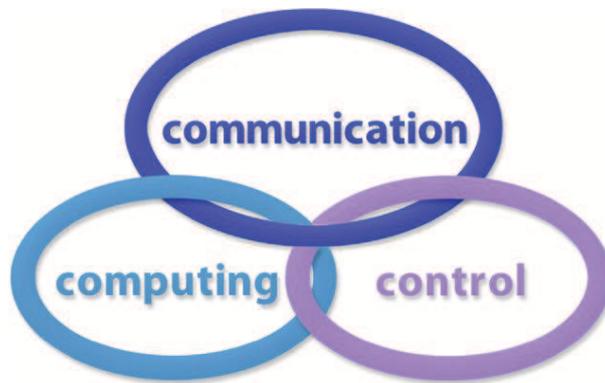


INTERNATIONAL JOURNAL
of
COMPUTERS COMMUNICATIONS & CONTROL

ISSN 1841-9836



A Bimonthly Journal
With Emphasis on the Integration of Three Technologies

Year: 2017 Volume: 12 Issue: 4 Month: August

This journal is a member of, and subscribes to the principles of, the Committee on Publication Ethics (COPE).



<http://univagora.ro/jour/index.php/ijccc/>

CCC Publications

Copyright © 2006-2017 by Agora University

BRIEF DESCRIPTION OF JOURNAL

Publication Name: International Journal of Computers Communications & Control.

Acronym: IJCCC; **Starting year of IJCCC:** 2006.

ISO: Int. J. Comput. Commun. Control; **JCR Abbrev:** INT J COMPUT COMMUN.

International Standard Serial Number: ISSN 1841-9836.

Publisher: CCC Publications - Agora University of Oradea.

Publication frequency: Bimonthly: Issue 1 (February); Issue 2 (April); Issue 3 (June); Issue 4 (August); Issue 5 (October); Issue 6 (December).

Founders of IJCCC: Ioan DZITAC, Florin Gheorghe FILIP and Misu-Jan MANOLESCU.

Indexing/Coverage:

- Since 2006, Vol. 1 (S), IJCCC is covered by Thomson Reuters/Clarivate Analytics and is indexed in ISI Web of Science/Knowledge: Science Citation Index Expanded.

2017 Journal Citation Reports® Science Edition (Thomson Reuters, 2016):

Subject Category: (1) Automation & Control Systems: Q4(2009, 2011, 2012, 2013, 2014, 2015), Q3(2010, 2016); (2) Computer Science, Information Systems: Q4(2009, 2010, 2011, 2012, 2015), Q3(2013, 2014, 2016).

Impact Factor/3 years in JCR: 0.373(2009), 0.650 (2010), 0.438(2011); 0.441(2012), 0.694(2013), 0.746(2014), 0.627(2015), **1.374(2016)**.

Impact Factor/5 years in JCR: 0.436(2012), 0.622(2013), 0.739(2014), 0.635(2015), **1.193(2016)**.

- Since 2008 IJCCC is indexed by Scopus: **CiteScore 2016 = 1.06**.

Subject Category:

(1) Computational Theory and Mathematics: Q4(2009, 2010, 2012, 2015), Q3(2011, 2013, 2014, 2016);

(2) Computer Networks and Communications: Q4(2009), Q3(2010, 2012, 2013, 2015), Q2(2011, 2014, 2016);

(3) Computer Science Applications: Q4(2009), Q3(2010, 2011, 2012, 2013, 2014, 2015, 2016).

SJR: 0.178(2009), 0.339(2010), 0.369(2011), 0.292(2012), 0.378(2013), 0.420(2014), 0.263(2015), 0.319(2016).

- Since 2007, 2(1), IJCCC is indexed in EBSCO.

Focus & Scope: International Journal of Computers Communications & Control is directed to the international communities of scientific researchers in computer and control from the universities, research units and industry.

To differentiate from other similar journals, the editorial policy of IJCCC encourages the submission of original scientific papers that focus on the integration of the 3 "C" (Computing, Communication, Control).

In particular the following topics are expected to be addressed by authors: (1) Integrated solutions in computer-based control and communications; (2) Computational intelligence methods (with particular emphasis on fuzzy logic-based methods, ANN, evolutionary computing, collective/swarm intelligence); (3) Advanced decision support systems (with particular emphasis on the usage of combined solvers and/or web technologies).

IJCCC EDITORIAL TEAM

Editor-in-Chief: Florin-Gheorghe FILIP

Member of the Romanian Academy
Romanian Academy, 125, Calea Victoriei
010071 Bucharest-1, Romania, ffilip@acad.ro

Associate Editor-in-Chief: Ioan DZITAC

Aurel Vlaicu University of Arad, Romania
St. Elena Dragoi, 2, 310330 Arad, Romania
ioan.dzitac@uav.ro

&

Agora University of Oradea, Romania
Piata Tineretului, 8, 410526 Oradea, Romania
rector@univagora.ro

Managing Editor: Mişu-Jan MANOLESCU

Agora University of Oradea, Romania
Piata Tineretului, 8, 410526 Oradea, Romania
mmj@univagora.ro

Executive Editor: Răzvan ANDONIE

Central Washington University, U.S.A.
400 East University Way, Ellensburg, WA 98926, USA
andonie@cwu.edu

