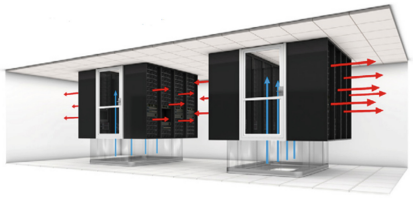


Fig. 7. Cold aisle container

5 Comparative Analysis of Metric PUE: Standard Value Vs. Real Values Obtained

Tables 2 and 3, show the results of the calculations of efficiency levels, according to the Green Grid, taking into account the formula (1). Data centers in this study are relatively small, with few equipment. The main data center is located in Soledad.

Table 2. PUE in Barranquilla's Data Center

Quantity	Device	Description	Unit consumption	Total consumption
1	Switch	Huawei S2300	12.8 W–38 W	38
2	Switch	Catalyst 2960	464 W–870 W (with POE)	1740
1	Switch	Catalyst 3560	449 W	449
1	Transceiver	RAISECOM RC001	36 W–72 W DC	72
1	Router	Cisco 3925	85–400 W	400
1	Conditioned air	ComfortStar	5290 W	5290
Total				7989
PUE Barranquilla				**2,95998518**

Table 3. PUE in Soledad's Data Center

Quantity	Device	Description	Unit consumption	Total consumption
2	Firewall	SonicWALL 4500	66 W–180 W	360
2	Switch	Cisco 2960	464 W–870 W (with POE)	1740
1	Switch	Quidway S3300	100 V–264 V	264
1	Switch	TpLink TL-SG3424	100 V–240 V	240
1	Switch	Cisco 2960	464 W–870 W (with POE)	870
1	Router	Cisco 2900	80 W–360 W	360
3	Servers	HP. ProLiant DL380p Gen8	460 W, 750 W–1200 W	3600
1	Conditioned air	ComfortStar	5290 W	5290
1	UPS	Galleon X9B 6 K	5400 W	5400
Total				18124
PUE Soledad				**2,437987624**

For this exercise, GAPS were calculated, or percentage differences between PUE standard and PUE real values obtained in data centers, using the formula (2):

$$GAP(\%) = \frac{PUE_{(real\ value\ obtained)} - PUE_{(standard)}}{PUE_{(standard)}} \times 100 \tag{2}$$

Table 4 shows the results of the comparative analysis between the standard value of the PUE metric and the actual values obtained in data centers, where GAPS was higher than 100% in relation to the high efficiency standard value established by PUE, which implies the establishment of priority improvement plans to reduce the PUE value and increase the level of efficiency with respect to the Green Grid criterion.

Table 4. Comparative analysis of metric PUE: standard value vs. real values obtained

Location	PUE real value obtained	Level of efficiency	PUE value standard	GAP
Soledad	2,43	Average	1,2	102,50%
Barranquilla	2,95	Very inefficient	1,2	145,83%

6 Conclusions

This research is a very good start point to validate how since data centers we can contribute to improving the environment. We found that in ITSA university institution, both data centers, it is necessary to work to change our data center and make them more friendly to the environment. There are several metrics, but we took a basic PUE measure in both data centers obtaining values that are not efficient. In Barranquilla, PUE value shows us that data center is very inefficient (according to Table 1 – levels of efficiency), and in Soledad the data center is inefficient.

The gaps between real values and ideal values are really considerable, and they invite us to take immediate corrective measures. This study seeks to make our academic and administrative community aware of the impact of a data center that is not in line with international guidelines to be environmentally friendly and that optimization decisions can be made. From the IT department it can develop a future efficient design to build a green data center and put our grain of sand to improve the world we live in.

This paper hopes to be a contribution to the scientific literature related to the design of green data centers, due to the impact this issue has on the environment.

References

1. Bin, N., Nor, M., Hasan, M., Selamat, B.: Green data center frameworks and guidelines review **6**, 338–343 (2014)
2. Wang, L., Khan, S.U.: Review of performance metrics for green data centers: a taxonomy study. J. Supercomput. **63**(3), 639–656 (2013)
3. Bilal, K., Malik, S.U.R., Khalid, O., Hameed, A., Alvarez, E., Wijaysekara, V., Irfan, R., Shrestha, S., Dwivedy, D., Ali, M., Shahid Khan, U., Abbas, A., Jalil, N., Khan, S.U.: A taxonomy and survey on green data center networks. Futur. Gener. Comput. Syst. **36**, 189–208 (2014)
4. Lei, H., Wang, R., Zhang, T., Liu, Y., Zha, Y.: A multi-objective co-evolutionary algorithm for energy-efficient scheduling on a green data center. Comput. Oper. Res. **75**, 103–117 (2016)
5. L. B. B. PUE: Data Center Energy Efficiency–Looking Beyond PUE (2011)
6. Johnsson, L.: Overview of data centers energy efficiency evolution, pp. 1–50, December 2011
7. Jin, Y., Wen, Y., Chen, Q., Zhu, Z.: An empirical investigation of the impact of server virtualization on energy efficiency for green data center. Comput. J. **56**(8), 977–990 (2013)
8. Guan, X., Choi, B.Y., Song, S.: Energy efficient virtual network embedding for green data centers using data center topology and future migration. Comput. Commun. **69**, 50–59 (2015)
9. Pendelberry, S.L., Thurston, M., Strasenburgh, J., Stein, R.: Case study - the making of a green data center. In: IEEE International Symposium on Sustainable Systems and Technology (2012)
10. Practice, B.: Google's Green Data Centers: Network POP Case Study (2011)

11. Agustí-Torra, A., Raspall, F., Remondo, D., Rincon, D., Giuliani, G.: On the feasibility of collaborative green data center ecosystems. Ad Hoc Netw. **25**, 565–580 (2015)
12. Revista Electroindustria - Tecnologias para Data Centers: Las claves para una energía confiable y eficiente. ElectroIndustria (2014). http://www.emb.cl/electroindustria/articulo.mvc?xid=2230
13. Gowri, K.: Desktop Tools for Sustainable Design. American Society of Heating, Refrigerating and Air-Conditioning Engineers (2005)
14. Bort, J.: Facebook Stores 240 Billion Photos - Business Insider. Business Insider (2013). http://www.businessinsider.com/facebook-stores-240-billion-photos-2013-1. Accessed 15 Jan 2017
15. Lintner, W., Tschudi, B., VanGeet, O.: Best Practices Guide for Energy-Efficient Data Center Design. U.S. Department Energy, pp. i–24, March 2011
16. Grid, T.G.: The Green Grid Dara Center Power Efficiency Metrics: PUE and DCiE (2007)
17. Judge, J., Pouchet, J., Ekbote, A., Dixit, S.: Reducing data center energy consumption. ASHRAE J. **50**, 1–2 (2008)
18. Emerson: Energy Logic: Reducing Data Center Energy Consumption by creating Savings that cascade across Systems (2011)
19. Niles, S.: Virtualization: optimized power and cooling to maximize benefits. Am. Power Convers. White Paper (2008)
20. Ton, M., Fortenbery, B., Tschudi, W.: DC power for improved data center efficiency. Power, January 2007
21. Ray, P.P.: A green initiative to data centers: a review **1**(4), 333–339 (2012)
22. Kleyman, B.B., Carey, R.: Data center optimization: a guide to creating better efficiency and improving rack heat density in air cooled facilities (2016)
23. Dennis, B.: Fundamentals of Power and Cooling Efficiency Zones. Green Grid (2009)

Game-Based Learning: How to Make Math More Attractive by Using of Serious Game

Marián Hosťovecký(✉) and Martin Novák

Department of Applied Informatics and Mathematics, University of SS. Cyril and Methodius, Trnava, Slovak Republic
marian.hostovecky@ucm.sk, martin.novak@google.com

Abstract. The dynamics of change in the field of information technology open the doors to use of new methods of education. Communication bandwidth (networks), computers, laptops, tablets and mobiles (hardware) and new generation of operating systems, program languages and game engines (software) offer new possibilities. Devices and networks are still improved, getting faster and prices fall. These attributes started a new area: game-based learning. In our paper we discuss how to make math more attractive by serious game. Math is subject which is essential in many fields, including natural science, engineering, medicine etc. Generally, math is not very popular with pupils. Question is: how could we change it, but answer is actually not easy, but one of the possibilities is serious games for math. This is the reason why we have decided design and development a serious game focused on math - for reader's attention and math word problems. A multiplatform game engine Unity 3D was used for development of the game and Blender's tool for real time projects for simulations. We describe steps of design and development from graphical environment, design of the buildings to music and sounds.

Keywords: Design · Programming · Science education · 3D serious game · Game-based learning

1 Introduction

Serious games are not just developed and employed exclusively for entertainment purposes, but have successfully being incorporated as learning and training tools for a broad range of different areas [1, 2]. Following the rise of digitalisation, games can be composed with the help of the computer, or can even be adopted for and with the computer, in order to "learn by playing" [3].

Serious games present themselves as one of the more interesting and also promising means of improving cognitive abilities, particularly with pupils [4]. Recently, research has consistently shown that several aspects in cognition such as visual short-memory, multitasking and spatial cognition can be enhanced by game play [5–7]. Substantial research on the design and effects of digital educational games [8–10] has been carried out. In order to utilize educational games, it seems necessary to include instructional

R. Silhavy et al. (eds.), *Cybernetics and Mathematics Applications in Intelligent Systems*, Advances in Intelligent Systems and Computing 574, DOI 10.1007/978-3-319-57264-2_35

features that foster appropriate cognitive processing while at the same time not decreasing players' intrinsic motivation [11].

Given the obvious appeal of computer-based games more generally – the computer game industry is growing much faster than the U.S. economy as a whole and 97% of students aged 12 through 17 play video games regularly – it is easy to understand and embrace the enthusiasm about and promise of computer games as a way to engage kids and lead to meaningful learning [12].

The body of academic literature on web-based games dedicated to increasing mathematical knowledge is still in its infancy compared with other well-developed research streams in education [13].

Unlike traditional learning, where the teacher gives students just theoretical knowledges, by using of problem-solving learning a teacher becomes students some problem tasks. Teacher motivates pupils or students, directs the search for new ways of solving the tasks which pupils acquire new knowledge and new methods of operation. Student discovers itself solve the problem. There is a team of investigators composed of two to five members. This kind of education supports thinking skills and students can apply theoretical knowledges into practical plane.

Problem-based learning as a method of instruction stands firm within the rationalist and, hence, is strongly influenced by cognitive psychology [14]. Another definition of problematic teaching says: "teacher not convey pupils in the final form of knowledge, places for pupils tasks containing unfamiliar to them knowledge and methods of operation, motivates them, directs the search for ways and means to solve problems [15].

Research shows that students in problem-based curricula are indeed learning facts and concepts and the skills needed for critical problem solving and self-learning [14–16]. Other research shows that students participating in these kinds of learning activities are more motivated to learn, that what they learn is more usable than the knowledge learned by students carrying out rote activities, and that they tend to better learn higher order thinking skills than do students in other learning situations [16–18].

Problem solving education should be one of the main goal how to develop this kind of learning, how to support analytical and critical skills. It is important to integrate a game to learning process but very important what kind of a game. A game should be effective beneficial for a students. In last twenty year is very untypical to integrate a game to learning. But in every age of pupils or students is important develop the specific skills.

According to the author Richard Rouse when drafting the game it is important to focus on three main areas that may seem unrelated at the first sight. They are:

- technology,
- gameplay,
- story.

2 Background of the Serious Game

The main goal of our interest was to create a serious game focused on support math skills for pupils of second level of primary schools. Serious game is focused on math word problems - primary goal and the reader's attention – secondary goal. It was

necessary to apply the taxonomy of education, which is often suitable for building cognitive tasks – Niemierko's taxonomy in our country (in Slovakia). Niemierko's taxonomy consists of four levels:

- *Remembering* – recognize or recall information, similar to knowledge in Bloom's taxonomy.
- *Understanding* – demonstrate that the student has sufficient understanding to organize and arrange material mentally, similar to comprehension in Bloom's taxonomy.
- *Specific transfer* – application of acquired information according presented patterns.
- *Non-specific transfer* – creative application of acquired information.

Our game called "The Young Seeker" is based on the single player adventure. Adventure is characterized by the solution of partial tasks during the action. After starting the game the player is shown tip, which is based on the way of commanding figure of the boy. After becoming aware of the keyboard shortcuts, the player starts the game. The beginning of the game takes place in front of the orphanage. It is surrounded by the tree alleys, rocky peaks, and by the fence around the building. The player can look through the orphanage and its surrounding area, as well as, the building itself. The player can move on the pavements and paths that are modeled in the surrounding area.

The plot starts with the player standing in front of the building (orphanage). After figure of boy enters the building he asks the player for his help. The story begins. The boy tells a player the beginning of the story. He finds out that there are paranormal phenomena in the orphanage and the goal is to solve the various subtasks so that the player gets four parts of the diamond. After they are obtained paranormal phenomena in the orphanage ends and the player gets to the next level. During the dialogues with the characters in the orphanage (boy, director of the orphanage, other children), the player must carefully read and consequently solve tasks. The game is quite difficult to find not only particular parts of diamonds, but also to answer the various puzzles that the player gets during the game. If the player does not answer at least half of the questions correctly, the player gets at the beginning of the game, in order to be able to answer individual questions. In the game there are also various Slovak proverbs, the player has to react in accordance with their sense. It strengthens not only the national awareness and national sayings and proverbs, but also the ability to solve various independent tasks (so called Property connectedness).

Overall, the story takes place in the recent past, partly to induce a sense of retro style. Therefore, the individual rooms of the orphanage and an environment are adapted to this period (typical character walls from 60's, 70's, wallpapers, pictures, curtains, chandeliers, staircases, tiles in the bathroom). Throughout the game playing music typical of this period is playing in the background.

During each part of the game, the player has available a list of individual tasks, respectively the task that has to be done right now. Helpdesk is activated via the key "J". Dialogues that are conducted with various people during the level are activated by running the "E" key.

3 "The Young Seeker" – Anatomy of the Serious Game

When designing serious game we remembered the target group that is pupils of Slovak's primary schools. It was necessary to adjust not only to the game as such, but also its design, playfulness, as well as graphic elements of the game and also the player.

3.1 Stage Creation

The first stage was to choose the genre of the game. At this stage, we have set the goal to choose the genre for the game, at the same time, to choose the way in which our story would be interpreted to the player (pupil). Among the fundamental objectives that were set was also the robustness of the game. We have therefore chosen that the game will be designed in 3D and at the same time we want to use the maximum of its potential. That is why we chose that the player will be able to move along the whole area without restrictions, thereby achieving the maximum impact of the environment and stronger overall impression of playing the game. Our goals and requirements were the biggest factor in choosing the genre of our game. With respect to the primary aspects of the game, we could only focus on a few genres that were dominated by adventures. It seemed to be the most favorable choice for our work. At the same time, however, during the game design we decided not to proceed in stereotypical manner and expand the genre of adventure game with elements from other genres, which would contribute to better impression of the game. Among the elements that would support the intended atmosphere of the story are the possibility of moving across the whole stage and a free camera. Our game can be characterized as a single player adventure, in which the player as the main hero of the story overcomes obstacles created by the environment and by the enemy. Here are the elements of the adventure, as the player is forced through interaction with the environment to solve tasks that divide him from further advancement in the game. These tasks usually involve finding hidden objects, solving puzzles and knowledge tasks etc.

The second stage was the game proposal. On that basis, there is following goal in the game: consistent solving of problems that are brought by the game based on interviews and "keys" to achieve further progress of the level story line. More detailed characteristics and description of the game objective can be found in Subsect. 3.2.

The third stage was the design of the structure of the game. After launching the game, player has the option to choose "New Game". The game can be stopped at any time, saved, temporarily stopped and continued. It depends on the decision of the player. More detailed characteristics of the design structure such as models, game interface and sound are described in the following subsection. Each of these attributes was designed so that they together constitute an integrated part and the structure of the game itself.

3.2 Structure of the Design

When we were creating design of the game we had to think especially about the impression that the game will leave. Since the game takes place in recent past, we had to adapt the entire visage of scenes and objects.

Models

Creating objects is among the most basic activities in any development environment and its definition is largely unchanged, although in each area it might look different. In the game engine the object is any object at the scene that can be handled, modified its appearance and behavior towards surrounding or another object.

Unity offers several possibilities to create objects that are the equivalent in practical use and their utilization depends on the user's habits. One way is through the button GameObject on the toolbar that displays the offer of the objects types that can be created. The basic types of objects are 3D objects, 2D objects, Audio, User interface and Light. The other way is to use the Create button at the object Hierarchy. Created objects are visible on the Scene and it is possible to work with them right away.

In terms of the time potential we decided that the proposal of the models design will be combined with the way of their creation. We modeled some models of the objects by ourselves (building, paintings, lamps). The reason was that we had special requirements for objects and it was necessary to model them according our perception. The second option was that we took the opportunity of free models that are available at Assets store. These were adapted according to the scene. There were less realistic, realistic but also fairytale animated models of objects, that allowed us to select from a wide range offer, but we had to take into account the intended dramatic impression of the story, so we chose the most realistic models. Assets store offers in addition to models also other content such as: animations, sounds, or scripts etc. Some of these attributes were also used and adjusted to our needs.

Gaming interface

Gaming interface or Graphic User Interface (GUI) plays an important role in presenting the various elements of the game to the user through images, text and animations; by their interactivity and imagination simplify orientation in propagated opportunities. Unity offers several alternatives for processing the interface that can be used not only as a variety of menu still defined on the screen, as well as interactive tools deployed on the playing area. GUI design elements were created with help of external software for editing 2D graphics. We used mainly freely available programs like Inkscape and Gimp in which we drew pictures to GUI representing key objects in the game.

When creating the menu, we have used a new system of Unit 5, which allows intuitive editing of individual windows and modifying their appearance and functionality by adding components. In this way it is thus possible menu item, which normally acts as a text, adjust so that the component of the button is added. Using this principle, we have compiled a minimalistic main menu, as well as other windows (Figs. 1 and 2).

Fig. 1. Introduction view of the "Young Seeker"

Fig. 2. Looking for a persons for conversation in sleeping room

Sound

In the next section we applied the soundtrack to particular parts of the game. The basis was to find the appropriate music in order to create the desired atmosphere of the game. In the Unity the work with sound is very intuitive, since the use of sound reflects its perception from the real world. It is therefore necessary to determine the source of the sound and the object that it hears, the audience. The sound source can be any object in the scene and the audience is mostly camera, which is essentially our eyes and ears. Unity has functions to convert the sound intensity according to the distance of the source from the listener, which effectively simulates the perception of 3D area. Individual sounds can be assigned their priority in order not to mix together and not to create meaningless noise. This feature can be used to separate less important sounds, such as background music, wind, etc., from more significant sounds, such as the voice of the character during dialogue, collection of the key objects etc.

3.3 Development Tools

Currently the developer of the games, who decides to design a game on any platform, can choose from several engines from different producers, based on the popularity of the game or by the offer that engine has to offer. Among the most famous and at the same time most used game engines belong Unreal Engine 4, Unity 3D, Source 2, CryEngine 3 and Frostbite 3.

Unity 3D - A suitable tool for game development after analyzing software options was the Unity 3D. The reason why we have decided to implement in our game just Unity 3D, is its ease of operation and compatibility with a wide variety of formats for different types of media such as graphics, audio, video and text. In addition, the "Unity Technologies carried out a coup in the gaming industry through Unity, advanced and emerging platform for creating games and interactive 3D and 2D experiences such as learning simulations and medical and architectural views through different platforms such as mobile, PC, web, console and other" [19]. Unity 3D like most of today's software WYSIWYG (What You See Is What You Get) has its own environment divided into several working windows that can be adapted to suit our needs by their enlargement,

reduction, or changing location in order to be most natural during working. Other reasons for our choice is that the software has been around for several years and it is connected to plenty of means that serve to educate users about working in this environment, whether there are training videos or entire online documentation of scripts and programming methods useful in the creation of any content. Furthermore, we are interested in the possibility of downloading different content from Assets store intended for further development, which clearly ease the development of the game itself.

Programming languages - Unity generally supports several programming languages that are used in programming scripts. These scripts contain encoded logic of objects' behavior that are present in the game. Among the programming languages that are supported by Unity are:

- C#;
- Java Script;
- Boo;

For C# and Java Script exists online documentation of all programming methods, in which there can be found their methods of using, properties, and parameters used when working with them. Each programming script can be saved to a separate file and managed in the Project window like any other file.

Furthermore, it is necessary to assign it to the object, but usually to the object for which the script was dedicated and whose behavior is to modify. This is not a requirement, since it is possible for the objects and their components to be referenced directly in the programming script.

3.4 Math Example of Word-Solving Task

Our game is focused on math word problems and logical thinking. While playing the game, pupil proceeded by first completed several interviews with specific characters in the game. Always it is necessary the rank ok interview with the persons (boys, girls, teacher…). Therefore, if a pupil-player wrong order came to the person or the right one in order to display the messages, "You cannot start a conversation, please find another person". After completing the interviews followed the search objects in the game such: diamond, gold, shields, etc. The game stops and then displays the question at each object. The question always was related to interviews before. Pupils should answer for this question. Questions are focused on simple verbal math problems. We choose as example one specific math problem: Player completes talks with two girls (sisters). Bianca & Mary saved a different amount of money: Bianca has got 60 €, Mary 140 €. Pupil should remember these amounts. After that shows a question: "How many euro saved sisters together?" If a student answered correctly, got a second question, if a student answered incorrectly, game player goes back to the conversation with the first sister and read the conversation again. Second question is: "Bianca and Mary have saved 200 € together. Bianca wants to go on a trip and wants to take a fifth of their savings and Mary quarter. They will have a total of 50 € on a trip. How many euro saved Mary and how many Bianca?"

In this case we are watching two factors:

- readers' attention;
- math word problems.

The answer is unlimited in time. Pupil can choose the correct answer from one of the five options. It was always the only one correct answer. On the first question a student has got three choices (easier question), in the second questions student has got a five options. The game is designed for fixing phase of the learning process. The aim is to fix the curriculum after completion of the exposure part of the learning process. The first level of the game contains 10 mathematical word problems for pupils of 5th year of primary schools. A choice word problem has consulted with the sixth primary school mathematics teachers.

Among the most frequently detected shortcomings of logical character are wrongly activated triggers whose presence on the scene meant that players were able to skip part of the story. This did not meet certain prior conditions for progress in the story line, which meant that the story could not be further completed. Such errors were easy to locate and eliminate because their activation was mentioned in the script, but they had to be switched off by that time. It was enough to deactivate them; thereby we prevented skipping the story.

It seemed to some players that commanding of the character was clumsy. In order to contribute to a better gameplay rotation of the character should accelerate, but the rate of movement of the character should slow down. By modifying these parameters, we met their criteria, but such adjustments usually depend on individuals and adjusting these figures can disadvantage other players. However, we can say that after gaining certain amount of experience in working with Unity, it could be possible to make other solutions to movement of the character that would be universal and would not create any more problems.

4 Conclusions and Future Work

In the theoretical part we dealt with the characteristics of computer games, serious game. We were creating a didactic computer game (serious game) for fix math skills in fifth level of primary schools.

In practical part, we have created simple and funny game that meets the didactic aim of mastering the curriculum. Testing showed the justness of that argument. After analyzing theoretical documents of computer graphics, we were able to create 3D models of objects in the environment, which we then used in constructing scenes of the game. That preceded the drafting of games, specifying its attributes based on our choice, and determination of the genre that would be suitable for such purposes. Analysis of freely available game engines followed, which purpose was to compare the functional characteristics of the engines and choose the best for our work.

Nowadays, with the spread of information and communication technologies, these games should be implemented into school curriculum even more. Pupils would be involved more into the process of education. The game is designed so that it can be

further developed for objects and settings. The game could offer number of illustrative examples focused on various topics of problem-solving education. Very important will be obtain some results about impact of communication and opinions among pupils and teachers during use of the game. It will be important to find out what kind of role or activities will have a teacher and how games can help to pupils in finding employment [20–23, 26].

In addition there is possible to integrate to the serious game a special kind of technical movement as motion capture technology [24] which can improve movements and graphics of serious games. Besides that it would be great if the grid or cloud systems can be used to dynamically allocate resources during the game design in order to minimize total maintenance costs [25].

At the end of this paper we want to say:

- Learning with the game, is possible to obtain a higher cognitive results of pupils with support of appropriate methodological-didactic approaches;
- Higher motivation and interest;
- Development problem-solving;
- Friendly atmosphere during class.

Acknowledgements. This work was supported in part by the University of SS. Cyril and Methodius in Trnava under the institutional project FPPV FPV 2/2014.

References

1. Breuer, J., Bente, G.: Why so serious? On the relation of serious games and learning. Eludamos: J. Comput. Game Cult. **4**(1), 7–24 (2010)
2. Ritterfeld, U., Cody, M.J., Vorderer, P.: Serious Games: Mechanisms and Effects. Routledge, New York (2009)
3. Ritterfeld, U.: Beim Spielen lernen? Ein differenzierter Blick auf die Möglichkeiten und Grenzen von Serious Games. Comput. + Unterr. **84**, 54–57 (2011)
4. Castellar, E.N., All, A., de Marez, L., Looy, J.V.: Cognitive abilities, digital games and arithmetic performance enhancement: a study comparing the effects of a math game and paper exercises. Comput. Educ. **85**, 123–133 (2015)
5. Štubňa, J.: Selected determinants influencing on student motivation in creating a relationship towards of science. Acta Humanica **13**(1), 68–77 (2016)
6. Štubňa, J.: The importance of educational game as a part of teacher students preparation in the context of developing intersubject relationships in science subject. Acta Humanica **13**(2), 39–45 (2016)
7. Bavelier, D., Green, C.S., Pouget, A., Schrater, P.: Brain plasticity through the life span: learning to learn and action video games. Annu. Rev. Neurosci. **35**, 391–416 (2012). Hyman, S.E. (ed.)
8. Clark, D.B., Tanner-Smith, E.E., Killingsworth, S.S.: Digital games, design, and learning. A systematic review and meta-analysis. Rev. Educ. Res. **86**(1), 79–122 (2016)
9. Romero, M., Usart, M., Ott, M.: Can serious games contribute to developing and sustaining 21st century skills? Games Cult. **10**(2), 148–177 (2015)

10. Wouters, P., van Nimwegen, C., van Oostendorp, H., van der Spek, E.D.: A metaanalysis of the cognitive and motivational effects of serious games. J. Educ. Psychol. **105**(2), 249–265 (2013)
11. Hawlitschek, A., Joeckel, S.: Increasing the effectiveness of digital educational games: the effects of a learning instruction on students' learning, motivation and cognitive load. Comput. Human Behav. **72**, 79–86 (2017). http://www.sciencedirect.com/science/article/pii/S0747563217300523
12. McLaren, B.M., Adams, D.M., Mayer, R.E., Forlizzi, J.: A computer-based game that promotes mathematics learning more than a conventional approach. Int. J. Game-Based Learn. (IJGBL) **7**(1), 36–56 (2017)
13. Fellnhofer, K.: All-in-one: impact study of an online math game for educational purposes. Int. J. Technol. Enhanced Learn. **8**(1), 59–76 (2016). http://www.inderscience.com/info/inarticletoc.php?jcode=ijtel&year=2016&vol=8&issue=1
14. Norman, G.T., Schimdt, H.G.: The psychological basis of problem-based learning: a review of the evidance. Acad. Med. **67**, 557–565 (1992)
15. Petlák, E.: Všeobecná didaktika. Iris, Bratislava (1997). ISBN: 80-88778-49-2
16. Hmelo, C.E.: Problem-based learning: development of knowledge and reasoning strategies. In: Proceedings of the Seventeenth Annual Conference of the Cognitive Science Society. Erlbaum, Hillsdale (1995)
17. Blumenfeld, P., Soloway, E., Marx, R., Krajcik, J., Guzdial, M., Palincsar, A.: Motivation project-based learning: sustaining the doing, supporting the learning. Educ. Psychol. **26**(3&4), 369–398 (1991)
18. Kolodner, J.L., Hmelo, C.E., Narayanan, N.H.: Problem-based learning meets case-based reasoning. In: Proceedings of the 1996 International Conference on Learning Sciences, pp. 188–195 (1996)
19. Unity Technologies. What is Unity 5? (2016). https://unity3d.com/unity
20. Toman, J., Michalík, P.: Possibilities of implementing practical teaching in distance education. Int. J. Mod. Educ. Forum **2**(4), 77–83 (2013)
21. Mišútová, M., Mišút, M.: Impact of ICT on the quality of mathematical education. In: Proceedings of the 6th International Multi-Conference on Society, Cybernetics and Informatics, IMSCI 2012, pp. 82–86 (2012)
22. Mišút, M., Pribilová, K.: Communication impact on project oriented teaching in technology supported education. In: Elleithy, K., Sobh, T. (eds.) Innovations and Advances in Computer, Information, Systems Sciences, and Engineering. LNEE, vol. 152, pp. 559–567. Springer, New York (2013). doi:10.1007/978-1-4614-3535-8_47
23. Gregáňová, R., Országhová, D.: K novým kompetenciám učiteľa matematiky v kontexte elektronického vzdelávania. In: Zborník vedeckých príspevkov z medzinárodnej vedeckej konferencie "The 6rd International Conference APLIMAT", pp. 337–343. STU, Bratislava (2007). ISBN: 978-80-969562-8-9
24. Ölvecký, M., Gabriška, D.: Motion capture as an extension of web-based simulation. Appl. Mech. Mater. **513**, 827–833 (2014)
25. Šimon, M., Huraj, L., Siládi, V.: Analysis of performance bottleneck of P2P grid applications. J. Appl. Math. Stat. Inf. **9**(2), 5–11 (2013). ISSN: 1336-9180. (IET Inspec)
26. Hlaďo, P.: Quality of lifelong learning in the Czech Republic: paradigms, development and perspectives in the European context. In: Tamášová, V. (ed.) Quality in the Context of Adult Education and Lifelong Education. Dubnica Institute of Technology, Dubnica nad Váhom, pp. 44–51 (2013). ISBN: 978-80-89400-53-9

Intelligent Telemetry Data Analysis of Small Satellites

Vadim Skobtsov[1(✉)], Natalia Novoselova[1], Vyacheslav Arhipov[1], and Semyon Potryasaev[2,3]

[1] United Institute of Informatics Problems of National Academy of Sciences of Belarus, Minsk, Belarus
{vasko_vasko,novos65}@mail.ru, arkhipau@gmail.com
[2] St. Petersburg Institute of Informatics and Automation of Russian Academy of Sciences (SPIIRAS), St. Petersburg, Russia
[3] St. Petersburg National Research University of Information Technologies, Mechanics and Optics (ITMO), St. Petersburg, Russia
spotryasaev@gmail.com

Abstract. The paper presents intelligent telemetry data analysis software module and methods of onboard equipment of small satellites. The suggested software module consists of feature selection, data preprocessing, clustering and predicting software components. The software components are based on the genetic algorithm based feature selection method, dynamic streaming clustering method, neural Kohonen self-organizing map and image processing based clustering and predicting methods for telemetry data of onboard equipment of small satellites. The computational experiments and testing of developed methods and software tools were performed on the processed telemetry data from the navigation device of onboard equipment of a small satellite and showed enough high efficiency and good results.

Keywords: Intelligent telemetry data analysis · Data preprocessing · Feature selection · Clustering · Predicting · Kohonen self-organizing map · Image segmentation · Satellite onboard equipment · Failures

1 Introduction

One of the important tasks of safety evaluation of the complex technical systems such as small satellites is the evaluation of their reliability.

A huge amount of information received and partly accumulated in specialized databases at the design, testing and operation of the onboard equipment (BE) of small satellites (SS) can be effectively used to improve the reliability evaluation process of BE SS and its individual components. The operational data of BE SS including telemetry data is irregular heterogeneous data. The intellectual data analysis (data mining) methods and software tools are strong and effective instruments for the processing and analysis of this type of data. Therefore the development of intelligent data analysis and predicting methods and according software tools for BE SS telemetry data is actual problem.

In our paper we suggest the software module and methods of intelligent BE SS telemetry data analysis which allow to extract the useful information and construct the cluster (state) structures and prognostic estimation of reliability and operability of BE SS.

R. Silhavy et al. (eds.), *Cybernetics and Mathematics Applications in Intelligent Systems*, Advances in Intelligent Systems and Computing 574, DOI 10.1007/978-3-319-57264-2_36

2 The Structure and Functionality of Intelligent BE SS Telemetry Data Analysis Software Module and Methods

Intellectual data analysis software module was designed for data mining of the BE SS telemetry, selection of potential states of the analyzed devices by clustering and predictive estimating the average number of failures across the entire data set and within the clusters-states of the device. Bellow the structural and functional flowchart of the BE SS telemetry data analysis software module is provided (Fig. 1).

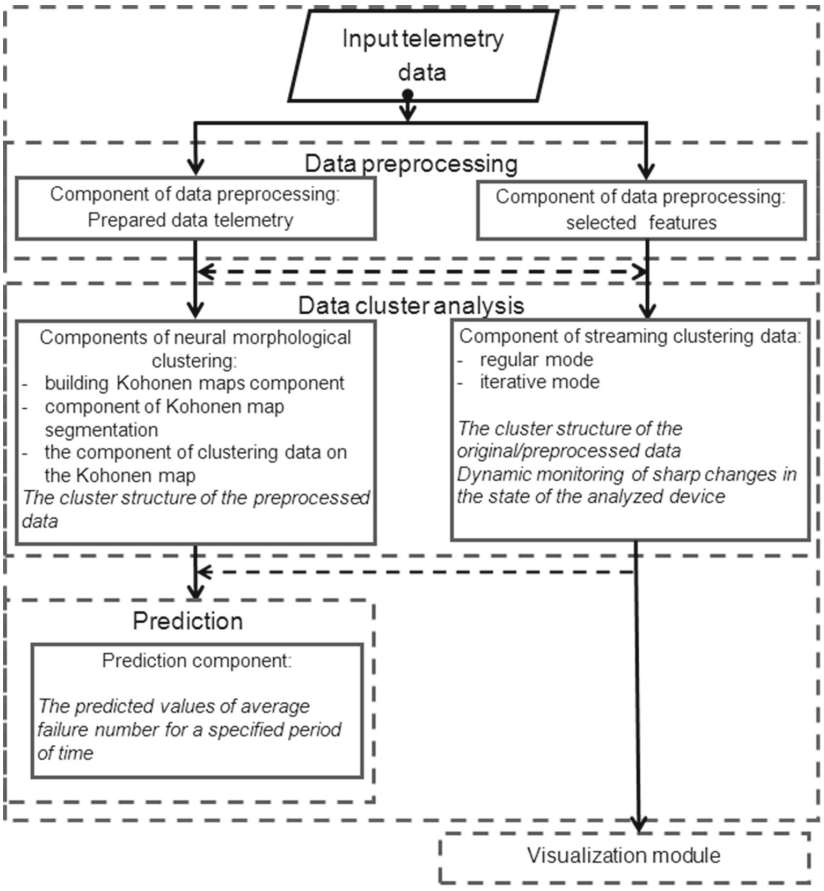


Fig. 1. Structural and functional flowchart of the BE SS telemetry data analysis software module

Unit of data preprocessing includes following components:

– component of data preprocessing, forming the table of vectors prepared on the basis of telemetry vectors and required for component of neural morphological clustering;
– second component of data preprocessing allows to select informative features for the streaming cluster analysis with developed genetic algorithm.

The first three components of the cluster analysis unit perform:

- generating Kohonen map based on the preprocessed telemetry data of the analyzed device;
- clustering visual space data by threshold segmentation of inter neuronal distances map;
- clustering the preprocessed telemetry data set through marking the data points.

The fourth component of the unit – streaming clustering, implements the streaming dynamic data clustering based on the two-level hierarchical approach (online/offline) of micro/macro clustering.

In the block of prediction, which consists in fact only one component of prediction, it is executed the evaluation of the forecast values of average failure number for a specified period of time.

3 Basic Methods and Software Components

3.1 Feature Selection Method and Software Component for Multidimensional Data Using a Genetic Algorithm

Suggested feature selection algorithm is based on the using of genetic algorithm (GA) for solving the problem of multi-criteria optimization. As a basic GA multi-criteria optimization algorithm PESA-II is used [1]. The general flowchart of the proposed feature selection method is shown in Fig. 2.

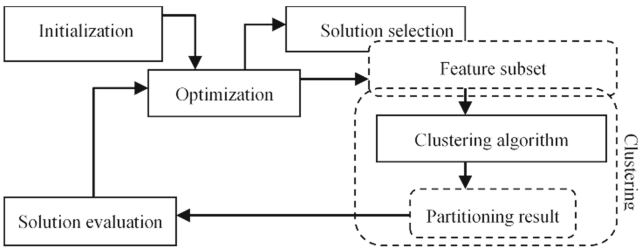


Fig. 2. The general flowchart of the proposed feature selection method

The following encoding individuals (solutions) was suggested. The subset of selected features and cluster number are encoded in the individual. The binary characteristic vector is used for feature subset encoding. The cluster number is encoded with the grey 4-bit code. For the first part of the individual GA uses uniform crossover and bitwise mutation. For the second part of the individual the bitwise mutation operation is only applied as recombination.

Developed genetic algorithm of feature selection solves the two-criteria optimization problem. The first optimization criterion evaluating is the quality index of data objects partitioning S - Silhouette Width [2]. To calculate the index only data objects are used without the cluster centers:

$$S = \frac{1}{n}\sum_{i=1}^{n}\frac{b_i - a_i}{\max\left(b_i, a_i\right)}, \tag{1}$$

where a_i – the average distance between object i and the all others objects of current cluster, b_i – the average distance between object i and the all objects of the closest cluster, n – the number of data objects.

As a second optimization criterion the power of selected feature subset d_U is used. For evaluating every GA individual the clustering with k-means algorithm is applied, which computational complexity is proportional to object number. So the following two-criteria optimization problem is solved $f_1 = \mathrm{S} \to \max$ $f_2 = d_U \to \max$.

The solution of these optimization problem is two-dimensional Pareto front of optimal solutions which accords to feature subset of different dimension.

The developed method was implemented as feature selection software component which is combination of command line computational application FeatureSel.exe and interface application FSGUI.exe. The main parameters of feature selection software component are: INPUT – the name of data file, POPSIZE – population size, NUMGEN – generations number, PROBCROSS – crossover probability, FMUT – mutation probability for feature part of individual, CMUT – mutation probability for cluster number part of individual.

The example of feature selection software component work is represented on Fig. 3. A component can be used via the interface (Fig. 3) and from the command line. In the second case the parameters are entered from a parameter file whose name is specified in the command line as a parameter *FeatureSel.exe param.txt*.

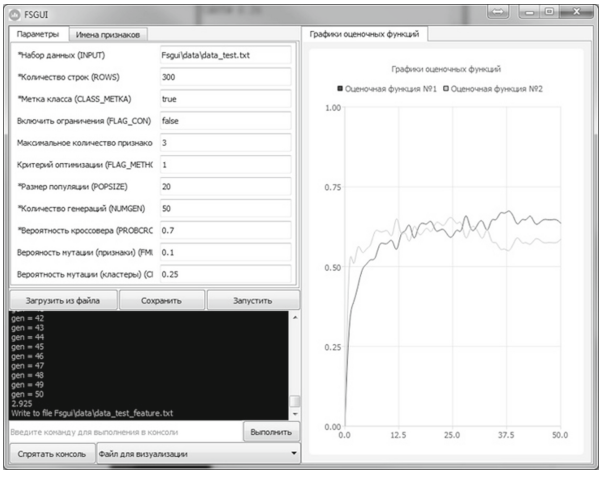


Fig. 3. The example of feature selection software component work

3.2 Streaming Clustering Method and Software Component

Method for dynamic streaming data clustering is a two-level combination of clustering algorithms for data streams, with the possibility to save the intermediate results of clustering in the form of microclusters [3, 4].

Definition. Microcluster is a cortege $\left(\overline{CF2^x}, \overline{CF1^x}, CF2^t, CF1^t, n\right)$, where $\overline{CF2^x}$ corresponds to the d-dimensional vector, which stores the sum of the squares of the data objects values, p-th element of the vector is defined as $\sum_{j=1}^{n}\left(x_{i_j}^p\right)^2$; $\overline{CF1^x}$ stores the sum of the data objects values, p-th element of the vector is defined as $\sum_{j=1}^{n} x_{i_j}^p$; $CF2^t$ and $CF1^t$ stores the sum of the squares and sum of time points $T_{i_1}, \ldots, T_{i_n}$ accordingly; n – number of data objects.

Generation of microclusters is performed in the online clustering stage. This automatic process is designed to collect statistics with sufficient temporal and spatial detail, which can be used effectively in the offline stage of clustering to allocate macroclusters and to analyze the dynamical evolution of the cluster structure.

This approach allows to recover the global cluster structure of the data by constructing macroclusters in an arbitrary time according to the stored data of microclusters. The offline stage of macroclustering consists of following steps:

(1) To set the time interval h and the number of macro clusters k. Consider two saved frames of microclusters in the current time t_c and the time $t_c - h$.
(2) To determine the microclusters $N(t_c, h)$ corresponding to the time interval h.
(3) To make reclustering of the microclusters $N(t_c, h)$ using a high-level algorithm of modified K-means.
(4) To select the initial centers of macroclusters from the selected microclusters with a probability proportional to their number.
(5) Calculate the distance for each microcluster to the centers of macroclusters.
(6) To modify the centers of macroclusters according to the calculated distances.
(7) Stop condition is satisfied? Yes – stop. No – goto the step 5.

The developed method was implemented as streaming clustering software component which is also combination of command line computational application StreamingClust.exe and interface application SCGUI.exe. The main parameters of streaming clustering software component are: INPUTTRAIN – the name of data file, FLAG_INCREMENT – flag to enable/disable the iterative dynamic clustering, TIMEWINDOW – the value of time interval, MAXKERNEL – number of microclasters, kMACRO – number of macroclasters.

The streaming clustering software component was designed to identify groups of objects telemetry data with similar characteristics and allows:

(1) Dynamically cluster big telemetry data by phased loading into the computer's memory from a file on disk, or by reading information from a data source;
(2) Perform clustering in two modes: static clustering (offline) - macroclasters, dynamic iterative clustering (online) – microclasters;
(3) Save results of online clustering as a set of microclusters;

(4) Save results of offline clustering as a set of macroclusters;
(5) Evaluate the clustering results and visualize it with graphic (Fig. 4) that allow to monitor state transitions of analyzed device or system;

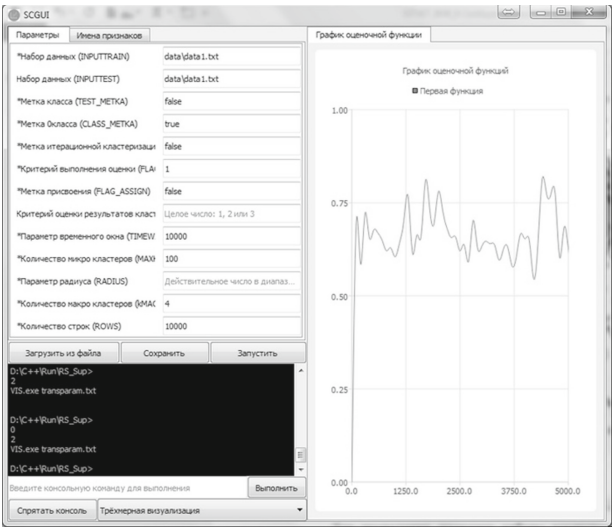


Fig. 4. The example of streaming clustering software component work

(6) Make 2-D and 3-D visualization of clustering results (Fig. 5), in this case additional developed component of visualization is used.

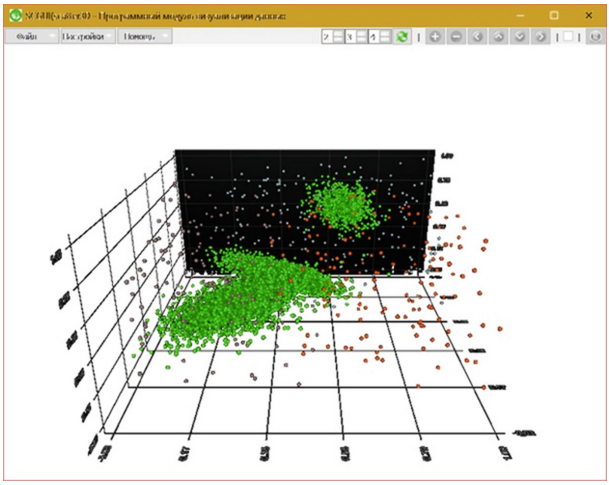


Fig. 5. Example of 3-D visualization of clustering results

The example of streaming clustering software component work is represented on Fig. 4. A component can be used via the interface (Fig. 4) and from the command line also. The parameters entered in the way same to previous component.

3.3 Neural Kohonen Self-Organizing Map Based Methods and Software Components Group

Kohonen self-organizing map (SOM) based software components group consists of following components:

- component of data preprocessing, forming the table of vectors prepared on the basis of telemetry vectors and required for component of neural clustering and predicting;
- generating Kohonen SOM component;
- Kohonen SOM based data space clustering component of the Kohonen map;
- Kohonen SOM based dataset clustering component;
- component of predicting value of average failure number for a specified period of time.

Due to restrictions on the length of the paper, we are forced to skip the description of the methods of generating Kohonen SOM [5], and clustering based on known algorithms of segmentation and image filtering, as well as the method of watershed.

Method and component of data preprocessing. The component was designed to transform the initial telemetry data for better clustering and predicting results: on the one hand the smoothing of the noise, on the other hand strengthening and stabilization of transitions indicating a change in system state. It was established experimentally that the data preprocessed by this component, showing a greater stability from the standpoint of continuous residence in certain clusters: increasing the average time of continuous residence in the cluster, reduces the probability of transition to another cluster (Fig. 6). It gives up the better quality of predicting.

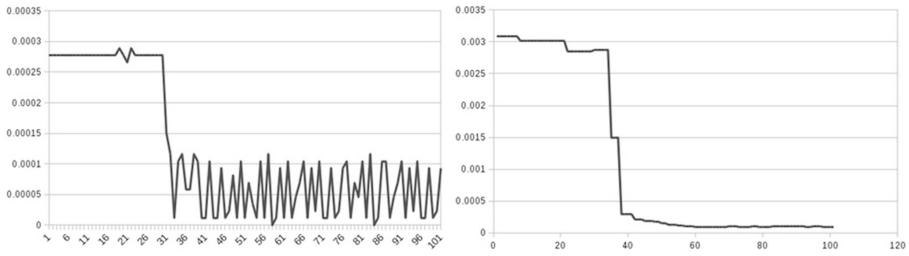


Fig. 6. Example of telemetry data preprocessing

The component implements the following transformation rules of input data channels:

- identical transformation (NONCH);
- the last nonzero variation (LSTCH);

– smoothed with a Butterworth filter of the 1st order of the absolute value of the signal second derivative (SECDR), $\alpha \in (0, 1)$ – smoothing parameter;

$$y_i = \begin{cases} 0.5(x_1 + x_3) - x_2, i = 1, 2 \\ y_{i-1} + \alpha(0.5(x_{i-1} + x_{i+1}) - x_i - y_{i-1}), i > 2 \end{cases}; \quad (2)$$

– raising to a power (POW);
– the average of the set of several local minima increments of the original data (INF);
– the average of the set of several local maxima increments of the original data (SUP).

The developed data preprocessing method was implemented as component of data preprocessing. It is also combination of command line computational application _dataprep.exe and interface application Kohonen.exe. All command line computational application of neural Kohonen SOM based software components group have common user interface which is provided by Kohonen.exe application (Fig. 7). For running corresponding component it is necessary to activate correspondent window tab (Fig. 7). The main parameters of data preprocessing software component are: INPUT – the name of data file, IN_SPACE – input data vector dimension, LEN – input data vectors number, OUT_SPACE – output data vector dimension, the set of transformation rules in right section of interface window (Fig. 7). A component can be used via the interface (Fig. 7) and from the command line like all developed tools.

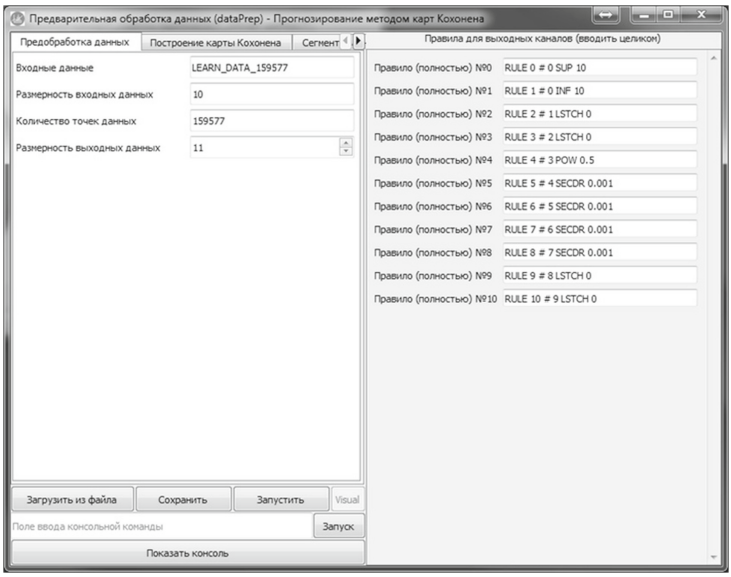


Fig. 7. Neural Kohonen SOM based software components group interface

The method and component of predicting. In the first stage of the predicting methods it is generated a list of values of the failure numbers for all possible periods that are defined on a given data sequence. In the next phase, the values of average failure number

for a specified period of time are calculated, as for a full list and sublists based on the values of the cluster labels. The algorithm pseudocode is below.

Here i – data point index; N – maximum index value; L(i), T(i), S(i) – cluster label, time and validity status for point i; P – time period; Q – the list of values of failure numbers without cluster label; Q_L – the list of values of failure numbers with cluster label; Flag – indicator of data set end.

```
Q <- Empty; Curr_I <- 0; Flag <- False
While (not(Flag))
  Work_I <- Curr_I+1
Start_t <- T(Work_I)
Work_n <- 0
  While (T(Work_I) - Start_t  < P and Work_I <= N)
    If(S(Work_I - 1) = 1 and S(Work_I) = 0)
      Work_n <- Work_n + 1
    End_if
    Work_I <- Work_I + 1
  End_while
  if ( Work_I <= N )
  Q <- Q + Work_n
  Curr _I <- Curr_I+1
else
  Flag <- true
End_if
  End_while
Out <- LABEL 0 #
Out <- Length (Q)
Out <- Average_value(Q)
For (Curr_L <- 1:MAX(L))
  Q_L <- Empty
  For (Curr_I <- 0: Length (Q))
    If (L(Curr_I) = Curr_L)
      Q_L <- Q_L + Q(Curr_I)
    End_if
  End_for
  Out <- LABEL Curr_L #
Out <- Length (Q_L)
If (Length (Q_L) > 0 )
    Out <- Average_value (Q_L)
    Else
    Out <- Average_value (Q)
    End_if
  End_if
```

The developed method of predicting was implemented as corresponding component. It is also combination of command line computational application _calcparams.exe and

interface application Kohonen.exe. The main parameters of predicting software component are: INPUT – the name of data file, LEN – input data points number, PERIOD – period of time (specified in seconds), which is used to calculate the predictive or current value of average number of failures; TIME_STATUS – file name of modes labels (correct or failure). The program output is a text file containing the list of the values of average failure number for a specified period of time (PERIOD), the number of analyzed time periods for the whole data set and for every cluster.

4 Computational Experiments

As data set to demonstrate the efficiency of the developed software module and methods of intellectual BE SS telemetry data analysis the processed telemetry data from the navigation device of onboard equipment of a small satellite are used. The number of vectors in telemetry data set during two month of performance is equal to 331086. Every telemetry vector is 10-dimensional vector where the first parameter is binary label of correct or failure performance mode of device, and the other nine parameters are real valued numbers.

The whole data set is divided into two parts: training data set of the first month and test data set of the second month of navigation device performance. The dimension of the training data set is159 577 time points (telemetry vectors) and the test data set is 171 509 time points.

At the first stage the telemetry data from the training data set were preprocessed, the Kohonen SOM was generated for preprocessed data set. The preprocessed data set were carried out the clustering (data space image segmentation clustering and dataset clustering) based on generated Kohonen SOM. For obtained clustering results there was calculated the list of the values of average failure number for a specified period of time for the whole data set and for every cluster. At the second stage the list of the values of average failure number for a specified period of time was calculated for the test data set based on the Kohonen SOM of training data set. As time periods of predicting were considered 3600 s, 86400 s and 604800 s.

Statistical tests for the coincidence of the calculated values for training and test data set are significant with a confidence probability level of the value about 0.9.

5 Conclusion

In the paper we considered developed intelligent telemetry data analysis software module and methods of onboard equipment of small satellites. The suggested software module are the set of data analysis software components which cover a wide spectrum of data mining stages and approaches: feature selection, data preprocessing, clustering and predicting software tools. The software components are based on the genetic algorithm based feature selection method, dynamic streaming clustering method based on the two-level hierarchical approach (online/offline) of micro/macro clustering, neural Kohonen SOM and image processing based data preprocessing, clustering and predicting methods for telemetry data of onboard equipment of small satellites. Note

that streaming clustering method and component is applicable not only for generating cluster structure of data but for dynamic monitoring of state transition of analyzed system. And the neural Kohonen SOM based software and methods group is efficient in clustering and predicting average failure rate value that was proved by the computational experiments and testing on the processed telemetry data from the navigation device of onboard equipment of a small satellite.

Acknowledgments. The research described in this paper is partially supported by the Russian Foundation for Basic Research (grants **15-07-08391, 15-08-08459, 16-07-00779, 16-08-00510, 16-08-01277, 16-29-09482-ofi-i, 17-08-00797, 17-06-00108, 17-01-00139, 17-20-01214**), grant **074-U01** (ITMO University), project **6.1.1** (Peter the Great St. Petersburg Politechnic University) supported by Government of Russian Federation, Program STC of Union State "Monitoring-SG" (project **1.4.1-1**, project **6МСГ/13-224-2**), state order of the Ministry of Education and Science of the Russian Federation №2.3135.2017/K, state research 0073–2014–0009, 0073–2015–0007, International project ERASMUS +, Capacity building in higher education, № 73751-EPP-1-2016-1-DE-EPPKA2-CBHE-JP, Innovative teaching and learning strategies in open modelling and simulation environment for student-centered engineering education.

References

1. Corne, D.W., Jerram, N.R., Knowles, J.D., Oates, M.J.: PESA-II: region-based selection in evolutionary multiobjective optimization. In: Proceedings of the Genetic and Evolutionary Computation Conference, pp. 283–290. Morgan Kaufmann Publishers, San Francisco (2001)
2. Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. **20**(1), 53–65 (1987)
3. Aggarwal, C.C.: A framework for diagnosing changes in evolving data streams. In: Proceedings of ACM SIG-MOD Conference (2003)
4. Aggarwal, C. (ed.): Data Streams Models and Algorithms. Springer, Heidelberg (2007)
5. Kohonen, T.: Self-Organizing Maps. Springer, Heidelberg (2001)

A Static Calibration of MEMS Accelerometers

Martin Sysel(✉)

Faculty of Applied Informatics, Tomas Bata University in Zlín, Zlín, Czech Republic
Sysel@fai.utb.cz

Abstract. The paper describes micro electro mechanical systems (MEMS) accelerometers and their calibration making reliable and accurate measurements. The first part discusses the physics of acceleration and accelerometers. The next part describe one of the calibration techniques. The final section shows static calibration of a 3D digital linear acceleration sensor in LSM303D.

Keywords: MEMS · Accelerometer · Calibration

1 Introduction

An accelerometer is an electromechanical device that measures acceleration forces. These forces may be static, like the gravity, or they could be dynamic - caused by moving or vibrating. Three-axis accelerometers supplied for the consumer market are typically calibrated by the sensor manufacturer using a six-element linear model comprising a gain and offset in each of the three axes. This factory calibration will change slightly as a result of the thermal stresses during soldering of the accelerometer to the circuit board. Rotation of the accelerometer package relative to the circuit board and misalignment of the circuit board to the final product will also add small errors. The original 6-parameter (gain and offset in each channel) factory calibration can be recomputed 12-parameter calibration to correct outputs. Own calibration improve accuracy for high-accuracy applications.

2 Accelerometer

2.1 Physical Principles and Structure

An accelerometer is an electromechanical device that measures acceleration forces. These forces may be static, like the gravity, or they could be dynamic - caused by moving or vibrating. There are many types of accelerometers and there are many different ways to make an accelerometer. Some accelerometers contain microscopic crystal structures that get stressed by accelerative forces use the piezoelectric effect. Another sensing changes in capacitance. Capacitive sensing has excellent sensitivity.

Typical MEMS accelerometer is composed of movable proof mass with plates that is attached through a mechanical suspension system to a reference frame, as shown in Fig. 1. A MEMS accelerometer differs from integrated circuits in that a proof mass is

R. Silhavy et al. (eds.), *Cybernetics and Mathematics Applications in Intelligent Systems*,
Advances in Intelligent Systems and Computing 574, DOI 10.1007/978-3-319-57264-2_37

machined into the silicon. Any displacement of the component causes this mass to move slightly according to Newton's second law, and that change is detected by sensors. Usually the proof mass disturbs the capacitance of a nearby node; that change is measured and filtered. Movable plates and fixed outer plates represent capacitors. The deflection of proof mass is measured using the capacitance difference [7]. The free-space (air) capacitances between the movable plate and two stationary outer plates C_1 and C_2 are functions of the corresponding displacements [4].

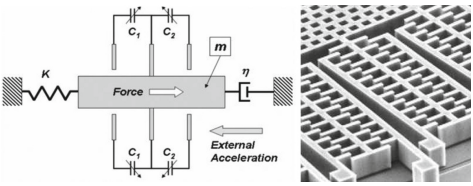


Fig. 1. Schematic and internal structure of a capacitive accelerometer [4]

The most important specification is the number of axis. The MEMS proof mass can measure only one parameter in each available axis, so a one axis device can sense acceleration in a single direction. Three axis units return sensor information in the X, Y, and Z directions.

Accelerometer sensors measure the difference between any linear acceleration in the accelerometer's reference frame and the earth's gravitational field vector. The earth's gravitational field is defined by a force vector that points directly down towards the earth's core. In the absence of linear acceleration, the accelerometer output is a measurement of the rotated gravitational field vector and can be used to determine the accelerometer pitch and roll orientation angles. The orientation angles are dependent on the order in which the rotations are applied. The most common order is the aerospace sequence of yaw then pitch and finally a roll rotation [2].

Three-axis accelerometers supplied for the consumer market are typically calibrated by the sensor manufacturer using a six-element linear model comprising a gain and offset in each of the three axes. This factory calibration will change slightly as a result of the thermal stresses during soldering of the accelerometer to the circuit board. Rotation of the accelerometer package relative to the circuit board and misalignment of the circuit board to the final product will also add small errors. The original factory accelerometer calibration is adequate for the majority of consumer applications. However, own calibration improve accuracy for high-accuracy applications.

2.2 Accelerometer Calibration

The relation between corrected accelerometer output and accelerometer raw measurements [1, 7] can be expressed as

$$\begin{bmatrix} A_x \\ A_y \\ A_z \end{bmatrix} = \left[Am_{xyz}\right]_{3x3} \begin{bmatrix} \frac{1}{S_x} & 0 & 0 \\ 0 & \frac{1}{S_y} & 0 \\ 0 & 0 & \frac{1}{S_z} \end{bmatrix} \cdot \begin{bmatrix} R_x - O_x \\ R_y - O_y \\ R_z - O_z \end{bmatrix} \tag{1}$$

Where:

A is the corrected reading in X, Y and Z axes.
Am is the 3×3 misalignment matrix between sensor axes and device axes.
S is the sensitivity of each channel.
R is the raw data from the accelerometer.
O is the accelerometer's zero-g level.

The sensitivities and offsets are constants, so the matrix can be simplified to:

$$\begin{bmatrix} A_x \\ A_y \\ A_z \end{bmatrix} = \begin{bmatrix} C_{11} & C_{12} & C_{13} \\ C_{21} & C_{22} & C_{23} \\ C_{31} & C_{32} & C_{33} \end{bmatrix} \cdot \begin{bmatrix} R_x \\ R_y \\ R_z \end{bmatrix} + \begin{bmatrix} C_{10} \\ C_{20} \\ C_{30} \end{bmatrix} \tag{2}$$

The main goal of accelerometer calibration is to determine 12 parameters [5, 6] from C_{10} to C_{33}. This is typically done using the least squares method [1, 2].

The original 6-parameter (gain and offset in each channel) factory calibration can be recomputed to correct for thermal stresses introduced in the soldering process. Used 12 parameter linear calibration model can correct for accelerometer package rotation on the circuit board and for cross-axis interference between the accelerometer's x, y and z channels. Another recalibration, a 15 parameter model includes cubic nonlinearities in the accelerometer response can grow up accuracy. The orientation angles used for the recalibration must be carefully selected to provide the best calibration accuracy from the limited number of measurement orientations available [3].

It is possible rewrite (2) as

$$\begin{bmatrix} A_x & A_y & A_z \end{bmatrix} = \begin{bmatrix} R_x & R_y & R_z & 1 \end{bmatrix} \cdot \begin{bmatrix} C_{11} & C_{21} & C_{31} \\ C_{12} & C_{22} & C_{32} \\ C_{13} & C_{23} & C_{33} \\ C_{10} & C_{20} & C_{30} \end{bmatrix} \tag{3}$$

and then

$$Y = w \cdot X \tag{4}$$

Where:

Matrix X is calibration parameters.

Matrix w is raw data from the accelerometer collected at 6 stationary positions (when device lays on each side of the device and normalized gravity vector is zero in two axis). For better accuracy is possible add more stationary positions.

Matrix Y is the known normalized gravity vector.

Finally, calculation of 12 calibration parameters by least squares method can be written as

$$X = \left[w^T \cdot w\right]^{-1} \cdot w^T \cdot Y \tag{5}$$

3 Experimental Results

The results are measured on compact board ($0.4'' \times 0.9''$) which is carrier board for ST's LSM303D e-compass module. This board is connected to BeagleBone Black and uses I^2C bus for communication.

The LSM303D is a system-in-package featuring a 3D digital linear acceleration sensor and a 3D digital magnetic sensor. The LSM303D appears as a single unified I^2C device, and it also offers an SPI interface for additional flexibility. The LSM303D has many configurable options, including dynamically selectable sensitivities for the accelerometer and magnetometer, a choice of output data rates, and two independently programmable external inertial interrupt pins. A minimum of two logic connections are necessary to use the LSM303D in I^2C mode (this is the default mode): SCL and SDA. These pins are connected to built-in level-shifters that make them safe to use at voltages over 3.3 V; they should be connected to an I^2C bus operating at the same logic level as main power supply [8, 9].

The accelerometer sensor of LSM303D has 16-bit data output of 3 acceleration channels with linear acceleration full-scales of ±2 g/±4 g/±6 g/±8 g/±16 g selectable by the user [8, 9] (Fig. 2).

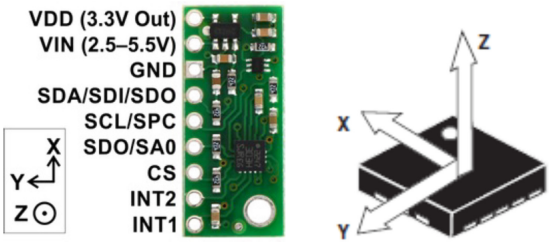


Fig. 2. LSM303D 3D compass and accelerometer carrier with voltage regulator and sensor axis orientation [8, 9]

3.1 I^2C Communication

With the CS pin in its default state (pulled up to VDD), the LSM303D can be configured and its readings can be queried through the I^2C bus. Level shifters on the I^2C clock (SCL) and data lines (SDA) enable I^2C communication with microcontrollers operating at the same voltage as VIN (2.5 – 5.5 V) [8]. In I^2C mode, the sensor's 7-bit slave address has its two least significant bits determined by the voltage on the SA0 pin. The carrier board

pulls SA0 to VDD through a 4.7 kΩ resistor, making the least significant bits 01 and setting the slave address to 0011101b by default. It is possible to use two LSM303D on the same bus by setting the least significant bits on the second sensor to 10 (the slave address is 0011110b). The I^2C interface on the LSM303D is compliant with the I^2C fast mode (400 kHz) standard [8, 9].

3.2 Sensor Characteristics

Linear acceleration sensitivity of the LSM303D is 0.061 mg/LSB for scale ±2 g. Linear acceleration typical zero-g level offset accuracy is specified in the sensor datasheet as ±60 mg for temperature 25 °C. Maximal delta from 25 °C of Linear acceleration zero-g level change vs. temperature is ±0.5 mg/ °C [9].

3.3 Experimental Sensor Calibration

The device is factory calibrated. The trimming values are stored inside the device by a non-volatile memory. Anytime the device is turned on, the trimming parameters are downloaded into the registers to be used during normal operation. This allows the user to use the device without further calibration. The content of the registers that are loaded at boot should not be changed. Their content is automatically restored when the device is powered up. But, own calibration is recommended for high-accuracy applications because it improve accuracy.

Table 1 shows the sign definition and optimal output of the sensor raw measurements at 6 stationary positions.

Table 1. Sign definition and expected sensor measurements

Stationary position	A_x	A_y	A_z
Z down	0	0	+1 g
Z up	0	0	−1 g
Y down	0	+1 g	0
Y up	0	−1 g	0
X down	+1 g	0	0
X up	−1 g	0	0

For better comparison of the measured output, two pieces of the LSM303D has been used; real measurements are shown in the Table 2. Both accelerometers were measured simultaneously but the calibration parameters were calculated for each accelerometer separately. The output of the accelerometers is acquired by a host computer at a rate of 200 Hz in the range ±2 g. Ambient temperature was 23 °C. The measured raw data was automatically adjusted by factory calibration. Nevertheless, in both accelerometers can be seen that output differs from the expected values listed in Table 1. Each presented value is calculated as an average value of 1000 measurements for noise avoidance.

Table 2. Real accelerometer sensor outputs (after factory calibration)

	Accelerometer 1			Accelerometer 2		
Stationary position	R_x	R_y	R_z	R_x	R_y	R_z
Z down	−0.0045	0.0457	0.9219	0.0024	0.0018	0.9188
Z up	0.0121	−0.0413	−1.0485	0.0153	0.0074	−1.0544
Y down	−0.0016	1.0038	−0.1132	0.0080	0.9743	−0.0684
Y up	0.0101	−0.9940	−0.0254	0.0102	−0.9764	−0.0788
X down	0.9614	−0.0138	−0.0671	0.9602	−0.0167	−0.0738
X up	−0.9890	0.0091	−0.0735	−0.9778	0.0179	−0.0761

Then we can apply the least squares method (5) for computation of 12 calibration parameters for the first accelerometer.

$$X_1 = \begin{bmatrix} C_{11} & C_{21} & C_{31} \\ C_{12} & C_{22} & C_{32} \\ C_{13} & C_{23} & C_{33} \\ C_{10} & C_{20} & C_{30} \end{bmatrix} = \begin{bmatrix} 1.0252 & 0.0118 & -0.0029 \\ 0.0063 & 0.9992 & 0.0445 \\ 0.0083 & -0.0440 & 1.0130 \\ 0.0025 & -0.0045 & 0.0684 \end{bmatrix}$$

And calculated calibration parameters for the second accelerometer.

$$X_2 = \begin{bmatrix} 1.0318 & 0.0183 & -0.0014 \\ 0.0011 & 1.0253 & -0.0054 \\ 0.0067 & -0.0030 & 1.0135 \\ -0.0027 & -0.0013 & 0.0731 \end{bmatrix}$$

The calculated matrix X is substituted into (2) and we subsequently receive the corrected values of the accelerometers that are more accurate (shown in the Table 3). Both used accelerometers corrected through calibration give very similar results and their outputs are comparable.

Table 3. Corrected accelerometer sensor outputs after user calibration

	Accelerometer 1			Accelerometer 2		
Stationary position	A_x	A_y	A_z	A_x	A_y	A_z
Z down	0.0058	0.0005	1.0044	0.0060	0.0034	1.0043
Z up	0.0060	0.0005	−0.9956	0.0061	0.0034	−0.9956
Y down	0.0063	1.0034	−0.0016	0.0062	0.9976	−0.0015
Y up	0.0064	−0.9965	−0.0016	0.0062	−1.0024	−0.0015
X down	0.9875	−0.0040	−0.0029	0.9875	−0.0010	−0.0029
X up	−1.0120	−0.0039	−0.0028	−1.0120	−0.0010	−0.0028

Results show that the MEMS output, after the calibration procedure, is far more accurate with respect to the output obtained using factory calibration data.

It has been assumed that the accelerometer will be used at the same temperature as the final calibration has been performed; so, temperature dependence can be ignored.

However, the recalibration can be very simply extended and future work will include temperature dependence by performing the recalibration at two or more different temperatures.

4 Conclusion

The Paper shows that a standard low cost accelerometer LSM303D is after user calibration far more accurate than this one with standard factory calibration. The goal of the static accelerometer calibration is to determine 12 parameters so that with any given raw measurements at arbitrary positions, the corrected values can be computed. Two pieces of the LSM303D has been compared and both give very similar results and their outputs are comparable.

References

1. STMicroelectronics, AN3192 Application note: Using LSM303DLH for a tilt compensated electronicc compass, p. 34 (2010). https://www.pololu.com/file/0J434/LSM303DLH-compass-app-note.pdf
2. Pedley, M.: AN3461 application note: tilt sensing using a three-axis accelerometer (2013). http://www.freescale.com/files/sensors/doc/app_note/AN3461.pdf
3. Pedley, M.: AN4399 application note: high precision calibration of a three-axis accelerometer (2013). http://www.freescale.com/files/sensors/doc/app_note/AN4399.pdf
4. Cacchione, F.: Mechanical characterisation and simulation of fracture processes in polysilicon MEMS, Ph.D. thesis. Politecnicio di Milano (2007). st.com/web/en/resource/technical/document/white_paper/phd_thesis.pdf
5. Stančin, S., Tomažič, S.: Time-and computation-efficient calibration of MEMS 3D accelerometers and gyroscopes. Sensors **14**(8), 14885–14915 (2014). http://www.mdpi.com/1424-8220/14/8/14885/pdf
6. Fang, B., Chou, W., Ding, L.: An optimal calibration method for a MEMS inertial measurement unit. Int. J. Adv. Robot. Syst. **11**(14), 57516 (2014). http://cdn.intechopen.com/pdfs-wm/46177.pdf
7. Ganssle, J.: A designer's guide to MEMS sensors (2012). http://www.digikey.com/en/articles/techzone/2012/jul/a-designers-guide-to-mems-sensors
8. Pololu, LSM303D 3D Compass and Accelerometer Carrier with Voltage Regulator. https://www.pololu.com/product/2127
9. STMicroelectronics, LSM303D - ultra compact high performance e-Compass 3D accelerometer and 3D magnetometer module (Datasheet), p. 54 (2012). https://www.pololu.com/file/0J703/LSM303D.pdf

A Survey of Optimization Techniques for Distributed Job Shop Scheduling Problems in Multi-factories

Imen Chaouch[1(✉)], Olfa Belkahla Driss[2], and Khaled Ghedira[3]

[1] COSMOS Laboratoy, Ecole Nationale des Sciences de l'Informatique, Université de la Manouba, Manouba, Tunisie
imen.chaouch@ensi.rnu.tn
[2] COSMOS Laboratoy, Ecole Supérieure de Commerce de Tunis, Université de la Manouba, Manouba, Tunisie
olfa.belkahla@isg.rnu.tn
[3] COSMOS Laboratoy, Institut Supérieur de Gestion de Tunis, Université de Tunis, Tunis, Tunisie
khaled.ghedira@isg.rnu.tn

Abstract. Distributed Job shop Scheduling Problem is one of the well-known hardest combinatorial optimization problems. In the last two decades, the problem has captured the interest of a number of researchers and therefore various methods have been employed to study this problem. The scope of this paper is to give an overview of pioneer studies conducted on solving Distributed Job shop Scheduling Problem using different techniques and aiming to reach a specified objective function. Resolution approaches used to solve the problem are reviewed and a classification of the employed techniques is given.

Keywords: Distributed scheduling · Job shop · Optimization method · Survey

1 Introduction

The manufacturing industry has undergone an important evolution these recent years due to the trend of globalisation. Owing to this evolution, there have been significant changes in the structure of production plants. Industrial companies are merging increasingly to distributed ones and thus, the structure of their shops changes from simple configurations to distributed ones. The study of production systems is complex given the large number of integrated entities and their interactions.

As scientists, engineers, and managers, we always have to take decisions. These decisions are becoming more complex with the evolution of the industrial environment and market requirement. An effective decision making process consists mainly of 4 steps summarized as follow:

R. Silhavy et al. (eds.), *Cybernetics and Mathematics Applications in Intelligent Systems*, Advances in Intelligent Systems and Computing 574, DOI 10.1007/978-3-319-57264-2_38

- Formulation of the problem: different variables of the problem are identified.
- Modelization of the problem: a mathematical model is built (optimization equations and constraints)
- Finding a solution: solving procedure is executed
- Implementation of the solution: the obtained solution is tested

Following the complexity of the problem, it may be solved by an exact method or an approximate method. Exact methods (Branch and X, Constraint Programming, Dynamic Programming, etc.) can be applied to small instances of difficult problems and guarantee an optimal solution. In complex problems, utilization of approximate methods is inevitable. They provide a good quality solutions in a reasonable time but there is no guarantee of achieving the optimal solution. Simulated Annealing, Tabu Search, Genetic Algorithm and Ant Colony Optimization are some examples of heuristics algorithms. We can find distributed scheduling in different studies dealing with various shop types such as distributed job shop [1,2], among many others, distributed flow shop [3–6] among many others and distributed parallel machines scheduling [7,8].

The scope of this paper is to provide a consolidated survey of various techniques that have been employed for problem resolution since its appearance in 2002. Numerous approaches have thus been investigated and these techniques are classified for ease of analysis. The rest of the paper is organized as follows. The DJSP is defined in Sect. 2. Section 3 review the related literature in the fields of DJSP. Section 4 discusses the studies of a shortcoming point of view and concludes the paper.

2 Problem Description

The Job shop Scheduling Problem (JSP) is one of the most important and complex problems in machine scheduling. A typical Job shop can be stated as a set of n jobs, which have to be processed on a set of m machines. Every job consists of a series of operations in a predetermined order on machines. There are various constraints on both jobs and machines. Each operation needs to be processed during an uninterrupted time of a fixed processing period and a given machine. A job can be processed by at most one machine at a time and a machine can process at most one job at a time. Furthermore, there are no precedence constraints among the operations of different jobs. In addition, it is assumed that a job does not visit the same machine twice. The aim is to determine the operation sequences on the machines in order to optimize a specified criteria such as makespan, maximum tardiness, total tardiness, etc. The most widely adopted criteria among literature is minimizing the makespan which can be set as the total time period it takes to complete all operations for all jobs.

To examine the performance of the most promising metaheuristics and compare their efficiency, benchmark problems are used. For the JSP, several researchers propose sets of benchmark problems including a number of instances with different dimensions and configurations. An instance consists of

Table 1. Benchmark problems of JSP

Problem	Processing time	Problem size		Proposed by
Abz	[50,100]	Abz 5	10×10	[9]
	[25,100]	Abz 6	10×10	
	[11,40]	Abz 7 to Abz 9	20×15	
Ft	[1,10]	Ft 06	6×6	[10]
		Ft 10	10×10	
		Ft 20	20×5	
Swv	[1,100]	Swv 1 to Swv 5	20×10	[11]
		Swv 6 to Swv 10	20×15	
		Swv 11 to Swv 15	50×10 (hard)	
		Swv 16 to Swv 20	50×10 (easy)	
Tai	[1,99]	Tai 01 to Tai 10	15×15	[12]
		Tai 11 to Tai 20	20×15	
		Tai 21 to Tai 30	30×15	
		Tai 31 to Tai 40	30×20	
		Tai 41 to Tai50	50×15	
		Tai 51 to Tai 60	50×20	
		Tai 61 to Tai 70	100×20	
La	[5,99]	La 01 to La 05	10×5	[13]
		La 06 to La 10	15×5	
		La 11 to La 15	20×5	
		La 16 to La 20	10×10	
		La 21 to La 25	15×10	
		La 26 to La 30	20×10	
		La 31 to La 35	30×10	
		La 36 to La 40	15×15	

the number of factories (f), the number of jobs (n), the number of machines (m), and the processing ($p_{j,i}$). The most popular benchmark problems are summarized in Table 1.

In a highly competitive economy, it is the quality of the service, that often makes the difference between one company to another. Thus, companies are on a continuous quest in order to respond to the requirements of a market in constant evolution and ensure full satisfaction to their customers. Job shops are constantly facing more challenges and an increased need to reduce on both their costs and time-to-market ([3]). In this trend, the phenomenon of decentralization appeared.

Distributed Job shop Scheduling Problem (DJSP) can be considered as an extension of the simple Job shop Scheduling Problem (JSP). It can be treated

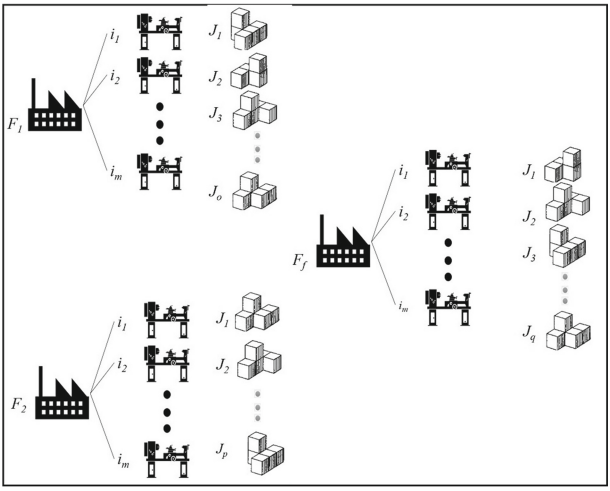


Fig. 1. An outline of a typical Distributed Scheduling problem

as a set of f factories, which are geographically distributed in different areas. Each factory contains m machines on which a certain number of jobs have to be processed, cf. Fig. 1. Distributed Scheduling problems in multi-factory production are much more complicated than classical scheduling problems [14] since two decisions have to be taken:

- allocation of jobs to suitable factories
- sequencing the operations on machines so that yield a feasible schedule aiming to minimize a predefined performance criterion.

Therefore, the DJSP is more difficult than the classical JSP due to the consideration of both factories selection and job scheduling. The JSP is strongly NP-hard [15] while the DJSP is much more complicated. Hence, it can be concluded that DJSP is also NP-hard. In distributed scheduling, makespan minimization becomes the minimization of maximum makespan among all factories.

3 Distributed Job Shop Scheduling: Literature Review

Scheduling problems have become a popular issue for researchers and industrialists in the last three decades, particularly the JSP since it is one of the most difficult tasks. [9,16–18] are pioneer researches in the literature that dealt with the JSP.

Due to the trend of globalization, the JSP has evolved from the classical JSP to the Distributed one and becomes increasingly, one of the most important issues to raise. In the literature, few researchers dealt with the DJSP and resolution methods employed are limited.

Genetic Algorithm (GA) is the first technique used to solve the DJSP. It is a computational method mainly proposed by Holland [19], based on the mechanics of biological genetic evolution. This method is commonly used to generate useful solutions to optimization, using technology inspired by the theory of evolution and biological processes, such as mutation, selection, and crossover operators.

The DJSP was studied for the first time by Jia et al. [20]. The authors proposed a web-based system to enable production scheduling with the utilization of the World Wide Web technology, in order to facilitate collaboration between geographically distributed plants. A Genetic Algorithm (GA) approach was adopted, involving a once gene crossover and twice gene mutation. To deal with the distributed scheduling problems, the genes in the GA must comprise the two dominant factors, i.e., the selected factory for every job and the operation processing sequence. The proposed system allows to the manufacturing system to choose to use the scheduling application locally, construct scheduling agents, or go to the web to make online scheduling, according to every factory's actual manufacturing situation and physical constraints.

In their next paper, to solve the DJSP, Jia et al. [21] presented a Modified Genetic Algorithm (MGA) in which two-step encoding method was used. The first one to encode the factory candidates and the second one, to affect jobs and operations. To test the performance of their MGA, the authors used benchmark instances (10 jobs/10 machines and 20 jobs/5 machines) proposed by [22]. In fact, MT1963 are instances for simple JS which are adapted to the DJS by distributing different jobs into the factories.

Later, in [2], authors refined their previous approach and proposed a GA integrated with Gantt Chart (GC) to derive the factory combination and schedule. The experimental results showed that the application of the GC is able to facilitate the chromosome evaluation procedures and thus improve the computational performance algorithm. In addition, the CPU time used for the GA–GC approach to solve the problem is 10 s. This computational time is 50% shorter than the time consumed by using the GA alone (15 s, without GC integration) developed by the authors in [21] to solve the same problem under a similar environment. In these three papers ([2,20,21]), authors proposed an encoding of chromosome, crossover mechanism and two mutation mechanisms for the DJS. The proposed approach was applied to the classical single-factory firstly, then to the DJSP. Numerical experiments were achieved and promising results were obtained.

Naderi and Azab [1] have mathematically formulated the DJSP by two different Mixed Integer Linear Programming models (MILP). The first model dealt with the problem as a sequencing decision while the second one dealt it as a positional one. Since the problem is NP-hard to solve, utilization of heuristics was inevitable. Hence, Three well-known heuristics were first adapted to the problem; these are Shortest processing Time first (SPT), Longest Processing Time first (LPT) and Longest Remaining Processing Time (LRPT). In addition, three Greedy Heuristics have been deployed (GH1, GH2 and GH3). The algorithms are greedy since at each step, different alternatives are generated and the best one is selected. In GH1, after assigning jobs to factories, the schedule

```
Procedure: Greedy heuristic 1

Assign jobs to facilities (job-facility assignment)
For k=1 to f do
    Determine random initial order of operations assigned to facility k
    For i=1 to q do   %q is the total number of operations assigned to facility k
        Take ith job number for the initial order
        For h=1 to i do
            If the resultant sequence is not redundant do
                Test inserting the job number into hth position
            Endif
        Endfor
        Select the best
    Endfor
Endfor
```

Fig. 2. The general outline of GH1 [1]

is built. Operations of each factory are iteratively inserted, one at a time, into a sequence to have at the final a complete permutation of operations. Decisions of assignment of job to factory and sequencing are sequentially taken (i.e., no interaction between the two decisions). Figure 2 shows the general outline of GH1. Unlike GH1, job-factory assignment and sequencing are interactively determined in GH2 and the jobs are initially sorted. Like GH2, GH3 interactively takes two decisions of job-factory assignment and sequencing. The main difference is the job factory selection rule. In GH2, each job is assigned to the factory with the lowest makespan. While in GH3, the job is assigned to the factory resulting in the lowest makespan after sequencing the operations of the job. That is, the assignments of jobs to all factories are tested and the best one is selected. The performance of the two proposed mathematical models and six heuristics (SPT, LPT, LRPT, GH1, GH2 and GH3) are evaluated and tested. Note that in their studies, Naderi and Azab used the PDI (Percentage Deviation Index) and the RPD to evaluate their models which can be calculated as shown in Table 3:

Two sets of instances were used. The first one, instances are generated by random processing times taken from a uniform distribution between 1 and 99. The second set, the instances of Taillard benchmark for job shops [12] are being used. With regards to the obtained optimal solutions, it is concluded that GH3 performs best with an average PDI of 1.21%. The second best heuristic is GH1 with an average PDI of 37.83%. Note that the PDI can be calculated as follows:

$$PDI = \frac{Alg - Min}{Max - Min} \times 100 \tag{1}$$

where

- Alg is the makespan obtained by any of the algorithm
- Min is the lowest makespan obtained for a given instance.
- Max is the highest makespan obtained for a given instance.

Finally, the redundancy mitigation mechanisms were applied and tested. In fact, the proposed algorithms suffer from a serious shortcoming, which is

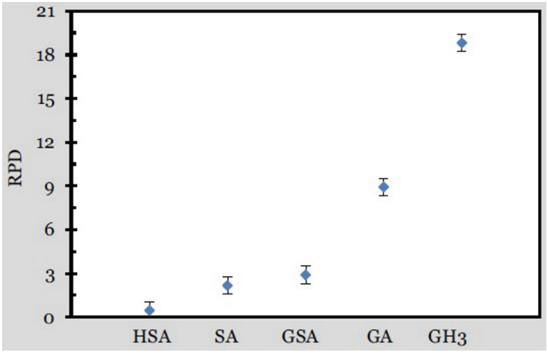


Fig. 3. The average RPD of the tested algorithms [23]

redundancy. That is, different permutations might represent the same schedule. For that reason, authors developed the redundancy mitigation mechanism which allow recognization and discard of redundant permutations. The results showed that more than 80% reduction in computational time with simple application of redundancy exclusion theorems.

Recently, Naderi and Azab [23] have differently treated th e same problem. First, authors have mathematically formulated the DJSP by a Mixed Integer Linear Programming models (MILP). And then, three different versions of simulated annealing have been designed and implemented. The first one, called SA, is implemented without any local search. The second one, called hybridized simulated annealing (HSA), is hybridized with local search type 1 which assume that the job number is first inserted into m random positions. Then, the position of one randomly selected job number is shifted to a random position. Finally, the third version, called greedy simulated annealing (GSA), employs the greedy local search 2. Unlike the local search type 1, the job number is added into permutation one by one. To insert the job number, it is assigned to all the possible positions and the best position is chosen. The performance of the proposed mathematical model and algorithms are evaluated and tested. To do the experiments, three sets of instances were generated. The first set is for parameter tuning, the second is for the experiment with small instances, and finally, the last one is for the experiment with larger instances. For the first set, there are 20 different combinations generated by random processing times taken from a uniform distribution between 1 and 99. For the second set, benchmark instances generated by [1] were used including 24 small instances. And finally, for the third set, 80 instances of [12] were tested with different levels of f (f= 2, 3, 4, 5), summing up 320 large instances. The solutions proposed in this article obtained a promising results and outperformed the other tested algorithms. Figure 3 shows the results of different experiments. The performance measure used in this research is Relative Percentage Deviation (RPD). It can be calculated as follows:

$$RPD = \frac{Alg - Min}{Min} \times 100 \tag{2}$$

where

– Alg is the makespan obtained by any of the algorithm
– Min is the lowest makespan obtained for a given instance.

4 Analysis of the Literature and Critics

The first systematic approach to scheduling problems was undertaken in the mid-1950s. Since then, thousands of papers on different scheduling problems have appeared in the literature [24]. Distributed Job shop Scheduling Problem is a fertile field for future research and literature still lacks works on this area. This can be clearly seen from this literature review, only 5 papers studying the DJS since 2002. As we can see from Table 2, almost all articles considered the minimizing of the makespan as objective function, except for [21] which have considered Multi-objective scheduling (Makespan, total tardiness). We have seen that several techniques have been employed to solve the DJSP, Approximate methods was dominant on solving the DJSP. This can be explained by the fact that the DJSP is NP-Hard to solve which makes inevitable the use of approximate methods. In the other hand, DJSP is a real world problem and in this case, approximate methods are more applicable than exact ones. All the studied papers consider the case of homogeneous network factories, which is an abstraction of a real world-manufacturing problem because it is rare and difficult to have identical factories or cells in the real world. Table 2 summarizes all the studies we have reviewed in this paper on solving DJSP and provides a clear classification of them in term of year of publication, objective function, employed techniques for resolution and factory type.

Table 2. Classification of Distributed Job shop Scheduling Problem papers

Papers	Methods	Objective function	Type of factory
[20]	Genetic Algorithm	Makespan + Total tardiness	Homogeneous
[21]	Genetic Algorithm	Makespan	Homogeneous
[2]	Genetic Algorithm	Makespan	Homogeneous
[1]	MILP + Heurictic Algorithms	Makespan	Homogeneous
[23]	MILP + Simulated Annealing	Makespan	Homogeneous

5 Conclusion and Research Opportunities

Globalization and market unpredictability have led many companies to enlarge their production and distribution network. Nowadays, it is essential to study the impact of this strategy on the production system and especially on scheduling which represents the first function to be influenced by such organization.

The importance of a distributed system has been strongly proved by researchers and practitioners in recent years. With factories situated in different geographical areas, factories can benefit from various advantages, such as their proximity to their suppliers and customers, a rapid and efficient adaptation to market changes and low production and distribution costs.

In this paper, we have reviewed and analyzed studies focusing on Distributed Job shop Scheduling Problem. The thematic area of distributed scheduling is very promising for research; however, we notice that there is a gap in the literature dealing with this issue. Studies are still focusing on single-manufacturing workshop, namely one factory or facility. In the field of DJSP, the problem is not yet deeply studied. For example, the maximum number of factories reached is 5 factories, which can be much more. Another thing, the authors considered that factories are identical, or in reality this is not always valid in the majority of cases. Several methods can be used for solving the problem which still not explored. Indeed, apart from the models developed in [1,23] only the genetic algorithms and Simulated annealing were employed. In addition, most of the authors considered the makespan as principal criteria to be minimized and used mainly approximate algorithms as a method of resolution.

According to the analysis of literature, future researches should be more focused on distributed scheduling and various methods of resolution should be proposed. Hybridization of algorithms has not been exploited in the Distributed Job shop Scheduling Problem, it can be a new area of research and also, it's interesting to explore larger instances in terms of number of factories. Another aspect would be interesting to study, is the case where factories are not identical since it is rarely the case in the reality. In fact, machines could have different technologies from one factory to another. And finally, it will be interesting to extend the problem and study the impact of the transportation time from factory to customers since it's a distributed problem.

References

1. Naderi, B., Azab, A.: Modeling and heuristics for scheduling of distributed job shops. Expert Syst. Appl. **41**(17), 7754–7763 (2014)
2. Jia, H.Z., Fuh, J.Y.H., Nee, A.Y.C., Zhang, Y.F.: Integration of genetic algorithm and gantt chart for job shop scheduling in distributed manufacturing systems. Comput. Ind. Eng. **53**(2), 313–320 (2007)
3. Naderi, B., Ruiz, R.: The distributed permutation flowshop scheduling problem. Comput. Oper. Res. **37**(4), 754–768 (2010)
4. Naderi, B., Ruiz, R.: A scatter search algorithm for the distributed permutation flowshop scheduling problem. Eur. J. Oper. Res. **239**(2), 323–334 (2014)
5. Rifai, A.P., Nguyen, H.-T., Dawal, S.Z.M.: Multi-objective adaptive large neighborhood search for distributed reentrant permutation flow shop scheduling. Appl. Soft Comput. **40**, 42–57 (2016)
6. Bargaoui, H., Driss, O.B., Ghédira, K.: Minimizing makespan in multi-factory flow shop problem using a chemical reaction metaheuristic. In: IEEE Congress on Evolutionary Computation, Vancouver, Canada, pp. 2919–2929 (2016)

7. Behnamian, J., Ghomi, S.M.T.F.: The heterogeneous multi-factory production network scheduling with adaptive communication policy and parallel machine. Inf. Sci. **219**, 181–196 (2013)
8. Hatami, S., Ruiz, R., Andrés-Romano, C.: Heuristics and metaheuristics for the distributed assembly permutation flowshop scheduling problem with sequence dependent setup times. Int. J. Prod. Econ. **169**, 76–88 (2015)
9. Adams, J., Balas, E., Zawack, D.: The shifting bottleneck procedure for job shop scheduling. Manage. Sci. **34**(3), 391–401 (1988)
10. Fisher, H., Thompson, G.L.: Probabilistic learning combinations of local job-shop scheduling rules. Ind. Sched. **3**, 225–251 (1963)
11. Storer, R.H., Wu, S.D., Park, I.: Genetic algorithms in problem space for sequencing problems. In: Fandel, G., Gulledge, T., Jones, A. (eds.) Operations Research in Production Planning and Control, pp. 584–597. Springer, Heidelberg (1993)
12. Taillard, E.: Benchmarks for basic scheduling problems. Eur. J. Oper. Res. **64**(2), 278–285 (1993)
13. Lawrence, S.: Resource constrained project scheduling: an experimental investigation of heuristic scheduling techniques (supplement). Graduate School of Industrial Administration (1984)
14. Chung, S.H., Chan, F.T.S., Chan, H.K.: A modified genetic algorithm approach for scheduling of perfect maintenance in distributed production scheduling. Eng. Appl. Artif. Intell. **22**(7), 1005–1014 (2009)
15. Garey, M.R., Johnson, D.S., Sethi, R.: The complexity of flowshop and jobshop scheduling. Math. Oper. Res. **1**(2), 117–129 (1976)
16. Colorni, A., Dorigo, M., Maniezzo, V., Trubian, M.: Ant system for job-shop scheduling. Belg. J. Oper. Res. Stat. Comput. Sci. **34**(1), 39–53 (1994)
17. Dell'Amico, M., Trubian, M.: Applying tabu search to the job-shop scheduling problem. Ann. Oper. Res. **41**(3), 231–252 (1993)
18. Davis, L.: Job shop scheduling with genetic algorithms. In: Proceedings of an International Conference on Genetic Algorithms and their Applications, vol. 140. Carnegie-Mellon University, Pittsburgh, PA (1985)
19. Holland, J.H.: Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence. MIT Press, Cambridge (1992)
20. Jia, H.Z., Fuh, J.Y.H., Nee, A.Y.C., Zhang, Y.F.: Web-based multi-functional scheduling system for a distributed manufacturing environment. Concurrent Eng. **10**(1), 27–39 (2002)
21. Jia, H.Z., Nee, A.Y.C., Fuh, J.Y.H., Zhang, Y.F.: A modified genetic algorithm for distributed scheduling problems. J. Intell. Manuf. **14**(3–4), 351–362 (2003)
22. Muth, J.F., Thompson, G.L.: Industrial Scheduling. Prentice-Hall, Upper Saddle River (1963)
23. Naderi, B., Azab, A.: An improved model and novel simulated annealing for distributed job shop problems. Int. J. Adv. Manuf. Technol. **81**, 693–703 (2015)
24. Allahverdi, A., Ng, C.T., Cheng, T.C.E., Kovalyov, M.Y.: A survey of scheduling problems with setup times or costs. Eur. J. Oper. Res. **187**(3), 985–1032 (2008)

Big Data Process Advancement

Roman Jasek(✉), Said Krayem, and Petr Zacek

Faculty of Applied Informatics, Tomas Bata University in Zlin, Zlin, Czech Republic
{jasek,zacek}@fai.utb.cz, drsaid@seznam.cz

Abstract. Information in this era is thriving to be maintained on a verity of sources. Data is available in different patterns and forms. Combining and processing all different types of datasets in a heterogeneity database is near to impossible, specifically, if the information is moving and changing on many different sources on a continuous basis. Information is represented in different modules and nowadays processing data from various sources can lead to critical risk assessment results. Big Data is a concept introduced to cover the use of different techniques serving the desired goals by processing the given information. Processing huge amount of data is a big challenge for a single machine to perform, in this paper we will discuss this idea and demonstrate a module of clustered machines to work as a single entity towards achieving the desired tasks while working on parallel cohesively.

The idea of a solution to combine different machines of different specification processing and power in a single cluster and then distributing input data of various data fairly to most powerful processing and well-designed data type machine in the cluster.

Distribution of input data and storing mechanism will depend on machine specification, data processing, the power of a machine, balance loading and data type.

We present our suggestion solving method by using Event-B based approach, the Key features of Event-B are the use of set theory as a modelling notation and we propose using the Rodin modelling tool for Event-B that integrates modelling and proving.

Keywords: Big data · Clustering · Parallel clustering · Distribution process · Distribution file system · Formal modelling · Event-B · Rodin

1 Introduction

Nearly everything in nature reached to a peak point of revolution, data is evolving around us and expanding rapidly over a short-period of time.

According to the Internet live stats, around 46%+ of the world's population have access to internet, and this state is increasing rapidly every second. Moreover, each individual user performs many different actions to acquire services, service providers and business owners are competing to move most if not all of their services to be available online (Table 1).

R. Silhavy et al. (eds.), *Cybernetics and Mathematics Applications in Intelligent Systems*,
Advances in Intelligent Systems and Computing 574, DOI 10.1007/978-3-319-57264-2_39

Table 1. Statistics for internet users (source [1])

Year	Internet users	Penetration (% of pop)	World population	Non-users (Internetless)	1Y user change	1Y user change	World pop. change
2016	3,424,971,237	46.10%	7,432,663,275	4,007,692,038	7.50%	238,975,082	1.13%
2015	3,185,996,155	43.40%	7,349,472,099	4,163,475,944	7.80%	229,610,586	1.15%
2014	2,956,385,569	40.70%	7,265,785,946	4,309,400,377	8.40%	227,957,462	1.17%
2013	2,728,428,107	38%	7,181,715,139	4,453,287,032	9.40%	233,691,859	1.19%
2012	2,494,736,248	35.10%	7,097,500,453	4,602,764,205	11.80%	262,778,889	1.20%

The above table shows a rapid increase in the number of users online, more than 11% of the human population have started to use the internet in the past 5 years. Moreover, each individual is engaging to either browse, operate or store data online, using a verity of different data hubs and services. Furthermore, Fig. 1 shows stats demonstrate a growth in the number of users using their smart devices or engaging in social media online.

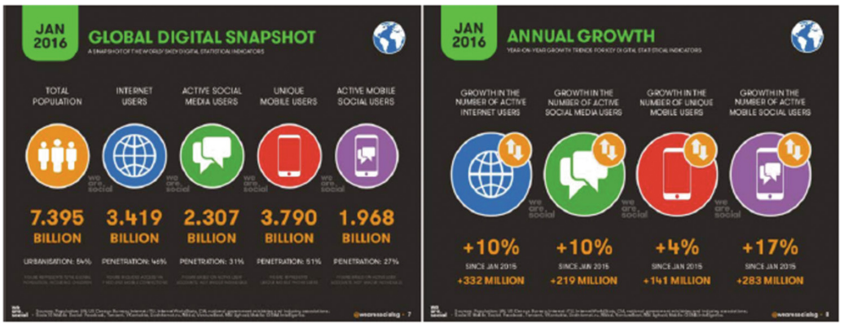


Fig. 1. Statistics on the growth of the number of users who use smart devices or engaging in social media online (source [2])

The data volume of global consumer the web usage, e-mails and data traffic from 2015 to 2020 (in petabytes per month):

Figure 2 shows a forecast for the internet traffic through the web, email, instant messaging, and other data traffic use, excluding file-sharing, from 2015 to 2020. In 2017, the global consumer internet traffic in this segment is projected to reach 11,061 petabytes per month.

Not only that data is being generated rapidly, its complexity and cohesion are magnified and exploited to insure security and accountability. In data science, it's a common practice to have a record/registration for every last information. It's like a gold mine, the more you dig deeper the more pure and valuable gold you get. So, breaking down your data into more levels can help you gain more accurate & reliable information in time.

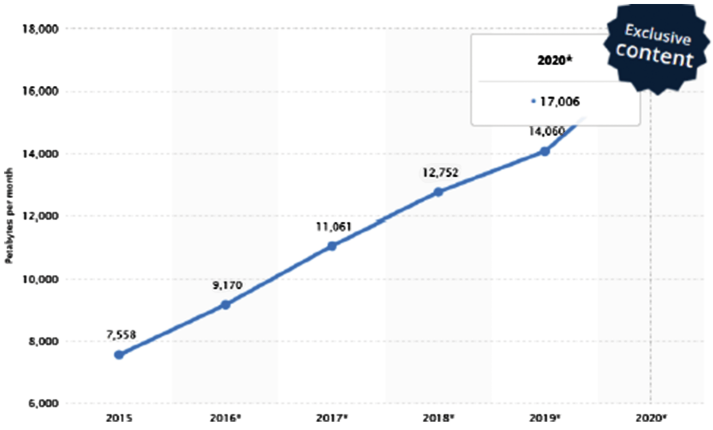


Fig. 2. Forecast for internet traffic through web from 2015 to 2020 (source [3])

2 Big Data

The term big data [1, 2] describes in itself the ability to work with massive data. Targeting much more complex, varied non-structured or structured sources and data types. Resulting from remarkable analysis. Visualizing and storing abilities within acceptable timeframes. Moreover, having such a complex, expandable module should help explore the data from different perspectives, applying different scenarios to reveal hidden patterns or correlations. Hence, it can't be stretched enough the importance of big data analysis.

In theory, the term Big Data used to describe any type of digital data that can be viewed, processed or stored, in any amount and form. Resulting from accurate information within a limited timeframe.

2.1 Characteristics of Big Data

Although as mentioned above any type of digital data can be processed, research shows that the term Big Data is characterized into 3 main categories: Volume, Variety and Velocity [4, 5].

Volume: With the implementation and expansion of technological equipment's, the amount of data is increasing rapidly and might reach zettabytes of row data format.

Variety: Any type of data can be processed, from legacy data format, structured, non-structured, XML, JSON, images, videos or even archived files.

Velocity: The speed of data obtained from many different data sources. It can be a simple data source, streaming data source or even a one-time batched data source.

Big Data also can be characterized over many other factors, such as: Variability, Veracity and Validity.

Variability: When the data is in motion, the inconsistency of information can be a factor to build an analytical metric.

Veracity: The quality of the given data can rapidly affect the accuracy of the outputted information.

Validity: The authenticity of the data can indicate analytical metrics to the security and verification of the data source.

Any other number of characteristics also can be attributed to defining the information being processed in Big Data.

Figure 3 drawing shows some important information regarding big data characteristics:

Fig. 3. Big data characteristics (source [11])

2.2 Importance and Benefits of Big Data

With the growth of data comes the responsibility of analyzing it accurately and within a certain time concluding reliable information. Although Big Data as a term may give an impression of processing a big amount of data, what we do with this data to reach to metrics, statistical information is the key benefit.

Below is a list of examples illustrating some key benefits of Big Data [4, 8]:

- Accurate Cause of issues: Gathering & analyzing real-time log & error files out of running production systems can lead to an accurate case of the issue.
- Risk Alerts: Accepting trending data from multiple health service centres, can help on predicting epidemics, and notifying the CDC if necessary.
- Test different scenarios on life data: With the advantages that big data provides, it allows you to reshape the method of accessing your data in order to perceive different perspectives of information. This can also lead to a powerful tool for testing the implementation of different scenarios into your data.
- Metrics that can lead to real-time change: Allowing you to perform real time analyzing your data, it can help you predict bottleneck and risks. Moreover, you can develop a solution that can be implemented on real life business to avoid the predicted risks.

2.3 Challenges and Problems with Big Data

Everything has limits and constraints in which it can work or function as expected. Due to the nature of how Big Data methodology works there are few main limitations and key points it relays on in order to success, listed below some points [4, 10]:

1. Reliable sources: big data can only provide accurate information if the data it relies on are correct and not fake.
2. Stability of streaming data: if your system relies on other moving sources of data, then it can be effected if you stopped receiving any of the data for any reason.
3. Size: with time, as the amount of data grows big data needs to implement a smart methodology to be able to handle the increasing amount of data. The amount of time requires to process a specific analysis can affect the performance of your system. Any system that implements a solution for working with big data needs to account for that and more. Furthermore, the system itself has upgrades, changes, fixes and configuration that need to be maintained and deployed according to the optimal methodology.
4. Speed: Processing some information require more time than others, and the value of processing a traffic of changing data to gain the desired information in time is high.

3 Excepted Problems

This project is aimed to demonstrate an overview of the importance on data growth. Processing and maintaining data has become a challenge, though it might appear maintainable for the time being. However, calculating the necessary equipment and resources to process zettabytes of information is not always accurate. Processing data is not easily addressed, calculating the number of rows in a data file contains about 15 million records require less or no processing from the CPU [6]. On the other hand, processing a few million records of data using complex algorithms to extract some specific statistics can take a great deal of the CPU processing time and power.

From the above, it's very clear that a single big machine with an advance and expensive hardware can never be enough or affordable to operate this amount of rapidly

increasing data. And the need for a mechanism that moderates the workload and efficiently handles the intended operations.

4 Recommended Solution

Looking around us, nature has some amazing complex systems. Systems such as bee hives, a set of bees working together to achieve a higher vision. Each member only performing a set of simple tasks, combined with other operated tasks resulting on much higher goal.

The same idea can be applied to computer systems, a set of machines working together to perform the same operation, with different data targets. However, this technology requires a higher coordination technique and intelligence business solution to handle the amount of processing of the tasks between the intended machines to achieve the main goal of the operation.

4.1 Requirement to Implement This Technology

1. In such a cluster of machines, there must always be one leader machine at a time, we refer to as "Master".
2. One or more members of the cluster, we refer to as "Nodes", to perform the operations assigned to it via the master. A master can act as both master and node at the same time if necessary.
3. A secure, reliable and continuous communication channel between the master and all the nodes.
4. A single method of communication between all the nodes, with the same language or form of encryption methodology.

4.2 Requirement to Implement This Technology

To implement such a technology, there must be a baseline of expectations that should be met, and these are the main characteristic:

1. The Master node should be the most powerful, reliable and secure machine in the cluster, its ability to perform administrative and comparative operation continuously to detect and prevent any failure.
2. This cluster must be expandable, variable allowing both growth and shrink on the number of nodes as needed.
3. The cluster should support working with different operating systems.
4. The cluster should support working with different processing speed, data storage verities and memory limitations.

5 Event-B and Rodin

5.1 Event-B

Event-B is a formal method for system-level modelling and analysis, especially complex systems, Key features of Event-B are the use of set theory as a modelling notation, the use of refinement to represent systems at different abstraction levels and the use of mathematical proof to verify consistency between refinement levels [7].

In Event-B we have two kinds of components.

- A context describes the static elements of a model. It has the following components:
  - Sets: User-defined types can be declared in the SETS section.
  - Constants: We can declare constants here. The type of each constant must be declared in the axiom section.
  - Axioms: The axiom section contains a list of predicates (called axioms). These axioms define rules that will always be the case for given elements of the context.
  - Theorems: Axioms can be marked as theorems. If this is the case, we are declaring that the predicate can be proved by using the axioms that have been written before this theorem.
  - Extends: A context may extend an arbitrary number of other contexts. When we extend another context A, we can then use all constants and axioms declared in A and also add new constants and axioms [13].
- A machine describes the dynamic behaviour of a model. It has the following components:
  - Refines: A machine has the option of refining another one.
  - Sees: We can use the context's sets, constants and axioms in a machine by declaring it in the Sees section.
  - Variables: The variables' values are determined by an initialisation event and can be changed by events. The type of each variable must be declared in the invariant section.
  - Invariants: These are predicates that should be true for every reachable state.
  - Events: An event can assign new values to variables. The guards of an event specify the conditions under which it can be executed. The initialisation of the machine is a special case of an event [13].

We can see the relationship between Machine and Context by Fig. 4.

Besides its state, a machine contains a number of Events which show the way it may evolve. Each event is composed of a guard and an action. The guard is the necessary condition under which the event may occur. The action, as its name indicates, determines the way in which the state variables are going to evolve when the event occurs.

An event may be executed only when its guard holds. Events are atomic and when the guards of several events hold simultaneously, then at most one of them may be executed at any one moment. The choice of event to be executed is non-deterministic. Practically speaking, an event, named evt, is presented in one of the three following simple forms (Fig. 5):

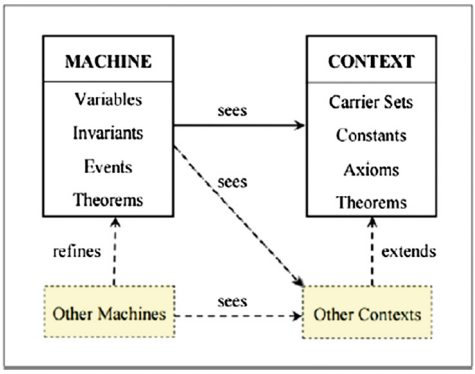


Fig. 4. Machine and context relationship (source [14])

evt = begin *S(v)* end
evt = when *P(v)* then *S(v)* end
evt = any *t* where *P(t, v)* then *S(t, v)* end

Fig. 5. Three possible forms of an event (source [15])

where P(…) is a predicate denoting the guard, t denotes some variables that are local to the event, and S(…) denotes the action that updates some variables. The variables of the machine containing the event are denoted by [15].

5.2 Rodin

The Rodin Platform is an open and extensible tool for Event-B specification and verification. It contains a database of modelling elements used for constructing system models such as variables, invariants and events. It is accompanied by various useful plug-ins such as a proof-obligation generator, provers, model-checkers, UML transformers, etc. [14]

6 Formal Modelling of Cluster in Event-B

To formalize Cluster model, we have used the Event-B modelling language because it supports the refinement approach that helps to verify the correctness of the system in an incremental way.

The Cluster consists the main machine, we refer to as “Master” and One or more of servers with different processing speed, data storage and memory, we refer them to as “Nodes”.

Initially, we define the Cluster on Master. Then, the Nodes can send a request to join the Cluster. To accept, Master check a set of things. In the case of acceptance, it registers information of processor, storage, memory, IP address and power of processing of

Nodes. Then the Node starts receiving files from Master those sending to the cluster. When Client sends request to get a set of information or statistics, the Master pass request to all Nodes in Cluster. The Node process request and send the result to Master which in turn collecting results from all Node and return to those Client who requests.

In some cases, there is a need to delete a set of files or would like to remove Nodes from Cluster.

Figure 6 drawing shows some components of Cluster.

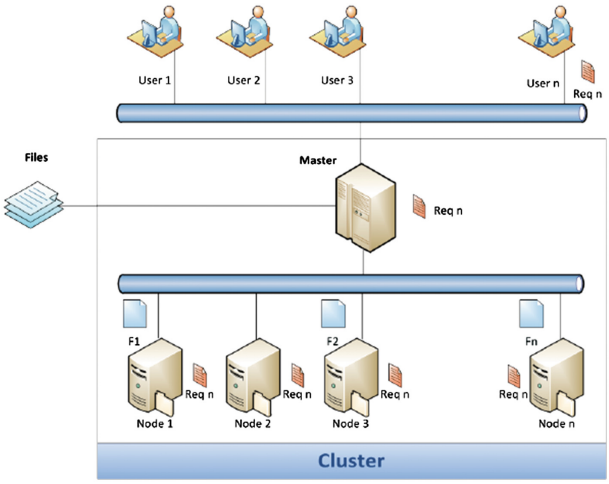


Fig. 6. Components of cluster (own work)

Abstract model: This is an initial model, which initialize Cluster and add Nodes to it, in addition to that add set of files to the Cluster and delete files or remove Nodes.

Refinement 1: This is an advance model, which distributed Cluster files fairly to nodes.

6.1 The Context and Abstract Model

Initially, we define Cluster into Master, then start adding Nodes to Cluster through Master as well as adding files to Cluster. In "Cluster_Ctx0" CONTEXT, we defined four SETS:

- CLUSTER: is represent a set of servers that consists of Master and all Nodes connected to it.
- IP_ADDRESSES: are represent a set of all IP addresses of Cluster members.
- NODES: are represent a set of servers with the role of storing and processing data into Cluster.
- FILES: are represent all files inside the Cluster.

And one CONSTANTS: MASTER, this is the primary server in the Cluster and there is only one master in Cluster, the role of Master is controlling and communication between Nodes and Clients.

It also has been defined three AXIOM which represents postulates to the definition of Cluster.

See Fig. 7:

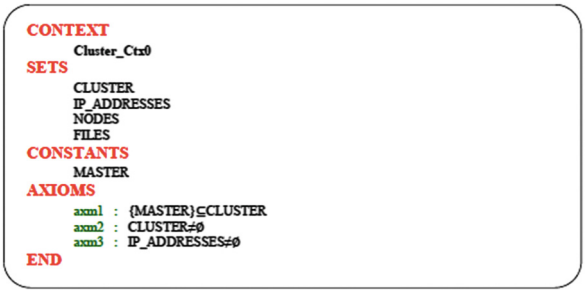


Fig. 7. "Cluster_Ctx0" CONTEXT (own work)

In "Cluster_Mac0" MACHINE, we defined seven VARIABLES which defined by INVARIANTS:

- cluster: is denote to a number of servers which join to Cluster. Initially, the Master belongs to cluster.
- nodes: is denoted by a number of servers which join to Cluster, that is a subset of NODES.
- ip_addresses: is denoted to IP addresses of servers which join to Cluster, that is a subset of IP_ADDRESSES.
- cluster_ips: represent a relationship between nodes and ip_addresses.
- files: is denoted to number of files which added to Cluster, that is a subset of FILES.
- cluster_files: represent a relationship between files and cluster_files.
- cluster_filesizes: represent a relationship between nodes and Natural number N.

See Fig. 8(a) and (b):

The event RequestIn specifies a set of required conditions for adding Nodes to Cluster. The first guard states that node (n) should belong to a set of servers that maintain the Cluster (NODES). The second guard states that node (n) should not be a member of nodes that belongs to the Cluster (nodes). The third guard states that IP address of the node (ip) should belong to a set of IP addresses that maintain the Cluster (IP_ADDRESSES). The final guard states that IP address of the node (ip) should not be a member of IP addresses that belongs to the Cluster (ip_addresses). If all guards are satisfied then it will add node (n) and IP address (ip) of node to a set of servers (nodes) and IP addresses (ip_addresses) of nodes which defined currently in Cluster, then define relationship between node which added and own IP address (cluster_ip(n)). Finally, we specify the file size for the node (cluster_filesizes(n)) which added is zero.

See Fig. 9:

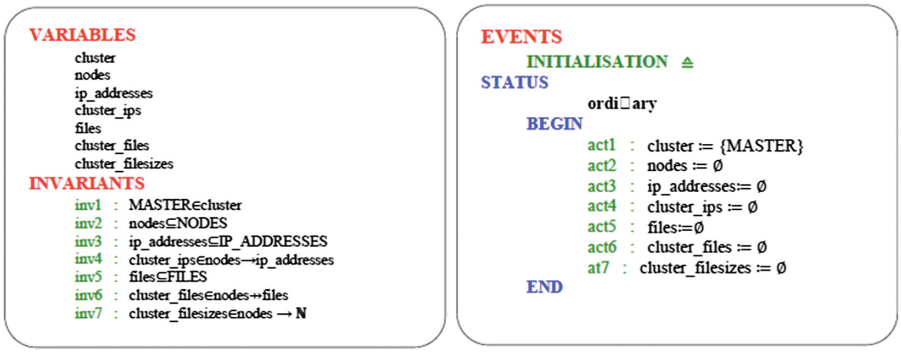


Fig. 8. **(a)** Variables and invariants of "Cluster_Mac0" (own work), **(b)** Initialization of "Cluster_Mac0" (own work)

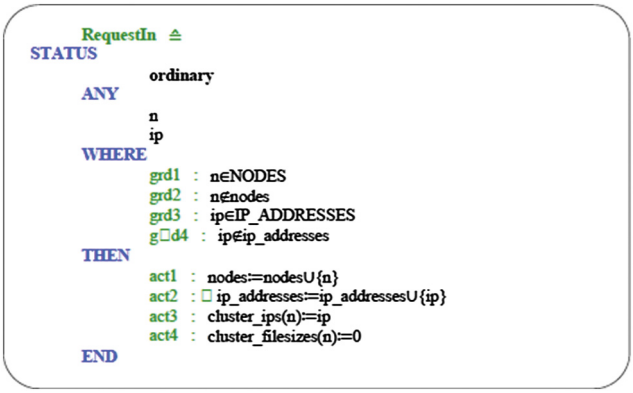


Fig. 9. A specification of Events RequestIn in "Cluster_Mac0" (own work)

The event RequestOut specifies a set of required conditions for removing Nodes from Cluster. The first guard states that nodes (n) should belong to a set of servers that added before to Cluster (nodes). The final guard states that IP address of node (ip) should belong to a set of IP addresses of nodes which added to Cluster (ip_addresses). If all guards are satisfied then it will remove node (n) and IP address (ip) of the node from a set of servers (nodes) and remove IP addresses (ip_addresses) of nodes which defined currently in Cluster, then remove relationship between removed node and IP address (cluster_ip(n)), files and file size associated with it (cluster_filesizes(n)).

See Fig. 10:

The event addFile specifies a set of required conditions for adding files to Cluster. The first guard states that nodes (n) which we want to add a file to it should belong to a set of servers that added before to Cluster (nodes). The second guard states that file should belong to a set of files which will be added to Cluster (FILES). The third guard states that file (f) should not be a member of Cluster files that added to the Cluster (files). The final guard states that size of the file (fz) should be known and belongs to Natural

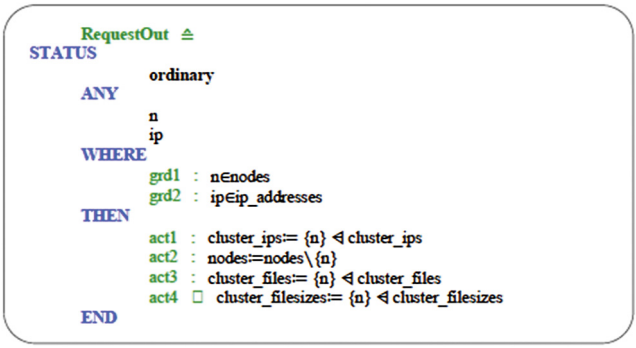


Fig. 10. A specification of Events RequestOut in "Cluster_Mac0" (own work)

number N. If all guards are satisfied then it will be defined relationship between node (n) and file (f) which added to it (cluster_files(n)), and sum the file size (fz) to total files sizes those stored in this node (cluster_filesizes(n)), then add file (f) to a set of files which defined currently in Cluster (files).

See Fig. 11:

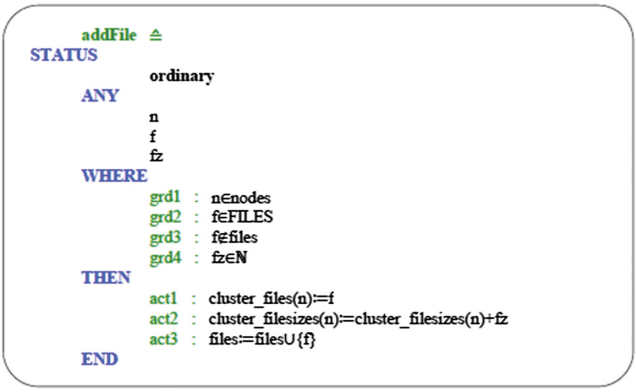


Fig. 11. A specification of Events addFile in "Cluster_Mac0" (own work)

The event removeFile specifies a set of required conditions for a remove files from Cluster. The first guard states that nodes (n) which we want to remove the file from it should belong to a set of servers that added before to Cluster (nodes). The second guard states that file should be a member of Cluster files that added to the Cluster (files). The third guard states that size of the file (fz) should be known and belongs to Natural number N. The final guard states that the relationship between the nodes and files (cluster_files) after removing node (n) belongs to the relationship between nodes and files except the file which to be removed. If all guards are satisfied then it will remove relationship between node (n) and file (f) which add to it (cluster_files(n)) and subtract file size (fz)

to total files sizes those stored in this node (cluster_filesizes(n)), then remove file (f) from a set of files which defined currently in Cluster (files).

See Fig. 12:

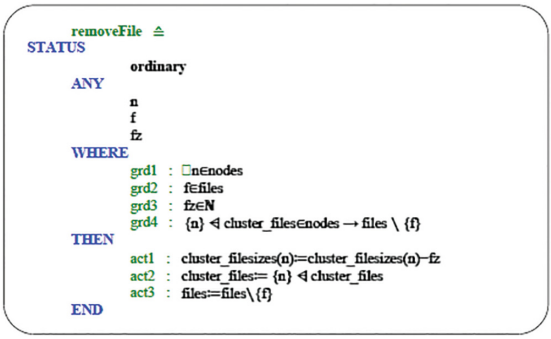


Fig. 12. A specification of Events removeFile in "Cluster_Mac0" (own work)

6.2 Refinement 1

We work to develop a solution to distribute files fairly on nodes maintained by Cluster, the distribution based on the power of nodes to achieve balanced performance according to power. Consequently, when the request received by the master, it sends to all nodes, then every node process to it in approximate same time and sends the result to master accordingly it will improve the performance of the Cluster.

In "Cluster_Mac1" MACHINE, we defined three more VARIABLES which defined by INVARIANTS also:

- nodes_powers: represent a relationship between nodes and Integer number Z. The power of node is calculated based on specification of CPUs, number of core, size of cache, size of memory etc., The complexity of the calculation process of power based on these criteria we can achieve it by easy way by sending a group of files to nodes when it adding to Cluster and request a set of process on it, whichever node take less processing time the power of node is high.
- nodes_priorities: represent a relationship between nodes and Integer number Z. The priority is an indicator to send files to nodes in the Cluster, as power node increased the priority increased and As files size on node increased then the priority decreased.
- best_priority: is denoted to Integer number Z. this is represent high priority node on Cluster. Consequently, it will be sent incoming files to the node associated with it.

See Fig. 13:

The event RequestIn we introduce one extra guard. The extra guard states that power of node (pw) should be known and belongs to Integer number Z. If all guards are satisfied then we specify the power of added node (nodes_power(n)) is equal to (pw) and priority of added node (nodes_priorities(n)) is equal to10*(pw).

See Fig. 14:

```
inv8  : nodes_powers∈nodes → ℤ
inv9  : nodes_priorities∈nodes → ℤ
inv10 : best_priority∈ℤ
```

```
act8  : nodes_powers ≔ ∅
act9  : nodes_priorities:=∅
act10 : best_priority:=0
```

Fig. 13. Some invariants and Initialisation of "Cluster_Mac1" "Cluster_Mac1" (own work)

```
RequestIn ≙
  extended
STATUS
        ordinary
REFINES
        RequestIn
ANY
        n
        ip
        pw
WHERE
        grd1 : n∈NODES
        grd2 : n∉nodes
        grd3 : ip∈IP_ADDRESSES
        grd4 : ip∉ip_addresses
        grd5 : pw∈ℤ
THEN
        act1 : nodes:=nodes∪{n}
        act2 : ip_addresses:=ip_addresses∪{ip}
        act3 : cluster_ips(n):=ip
        act4 : cluster_filesizes(n):=0
        act5 : nodes_powers(n):=pw
        act6 : nodes_priorities(n):=10*pw
END
```

Fig. 14. A specification of Events RequestIn in "Cluster_Mac1" (own work)

The event RequestOut we did not introduce any extra guard. But If all guards are satisfied then remove the relationship between removed node and power associated with it (nodes_power(n)), removed node and priority associated with it (nodes_priorities(n)).

See Fig. 15:

```
RequestOut ≙
  extended
STATUS
        ordinary
REFINES
        RequestOut
ANY
        n
        ip
WHERE
        grd1 : n∈nodes
        grd2 : ip∈ip_addresses
THEN
        act1 : cluster_ips:= {n} ⩤ cluster_ips
        act2 : nodes:=nodes \ {n}
        act3 : cluster_files:= {n} ⩤ cluster_files
        act4 : cluster_filesizes:= {n} ⩤ cluster_filesizes
        act5 : nodes_powers:= {n} ⩤ nodes_powers
        act6 : nodes_priorities:= {n} ⩤ nodes_priorities
END
```

Fig. 15. A specification of Events RequestOut in "Cluster_Mac1" (own work)

The event CalcPriorityForAllNodes specifies a set of required conditions to calculate priority for all node in the Cluster. The first guard states that nodes (n) to be calculated priority should belongs to a set of servers that added before to Cluster (nodes). The second guard states that power of node (pw) should be known and belongs to Integer number Z. The final guard states that files sizes (fz) should be known and belongs to Natural number N. If all guards are satisfied then the priority for the node (n) is equal to power (pw) minus sizes of files stored on the node (n).

See Fig. 16:

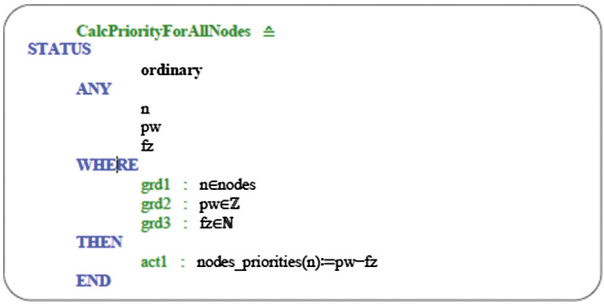


Fig. 16. A specification of Events CalcPriorityForAllNodes in "Cluster_Mac1" (own work)

The event GetBestPriority specifies a set of required conditions for getting the best priority on the Cluster. The first guard states that nodes (n) to be calculated priority should belong to a set of servers that added before to Cluster (nodes). The final guard states that priority for the node (n) greater than or equal to any other priority in the Cluster. If all guards are satisfied then the best priority is equal to priority for the node (n).

See Fig. 17:

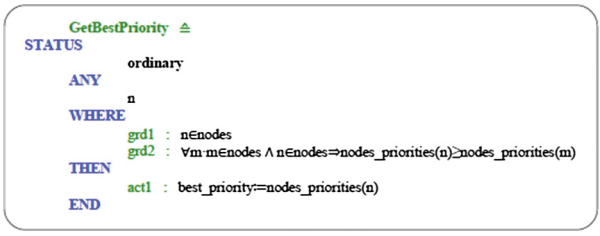


Fig. 17. A specification of Events GetBestPriority in "Cluster_Mac1" (own work)

7 Proof and Proof Statistics

After creating the models in Rodin, we can expand the "Cluster_Mac0" machine and "Cluster_Mac1" machine in the Event-B Explorer, and then expand the proof obligations

section, we can see 16 proofs for "Cluster_Mac0" and 23 proofs for "Cluster_Mac1" that have been completed, (a completed proof is indicated by a green mark), see Fig. 18.

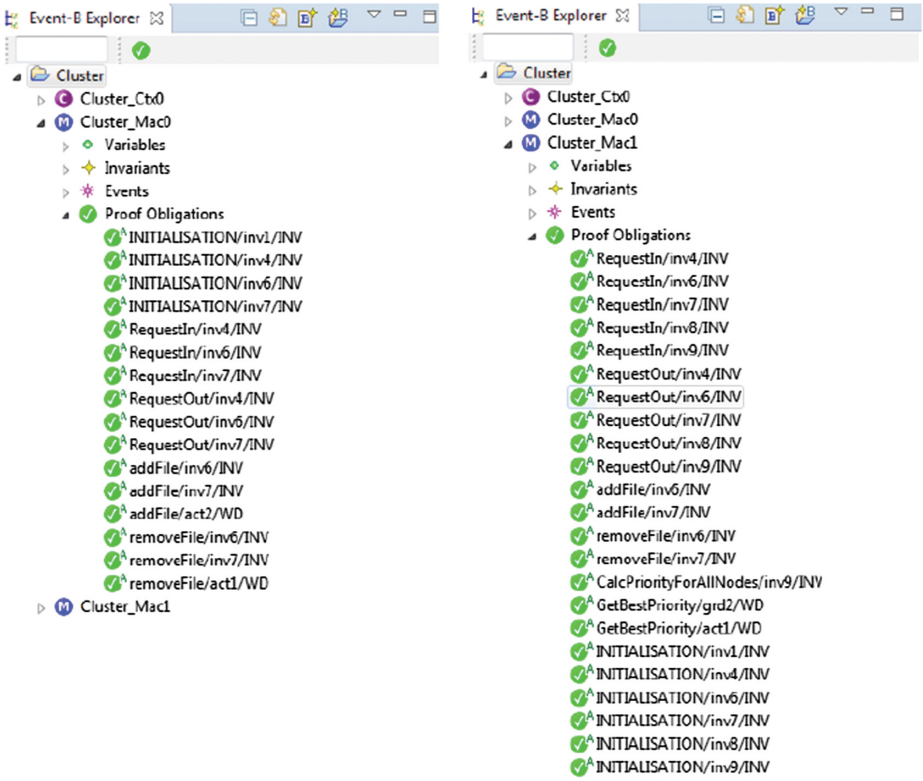


Fig. 18. Proof obligations of cluster model (own work)

Rodin automatically generates proof obligations (often abbreviated as PO) for properties that need to be proven. Each proof obligation has a name that identifies where the proof obligation was generated, e.g. RequestIn/inv6/INV [13, 17].

In Table 2, we can see proof statistics for Cluster model using the Rodin3.2 platform, the statistics give us the proof obligations generated and discharged by the Rodin (auto), and we can see that there is not interactively proved. The finial development of the Cluster results in 39 POs that are proved automatically by the Rodin, Fig. 18 show proof obligations of the Abstract model and the first refinement of Cluster model.

Table 2. Proof statistics of cluster model (own work)

Model	Automatic (%)	Manual (%)	Total (%)
Abstract model	16 (100%)	0 (0%)	16 (100%)
First refinement	23 (100%)	0 (0%)	23 (100%)
Total (%)	**39 (100%)**	**0 (0%)**	**39 (100%)**

8 Future Work

In these paper, we discuss the process, Master send the file to the node while adding it to the cluster to calculate process time for assigning priority to the cluster without specifying the type of files. In future work, we will make this process of sending files with the type of files like image files and text files. And request the nodes to process to these two types of files while adding it to cluster, and record the process time for both types of files. According to these process, the least time taking node give high priority to processing that type of file.

Master in the cluster must have fault tolerance machine like slave machine, in case the master fails down the slave promote itself as master so that the cluster will not stop all nodes member of cluster start communication with the new master.

Acknowledgement. This work was supported by the Ministry of Education, Youth and Sports of the Czech Republic within the National Sustainability Programme project No. LO1303 (MSMT-7778/2014) and by the European Regional Development Fund under the project CEBIA-Tech No. CZ.1.05/2.1.00/03.0089. Also supported by grant No. IGA/CebiaTech/2017/007 from IGA (Internal Grant Agency) of Thomas Bata University in Zlin.

References

1. http://www.internetlivestats.com
2. http://www.wearesocial.com
3. https://www.statista.com
4. Mall, N.N., Shikha, Rana, S.: Overview of Big Data and Hadoop, Department of Computer Science & Engineering, IIMT College of Engineering, Greater Noida (2016). ISSN: 2454-1362
5. Abbasi, A., Sarker, S., Chiang, R.H.L.: Big Data Research in Information Systems: Toward an Inclusive Research Agenda, McIntire School of Commerce, University of Virginia, USA (2016)
6. Shirkhorshidi, A.S., Aghabozorgi, S., Wah, T.Y., Herawan, T.: Big data clustering: a review. In: Murgante, B., et al. (eds.) ICCSA 2014. LNCS, vol. 8583, pp. 707–720. Springer, Cham (2014). doi:10.1007/978-3-319-09156-3_49
7. http://www.event-b.org
8. Megahed, F.M., Jones-Farmer, L.A.: Statistical perspectives on "big data". In: Knoth, S., Schmid, W. (eds.) Frontiers in Statistical Quality Control 11. FSQC, pp. 29–47. Springer, Cham (2015). doi:10.1007/978-3-319-12355-4_3
9. Havens, T.C., Bezdek, J.C., Palaniswami, M.: Scalable single linkage hierarchical clustering for big data. In: 2013 IEEE Eighth International Conference on Intelligent Sensors, Sensor Networks and Information Processing, pp. 396–401. IEEE (2013)
10. Kaur, A.: Big Data: A Review of Challenges, Tools and Techniques, Department of Computer Science and Applications, Khalsa College, Amritsar, Punjab, India (2016)
11. http://www.kapowsoftware.com
12. Williams, P., Soares, C., Gilbert, J.E.: A clustering rule based approach for classification problems. Int. J. Data Warehous. Min. **8**(1), 1–23 (2012)

13. Jastram, M., Butler, M.: Rodin User's Handbook: Covers Rodin v.2.8, CreateSpace Independent Publishing Platform, USA (2014). https://www3.hhu.de/stups/handbook/rodin/current/pdf/rodin-doc.pdf. ISBN 10: 1495438147, ISBN 13: 9781495438141
14. Damchoom, K., Butler, M., Abrial, J.-R.: Modelling and proof of a tree-structured file system in Event-B and Rodin. In: Liu, S., Maibaum, T., Araki, K. (eds.) ICFEM 2008. LNCS, vol. 5256, pp. 25–44. Springer, Heidelberg (2008). doi:10.1007/978-3-540-88194-0_5. http://www.ensiie.fr/~dubois/PR_2010/TreeFileSysICFEM2008.pdf
15. Abrial, J.-R., Butler, M., Hallerstede, S., Voisin, L.: An open extensible tool environment for Event-B. In: Liu, Z., He, J. (eds.) ICFEM 2006. LNCS, vol. 4260, pp. 588–605. Springer, Heidelberg (2006). doi:10.1007/11901433_32
16. Hoang, T.S., Furst, A., Abrial, J.-R.: Event-B Patterns and Their Tool Support. http://e-collection.library.ethz.ch/eserv/eth:5538/eth-5538-01.pdf
17. Abrial, J.-R., Butler, M., Hallerstede, S., Hoang, T.S., Mehta, F., Voisin, L.: Rodin: An Open Toolset for Modelling and Reasoning in Event-B (2009). http://deploy-eprints.ecs.soton.ac.uk/130/1/main.pdf

Proving the Effectiveness of Negotiation Protocols KQML in Multi-agent Systems Using Event-B

Ammar Alhaj Ali, Roman Jasek(✉), Said Krayem, and Petr Zacek

Faculty of Applied Informatics, Tomas Bata University in Zlin, Zlin, Czech Republic
ammar282n@hotmail.com, {jasek,zacek}@fai.utb.cz,
drsaid@seznam.cz

Abstract. Multi-Agents Systems (MAS) provide a good basis to build complex systems and in MAS a negotiation is a key form of interaction that enables agents to arrive at a final agreement. We present an event-B based approach to reasoning about a negotiation protocols in multi-agent systems (MAS), Key features of Event-B are the use of set theory as a modeling notation and it is a formal method that can be used in the development of reactive distributed systems and we propose using the Rodin modeling tool for Event-B that integrates modeling and proving.

Keywords: Multi-agent systems · Negotiation protocols · KQML · Event-B · Rodin

1 Introduction

In systems composed of multiple autonomous agents, negotiation is a key form of interaction that enables groups of agents to arrive at a mutual agreement regarding belief, goal or plan. Particularly because the agents are autonomous and in many cases, they have self-motivated as well as agents must influence others to convince them to act in certain ways. The potential benefits of agent negotiation include saving time and money, efficiency for computationally intense negotiations searching for optimal results, and the ability to incorporate multiple negotiation strategies for changing environments [1].

In this paper, we used Event-B to model and prove the effectiveness of negotiation protocols KQML.

2 Communication and Negotiation in MAS

Multi-agent systems are systems composed of multiple interacting computing elements, known as agents. Agents are computer systems with two important capabilities:

1. First, they are at least to some extent capable of autonomous action of deciding for themselves what they need to do in order to satisfy their design object lives.
2. Second, they are capable of interacting with other agents not simply by exchanging data, but by engaging in analogues of the kind of social activity that we all engage in every day of our lives: cooperation, coordination, negotiation, and the like [2].

R. Silhavy et al. (eds.), *Cybernetics and Mathematics Applications in Intelligent Systems*,
Advances in Intelligent Systems and Computing 574, DOI 10.1007/978-3-319-57264-2_40

Negotiation is the core of many MAS in construction because it is often unavoidable between different project participants with their particular tasks and domain knowledge whilst they interact to achieve their individual objective as well as the group goals. Furthermore, the importance of negotiation in MAS is likely to increase.

One reason is the growth of fast and inexpensive standardized communication infrastructures, over which separately designed agents belonging to different organizations, can interact in an open environment in real-time, and safely carry out transactions. Secondly, there is an industrial trend to be able to respond to larger and more diverse orders. Such ventures can realize the resource allocation efficiently and are easier to adapt to a dynamically changing economic environment [1].

3 KQLM

The Knowledge and Query Manipulation Language (KQML) represents the most widely used protocol for communication in multi-agent systems. KQML was developed as part of the American Knowledge Sharing Efforts (KSE) project at the University of Maryland.

KQML defines both a message format and a message transmission system that provides a general frame for the communication and cooperation in multi-agent systems.

In particular, KQML provides a group of protocols for identification, connection establishment, and message exchange. The semantic content of a message is not specified in more detail in KQML. Because the standard is open, various languages can be used to exchange knowledge and can be integrated into a KQML message [3].

KQML incorporates some ideas from speech-act theory, especially the idea of performatives. It has a built-in set of performatives, such as ACHIEVE, ADVERTISE, BROKER, REGISTER, and TELL [4], see Table 1 for some KQML performatives [1].

Table 1. Some of KQLM performative (own work)

Category	Performatives	Description
Basic query	Ask-if	A wants to know what B believes regarding the truth status of the content
	Ask-one	A wants one of B's answer to question the content
	Ask-all	A wants to know all B's responses that would make the content true of B (the response will be a collection of expressions)
	Evaluate	A wants B to evaluate the content
Multi-response query	stream-all	like ask-all, but the responses are to be delivered one by one
	eos	end of a stream of responses to an earlier query
Response	error	A states to B that B's message was not processed by A
	sorry	A states to B that B's message was processed by A, but no reply can be provided

In Figs. 1 and 2 are examples of a KQML message.

1. Performative: (Ask-one and Tell) determine the kinds of interactions between KQML-speaking agents.
2. Sender: The sender (Agent A in Ask-one message & Agent B in Tell message).
3. Receiver: The Receiver (Agent B in Ask-one & Agent A in Tell).

4. Language: The name of the representation language of content.
5. Ontology: The ontology (BookStore) provides additional information for the interpretation of the content.
6. Content: content, message or question…etc.
7. Reply-with & in-reply-to: the message will be identified by ID [6].

```
(Ask-one
: Sender        AgentA
: Receiver      AgentB
: reply-with    id1
: ontology      BookStore
: language      prolog
: content       (price ISBN 2541? Price)
)
```

Fig. 1. Agent A asks for the price (source [5])

```
(Tell
: Sender        AgentB
: Receiver      AgentA
: in-reply-to   id1
: ontology      BookStore
: language      prolog
: content       (price ISBN 2541, 24.50)
)
```

Fig. 2. Agent B sends a reply to Agent A (source [5])

4 Event-B & Rodin

Event-B is a formal method for system-level modelling and analysis, especially complex systems, Key features of Event-B are the use of set theory as a modelling notation, the use of refinement to represent systems at different abstraction levels and the use of mathematical proof to verify consistency between refinement levels.

The Rodin Platform is an Eclipse-based IDE for Event-B that provides effective support for refinement and mathematical proof. The platform is open source, contributes to the Eclipse framework and is further extendable with plugins [7].

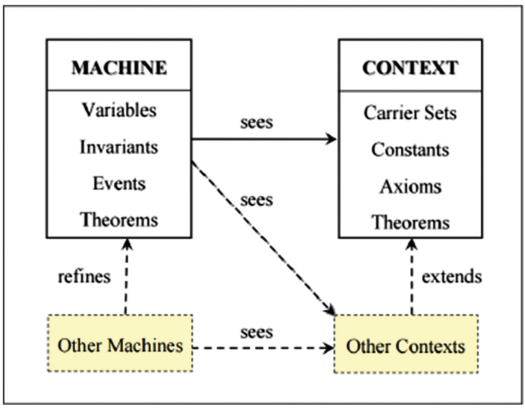


Fig. 3. Machine and Context relationship (source [8])

In Event-B we have two kinds of components.

1. Context: describes the static elements of a model.
2. Machine: describes the dynamic behaviour of a model.

Figure 3 sees relationships between Machine and Context.

4.1 Contexts

A context has the following components:

- Sets: User-defined types can be declared in the SETS section.
- Constants: We can declare constants here.
- Axioms: These axioms define rules that will always be the case for given elements of the context.
- Extends: A context may extend an arbitrary number of other contexts [9].

4.2 Machines

A machine has the following components:

1. Refines: the machine has the option of refining another one.
2. Sees: We can use the context's sets.
3. Variables: The variable's values are determined by an initialization event and can be changed by events.
4. Invariants: These are predicates that should be true for every reachable state.
5. Events: An event can assign new values to variables and the guards of an event specify the Conditions under which it can be executed [10].

4.2.1 Events

The event is presented in one of three possible forms, see Fig. 4. Where S() is generalized substitutions of the variable, G() represents a guard of event Evt, and t is a local variable [8, 9].

$$\text{Evt} \triangleq \text{begin } S(v) \text{ end}$$
$$\text{Evt} \triangleq \text{when } G(v) \text{ then } S(v) \text{ end}$$
$$\text{Evt} \triangleq \text{any } t \text{ where } G(t, v) \text{ then } S(t, v) \text{ end}$$

Fig. 4. Three possible forms of an event (own work)

The guards specify when an event is allowed to occur; the event can only be executed if the values of the machine's variables and parameters match the values listed in the guard.

If this is the case, we say that the event is enabled. The actions describe what changes will then be applied to the variables [10].

5 Formal Modelling of KQLM Performatives in Event-B

5.1 Ask-one

In Ask-one case the sender (Agent A) asks receiver (Agent B) for a response to its query that is contained in the message (m), see Fig. 5.

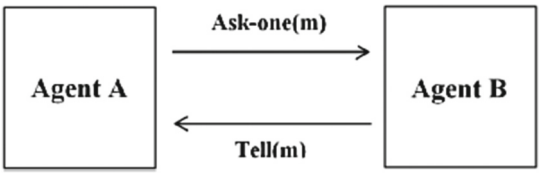


Fig. 5. Ask-one & Tell performatives (own work)

The receiver (Agent B) will respond to sender's request by Tell (m), in another word (Agent A) is interested in receiving exactly one response.

Formally, we will use two variables to represent the state of the process, AskOne to denote the number of requests that have been sent, and Tell to denote the number of responses that have been sent.

VARIABLES
AskOne
Tell
AskOneChannel
TellChannel
INVARIANTS
inv1 : AskOne ∈ ℕ
inv2 : Tell ∈ ℕ
inv3 : AskOne=Tell ∨ AskOne = Tell +1
inv4 : AskOneChannel ∈ BOOL
inv5 : TellChannel ∈ BOOL

Fig. 6. Variables and invariants of KQML_AskOne machine (own work)

As well as, we will use two variables two to represent the state of channels, AskOneChannel to denote there is one message on the channel in Ask-one state and TellChannel to denote there is one message on the channel in Tell state.

By invariant we will specify AskOne & Tell as natural numbers (inv1& inv2) and AskOneChannel & TellChannel as a Boolean (inv4& inv5), and the number of sent messages is identical or greater than the number of received messages by exactly 1, see Fig. 6.

At first, there is no message have been sent or received on the channel, see Fig. 7.

```
INITIALISATION ≙
EGIN
        act1  :  AskOne ≔ 0
        act2  :  Tell ≔ 0
        act3  :  AskOneChannel ≔ FALSE
        act4  :  TellChannel ≔ FALSE
END
```

Fig. 7. Initial values of variables in KQML_AskOne machine (own work)

We will define two events in our model. An event Agent_A_AskOne that will send AskOne from the agent A to the agent B, it will start when the number of AskOne and the number of Tell are identical and increase the number of AskOne by 1, see Fig. 8.

```
Agent_A_AskOne  ≙
WHEN
        grd1  :  AskOne = Tell
        grd2  :  AskOneChannel = FALSE
THEN
        act1  :  AskOneChannel ≔TRUE
        act2  :  AskOne ≔AskOne+1
END
```

Fig. 8. A specification of Event Agent_A_AskOne (own work)

An event Agent_B_tell that will send Tell from agent B to agent A, the guard of this event is the number of AskOne and Tell are different, see Fig. 9.

```
    Agent_B_tell  ≙
STATUS
        ordinary
    WHEN
        grd1  :  AskOne ≠ Tell
        grd2  :  TellChannel = FALSE
    THEN
        act1  :  Tell ≔ Tell + 1
        act2  :  TellChannel ≔ TRUE
    END
```

Fig. 9. A specification of Event Agent_B_tell (own work)

5.2 Stream-all

In stream-all performative asks that a set of answers be turned into a set of replies and eos purposes to notify Agent A that there are no more, see Fig. 10.

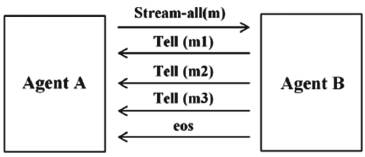


Fig. 10. Stream-all & eos performatives (own work)

In CONTEXT, we will use set Messages of possible messages, and we will define constant eosMsg that represents the end of stream message, see Fig. 11.

```
CONTEXT
    KQML_StreamAll_Ctx
SETS
    Messages
CONSTANTS
    eosMsg
AXIOMS
    Axm1  :  eosMsg ∈ Messages
END
```

Fig. 11. A specification of KQML_StreamAll_Ctx Context (own work)

```
MACHINE
    KQML_StreamAll_Mac
SEES
    KQML_StreamAll_Ctx
VARIABLES
    SendSN
    RecvSN
    Received_Messages
    Sent_Messages
    StreamAllSent
    StreamAllReceived
INVARIANTS
    inv1  :  SendSN∈ℕ
    inv2  :  RecvSN∈ℕ
    inv3  :  Received_Messages ∈ ℙ(Messages×ℕ)
    inv4  :  Sent_Messages ∈ ℙ(Messages×ℕ)
    inv5  :  StreamAllSent ∈ BOOL
    inv6  :  StreamAllReceived ∈ BOOL
```

Fig. 12. Variables and invariants of KQML_StreamAll_Mac machine (own work)

In MACHINE, the variable Sent_Messages is a set of messages sent by agent B and Received_Messages is set of messages received by agent A.

SendSN & RecvSN represent the current sender sequence number and the current receiver sequence number.

StreamAllSent & StreamAllReceived represent states of StreamAll message if it sent by agent A and received by agent B, see Fig. 12.

Initially, there is no message have been sent or received, see Fig. 13.

```
INITIALISATION  ≙
BEGIN
    act1     SendSN:=0
    act2  :  RecvSN:=0
    act3  :  Sent_Messages :=∅
    act4  :  Received_Messages:=∅
    act5  :  StreamAllSent := FALSE
    act6  :  StreamAllReceived := FALSE
END
```

Fig. 13. Initial values of variables in of KQML_StreamAll_Mac machine (own work)

Events of our model are as follows:

- AgentA_Send_StreamAll: for send StreamAll from agent A to agent B.
- AgentB_Receive_StreamAll: for receive StreamAll by agent B, see Fig. 14.

```
AgentA_Send_StreamAll ≙
WHEN
    grd1 : StreamAllSent =FALSE
THEN
    act1 : StreamAllSent ≔ TRUE
END
AgentB_Receive_StreamAll ≙
WHEN
    grd1 : StreamAllReceived=FALSE ∧ StreamAllSent=TRUE
THEN
    act1 : StreamAllReceived ≔TRUE
END
```

Fig. 14. A specification of Events AgentA_Send_StreamAll & AgentB_Receive_StreamAll (own work)

AgentB_Send_Tell: for send Tell(m1), Tell(m2) … Tell(mn) from agent B to agent A, we will send a message by adding it to Sent_Messages then increases SendSN by 1, the guards of this event are StreamAll message should be Received and MSG message has not been sent, see Fig. 15.

```
AgentB_Send_Tell ≙
ANY
    MSG
WHERE
    grd1 : StreamAllReceived =TRUE
    grd2 : MSG↦SendSN ∉ Sent_Messages
THEN
    act1 : Sent_Messages≔Sent_Messages ∪ {MSG↦SendSN}
    act2 : SendSN≔SendSN+1
END
```

Fig. 15. A specification of Event AgentB_Send_Tell (own work)

```
AgentA_Receive_Tell ≙
ANY
    MSG
WHERE
    grd1 : MSG↦RecvSN ∈ Sent_Messages
    grd2 : MSG↦RecvSN ∉ Received_Messages
    grd3 : RecvSN≠SendSN
    grd4 : MSG ≠ eosMsg
THEN
    act1 : Received_Messages≔Received_Messages ∪ {MSG↦RecvSN}
    act2 : Sent_Messages≔Sent_Messages \ {MSG↦RecvSN}
    act3 : RecvSN≔RecvSN+1
END
```

Fig. 16. A specification of Event AgentA_Receive_Tell (own work)

AgentA_Receive_Tell: in this event, Agent A will receive messages by add a message to Received_Messages and remove it from Sent_Messages then increases RecvSN by 1, the guards of this event are MSG message has been sent and RecvSN≠SendSN and message not equal eos, see Fig. 16.

6 Proof

After creating a model in Rodin, we can expand the KQML_StreamAll_Mac machine in the Event-B Explorer, and then expand the proof obligations section, we can see nine proofs have been completed, (a completed proof is indicated by a green mark), see Fig. 17, and in same steps we can see the proof obligations section of KQML_AskOne machine.

Rodin automatically generates proof obligations (often abbreviated as PO) for properties that need to be proven. Each proof obligation has a name that identifies where the proof obligation was generated, e.g. AgentB_Send_Tell/inv1/INV [9, 12].

7 Conclusion

Communication and negotiation are very important characteristics of multi-agent systems; and in this paper, we have presented some of the basic concepts in KQML also we have presented verification of interaction protocols in multi-agent system using KQML and Event-B.

Our final model is translated into the Event-B notation to verify required properties, so we can say, event-B allows us to define a kind of modeling methodology by writing the correct mathematical notions; wherefore we can apply event-B in modeling many different complex projects, but we should choose carefully invariants and variables to ease effort of proof.

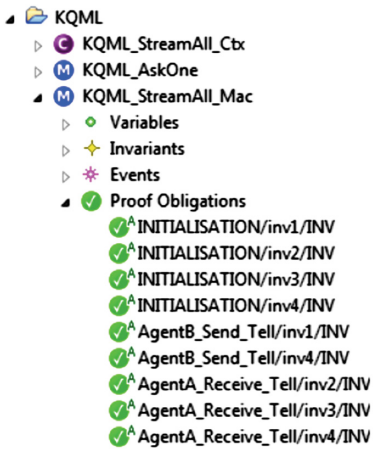


Fig. 17. Proof obligation

Thus, we can combine the event-B and protocol model to be able to use features of event-B to model negotiation protocols,

As well as the Rodin tool offers a reactive environment for constructing and analyzing models as do most modern integrated development environments, and provides integration between modelling and proving whereas this is an important feature for the developers to focus on the modelling task without switch between different tools to check proving in the same time.

Acknowledgement. This work was supported by the Ministry of Education, Youth and Sports of the Czech Republic within the National Sustainability Programme project No. LO1303 (MSMT-7778/2014) and by the European Regional Development Fund under the project CEBIA-Tech No. CZ.1.05/2.1.00/03.0089. Also supported by grant No. IGA/CebiaTech/2017/007 from IGA (Internal Grant Agency) of Thomas Bata University in Zlin.

References

1. Anumba, C., Ren, Z., Ugwu, O.O.: Agents and Multi-agent Systems in Construction, Routledge, United Kingdom (2005). ISBN 9781134242665
2. Wooldridge, M.: An Introduction to Multiagent Systems. Wiley, Hoboken (2002). ISBN 0-471-49691-X, Department of Computer Science, University of Liverpool, UK
3. Brenner, W., Zarnekow, R., Wittig, H.: Intelligent Software Agents: Foundations and Applications. Springer, Heidelberg (2012). ISBN 9783642804847
4. Shoham, Y., Leyton, K.: Multiagent Systems Algorithmic, Game-Theoretic, and Logical Foundations. Cambridge University Press, New York (2009). ISBN 9783642804847, UK
5. The Agent's Language. University of Osnabrück, Germany. http://www-lehre.inf.uos.de/~milic/Coxi/Communication_in_MAS.pdf
6. Srinivasan, D.: Innovations in Multi-agent Systems and Application. Springer, Heidelberg (2010). ISBN 9783642144356, Germany
7. http://www.event-b.org/
8. Damchoom, K., Butler, M., Abria, J.-R.: Modelling and Proof of a Tree-Structured File System in Event-B and Rodin (2008). http://www.ensiie.fr/~dubois/PR_2010/TreeFileSysICFEM2008.pdf
9. Abrial, J.-R., Butler, M., Hallerstede, S., Hoang, T.S., Mehta, F., Voisin, L.: Rodin: An Open Toolset for Modelling and Reasoning in Event-B (2009). http://deploy-eprints.ecs.soton.ac.uk/130/1/main.pdf
10. Jastram, M., Butler, M.: Rodin User's Handbook: Covers Rodin v.2.8. CreateSpace Independent Publishing Platform (2014). ISBN 10: 1495438147, ISBN 13: 9781495438141, USA. https://www3.hhu.de/stups/handbook/rodin/current/pdf/rodin-doc.pdf
11. Finin, T., Labrou, Y., Mayfield, J.: KQML as an agent communication language. http://www.cs.umbc.edu/kqml/papers/kqmlacl.pdf
12. Hoang, T.S., Furst, A., Abrial, J.-R.: Event-B Patterns and Their Tool Support. http://e-collection.library.ethz.ch/eserv/eth:5538/eth-5538-01.pdf

Correlation Analysis of Decay Centrality

Natarajan Meghanathan(✉)

Jackson State University, Jackson, MS, USA
natarajan.meghanathan@jsums.edu

Abstract. The decay centrality (DEC) metric for a vertex weighs the distance of the vertex to the rest of the vertices on the basis of a decay parameter ($0 < \delta < 1$). In this paper, we analyze a suite of 48 real-world networks and compute the DEC values for δ values ranging from 0.01 to 0.99 for each of these networks. We explore the presence of a particular or range of δ values within which there is a very strong positive correlation (Pearson's correlation coefficient of 0.8 or above) between DEC and each of the four commonly studied centrality metrics: degree centrality (DEG), eigenvector centrality (EVC), betweenness centrality (BWC) and closeness centrality (CLC). We observe 0.01 to be the most appropriate δ value for which there exists a very strong positive correlation between DEC and each of DEG, EVC and BWC for at least 50% of the networks.

Keywords: Decay centrality · Correlation analysis · Decay parameter · Real-world network graphs · Centrality metrics

1 Introduction

Centrality metrics [1] quantify the topological importance of a node in a network. The centrality metrics available for complex network analysis differ on the basis of the importance given to the node and its neighbors (degree-based metrics) or the shortest paths emanating or going through the node (shortest path-based metrics). Among the plethora of centrality metrics that exist in the literature, the four commonly studied centrality metrics are the degree-based degree centrality (DEG) [1] and eigenvector centrality (EVC) metrics [2], and the shortest path-based betweenness centrality (BWC) [3] and closeness centrality (CLC) metrics [1]. The degree centrality of a node is a measure of the number of neighbors of the node. The eigenvector centrality of a node is a measure of the degree of the node as well as the degree of its neighbors, and is computed using the power-iteration algorithm [6]. The betweenness centrality of a node is a measure of the fraction of the shortest paths between any two nodes that go through the node and the closeness centrality of a node is a measure of the distance (typically the number of hops) from the node to the rest of the nodes in the network. For a graph of V vertices and E edges, the time-complexity to compute the DEG, CLC, EVC and BWC metrics are respectively $\Theta(V^2)$, $\Theta(V^2 + VE)$, $\Theta(V^3)$ and $\Theta(V^2E)$. The BWC and EVC metrics are typically considered computationally-heavy metrics [15], whereas DEG and CLC are considered computationally-light metrics. For more details about the procedure to compute the centrality metrics, the interested reader is referred to [15].

R. Silhavy et al. (eds.), *Cybernetics and Mathematics Applications in Intelligent Systems*, Advances in Intelligent Systems and Computing 574, DOI 10.1007/978-3-319-57264-2_41

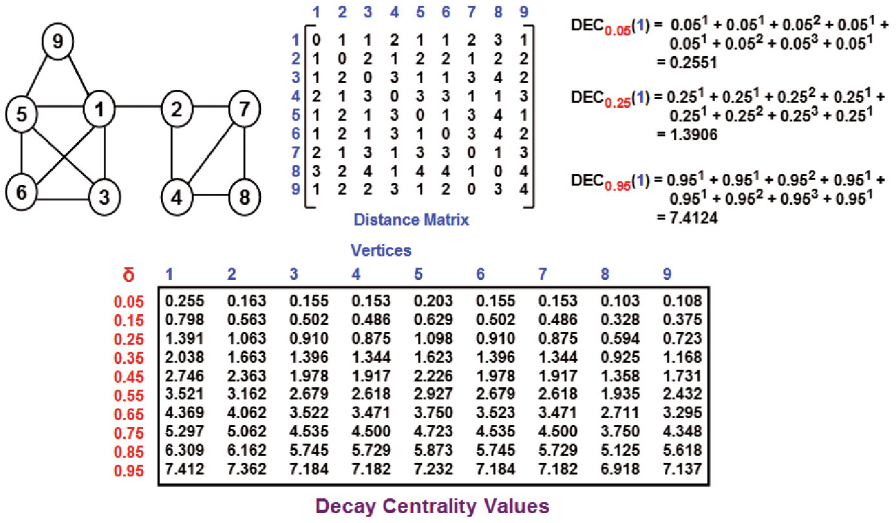
Throughout the paper, the terms 'node' and 'vertex', 'link' and 'edge' are used interchangeably. They mean the same.

In this paper, we explore a parameter-driven centrality metric called the decay centrality (DEC) metric [4] and analyze the correlation of the decay centrality of the vertices (when computed under different values of the decay parameter δ) with each of the above four commonly studied centrality metrics. We adopt the ordinal range of values proposed by Evans [5] and consider two centrality metrics to exhibit a very strong positive correlation if the Pearson's correlation coefficient [6] computed on the basis of the values incurred for the two metrics is 0.8 or above. We consider a suite of 48 real-world networks for the correlation study and employ the widely used Pearson's correlation coefficient (denoted PCC) as the correlation measure. Our objective is to explore whether there exists a particular δ value or a smaller range of δ values within which we observe a very strong correlation between DEC and a majority of the above four commonly studied centrality metrics. The results of our correlation study are very encouraging and informative. We observe DEC to exhibit a very strongly positive correlation with DEG, EVC and BWC for lower values of δ, especially when $\delta = 0.01$ for at least 50% of the real-world networks. On the other hand, we observe a very strongly positive correlation between DEC and CLC when $\delta = 0.99$ for all the 48 real-world networks.

The rest of the paper is organized as follows: Sect. 2 reviews the decay centrality (DEC) metric and illustrates its computation on an example graph for different values of the decay parameter δ. Section 3 presents an overview of the real-world network graphs and presents in detail the results of the correlation study for DEC vs. DEG, EVC, BWC and CLC for the real-world networks. Section 4 discusses related work and highlights our contributions. Section 5 concludes the paper by summarizing the results of the correlation study.

2 Decay Centrality

Decay centrality (DEC) is a measure of the closeness of a node to the rest of the nodes in the network [4]. However, unlike closeness centrality, the importance given to the distance (typically, in terms of the number of hops if the edges do not have weights) is weighted in terms of a parameter called the decay parameter δ ($0 < \delta < 1$). The formulation for computing the decay centrality of a vertex v_i for a particular value of the decay parameter δ is [4]: $\text{DEC}(v_i) = \sum_{v_i \neq v_j} \delta^{d(v_i, v_j)}$ where $d(v_i, v_j)$ is the distance from node v_i to node v_j (computed using the Breadth First Search algorithm [7]). The decay parameter δ essentially controls how important is a node v_j to a node v_i ($v_i \neq v_j$) that are at a distance $d(v_i, v_j)$ from each other. Nodes that have a higher decay centrality are more likely to be nodes that have several neighbors as well as be much closer to the rest of the nodes in the network [4]. For a graph of V vertices and E edges, the time-complexity to compute the DEC values for the vertices for a particular δ value would be $\Theta(V^2 + VE)$. Figure 1 presents the decay centrality of the vertices in an example graph for different values of the decay parameter δ. We also illustrate sample calculations of the decay centrality of vertex 1 for three different values of δ.



	1	2	3	4	5	6	7	8	9
1	0	1	1	2	1	1	2	3	1
2	1	0	2	1	2	2	1	2	2
3	1	2	0	3	1	1	3	4	2
4	2	1	3	0	3	3	1	1	3
5	1	2	1	3	0	1	3	4	1
6	1	2	1	3	1	0	3	4	2
7	2	1	3	1	3	3	0	1	3
8	3	2	4	1	4	4	1	0	4
9	1	2	2	3	1	2	0	3	4

δ	1	2	3	4	5	6	7	8	9
0.05	0.255	0.163	0.155	0.153	0.203	0.155	0.153	0.103	0.108
0.15	0.798	0.563	0.502	0.486	0.629	0.502	0.486	0.328	0.375
0.25	1.391	1.063	0.910	0.875	1.098	0.910	0.875	0.594	0.723
0.35	2.038	1.663	1.396	1.344	1.623	1.396	1.344	0.925	1.168
0.45	2.746	2.363	1.978	1.917	2.226	1.978	1.917	1.358	1.731
0.55	3.521	3.162	2.679	2.618	2.927	2.679	2.618	1.935	2.432
0.65	4.369	4.062	3.522	3.471	3.750	3.623	3.471	2.711	3.295
0.75	5.297	5.062	4.535	4.500	4.723	4.535	4.500	3.750	4.348
0.85	6.309	6.162	5.745	5.729	5.873	5.745	5.729	5.125	5.618
0.95	7.412	7.362	7.184	7.182	7.232	7.184	7.182	6.918	7.137

Fig. 1. Decay centrality of the vertices for an example graph

3 Correlation Analysis of Real-World Network Graphs

In this section, we first introduce the 48 real-world networks analyzed in this paper. Table 1 lists the network ID #, name and the values for the spectral radius ratio for node degree (λ_{sp}). All the real- world networks are modeled as undirected graphs. The spectral radius ratio for node degree [8] for the real-world network graphs analyzed in this paper ranges from 1.01 to 5.51 (indicating that the real-world network graphs analyzed range from random networks [9] with smaller variation in node degree to scale-free networks [10] of larger variation in node degree). The networks considered cover a broad range of categories (as listed below along with the number of networks in each category): Acquaintance network (12), Friendship network (9), Co-appearance network (6), Employment network (4), Citation network (3), Collaboration network (3), Literature network (3), Political network (2), Biological network (2), Game network (2), Transportation network and Trade network (1 each).

We now present the results of our correlation analysis for DEC vs. DEG, EVC, BWC and CLC (in Tables 2, 3, 4, 5). The δ values considered are 0.01,. . ., 0.99, in increments of 0.01. We do not consider δ values of 0 and 1 as DEC_0 would be zero for all vertices and DEC_1 for a vertex would correspond directly to the number of vertices in the component to which the vertex belongs to. The four tables (Tables 2, 3, 4 and 5 for the DEC-DEG, DEC-CLC, DEC-EVC and DEC-BWC correlations respectively) present the PCC values for DEC_δ ($\delta = 0.01$ and $\delta = 0.99$) vs. the centrality metrics as well as the largest PCC value for DEC_δ with the centrality metrics and the corresponding value of δ (referred to as δ*). We highlight the cells for the PCC values if the values are at least 0.80 (the minimum PCC for very strongly positive correlation).

Table 1. Real-world networks used in the correlation analysis

#	Network description	λ_{sp}	#	Network description	λ_{sp}
1	Word adjacency network	1.73	25	Les miserables network	1.70
2	Anna Karnenina network	2.48	26	Macaque dominance network	1.82
3	Jazz band network	1.45	27	Madrid train bombing network	1.04
4	C. Elegans neural network	1.68	28	Manufact. Comp. Empl. network	1.95
5	Centrality literature network	2.03	29	Soc. Net. journal co-authors	1.12
6	Citation graph drawing network	2.24	30	Author facebook network	3.48
7	Copperfield network	1.83	31	Mexican political elite network	2.29
8	Dolphin network	1.40	32	ModMath network	1.23
9	Drug network	2.76	33	Net. science co-author network	1.59
10	Dutch literature 1976 net.	1.49	34	US politics books network	5.51
11	Erdos collaboration network	3.00	35	Primary school contact network	1.42
12	Faux Mesa high school network	2.81	36	Prison friendship network	1.22
13	Friendship in hi-tech firm	1.57	37	San Juan Sur family network	1.32
14	Flying teams cade network	1.21	38	Scotland corp. interlock net.	1.29
15	US football network	1.01	39	Senator press release network	1.94
16	College dorm fraternity net.	1.11	40	Soccer world cup 1998 net.	1.47
17	Graph drawing 1996 network	2.38	41	Sawmill strike comm. network	1.45
18	Marvel universe network	2.54	42	Taro exchange network	1.22
19	Graph glossary network	2.01	43	Teenage female friend network	1.06
20	Hypertext 2009 network	1.21	44	UK faculty friendship network	1.59
21	Huckleberry coappearance net.	1.66	45	US airports 1997 network	1.35
22	Infectious socio-patterns net.	1.69	46	Residence hall friend network	3.22
23	Karate club network	1.47	47	Windsurfers beach network	1.27
24	Korea family planning network	1.70	48	World trade metal network	1.22

We observe the decay centrality metric to exhibit a very strongly positive correlation with degree centrality for all the 48 real-world networks (i.e., 100% of the networks) when $\delta = 0.01$. For $\delta = 0.99$, we observe a very strongly positive correlation between DEC and DEG for only 13 of the 48 real-world networks. The median of the δ^* values for which we observe the maximum PCC between DEC and DEG is also 0.01. On the other hand, we observe the decay centrality metric to exhibit a very strongly positive correlation with closeness centrality for all the 48 real-world networks (i.e., 100% of the networks) when $\delta = 0.99$. The median of the δ^* values for which we observe the maximum PCC between DEC and CLC is 0.89, close to 0.99. For $\delta = 0.01$, we observe a very strongly positive correlation between DEC and CLC for only 21 of the 48 real-world networks. Hence, it is very clear that 0.01 and 0.99 could be respectively considered the preferred δ values to observe a very strongly positive correlation for DEC vs. DEG and DEC vs. CLC metrics, and the probability of observing such a correlation is 100%.

For all the 48 real-world networks, we observe an overall trend of the PCC(DEC, DEG) values to decrease with increase in δ and the PCC(DEC, CLC) values to increase

Table 2. Correlation analysis of decay centrality vs. degree centrality

#	PCC(DEG, $DEC_{0.01}$)	PCC_{Max} ($DEGDEC_{\delta*}$)	PCC(DEG, $DEC_{0.99}$)	δ*	#	PCC(DEG, $DEC_{0.01}$)	PCC_{Max}(DEG, $DEC_{\delta*}$)	PCC(DEG, $DEC_{0.99}$)	δ*
1	**1.000**	**1.000**	0.733	0.01	25	**1.000**	**1.000**	0.712	0.01
2	**1.000**	**1.000**	0.691	0.01	26	**1.000**	**1.000**	**1.000**	0.04
3	**1.000**	**1.000**	0.740	0.01	27	**1.000**	**1.000**	0.352	0.01
4	**0.999**	**0.999**	0.615	0.01	28	**1.000**	**1.000**	**1.000**	0.02
5	**1.000**	**1.000**	0.290	0.01	29	**1.000**	**1.000**	0.232	0.01
6	**1.000**	**1.000**	0.506	0.01	30	**1.000**	**1.000**	−0.093	0.01
7	**1.000**	**1.000**	0.775	0.01	31	**1.000**	**1.000**	**0.840**	0.01
8	**1.000**	**1.000**	0.693	0.01	32	**1.000**	**1.000**	0.737	0.01
9	**1.000**	**1.000**	0.618	0.01	33	**1.000**	**1.000**	0.248	0.01
10	**1.000**	**1.000**	0.790	0.01	34	**1.000**	**1.000**	0.564	0.01
11	**1.000**	**1.000**	0.275	0.01	35	**1.000**	**1.000**	**0.923**	0.01
12	**1.000**	**1.000**	0.632	0.01	36	**1.000**	**1.000**	**0.821**	0.01
13	**1.000**	**1.000**	0.419	0.01	37	**1.000**	**1.000**	0.630	0.01
14	**1.000**	**1.000**	0.796	0.01	38	**1.000**	**1.000**	0.386	0.01
15	**0.997**	**0.997**	0.288	0.01	39	**1.000**	**1.000**	**0.848**	0.01
16	**1.000**	**1.000**	**0.998**	0.01	40	**1.000**	**1.000**	**0.855**	0.01
17	**0.999**	**0.999**	0.421	0.01	41	**1.000**	**1.000**	0.713	0.01
18	**1.000**	**1.000**	0.310	0.01	42	**1.000**	**1.000**	0.585	0.01
19	**1.000**	**1.000**	0.374	0.01	43	**1.000**	**1.000**	0.377	0.01
20	**1.000**	**1.000**	**0.997**	0.01	44	**1.000**	**1.000**	**0.855**	0.01
21	**1.000**	**1.000**	0.248	0.01	45	**1.000**	**1.000**	0.642	0.01
22	**1.000**	**1.000**	0.610	0.01	46	**1.000**	**1.000**	**0.856**	0.01
23	**1.000**	**1.000**	0.710	0.01	47	**1.000**	**1.000**	**0.960**	0.01
24	**1.000**	**1.000**	0.486	0.01	48	**1.000**	**1.000**	**1.000**	0.01

with increase in δ (also corroborated through the median of the δ* values of 0.01 and 0.89 for the DEC-DEG and DEC-CLC correlations respectively). Hence, if we observe a very strongly positive correlation between DEC-DEG for $\delta = 0.99$, it implies the DEC-DEG correlation could be considered to be very strongly positive through the entire δ-space ($0 < \delta < 1$). Likewise, if we observe a very strongly positive DEC-CLC correlation for $\delta = 0.01$, it implies the DEC-CLC correlation could be considered to be very strongly positive through the entire δ-space.

To understand this phenomenon further, we plot the distribution of the PCC ($DEC_{0.99}$, DEG) values with the PCC($DEC_{0.01}$, CLC) values in Fig. 2. We observe the data points to more or less fall along the diagonal line, indicating that real-world networks exhibited a similar level of $DEC_{0.99}$-DEG correlation vis-a-vis the $DEC_{0.01}$-CLC correlation. Real-world networks that tend to exhibit a very strongly positive DEC-DEG correlation through the entire δ-space ($0 < \delta < 1$) are also more likely to exhibit a very strongly positive DEC-CLC correlation through the entire δ-space, and vice-versa. We also observe that real-world networks that exhibited very strongly positive correlation with respect to both DEC-DEG and DEC-CLC had λ_{sp} less than 1.5.

Table 3. Correlation analysis of decay centrality vs. closeness centrality

#	PCC(CLC, $DEC_{0.01}$)	PCC_{Max} (CLC, $DEC_{\delta*}$)	PCC(CLC, $DEC_{0.99}$)	$\delta*$	#	PCC(CLC, $DEC_{0.01}$)	PCC_{Max}(CLC, $DEC_{\delta*}$)	PCC(CLC, $DEC_{0.99}$)	$\delta*$
1	**0.854**	**0.998**	**0.976**	0.47	25	**0.811**	**0.989**	**0.971**	0.55
2	**0.858**	**0.998**	**0.958**	0.45	26	**0.992**	**0.992**	**0.992**	0.99
3	**0.867**	**0.994**	**0.961**	0.51	27	0.347	**1.000**	**1.000**	0.99
4	0.725	**0.996**	**0.984**	0.55	28	**0.982**	**0.982**	**0.982**	0.90
5	0.294	**1.000**	**1.000**	0.99	29	0.227	**0.999**	**0.999**	0.99
6	0.500	**1.000**	**1.000**	0.99	30	−0.171	**0.992**	**0.992**	0.99
7	**0.911**	**0.980**	**0.936**	0.45	31	**0.886**	**0.993**	**0.987**	0.78
8	0.722	**0.986**	**0.977**	0.77	32	0.737	**1.000**	**1.000**	0.99
9	0.611	**1.000**	**1.000**	0.99	33	0.255	**1.000**	**1.000**	0.99
10	**0.913**	**0.997**	**0.969**	0.52	34	0.589	**0.990**	**0.989**	0.92
11	0.265	**1.000**	**1.000**	0.99	35	**0.955**	**0.998**	**0.994**	0.56
12	0.628	**1.000**	**1.000**	0.99	36	**0.882**	**0.997**	**0.979**	0.59
13	0.415	**1.000**	**1.000**	0.99	37	0.686	**0.991**	**0.985**	0.75
14	**0.845**	**0.991**	**0.983**	0.65	38	0.387	**1.000**	**1.000**	0.99
15	0.361	**0.999**	**0.998**	0.89	39	**0.935**	**0.998**	**0.979**	0.44
16	**0.990**	**0.990**	**0.986**	0.91	40	**0.945**	**0.996**	**0.971**	0.43
17	0.535	**0.988**	**0.976**	0.75	41	0.791	**0.991**	**0.973**	0.78
18	0.317	**1.000**	**1.000**	0.99	42	0.625	**0.997**	**0.995**	0.88
19	0.377	**1.000**	**1.000**	0.99	43	0.330	**0.994**	**0.994**	0.99
20	**0.990**	**0.990**	**0.987**	0.91	44	**0.922**	**0.994**	**0.981**	0.50
21	0.254	**1.000**	**1.000**	0.99	45	**0.818**	**0.997**	**0.962**	0.51
22	0.734	**0.988**	**0.970**	0.75	46	**0.901**	**0.996**	**0.990**	0.54
23	0.780	**0.992**	**0.983**	0.75	47	**0.976**	**0.995**	**0.991**	0.88
24	0.474	**1.000**	**1.000**	0.99	48	**0.987**	**0.987**	**0.986**	0.91

We observe a very strongly positive correlation between DEC and BWC for 30 of the 48 real-world networks (about 62% of the networks) and for the 28 of these 30 networks (i.e., for slightly more than 58% of the 48 real-world networks), the very strongly positive correlation between DEC and BWC could be observed even for $\delta = 0.01$. The median of the PCC(DEC, BWC) values is about 0.81 when $\delta = 0.01$ and as well as for $\delta = \delta*$ (when we observe the maximum PCC value) for the different real-world networks. The median of the $\delta*$ values is also 0.01. We observe a very strongly positive correlation between DEC and BWC for only 6 of the 48 real-world networks (less than 15% of the networks) when $\delta = 0.99$. All of these 6 real-world networks also exhibited a very strongly positive correlation throughout the entire δ space ($0 < \delta < 1$) for DEC-DEG, DEC-EVC and DEC-CLC as well as had spectral radius ratio of node degree values less than 1.5. Only 2 of the 30 real-world networks appear not to exhibit very strongly positive DEC-BWC correlation when we use $\delta = 0.01$ instead of the $\delta*$ value (that is not 0.01). Hence, we could conclude that 0.01 could be the preferred δ value to observe a very strongly positive correlation between DEC and BWC for any chosen real-world network and the probability of observing such a correlation would be slightly above 58%.

With regards to eigenvector centrality, we observe a very strongly positive correlation between DEC and EVC for 37 of the 48 real-world networks (about 77% of the networks).

Table 4. Correlation analysis of decay centrality vs. eigenvector centrality

#	PCC(EVC, $DEC_{0.01}$)	PCC_{Max}(EVC, $DEC_{\delta*}$)	PCC(EVC, $DEC_{0.99}$)	$\delta*$	#	PCC(EVC, $DEC_{0.01}$)	PCC_{Max}(EVC, $DEC_{\delta*}$)	PCC(EVC, $DEC_{0.99}$)	$\delta*$
1	**0.963**	**0.982**	**0.830**	0.11	25	**0.848**	**0.849**	0.645	0.04
2	**0.941**	**0.970**	0.781	0.15	26	**0.994**	**0.994**	**0.994**	0.04
3	**0.902**	**0.902**	0.675	0.03	27	**0.926**	**0.939**	0.283	0.13
4	**0.880**	**0.906**	0.714	0.09	28	**0.957**	**0.957**	**0.957**	0.08
5	**0.965**	**0.973**	0.320	0.07	29	0.506	0.506	−0.065	0.01
6	**0.813**	**0.822**	0.325	0.08	30	−0.712	0.552	0.552	0.99
7	**0.936**	**0.945**	0.789	0.15	31	**0.910**	**0.946**	**0.905**	0.22
8	0.726	**0.813**	0.661	0.31	32	**0.879**	**0.887**	0.545	0.10
9	0.659	0.760	0.307	0.28	33	−0.504	0.054	0.054	0.99
10	**0.949**	**0.955**	**0.825**	0.12	34	0.671	0.673	0.390	0.07
11	**0.920**	**0.929**	0.189	0.06	35	**0.982**	**0.982**	**0.907**	0.01
12	0.563	0.610	0.287	0.22	36	**0.847**	**0.877**	0.783	0.22
13	**0.938**	**0.941**	0.378	0.07	37	0.673	0.772	0.644	0.26
14	**0.827**	**0.865**	0.786	0.23	38	0.319	0.319	0.068	0.01
15	0.751	0.751	0.240	0.01	39	**0.977**	**0.981**	**0.870**	0.08
16	**0.997**	**0.998**	**0.997**	0.07	40	**0.969**	**0.982**	**0.906**	0.18
17	**0.858**	**0.965**	0.725	0.20	41	0.784	**0.818**	0.742	0.25
18	−0.723	−0.333	−0.333	0.99	42	0.781	**0.819**	0.634	0.19
19	**0.862**	**0.921**	0.377	0.17	43	0.527	0.597	0.188	0.34
20	**0.994**	**0.994**	**0.993**	0.12	44	**0.946**	**0.955**	**0.881**	0.14
21	**0.940**	**0.961**	0.353	0.11	45	**0.960**	**0.971**	0.706	0.07
22	**0.899**	**0.926**	0.665	0.14	46	**0.893**	**0.895**	0.794	0.04
23	**0.922**	**0.975**	**0.866**	0.07	47	**0.983**	**0.988**	**0.974**	0.23
24	**0.932**	**0.933**	0.342	0.05	48	**0.983**	**0.984**	**0.983**	0.09

Among these 37 real-world networks, we observe a very strongly positive correlation between DEC and EVC for 34 networks (i.e., for about 70% of the 48 real-world networks) when $\delta = 0.01$ and a median PCC of 0.900 (very close to the median of 0.928 for the maximum PCCs observed with the $\delta*$ values). The median of the $\delta*$ values (for which we observe the maximum PCC between DEC and EVC) is 0.12, in the vicinity of 0.01. On the other hand, we observe a very strongly positive correlation between DEC and DEG as well as between DEC and EVC for about 14 of the 48 real-world networks (less than 30% of the networks) when $\delta = 0.99$. Hence, we could conclude that 0.01 could be the preferred δ value to observe a very strongly positive correlation between DEC and EVC for any chosen real-world network and the probability of observing such a correlation would be about 70%.

From Tables 4 and 5, we observe that 25 of the 48 real-world networks indeed exhibit a very strongly positive DEC-EVC as well as DEC-BWC correlation at $\delta = 0.01$. Considering all of the above analysis, we could conclude that $\delta = 0.01$ would be the most appropriate value to observe a very strongly positive correlation for DEC with each of DEG, EVC and BWC for at least 50% of the real-world networks. We also observe (from Fig. 2) the PCC($DEC_{0.01}$, BWC) values to be lower than that of PCC ($DEC_{0.01}$, EVC) values for 36 of the 48 real-world networks, indicating that there is a

Table 5. Correlation analysis of decay centrality vs. betweenness centrality

#	PCC(BWC, $DEC_{0.01}$)	PCC_{Max}(BWC, $DEC_{\delta*}$)	PCC(BWC, $DEC_{0.99}$)	δ*	#	PCC(BWC, $DEC_{0.01}$)	PCC_{Max}(BWC, $DEC_{\delta*}$)	PCC(BWC, $DEC_{0.99}$)	δ*
1	**0.909**	**0.909**	0.522	0.01	25	0.746	0.746	0.477	0.01
2	**0.887**	**0.887**	0.493	0.01	26	**0.935**	**0.935**	**0.935**	0.01
3	0.607	0.607	0.344	0.01	27	0.728	0.728	0.163	0.01
4	0.773	0.773	0.330	0.01	28	**0.885**	**0.885**	**0.885**	0.01
5	**0.819**	**0.819**	0.120	0.01	29	0.397	0.501	0.340	0.39
6	**0.803**	**0.803**	0.313	0.01	30	0.266	0.431	0.186	0.37
7	**0.807**	**0.807**	0.579	0.01	31	**0.892**	**0.892**	0.751	0.01
8	0.601	0.643	0.584	0.37	32	**0.840**	**0.840**	0.424	0.01
9	0.650	0.652	0.318	0.05	33	0.439	0.542	0.248	0.25
10	**0.801**	**0.801**	0.612	0.01	34	0.715	0.799	0.720	0.48
11	0.778	0.778	0.155	0.01	35	**0.838**	**0.838**	0.769	0.01
12	0.631	0.638	0.348	0.15	36	**0.851**	**0.857**	0.664	0.08
13	**0.815**	**0.815**	0.243	0.01	37	**0.818**	**0.858**	0.703	0.18
14	0.786	**0.803**	0.718	0.16	38	0.744	0.797	0.241	0.17
15	0.341	**0.828**	0.788	0.46	39	**0.831**	**0.831**	0.597	0.01
16	**0.857**	**0.857**	**0.845**	0.01	40	**0.903**	**0.903**	0.677	0.01
17	**0.956**	**0.962**	0.432	0.04	41	**0.855**	**0.892**	0.708	0.19
18	0.706	0.710	0.172	0.05	42	**0.866**	**0.945**	0.798	0.17
19	**0.929**	**0.929**	0.239	0.01	43	0.220	0.394	0.361	0.85
20	**0.829**	**0.829**	**0.817**	0.01	44	**0.801**	**0.801**	0.595	0.01
21	**0.823**	**0.823**	0.064	0.01	45	0.699	0.699	0.358	0.01
22	**0.466**	**0.466**	0.274	0.01	46	**0.840**	**0.840**	0.696	0.01
23	**0.918**	**0.918**	0.639	0.01	47	**0.895**	**0.895**	**0.838**	0.01
24	0.468	0.471	0.285	0.10	48	**0.907**	**0.907**	**0.906**	0.01

75% chance that the PCC values for DEC-EVC and DEC-DEG correlations at $\delta = 0.01$ would be at least the PCC value for DEC-BWC at $\delta = 0.01$.

Figure 3 plots the δ* values (the δ value for which we observe the maximum PCC) observed for the real-world networks for DEC vs. each of the four centrality metrics. The δ* values are plotted in the sorted order so that we could easily infer the median of

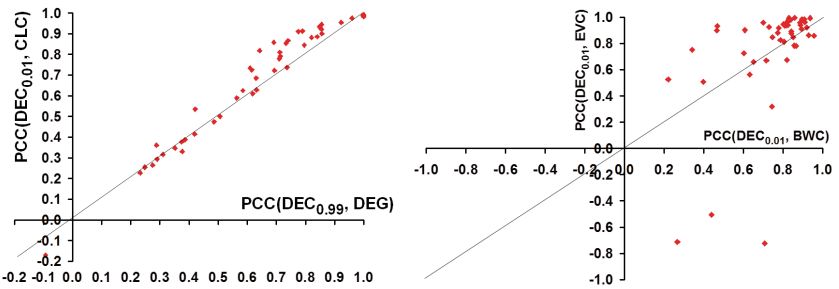


Fig. 2. Distribution of the PCC($DEC_{0.99}$, DEG) vs. PCC($DEC_{0.01}$, CLC) Values and Distribution of the PCC($DEC_{0.01}$, BWC) vs. PCC($DEC_{0.01}$, EVC) Values

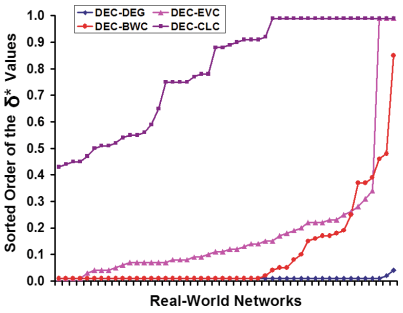


Fig. 3. Distribution of the Sorted δ^* Values for the Maximum PCC Values for the DEC vs. {DEG, EVC, BWC, CLC} Correlation

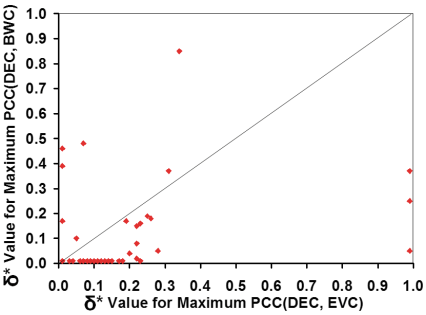


Fig. 4. Distribution of the δ^* Values for the Maximum PCC Values for the DEC-EVC vs. DEC-BWC Correlation

these values for the DEC correlation with each of the four centrality metrics. To corroborate our earlier conclusion, we could clearly observe the distribution of the δ^* values for the DEC-DEG, DEC-EVC and DEC-BWC metrics to be on the lower side and the distribution of the δ^* values for the DEC-CLC metrics to be on the larger side. Figure 4 plots the δ^* values for DEC-EVC correlation vs. the δ^* values for DEC-BWC correlation. We observe the δ^* values for the DEC-BWC correlation to be relatively larger for less than 15% (only 7 of the 48) of the real-world networks (corresponding to data points above the diagonal line).

4 Related Work and Our Contributions

Decay centrality has not been explored much in the literature for complex network analysis. To the best of our knowledge, ours is the first work to conduct a correlation study focusing on decay centrality, a parameter-driven centrality metric whose values change with the values for the decay parameter (δ). Most of the work (e.g., [11, 12]) on correlation studies (involving centrality metrics) were focused on the commonly studied centrality metrics such as the neighborhood-based degree centrality and eigenvector centrality and shortest path-based betweenness centrality and closeness centrality, all of which are not parameter-driven and remain the same for a particular network. The objective of such correlation studies has been typically to identify computationally-light alternatives (like DEG and its derivatives [13]) for computationally-heavy metrics (such as EVC and BWC) for both real-world networks and simulated networks of theoretical models [9, 10]. The focus of our paper is different from such typical correlation studies in the literature. We seek to explore a particular value or a smaller range of values for the decay parameter (δ) such that the decay centrality metric exhibits a very strong correlation with more than one centrality metric. The results of our correlation analysis indicate that $\delta = 0.01$ could be the appropriate value for use to determine decay centrality values that simultaneously exhibit very strong positive correlation with the DEG, EVC and BWC metrics.

The most related work to our research is a recent study [4] on random networks [9] for which a single threshold value of the decay parameter (referred here as δ_{thresh}) was observed to exist (for a particular operating condition) such that nodes with high degree centrality also had a high decay centrality computed for δ values less than δ_{thresh} and nodes with high closeness centrality also had a high decay centrality computed for δ values above δ_{thresh}. It was observed in [4] that for random networks: nodes with the largest values for degree centrality and closeness centrality are more likely to be nodes that also incur the largest values for decay centrality for almost all values of δ. In addition, nodes that had the largest decay centrality for a certain value of δ are more likely to be part of the set of nodes that had the largest degree centrality or the largest closeness centrality. The likelihood of all of the above was studied using multinomial logistic regression [14].

Our paper differs from the above work and is innovative on the following lines: We analyze real-world networks rather than the simulated random networks. We use the Pearson's correlation measure to study the correlation between the actual centrality values rather than multinomial logistic regression [14] to study the sets of vertices that had the largest values of centrality. We observe the decay centrality to exhibit very strong positive correlation with degree centrality, betweenness centrality and eigenvector centrality for lower values of the decay parameter δ (especially, $\delta = 0.01$) and very strong positive correlation with closeness centrality for larger δ values (especially, $\delta = 0.99$). Unlike the observation for random networks in [4], for each of the 48 real-world networks studied in this paper: we observe two different ranges of δ values in which the decay centrality metric exhibits very strong positive correlation with the degree centrality and closeness centrality metrics.

5 Conclusions and Future Work

The high-level contribution of this paper is a comprehensive correlation analysis of the decay centrality (DEC) metric vis-a-vis the four commonly studied centrality metrics such as degree (DEG), eigenvector (EVC), betweenness (BWC) and closeness (CLC) centralities. We analyzed a suite of 48 real-world networks of spectral radius ratio for node degree ranging from 1.01 to 5.51 (degree distribution ranging from random to scale-free). We observe the DEC-DEG, DEC-EVC and DEC-BWC correlations to be very strong for lower values of δ and the DEC-CLC correlation to be very strong for larger values of δ. More specifically, we observe $\delta = 0.01$ to be the most appropriate value to simultaneously observe a very strongly positive correlation (Pearson's correlation coefficient of 0.8 or above) for DEC with the three centrality metrics: DEG, EVC and BWC, and $\delta = 0.99$ to be the most appropriate value for observing a very strongly positive DEC-CLC correlation.

As we analyze networks of diverse degree distributions, we could consider the % of real-world networks for which we observe a very strongly positive correlation as the probability for observing a very strongly positive correlation for any real-world network. On these lines, there is a 100% probability of observing very strongly positive DEC-DEG and DEC-CLC correlations for $\delta = 0.01$ and $\delta = 0.99$ respectively. Also, at $\delta = 0.01$, we claim there is respectively a 58% and 70% chance of observing a very

strongly positive DEC-BWC and DEC-EVC correlation. With respect to impact of degree distribution, we observe real-world networks with a lower spectral radius ratio for node degree (i.e., lower variation in node degree) to more likely exhibit a very strongly positive correlation between DEC and one or more of the four centrality metrics. All the six real-world networks that exhibited a very strongly positive correlation between DEC and each of the four centrality metrics through the entire δ-space ($0 < \delta < 1$) as well as all the thirteen real-world networks that exhibited very strongly positive DEC-DEG and DEC-CLC correlations through the entire δ-space had spectral radius ratio for node degree values less than 1.5.

The results of our paper could lay the groundwork for using the decay centrality values of the vertices (obtained for lower values of δ) to predict the values for the computationally-heavy BWC and EVC metrics. As part of future work, we plan to investigate the suitability of using $DEC_{0.01}$ (decay centrality determined for $\delta = 0.01$) to obtain a network-wide ranking of the vertices or conduct a pair-wise ordering of the vertices in lieu of BWC and EVC.

Acknowledgments. The research is financed by the NASA EPSCoR sub award (#: NNX14AN38A) from University of Mississippi.

References

1. Freeman, L.C.: Centrality in social networks: conceptual clarification. Soc. Netw. **1**(3), 215–239 (1979)
2. Bonacich, P.: Power and centrality: a family of measures. Am. J. Sociol. **92**(5), 1170–1182 (1987)
3. Brandes, U.: A faster algorithm for betweenness centrality. J. Math. Sociol. **25**(2), 163–177 (2001)
4. Tsakas, N.: On decay centrality. arXiv:1604.05582 (2016)
5. Evans, J.D.: Straightforward Statistics for the Behavioral Sciences. Brooks Cole Publishing Company, Pacific Grove (1995)
6. Lay, D.C., Lay, S.R., McDonald, J.J.: Linear Algebra and its Applications. Pearson, London (2015)
7. Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C.: Introduction to Algorithms. MIT Press, Cambridge (2009)
8. Meghanathan, N.: Spectral radius as a measure of variation in node degree for complex network graphs. In: Proceedings of the 3rd International Conference on Digital Contents and Applications, Hainan, China, pp. 30–33 (2014)
9. Renyi, E.: On random graphs I. Publicationes Mathematicae **6**, 290–297 (1959)
10. Barabasi, A.-L., Albert, R.: Emergence of scaling in random networks. Science **286**(5439), 509–512 (1999)
11. Li, C., Li, Q., Van Mieghem, P., Stanley, H.E., Wang, H.: Correlation between centrality metrics and their application to the opinion model. Eur. Phys. J. B **88**(65), 1–13 (2015)
12. Meghanathan, N.: Correlation coefficient analysis of centrality metrics for complex network graphs. In: Silhavy, R., Senkerik, R., Oplatkova, Z.K., Prokopova, Z., Silhavy, P. (eds.) Intelligent Systems in Cybernetics and Automation Theory. AISC, vol. 348, pp. 11–20. Springer, Cham (2015). doi:10.1007/978-3-319-18503-3_2

13. Meghanathan, N.: A computationally-lightweight and localized centrality metric in lieu of betweenness centrality for complex network analysis. Vietnam J. Comput. Sci. **4**(1), 23–38 (2017). Springer
14. Greene, W.H.: Econometric Analysis. Pearson, London (2011)
15. Meghanathan, N.: Correlation coefficient analysis: centrality vs. maximal clique size for complex real-world network graphs. Int. J. Netw. Sci. **1**(1), 3–27 (2016)

Virtual Lab: An Adequate Multi-modality Learning Channel for Enhancing Students' Perception in Chemistry

Krishnashree Achuthan[1] and Smitha S. Murali[2(✉)]

[1] Amrita Center for Cybersecurity Systems and Networks, Amritapuri, India
[2] VALUE Virtual Labs, Amrita School of Engineering,
Amrita Vishwa Vidyapeetham, Amrita University, Amritapuri, India
smithasm@am.amrita.edu

Abstract. This paper investigates the instructional effectiveness of learning modalities towards enhancing learners' conceptual understanding of crystal field theory (CFT) using a multimedia rich platform such as Virtual Laboratory. The virtual laboratory in the present work has integrated modalities such as graphics, images, animations, videos and simulations for simultaneous demonstration of concepts related to CFT. This study aims to evaluate the impact of these modalities on the learning outcomes of visual, auditory and kinesthetic learners irrespective of their preferred learning modality. A case study of 524 undergraduate chemistry students from four higher educational institutes was carried out as part of the evaluation. Assessment of knowledge, conceptual understanding, application and analysis with and without the virtual lab platform was done using assessment quizzes. Results showed that students that underwent a combination of visual, auditory and kinesthetic learning modalities within virtual lab environment had significantly improved their understanding resulting in better performance. The study also characterizes the effectiveness of integrated modalities on the enhancement of learning amongst the three types of learners.

Keywords: Animation · Crystal field theory · Instruction styles · Learning modality · Simulation · Virtual lab

1 Introduction

The importance and influence of modality effects are crucial factors to learning and the development of information and communication technology (ICT) having had a significant impact on knowledge outcome [1, 2]. This leads educationalists to explore methods to improvising existing learning and teaching styles in addition to incorporating new technologies in education. Learning is a process of gaining knowledge; to strengthen or modifying the existing knowledge. Learning can also be thought of as a process of reinforcing or acquiring a skill by either being taught or practicing or experiencing a novel idea or thought construct. Learners perceive knowledge through different modality channels such as seeing, hearing or intuition [2]. Learning modalities are the sensory channels through which the learners perceive and store information. They indicate how effectively learners perceive an external stimulus during the process of learning [2].

R. Silhavy et al. (eds.), *Cybernetics and Mathematics Applications in Intelligent Systems*,
Advances in Intelligent Systems and Computing 574, DOI 10.1007/978-3-319-57264-2_42

Learners' understanding of concepts improve when new skills or information is explained using multi-modal channels such as audio, video, animations, simulations or models in comparison to simple text book reading. Studies [1–6] have shown that teaching through various modalities help students with their understanding and also enhance their ability to think creatively and discover new things. Modality becomes crucial when the subject of learning is complex. Nowadays different learning modality channels are available. A learner can choose his modality channel based on his or her learning style. The pathways through which a learner perceives information are broadly classified into three categories - visual, auditory and kinesthetic or sometimes known as learning modalities. Some learners use some or all three modalities to perceive and learn new information. According to VAK modality theory [1], only one or two of these modalities are used by a learner to perceive information. But in a classroom, all the three types of learners – visual, auditory and kinesthetic are present and the perception level of students may vary depending on the method of instruction. Most methods of instruction address visual and auditory learners and not the kinesthetic learners. Hence there continues to be a wide gap exists between the learning styles of students and the traditional teaching styles of instructors [2, 3, 6]. Thus it is important to identify each student's predominant learning style and incorporate activities pertaining to each domain, thereby optimizing holistic ways to teach students effectively. By understanding the most common learning styles, an instructor can easily recognize a student as a visual, auditory or kinesthetic learner. Table 1 shows different type of learners, their important characteristics and the preferred learning modalities [1–3].

Table 1. Learning styles and modalities

Category	Important characteristics	Learning modality
Visual learners	• Observe things rather than act or listen	Demonstrations, Videos, Animations, Graphics, Tables
	• Prefer reading during verbal activities	
	• They perceive and learn information via visual demonstrations	
Auditory learners	• Listen to verbal instructions rather than observing or engaging in activities	Audios, lectures
	• They perceive and learn information via verbal instructions and also prefer demonstrations with verbal explanation (audio) for learning	
Kinesthetic learners	• Doing things rather than observe or listen	Experimentation, simulation
	• They mainly perceive and learn information through the sense of touching that is by practicing or doing activities (real or simulation)	

2 Transformational Teaching Styles

In most of the educational institutions, traditional mode of socratic teaching is followed wherein a teacher usually delivers a lecture or sometimes uses some physical models to

present a concept that is difficult to students. Such demonstration based on physical models may be insufficient for teaching difficult phenomena and to identify learning deficiencies as they occur. This results in a gap between students' level of understanding and instructor's expectation and demands of curriculum. This increases cognitive and psychological challenges for both the learner and the instructor [2]. In many sciences, especially in chemistry, the concepts are abstract and do not have distinguishable examples within the macroscopic world for students to form cognitive maps [9]. In such cases, most of the students are unable to perceive those concepts and they learn via rote learning resulting in poor retention. Many studies [10, 11] have indicated that instructions using modalities increase the learners working memory capacity and therefore their learning ability. In this paper we are presenting such a multi modality learning platform "Virtual Lab" that integrates multiple learning modalities under a single platform and is shown in Fig. 1.

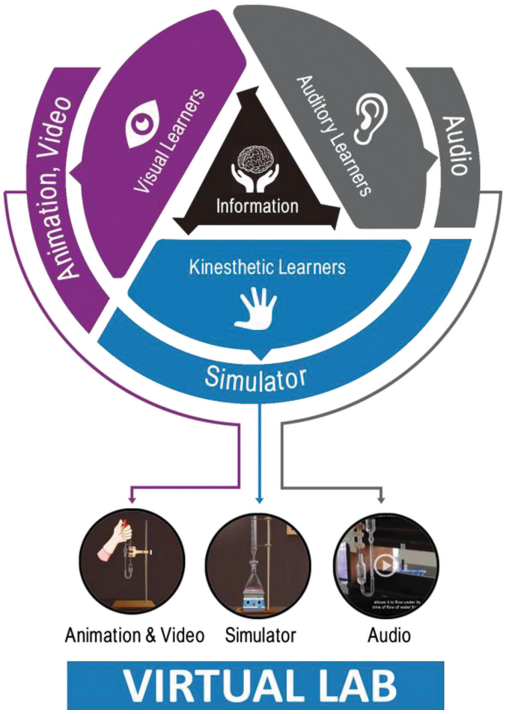


Fig. 1. Virtual lab as a multimodality learning channel

Amrita University developed VALUE initiative (Virtual Amrita Laboratories Universalizing Education), the virtual lab platform (vlab.amrta.edu) that provides students with access to theory and virtual experimentation, at all times and from anywhere in an interactive fashion and through the internet [12–15]. Studies have also been reported on the deployment of virtual labs in the regular curriculum enhanced the teaching and learning experiences and thereby reducing the major challenges faced by

traditional laboratory education mainly in remote areas [15, 16]. This work explores the effectiveness of integrated modalities within the virtual lab environments on various types of learners. The visualization in virtual lab can be brought about by use of graphics, 3 dimensional images, videos, animations and simulations to represent the colour change in a reaction, electron delocalization, orientation of orbitals, shape of molecules etc. As a case study, we have selected "Crystal Field Theory" in chemistry, a theory based experiment.

2.1 Crystal Field Theory

Crystal Field Theory (CFT) was first proposed by Hans Bethe in 1929 and was later developed by H. Bethe and John Hasbrouck van Vleck in 1930 for describing various spectroscopies of transition metal coordination complexes. The details below describe the importance of visualization in understanding the concepts underlying CFT. According to CFT, bonding between a central metal atom and its ligand arises from pure electrostatic force of attraction [7]. The transition metal which forms a central atom in

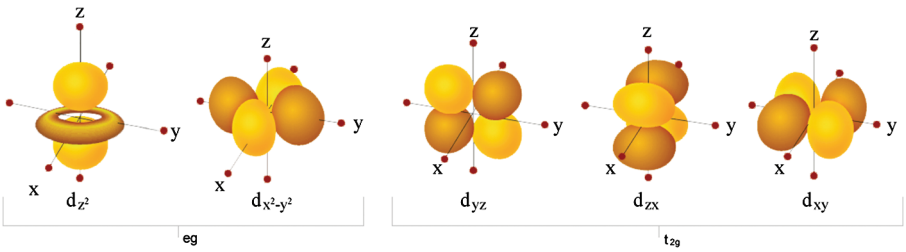


Fig. 2. Angular dependence functions of d-orbitals

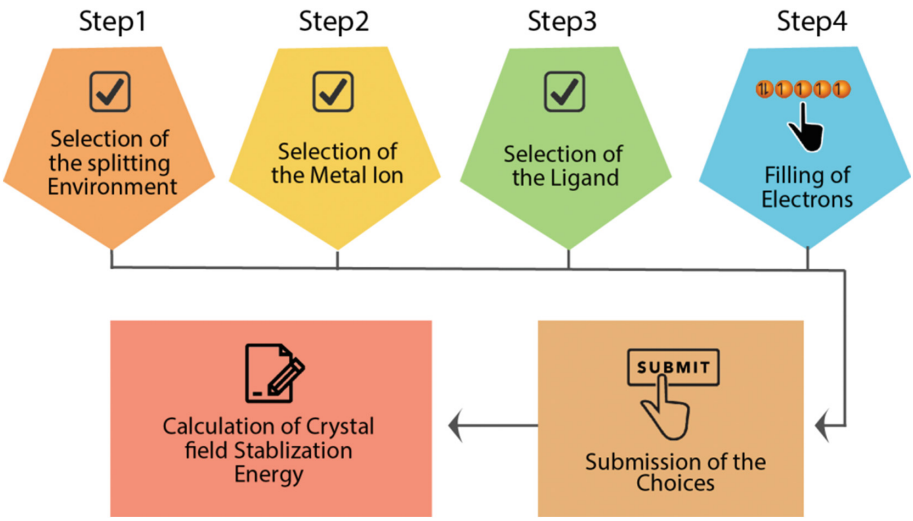


Fig. 3. Different steps in simulator for CFT

the complex is considered as a positive ion and is surrounded by oppositely charged species or neutral molecules known as ligands. In case of free metal atom/ion, all the 5d-orbitals have the same energy i.e., they are in a degenerate state. The angular dependence functions of d-orbitals are shown in Fig.2.

In the formation of coordination complexes, if all the ligands are approaching the central metal ion at an equal distance from each of the d-orbital, they will still remain in the degenerate state but the d-orbitals will differ in their orientations. This is due to the repulsive forces between the electrons on the central metal atom and the electrons on the ligands. As a result, the electrons reside the d-orbitals farthest away from the direction of approach of the ligand. The energy of the orbital lying in the direction of ligand is larger than that of the orbitals lying in between the ligands. Hence, the d – orbitals can be divided into two groups i.e. e_g and t_{2g} depending upon the nature of their orientations. The lobes of the e_g orbitals ($dx^2 - y^2$ and dz^2) oriented along the x, y and z axes while those of t_{2g} orbitals (dxy, dxz and dyz) oriented in region in between the coordinate axes. The conversion of five degenerate d-orbitals of the ion into two set of orbitals having different energy is called crystal field splitting. This concept forms the basis of crystal field theory [7, 8] Traditionally, the orientation of d-orbitals, orbital splitting patterns, identification of spin state (low spin/high spin), orbital splitting diagram and calculation of crystal field stabilization energy (CFSE) in CFT are difficult concepts to explain in the classroom environment without any modalities. The most effectual way to acquire practical knowledge in chemistry is through experimental and laboratory work [6]. Although there have been many studies that emphasize the effectiveness of visualizations on learning skills [9–11], there has not been detailed explorations on improving learning abilities in the area of CFT.

In virtual lab platform, crystal field theory is explained through in-depth theory and simulator. The interactive design of virtual lab simulator allows students to determine the number of electrons in e_g and t_{2g} orbitals and there by calculating the crystal field stabilization energy. In CFT, Octahedral and tetrahedral complexes are most common and important. In the octahedral environment, the energy of e_g orbitals is always greater than t_{2g} orbital ($e_g > t_{2g}$) and there exist a chance for the formation of both low spin & high spin complexes but in tetrahedral environment, e_g orbitals has lower energy than t_{2g} orbitals ($e_g < t_{2g}$) resulting in the formation of low spin complexes. In low spin complexes, electrons are filled according to aufbau principle however high spin complexes always deviated from aufbau principle [7, 8]. Usually students have had difficulties in understanding these concepts in a typical classroom environment. But in virtual lab, a step-by-step procedure is adopted in the simulator for explaining these phenomena. This was shown in Fig. 3. The first step is the selection of splitting; two choices are given- octahedral and tetrahedral. For example, if the user selected octahedral splitting, Fe^{2+} (d^6) as metal and Cl^- as ligand, then the simulator shows the orbital energy diagram of the corresponding complex $[FeCl_6]^{4-}$. It is followed by the filling of electrons in stage 3 of the orbital energy diagram. Virtual lab platform allows the user to fill the electrons by clicking on the energy levels and submit his choices. Simulator will show the feedback immediately. This was shown in Fig. 4. Once the electrons are filled, the CFSE can be calculated by counting the number of electrons in t_{2g} and e_g

orbital. This entire process can be repeated until the student improves his conceptual understanding in the area without the help of an instructor. In addition to this, students are provided with quizzes and assignments to self assess their conceptual knowledge. By augmenting virtual lab, teachers can explain the concept of crystal field splitting; the filling of d- orbitals in different bonding environment and their orientation; calculation of CFSE effortlessly.

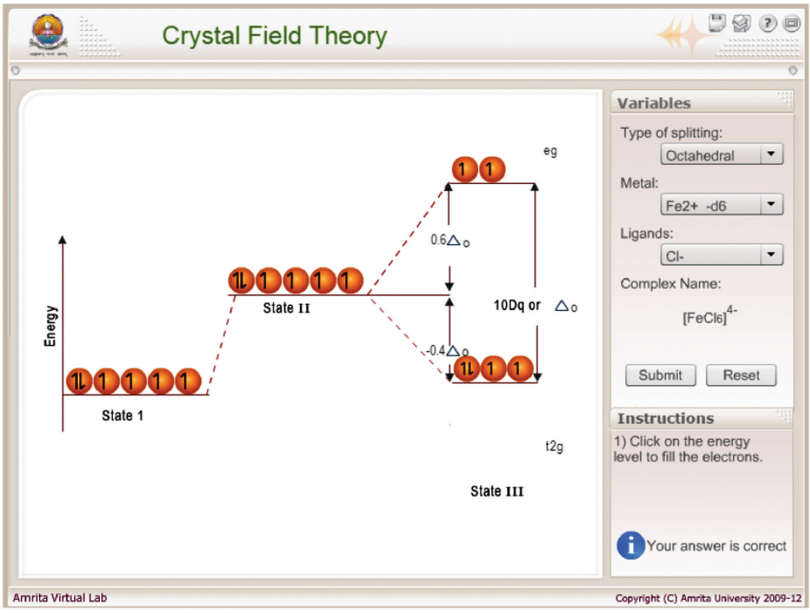


Fig. 4. Virtual lab simulator for crystal field theory

3 Research Method

The main objective of this paper was to investigate the instructional effectiveness of learning modalities in virtual labs in enhancing learners' conceptual understanding of theory/practical aspects in chemical sciences. For this, we used an experimental methodology that differs from the traditional empirical-analytical teaching. Our objective is to observe the impact on students' knowledge, when the chemical reactions or rearrangements are presented as animations, videos or simulations with the help of virtual laboratory. Most of the basic concepts related to the crystal field theory are taught as part of K-12 chemistry curriculum via textbook oriented instruction. This study was conducted amongst students (N = 524) pursuing undergraduate education in chemistry from four higher educational institutions (HEI 1, HEI 2, HEI 3 & HEI 4). The undergraduate program is usually 3 years long and the institutes chosen were active members of a program called the nodal center program [12]. Within the nodal center program, the institutes and teachers are trained to use ICT tools such as virtual lab as part of their

teaching and learning activities. The cohort chosen in the present study were in their 2nd year of study. The control and experimental groups were designed in such a way that both groups of students learnt crystal field theory in two different learning environments.

- Control group (TTI) that underwent only text book oriented instruction.
- Experimental group (TVI) underwent textbook followed by hands-on experience with virtual labs.

The distributions of students from four HEI are tabulated in Table 2. For the study, we designed a single-factor experiment with parallel classes of compared groups; experimental group and control group [17].

Table 2. Distribution of students

HEI	No. of students	TTI`	TVI
HEI 1	176	88	88
HEI 2	134	67	67
HEI 3	132	66	66
HEI 4	82	41	41
Total no. of students		262	262

A schematic of the experimental activities is described in Fig. 5. Both groups were subjected to a prior knowledge test to assess the basic knowledge in CFT. After the prior knowledge test, students were exposed to explanation of theory concepts and its interpretations in detail using symbolic illustrations. In TTI (control group), Teachers used one dimensional or two dimensional images in the text book to explain the theory in-depth. On the other hand, in the TVI (experimental group) teachers use blended pedagogic approach wherein learning modalities like animations, simulations and graphics in virtual laboratory are integrated with text book references for instruction.

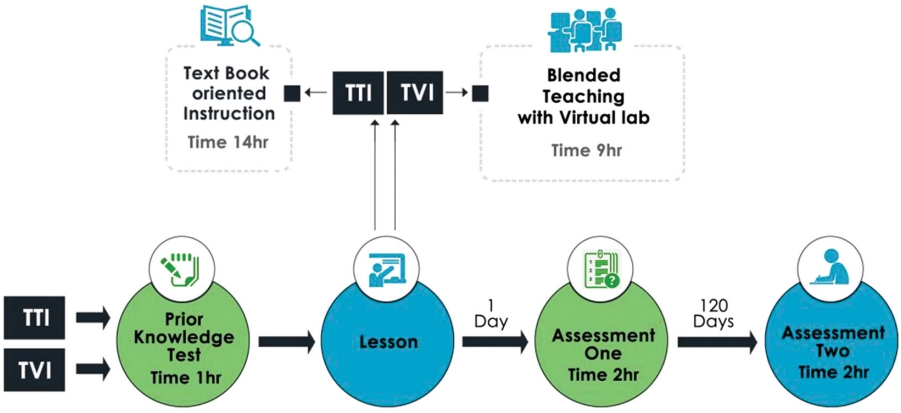


Fig. 5. Experimental study

In both cases, 12 h were allotted for the lecture session. TTI teachers take a total of 14 h that is 2 h extra for completing the lesson meanwhile TVI teachers completed the lesson within 9 h. Two assessments were also conducted in both groups; one at the end of the lecture session and the other at the end of the semester to compare the learning outcomes of TTI and TVI groups. Figure 6 shows the scheme of the examinations.

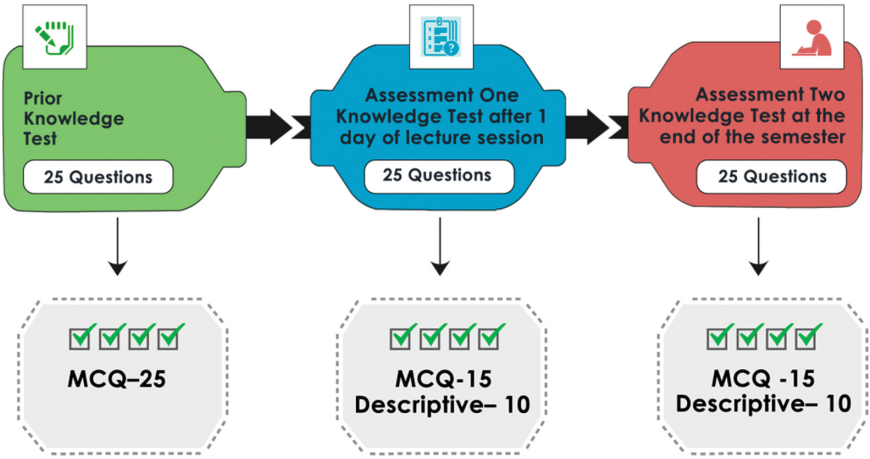


Fig. 6. Scheme of student assessment

Assessment 1 and 2 consisted of four levels of questions based on Bloom's taxonomy of the cognitive domain which includes knowledge, conceptual understanding, application and analysis [18]. For each assessment, that had a summative of 25 questions for assessing each domain (knowledge: question 1–7; conceptual understanding: question 8–13; application: question 14–19; and analysis: question 20–25). Cronbach's Alpha were calculated for measuring the internal consistency for prior knowledge test ($\alpha = 0.803$), assessment 1 ($\alpha = 0.819$) and assessment 2 ($\alpha = 0.854$). For comparison of the scores of TTI and TVI groups, a parametric t-test was used. The mean, standard deviation, t-values and p-values were calculated for each test. The 5% level of significance (p-value < 0.05) was used to denote the statistical differences [17].

4 Results and Discussion

4.1 Preferred Learning Modality Channel

A survey was conducted among the sample groups ($N = 524$) for knowing their preferred learning modality, whether they are a visual, auditory or kinesthetic learner. Students' response from the survey conducted in four HEI's is shown in Fig. 7. The result shows that 62% ($N = 320$) of the total samples were visual learners, 27% ($N = 141$) were kinesthetic learners and 11% ($N = 62$) were auditory learners. Students were distributed into two groups, TTI and TVI based on this survey result. Grouping was done in such a way that both groups had equal number of students with same modalities. That is, both

TTI and TVI groups have 160 visual learners, 72 kinesthetic learners and 30 auditory learners.

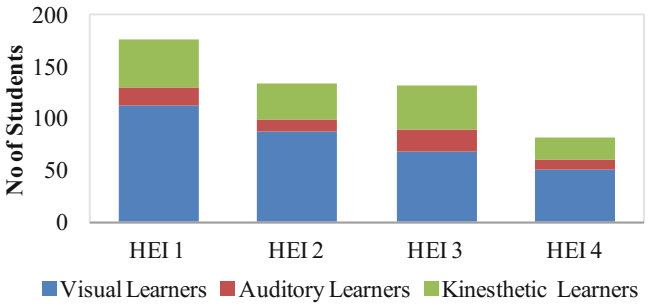


Fig. 7. Students' preferred learning modality

4.2 Prior Knowledge Test

To gauge the prior knowledge of students in crystal field theory, an examination was conducted to assess both i.e. the TTI and TVI groups. Students score in prior knowledge test is shown in Fig. 8. The results indicate that the students score in the prior knowledge test is not significantly different between the two groups. An independent two sample t-test was also conducted to analyze the differences between the scores of TTI and TVI to understand the impact of the prior knowledge of students in these groups. The t-test results shown in Table 3 verified no statistically differences in the groups amongst all institutions. This implies both group of students have similar levels of understanding of the basic concepts of CFT.

Table 3. Result of the t-test analysis TTI and TVI groups in prior knowledge test

Prior knowledge test	Group	N	Mean score	Standard deviation	T	P
HEI 1	TTI	88	44.77	26.03	0.55	0.57
	TVI	88	42.65	24.76		
HEI 2	TTI	67	42.74	21.92	0.25	0.80
	TVI	67	43.70	21.99		
HEI 3	TTI	66	44.46	22.19	0.71	0.47
	TVI	66	47.16	21.38		
HEI 4	TTI	41	37.82	23.84	0.19	0.84
	TVI	41	38.90	24.87		

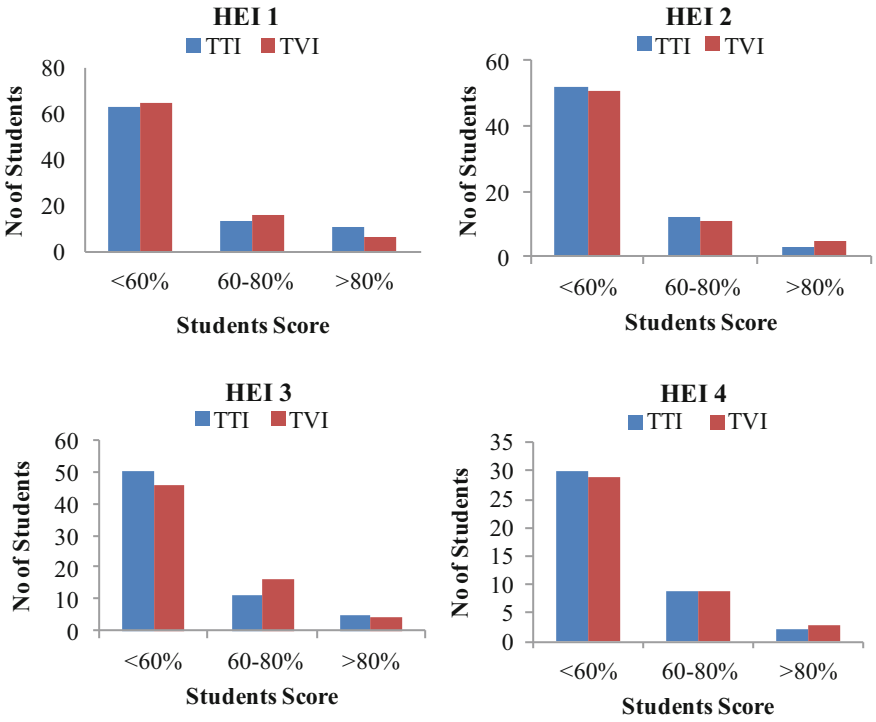


Fig. 8. Students' score in prior knowledge test

4.3 Assessment 1 - Knowledge Test After Lesson

In order to study the effectiveness of TTI and TVI instruction style, students in both student groups were assessed through a knowledge test at the end of the lecture session. Both student groups were provided with same set of questions which were intended to assess their knowledge, conceptual understanding and applications of CFT. Students' score from the assessment are shown in Fig. 9. Result shows that students that went through TTI instruction mode had secured a higher average score than students that went through TVI instruction mode amongst all institutions.

Table 4 gives the result of the t-test analysis which showed statistically a significant difference in the scores of TTI and TVI group ($P < 0.00$). Findings from this study suggest that the experimental group that had both text book learning and VL oriented instruction did better in comparison with those that had undergone only text book learning without the use of any modalities.

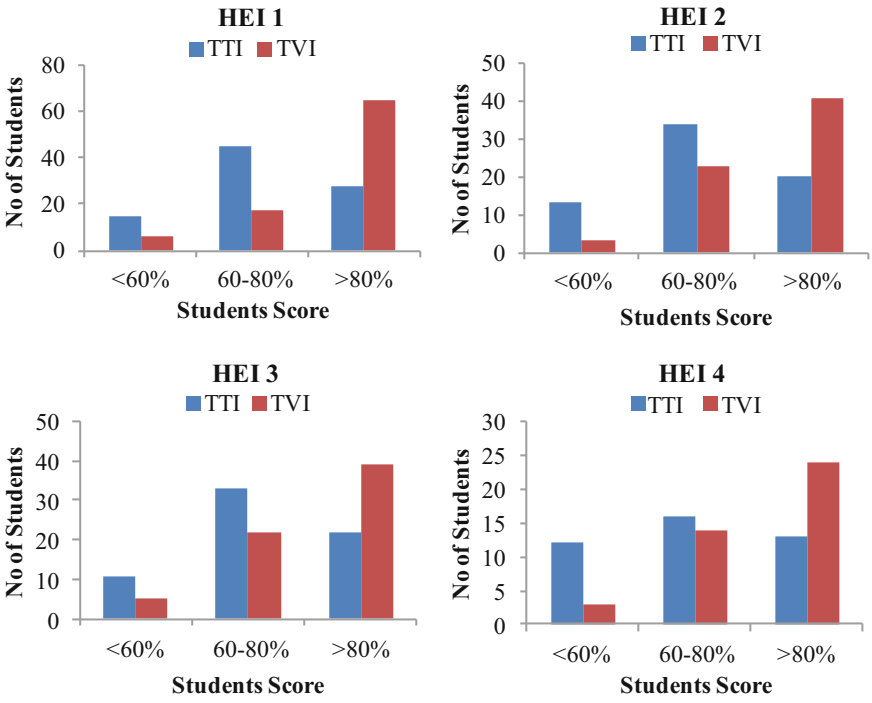


Fig. 9. Students' score in Assessment 1

Table 4. Result of the t-test analysis TTI and TVI groups in Assessment 1

Assessment 1	Group	N	Mean score	Standard deviation	T	P
HEI 1	TTI	88	69.17	20.98	5.325	0.0001
	TVI	88	83.96	15.46		
HEI 2	TTI	67	68.74	21.72	3.904	0.0001
	TVI	67	81.35	15.06		
HEI 3	TTI	66	70.98	20.64	2.893	0.0040
	TVI	66	80.09	15.07		
HEI 4	TTI	41	63.65	25.52	3.705	0.0001
	TVI	41	81.29	16.63		

4.4 Assessment 2 - Knowledge Test at the End of the Semester

In order to compare the learning outcomes of text book oriented instructional style (TTI style) and text book in combination with virtual lab oriented instructional style (TVI style), students in both groups were assessed through a knowledge test at the end of the semester. Results from the assessment are shown in Fig. 10.

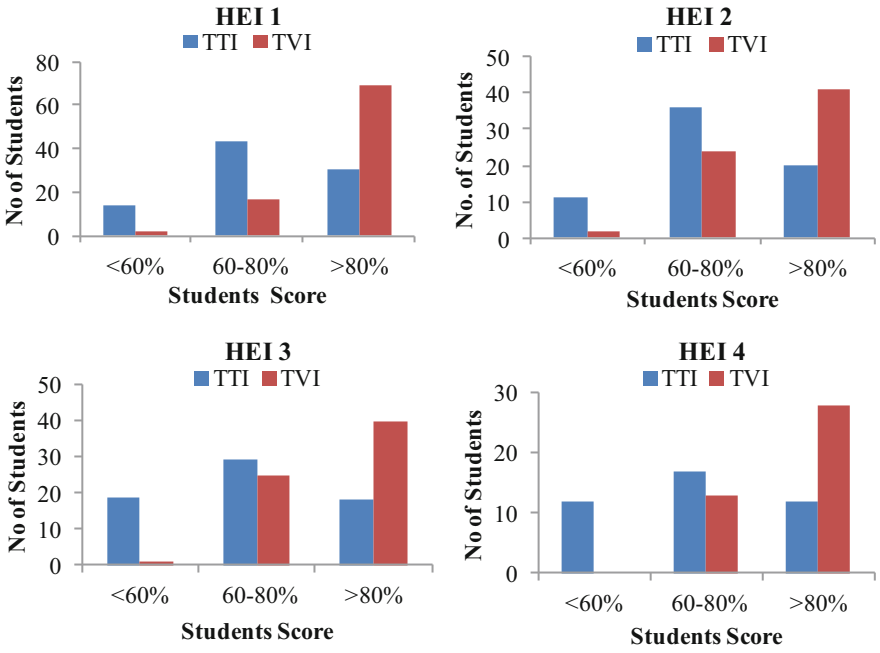


Fig. 10. Students' scores in Assessment 2

Students that went through TTI instructional method had secured a higher average score than students that went through TVI instructional method amongst all institutions. Table 5 shows the result of the t-test analysis which indicates statistically a significant difference in the scores of TTI and TVI groups (P = 0.000) amongst all institutions.

Table 5. Result of the t-test analysis TTI and TVI groups in Assessment 2

Assessment 2	Group	N	Mean score	Standard deviation	T	P
HEI 1	TTI	88	70.38	20.98	5.849	0.0001
	TVI	88	85.68	12.91		
HEI 2	TTI	67	72.44	16.89	3.535	0.00001
	TVI	67	81.88	13.83		
HEI 3	TTI	66	67.51	20.79	4.603	0.00003
	TVI	66	81.65	13.78		
HEI 4	TTI	41	62.41	25.54	4.409	0.00010
	TVI	41	82.75	14.82		

From the results, it is clear that TVI style of instruction has an advantage over TTI style of instruction. The results also indicate the impact of the blended modes of learning that includes use of images, animations, simulations and videos in virtual lab which enabled the TVI group of students for the better understanding of the chemical phenomena in-depth. From Figs. 7, 8, 9 and 10, it is clear that the impact of learning

modalities in virtual lab enhances the learning experience of visual, auditory and kinesthetic learner simultaneously. This was reflected in the performance of TVI group in assessment 1 and 2 in comparison with the TTI group. Out of 524 students, 262 students (TVI) were actively used the learning modalities in virtual lab for learning. Figure 11 shows the variation in the cumulative learning outcome of visual, auditory, and kinesthetic learners in TTI and TVI group.

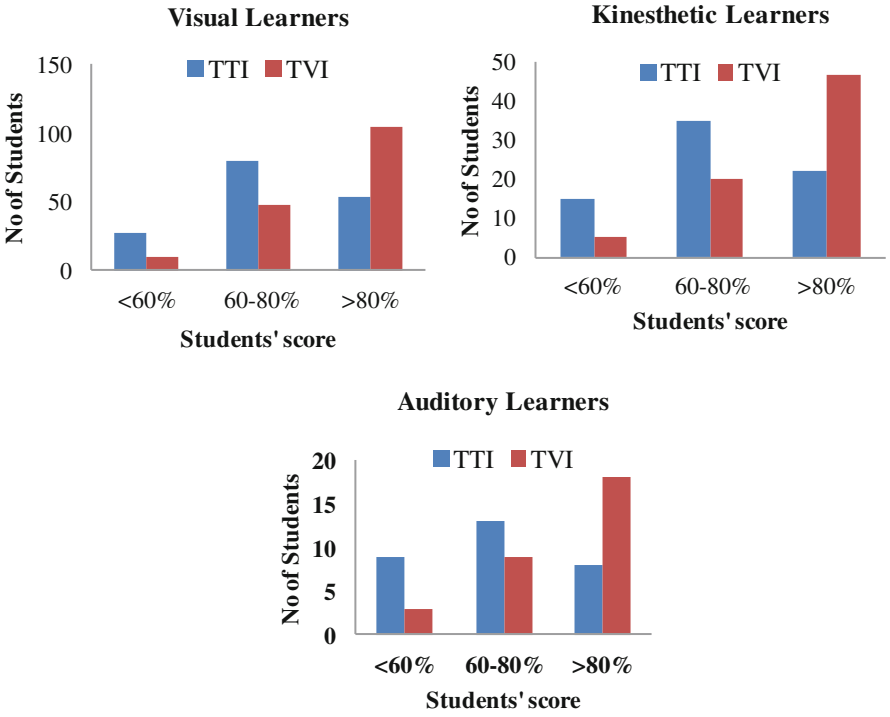


Fig. 11. Cumulative Comparison of TTI & TVI based on learning Modality

Findings from this study pointed out that the text book oriented instructional style (TTI) favors mostly the auditory learners while text book in combination with virtual lab oriented instruction (TVI) favors all the three type of learners – visual, auditory and kinesthetic irrespective of their learning modality channel. By augmenting virtual lab in teaching, any teachers or educators can provide a resourceful platform for the students where all the learning modalities were under a single learning channel. In other words, we can say that virtual lab act as a multimodality learning channel.

5 Conclusion

The impact of different learning modalities available in virtual lab in enhancing students' conceptual understanding was examined amongst undergraduate students learning

chemistry. Three tests were conducted in both groups. Result obtained in the prior knowledge test suggested that both student groups have almost same knowledge in CFT. But after instruction in two different styles i.e. blended learning with practical hands-on instruction using the virtual lab versus text book only instruction, a significant difference were found between two groups. Result of the later two tests shows that students in virtual lab with text book oriented instruction scored substantially higher marks than text book oriented instruction in terms of knowledge, conceptual understanding, application and analysis. Also virtual lab blended pedagogic approach was found to be more time saving than traditional pedagogy. This paper recommends the use of learning modalities in teaching as well as learning. Findings from our study show the effectiveness of virtual lab oriented instruction which helps students' to perceive complex theory aspect in an interesting and interactive way. Virtual lab serves as a multi-modality learning channel in which all the important learning modalities like images, animations, videos and simulator apt for a visual, auditory and kinesthetic learner under a single platform.

Acknowledgments. Our work derives direction and ideas from the Chancellor of Amrita University, Sri Mata Amritanandamayi Devi. The authors would like to acknowledge the contributions of faculty and staff at Amrita University whose feedback and guidance was invaluable.

References

1. Felder, R.M.: Learning and teaching styles in engineering education. Engr. Educ. **78**, 674–681 (1988)
2. Moreno, R., Mayer, R.E.: Cognitive principles of multimedia learning: the role of modality and contiguity. J. Educ. Philos. **91**(2), 358–368 (1999)
3. Baker, D.R.: A summary of research in science education. Sci. Educ. **75**(Pt. I), 288–296 (1991)
4. Hartley, J.R.: Learning from computer based learning in science. Stud. Sci. Educ. **15**(1), 55–76 (1988)
5. Eckhoff, E.C., Eller, V.M., Watkins, S.E., Hall, R.H.: Interactive virtual laboratory for experience with a smart bridge test. In: American Society for Engineering Education Annual Conference & Exposition (2002)
6. Herga, N.R.: Virtual laboratory in the role of dynamic visualisation for better understanding of chemistry in primary school. Eurasia J. Math. Sci. Technol. Educ. **12**(3), 593–608 (2016)
7. Medhi, O.K., Huheey, J.E., Keiter, E.A., Keiter, R.L.: Crystal field theory. In: Inorganic Chemistry: Principles of Structure and Reactivity, 4th edn., p. 428. Pearson Education India (2016)
8. Lee, J.D.: Crystal field theory. In: Concise Inorganic Chemistry, 5th edn., p. 202. Wiley (2008)
9. Falvo, D.A.: Animations and simulations for teaching and learning molecular chemistry. Int. J. Technol. Teach. Learn. **4**, 68–77 (2008)
10. Prof, A., Tatli, Z.: Virtual Chemistry laboratory: effect of constructivist learning environment. Turkish Online J. Distance Educ. **13**, 183–199 (2012)
11. Pyatt, K., Sims, R.: Learner performance and attitudes in traditional vs simulated laboratory experiences. In: ascilite 2007 (2007)
12. Raman, R., Achuthan, K., Nedungadi, P., Diwakar, S., Bose, R.: The VLAB OER experience: modeling potential-adopter student acceptance. IEEE Trans. Educ. **57**(4), 235–241 (2014)

13. Raman, R., Nedungadi, P., Achuthan, K., Diwakar, S.: Integrating collaboration and accessibility for deploying virtual labs using vlcap. Int. Trans. J. Eng. Manag. Appl. Sci. Technol. **2**(5), 547–560 (2011)
14. Achuthan, K., Murali, S.S.: A comparative study of educational laboratories from cost & learning effectiveness perspective. In: Silhavy, R., Senkerik, R., Oplatkova, Z., Prokopova, Z., Silhavy, P. (eds.) Software Engineering in Intelligent Systems. AISC, vol. 349, pp. 143–153. Springer, Cham (2015)
15. Diwakar, S., Kumar, D., Radhamani, R., Sasidharakurup, H., Nizar, N., Achuthan, K., …, Nair, B.: Complementing education via virtual labs: implementation and deployment of remote laboratories and usage analysis in south indian villages. Int. J. Online Eng. (iJOE) **12**(03), 8–15 (2016)
16. Murali, S.S., Achuthan, K., Diwakar, S.: Comparative study of laboratory education in disparate institutes of India. In: International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT), pp. 3678–3683. IEEE, March 2016
17. Oehlert, G.W.: A First Course in Design and Analysis of Experiments. The American Statistician, vol. 1 (2003). doi:10.1198/tas.2003.s210
18. Bloom, B.S., Engelhart, M.D., Furst, E.J., Hill, W.H., Krathwohl, D.R.: Taxonomy of Educational Objectives: The Classification of Educational Goals. Handbook I: Cognitive Domain. David McKay Company, New York (1956)

LDPC Binary Vectors Coding Enhances Transmissions and Memories Reliability

Tomas Knot(✉) and Karel Vlcek

Faculty of Applied Informatics, Tomas Bata University in Zlin, Nad Stranemi 4511, Zlín, Czech Republic
{knot,vlcek}@fai.utb.cz

Abstract. The paper interests in the research and implementation of memory coded information by modulation with highly effective concatenated codes, represented by LDPC (Low Density Parity Check) codes. Parameters optimization of coding is solved with respect to its implementation by semicustom integrated circuit of gate array and highly effective ARM processor created as SoC (System on Chip). Vendors offer a lot of types programmable circuits and software environments for this technique now. Basic modelling technique is model creation and simulation special architectures described by modeling in C/C++, and SystemC languages. The basic idea – the instruction set extension of the ARM processor – is realized by freely programmable gates as special "data flow" controlled execution unit and additional instruction decoder.

Keywords: System design · VHDL · Verilog · SystemC – AMS · Register transfer language · MEMS · SoC · HW/SW Co-design · 3D design · IP · OOP · GALS

1 Introduction

This paper encapsulates the particular technological trends that are specifically oriented on integrated circuits in mobile communication systems – coded digital modulation.

The semiconductor component integration has reached a level that allows the integration of system's ability with special functions in such a component. These requirements are accompanied by demands of RF subsystems, as well as other functions oriented not only on the comfortable processing of incoming messages but also the generation of automatic responses to the communication network. The complexity of these requirements increases – and especially in cases where economic aspects are included these issues.

Modern gate arrays are characterized by their large number of freely programmable gates. The gates can be embedded into the application in such a way so as to be able to fulfil specific functions in systems. The chips are also equipped with very powerful embedded processors, with internal architectures capable of performing the required computational algorithm fast enough. This alternative to the purposeful interconnection of system components are very important because it allows to use part of the application code (user programs) which have been developed and have proved themselves in many

R. Silhavy et al. (eds.), *Cybernetics and Mathematics Applications in Intelligent Systems*, Advances in Intelligent Systems and Computing 574, DOI 10.1007/978-3-319-57264-2_43

previous application tasks. The suggested procedure also allows to accelerate the application design process into its final form as components with specified functions and functionality targeted on covering customer requirements; thereby creating a customer-specific purpose processor.

The method for further increasing the computing power of embedded processors is described in the title of this text. It is an extension of the set of instructions for the optimized sequence of operations, which is a carefully considered way linked to the output signals of the additional decoder instructions.

2 The Implementation of Special Functions

When creating a purpose-oriented system for using in digital communications, it should be borne in mind that the embedded processor will be subjected to high requirements for calculation accuracy as well as the speed of the algorithms. For these reasons, the aim is to optimize processing power (using ARM processors).

The purpose is to have both hardware (HW) and corresponding software (SW) resources available, which will act upon the core computing system of the given operating system. Effective design methods are the concurrent use of co-design (HW/SW Co-Design).

A further requirement is that it will be possible to convert the application programme into a machine-compilation form. The activities are required and they are grouped into sub-units. It allows to create a special instruction whose design and optimization is reflected on the one hand in the arrangement of powerful computational units – and on the other on the functions requisite for decoding the specific instruction.

The requisite special instructions´ aim is to increase the efficiency of operations by used iterative decoding LDPC codes, (Low-Density Parity-Check Codes). A security coding is assured by LDPC codes by R. Gallager [5]. Nevertheless, the efficiency of coded modulation with LDPC codes enables one to achieve data rates, which – according to Shannon's Inequality Law [6], are very close to the theoretical boundaries of fault-free data transmissions:

$$\frac{E_b}{N_0} > \frac{2^{\frac{R}{B}} - 1}{\frac{R}{B}} = \frac{2^{\eta} - 1}{\eta} \tag{1}$$

where (1) describes the threshold, which must be respected in order to achieve spectral efficiency given the energy efficiency of the modulation of the message source. The left-hand side of the inequality is called its "energy efficiency"; while the right-hand side contains the term known as spectral efficiency.

This crucial inequality is best presented in a monograph [7]. Coded Modulation improves spectral efficiency by the so-called Code Gain Value. The use and implementation of LDPC codes as a part of a system with coded modulation, whose role and actual advantages reside in the use of an ARM processor with freely programmable gates – which are used to make an additional decoder using specially-designed instructions.

This method for modifying messages is the way to achieve the above-mentioned coding gains. This has to do with the application of the second sentence of Shannon´s Law; by adding redundancy to the message, this reduces the probability of errors in the transmission message.

2.1 Resolutions Aided by Special Instructions

It is necessary to use appropriate modelling techniques in order to describe an additional instruction decoder. It is appropriate for such purposes to choose a high abstraction description level that encapsulates the implement algorithms without any preferred method of implementation. The SystemC language is the most appropriate for this description level. The abilities of this language – that is to say, (C++, complemented by the RTL (Register Transfer Level) functionality level; as well as the possibility of TLM (Transaction Level Modelling) descriptions), not only meet the implementation requirements of algorithms as logical networks, and a programme routine implemented as an application program in the computational core of the relevant ARM processor. The embedding of special functionalities is based upon the existing schemata of ARM processor computing cores. An example of such a computing core is a controlled data-stream processor. The model is described in the dissertation [1] but also earlier [9].

The LDPC code security properties are very suitable for modern digital communications since they enable the setting of messaging mode conditions into a regime, which thus approaches the theoretical limit guaranteeing the assumption of the faultless transmission of all symbols in the message. The transmission conditions´ definition is also accompanied by properties that are a drawback – namely, the very extensive length of code words in the message. This is the motivation - and the real reason, for the introduction of special instructions into an ARM processor. In the study [4] can be found a breakdown in four different ways of decoding LDPC codes which is based on a theoretical analysis [4].

Figure 1 shows the overall framework solution that was the subject of analyses on the level of technologies like multi-chip solutions [9], existing at that time; which are now feasible on a single chip in the form of a processor with multiple cores [1], and where the computing cores are connected in a cascade [9].

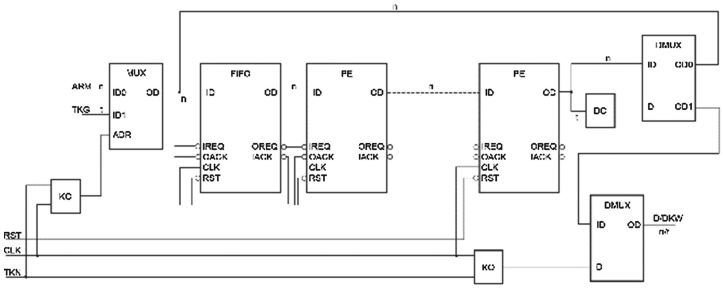


Fig. 1. Schema of a data-flow processor [1, 9]

The computational cores integrated into the cascades lead to the derivation of the solutions, which were used in the course of the formulation – and subsequently, in the simulation of problems in instruction design issues which supplement the instruction set of the ARM processor for matrix operations.

This solution offers the use of Autonomous Control Data-flow Principles, analyzed in detail for matrix algorithms for processing images [9]. In the course of resolving the decoding of LDPC codes, the situation is easier in view of the unchanged format of vectors; as well as more being more demanding with regard to the length of the processed vectors which in some cases exceed more than one hundred bits. The special arrangement of the units are needed for necessary solutions, particularly for the execution speed of decoding operations.

2.2 The Matrix Core Algorithm

Core matrix algorithms, which work with the respective registers, were selected as the new instructions for the process of decoding LDPC codes. These registers are made up of freely programmable gates on an ARM processor chip. Programmable gates have the character of "Coarse Grain" programmable logic devices. The gates are made up of logical units which they acquire the function of the appropriate logic gate during the programming process and its connection to the outputs and inputs of the surrounding logical network. The designer does not need to closely monitor this activity since the description of the higher programming language – SystemC, which assures the compilation process. The relevant decoder functions are assigned to the new instructions.

The complexity of the new instructions has a great influence on the efficiency of the calculations of each decoding procedure. In order to encourage the character of an ARM processor is better to use a Reduced Instruction Set (RISC) due to simple instructions in the design process.

This means that the new instruction design process will be submitted to a detailed simulation - and subsequently, whichever draft version that proves to be the better one, will be selected to extend the added instruction directly to the execution unit, which is also added. The proposed solution is determined the optimum between the use of existing instructions linked into a purposeful sequence, which also includes the management of the requisite registers needed to increase the parallelism of the requisite actions.

The HW/SW Co-design Method is the best method to use for this purpose. The actual process, which is utilized, is described for instance in [10]. Even in this case, it is possible to use tools, like Architecture Analysis and the (AADL) tool for the design of the language. The AADL tool is, in case of need, for the optimization of the design process.

2.3 Decoding Variants

When designing the properties of the additional instructions, it is advantageous to retain the setting of possible variant algorithms for encoding and decoding the LDPC codes. The suitability of the selection of these algorithms is also given by the type of digital modulation of the symbols in the reports. These parameters are determined by the application so it is necessary to respect the requirements of customers, while also providing

other effective functional verification tools for the control process of the mutual analogue and digital circuits [11] – including the ARM processors with the customers´ instructions and the RF circuitry block – with which the customer instructions cooperate. The preparation of code-words using the LDPC code can be performed by means of various methods but always has to do with the product of the operation between the information vector and LDPC matrix generating code. This operation is entrusted to the add-on technical unit composed of ALU cascades, whose activity is regulated by the data. The choice of this ALU control principle is discussed in detail and proposed in [9]. When designing customer circuits that expand the instruction set, it is adapted to matrix operations with the length of the LDPC code vectors.

Four algorithms used to decode LDPC codes so as to establish the activities carried out by the operations that extend the instruction set of the ARM processor were examined. The first of these, called the "MacKay Neal Algorithm" can be adapted to avoid cycles with a length of 4 in the algorithm; it is also possible to use a corresponding number of logical units connected in cascades.

The following diagrams, which are presented in [3], show the high efficiency of additional "coding gain" of achieving the spectral efficiency for given value of the energy efficiency. A coding gain is the measure in the difference between the Signal to Noise Ratio (SNR) levels between the non-coded system and coded system required to reach the same Bit Error Rate (BER) levels when used with the Error Correcting Code (ECC). For example, when the non-coded system with noise environment has a Bit Error Rate (BER) of 10^{-2} at the SNR level 4 dB, and the corresponding coded system has the same BER at an SNR of 2.5 dB, then we say the coding gain = 4 dB − 2.5 dB = 1.5 dB, due to used code (Fig. 2).

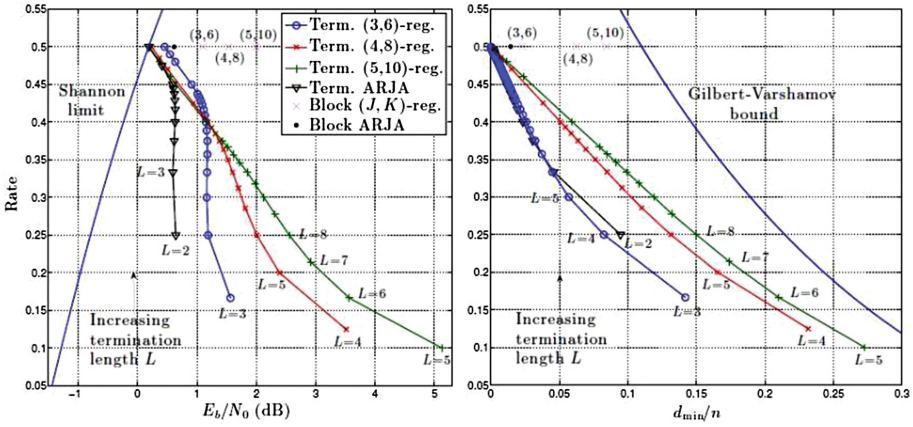


Fig. 2. Shannon limit and Gilbert Varshamov bound [3]

In Algorithm 1, the method is used to resolve four code-words with the assistance of calculations ongoing in the current four cycles. Here, it is possible to use a structure that solves the calculation in independent units in a controlled data stream

cascade. The input length of the code-words is n, the code rate is r, and the non-zero values are the values v and h: [3]

```
Algorithm 1 MacKay Neal LDPC Codes
1: procedure MN CONSTRUCTION (n, r, v, h) ◁ Required
length, rate and degree distributions
2: H = all zero n(1 ▬ r)   n matrix ◁Initialization
3:   = [];
4: for i = 1: max(v) do
5: for j = 1 : vi   n do
6:   = [ , i]
7: end for
8: end for
9: β = []
10: for i = 1 : max(h) do
11: for j = 1 : hi   m do
12: β = [β, i]
13: end for
14: end for
15:
16: for i = 1 : n do ◁ Construction
17: c = random subset of β, of size  i
18: for j = 1 :  i do
19: H(cj , i) = 1
20: end for
21:   =   ▬ c
22: end for
23:
24: repeat
25: for i = 1 : n ▬ 1 do ◁ Remove 4-cycles
26: for j = i + 1 : n do
27: if |H(:, i)   H(:, j)| > 1 then
28: permute the entries in the j-th column
29: end if
30: end for
31: end for
32: until cycles removed
33: end procedure
```

where, Vector α has the length n, and Vector β has the length m with i non-nullified elements in the rows of matrix H, weighted by i.

The class of algorithms used for decoding LDPC codes is can be called “algorithms solved while processing/transmitting messages”. It is so named because it can be described by the so-called Tanner Graph. The decoding algorithm described as the processing/transmission of messages is also known as Iterative Decoding.

The algorithm evaluates the message bits in both the forward and backward direction, until it attains the smallest deviation criteria – or, until this process is not stopped by other means.

The first step calculates the control nodes of the first bits of the message; the second step performs the same calculation shifted by one bit - while using in the binary incoming symbol of the message. This process continues, according to the values of control bits, until the entire vector is evaluated.

2.4 Decoding Algorithms´ Effectiveness

Which of the decoding methods mentioned above will prove to be most appropriate, is given by the application requirements, demands on speed, and decoding accuracy. The structure of processor elements linked in cascades was chosen as a supporting tool for implementing unit operations - thereby increasing the efficiency of the calculations required to encode and decode LDPC codes. There are a few reasons for this structural linking: when arranging the computing cores´ cascades, it is essential to handle multiple computing cores for their effective cooperation in the source code language. When performing concurrent operations, primitives defining parallel processing are not necessary, since, when it detects a concurrently running program, written in a sequential language, the compiler can generate specific implementation code for the system automatically.

The source program language is selected independently of the structure itself; finished programs can be transferred directly to a multiple core system – but, their concurrent operation is entrusted to the data flow, which includes the control "token". The length of the vector increases, despite this however, it can still be tuned by the user programs to the usual sequential computer.

The method called MacKay Neal LDPC Codes (Algorithm 1) is therefore the fastest, since more vectors can be evaluated. The method operates such that, during the decoding process the value of binary symbols is allocated for the four vectors concurrently. This may increase the probability of errors for certain values of symbol sequences in the message.

For Algorithm 1 [3], it is therefore advantageous to create an execution unit with multiple cores, which are arranged in cascades and controlled by means of the "data-flow" – that is to say, by using so-called "Tags"; these are automatically generated during source code compilation for the selected length of the cascade. Only the parameter that determines the length of the cascades of the implementing core units can be changed. The implementation method is a new method, which is useful as a basis for standardized applications that can be found on the references page [14].

3 Analogue Signals and RF Signals

FPGA applications in communications are caused by the need to increase the reliability of message transmission services. Apart from digital signals, analogue signals are also

necessary. The specific requirements in designing a SoC with these desired properties however, must always be treated if they are mixed signal systems.

In particular, it is essential to pay attention to the crystal oscillators on the chip - called PLL circuits, but equally - to all of the oscillators that serve as voltage-controlled sources of synchronization, not only for the control of serial - parallel and parallel - serial converters, but also all circuits for the renewal of synchronization sources´ signals.

Signals designated with the abbreviation RF (Radio - Frequency) are, in the design process, always treated preferentially. This is essential because high frequencies are also propagated in the SoC environment in a manner that is different from their propagation along metallic wiring. With regard to the microscopic structure on an integrated circuit chip, the method of distributing RF signals can rather be compared to a free-space environment in which however, much more frequently, the situation occurs that the RF signals become interference.

4 Implementation Base Specifications

A common problem of hybrid circuits is the differential comparators inputs on the SoC. They are designed to connect to the channel interface, known as the "differential signal". Some of the gate arrays designated as "Mixed Signal" FPGAs have analogue-to-digital (ADC) integrated peripheries and a digital-analogue converter (DAC) with analogue-signal treatment abilities which allow it to be considered as being suitable – from the construction point-of-view, for SoC implementation designs.

However, such components do not respect all of the differences between a numerical FPGA, which presents reality in a digital form, expressed by ones and zeros; and the internal connections - termed a "Field-Programmable Analogue Array", (FPAA), which uses analogue values of the variables and these are represented by the analogue expression of programmable system signals. For an FPAA however, the conditions used for simulation and verification purposes are much more complicated. The systems control whether Ohm's and Kirchoff's Laws - and the Law of Conservation of Energy are all satisfied. In such an environment, it is essential to respect all of these aspects above.

5 Further Development Work

Other work carried out in the framework of this task is oriented on the aims set by the 5G Access Networks project. The closest "work-in-progress" task is the implementation of so-called "green" applications. The diploma thesis [15] dealt with the principles of turbo-codes, but also provided useful tools for the calculation in the MATLAB environment.

This work [15] is devoted to the decoding of LDPC codes and compares four decoding methods: not only the Hard-Decision Algorithm and the Bit-Flipping Algorithm.

All of these procedures are, thanks to their high integration level, efficient in use and contribute to reliability of demanding applications – like for instance, driving cars without drivers. Since, for such applications, the principal activity is simply data

transmission. Based on these comparisons, a study has been elaborated that aims to assess the increased efficiency of coded digital modulation when using LDPC codes. This method is used to compare the size of the parity matrix of the LDPC codes, the requisite number of iterations of decoding, and the coding gain.

The effectiveness of appropriate modulation coding methods is evaluated depending upon its size. Subsequently, the demands upon their technical aspects are evaluated - with of course, regard to the possibility of the proven method being extended to an ARM processor´s instruction set. Other algorithms will be studied in further work.

6 Conclusion

Based on coded modulation, the implementation of the highly-reliable intensive data communications methods has a very wide range of applications for modern transmission resources.

These are linked with basic "decoding during message transmissions" and the specific processing of signal transmissions – not unlike the filtering of digital algorithms. These demands are resolvable, even during the course of real-time processing requirements, or with specialized computing units - whose control is given in line with the proposed SoC, which is implemented by extending the ARM processor instruction set. This ARM processor also has a specialized unit for the calculations required for a range of advanced digital modulation, which is controlled by the file extension instructions.

SoC design can be challenging. The cooperation of experts experienced in design solutions and the resolution of access to diagnostic data and the designing of tests for hybrid circuits in integrated circuit technology, including modelling disorders; as well as in the design of 3D circuits when interconnecting the circuit chips and the system.

The designer encounters the problems and issues relating to the connection of signals between co-functioning chips using the so-called "Connectors Through Silicon" - TSV (Through Silicon Via), which it is also necessary to include in signal path design optimization, despite the fact that it is more reliable, and manufacturing is more manageable using multiple chips in one package, than the implementation of a system with MEMS or RF components on a single chip [17]. In contrast, the integration of a single-chip system - with AMS (Analogue Mixed Signal) characteristics is exclusively due to their reliability and implementation costs; therefore, implementation on a common chip is preferable. However, it is a question of chip manufacturing economics and the support of their design for the given field of applications. With regard to the 5G project, this path is very promising while also encouraging the development of wireless communication technology.

Acknowledgments. The Author wishes to express their thanks for support from the Cebia – Tech Research Centre of Czech Republic, Project No.: CZ.1.05/2.1.00/03.0089.

References

1. Mitesh, L., Kothari, N.: Modeling and Simulation of ARM Processor Architecture Using SystemC. Lambert Acad. Publishing (2012). ISBN 978-3-659-12088-6J. Maxwell, C.: A Treatise on Electricity and Magnetism, 3rd edn., vol. 2. Oxford: Clarendon, pp. 68–73 (1892)
2. Franceschini, M., Ferrari, G., Raheli, R.: LDPC Coded Modulations. Springer (2009). ISBN 978-3-540-69455-7K. Elissa, Title of paper if known (unpublished)
3. Johnson, S.J.: Introducing Low-Density Parity-Check Codes, p. s.83 (2008). http://sigpromu.org/sarah/SJohnsonLDPCintro.pdf. Accessed 19 Jan 2017
4. Mackay, D., Neal, R.M.: Near shannon limit performance of low-density parity-check codes. IEE Electron. Lett. **32**(18), 1645–1646 (1996)
5. Gallager, R.G.: Low-density parity-check codes. In: MIT Press, Cambridge, p. s.90. http://www.rle.mit.edu/rgallager/documents/ldpc.pdf. Accessed 19 Jan 2017
6. Shannon, C., E.: A mathematical theory of communication. Bell System Tech. J. **27**, 379–423 (1948). doi:10.1002/j.1538-7305.1948.tb01338.x
7. Dobeš, J., Žalud, V.: Moderní radiotechnika. Praha: BEN - technická literatura, pp. 98–99 (2006). ISBN 80-730-0132-2
8. Vlček, K.: Multimedia Processing and Mobile Communication. http://www.utb.cz/file/54551_1_1/. Accessed 19 Jan 2017
9. Vlček, K.: Obrazový víceprocesorový systém, Disertační práce, VÚMS, ČVUT Praha (1988) (In Czech)
10. Extending ASSERT to HW/SW Co-design Approach. http://hwswcodesign.gmv.com/resources.htm. Accessed 19 Jan 2017
11. Vroom, W.: Komplexní funkční ověření smíšených obvodů s QUESTA ADMS, DPS, No. 4, pp. 26–28 (2016) (In Czech), ISSN 1805-5044
12. Berrou, C., Glavieux, A., Thitimajshima, P.: Near shannon limit error-correcting coding: turbo codes. In: Proceedings of the IEEE International Conference Communication, Geneva, Switzerland, pp. 1064–1070 (1993)
13. Benedetto, S., Divsalar, D., Montorsi, G., Pollara, F.: Serial concatenation of interleaved codes: performance analysis, design, and iterative decoding, pp. s.42–126 doi:10.1.1.30.4514. Accessed 19 Jan 2017
14. Efficient implementation of iterative receiver components. Rob Maunder's research pages Wireless communications. http://users.ecs.soton.ac.uk/rm/resources/implementation/. Accessed 19 Jan 2017
15. Kettner, J., Kódování – Turbo kódy. Zlín (2010). http://theses.cz/id/80kg4w/. Diploma Thesis. Univerzita Tomáše Bati ve Zlíně, Fakulta aplikované informatiky. Supervisor RNDr. Ing. Miloš Krčmář. Accessed 19 Jan 2017
16. Hrouza, O.: LDPC kódy: Diploma Thesis. Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií (2012). 80 s. Supervisor Ing. Pavel Šilhavý, Ph.D.
17. Vlček, K.: SystemC – nástroje a prostředí pro návrh systémů na čipech moderních rozsáhlých hradlových polí a polí se smíšenými signály In Czech. http://www.utb.cz/file/44257_1_1/. Accessed 19 Jan 2017
18. Vlček, K., Voráč, J., Mitrych, J.: Iterative ECC decoder with sparse matrices solution. In: 7th IEEE Workshop, DDECS 2004, 18–21 April, Stará Lesná, Slovakia, pp. 203–206 (2004). ISBN 80-969117-9-1

Author Index

R. Silhavy et al. (eds.), *Cybernetics and Mathematics Applications in Intelligent Systems*,
Advances in Intelligent Systems and Computing 574, DOI 10.1007/978-3-319-57264-2