Reviewing Editor: Horea OROS

University of Oradea, Romania
St. Universitatii 1, 410087, Oradea, Romania
horos@uoradea.ro

Layout Editor: Dan BENTA

Agora University of Oradea, Romania
Piata Tineretului, 8, 410526 Oradea, Romania
dan.benta@univagora.ro

Technical Secretary

Domnica Ioana DZITAC

R & D Agora, Romania
ioana@dzitac.ro

Simona DZITAC

R & D Agora, Romania
simona@dzitac.ro

Editorial Address:

Agora University/ R&D Agora Ltd. / S.C. Cercetare Dezvoltare Agora S.R.L.
Piata Tineretului 8, Oradea, jud. Bihor, Romania, Zip Code 410526
Tel./ Fax: +40 359101032

E-mail: ijccc@univagora.ro, rd.agora@univagora.ro, ccc.journal@gmail.com

Journal website: <http://univagora.ro/jour/index.php/ijccc/>

IJCCC EDITORIAL BOARD MEMBERS

Luiz F. Autran M. Gomes

Ibmec, Rio de Janeiro, Brasil
Av. Presidente Wilson, 118
autran@ibmecrj.br

Boldur E. Bărbat

Sibiu, Romania
bbarbat@gmail.com

Pierre Borne

Ecole Centrale de Lille, France
Villeneuve d'Ascq Cedex, F 59651
p.borne@ec-lille.fr

Ioan Buciu

University of Oradea
Universitatii, 1, Oradea, Romania
ibuciu@uoradea.ro

Hariton-Nicolae Costin

Faculty of Medical Bioengineering
Univ. of Medicine and Pharmacy, Iași
St. Universitatii No.16, 6600 Iași, Romania
hcostin@iit.tuiasi.ro

Petre Dini

Concordia University
Montreal, Canada
pdini@cisco.com

Antonio Di Nola

Dept. of Math. and Information Sci.
Università degli Studi di Salerno
Via Ponte Don Melillo, 84084 Fisciano, Italy
dinola@cds.unina.it

Yezid Donoso

Universidad de los Andes
Cra. 1 Este No. 19A-40
Bogota, Colombia, South America
ydonoso@uniandes.edu.co

Ömer Egecioglu

Department of Computer Science
University of California
Santa Barbara, CA 93106-5110, U.S.A.
omer@cs.ucsb.edu

Constantin Gaidric

Institute of Mathematics of
Moldavian Academy of Sciences
Kishinev, 277028, Academiei 5
Moldova, Republic of
gaidric@math.md

Xiao-Shan Gao

Acad. of Math. and System Sciences
Academia Sinica
Beijing 100080, China
xgao@mmrc.iss.ac.cn

Enrique Herrera-Viedma

University of Granada
Granada, Spain
viedma@decsai.ugr.es

Kaoru Hirota

Hirota Lab. Dept. C.I. & S.S.
Tokyo Institute of Technology
G3-49,4259 Nagatsuta, Japan
hirota@hrt.dis.titech.ac.jp

Gang Kou

School of Business Administration
SWUFE
Chengdu, 611130, China
kougang@swufe.edu.cn

George Metakides

University of Patras
Patras 26 504, Greece
george@metakides.net

Shimon Y. Nof

School of Industrial Engineering
Purdue University
Grissom Hall, West Lafayette, IN 47907
U.S.A.
nof@purdue.edu

Stephan Olariu

Department of Computer Science
Old Dominion University
Norfolk, VA 23529-0162, U.S.A.
olariu@cs.odu.edu

Gheorghe Păun

Institute of Math. of Romanian Academy
Bucharest, PO Box 1-764, Romania
gpaun@us.es

Mario de J. Pérez Jiménez

Dept. of CS and Artificial Intelligence
University of Seville, Sevilla,
Avda. Reina Mercedes s/n, 41012, Spain
marper@us.es

Dana Petcu

Computer Science Department
Western University of Timisoara
V.Parvan 4, 300223 Timisoara, Romania
petcu@info.uvt.ro

Radu Popescu-Zeletin

Fraunhofer Institute for Open
Communication Systems
Technical University Berlin, Germany
rpz@cs.tu-berlin.de

Imre J. Rudas

Óbuda University
Budapest, Hungary
rudas@bmf.hu

Yong Shi

School of Management
Chinese Academy of Sciences
Beijing 100190, China &
University of Nebraska at Omaha
Omaha, NE 68182, U.S.A.
yshi@gucas.ac.cn, yshi@unomaha.edu

Athanasios D. Styliadis

University of Kavala
Institute of Technology
65404 Kavala, Greece
styliadis@teikav.edu.gr

Gheorghe Tecuci

Learning Agents Center
George Mason University
U.S.A.
University Drive 4440, Fairfax VA
tecuci@gmu.edu

Horia-Nicolai Teodorescu

Faculty of Electronics and
Telecommunications
Technical University "Gh. Asachi" Iasi
Iasi, Bd. Carol I 11, 700506, Romania
hteodor@etc.tuiasi.ro

Dan Tufiş

Research Institute for Artificial Intelligence
of the Romanian Academy
Bucharest, "13 Septembrie" 13, 050711, Romania
tufis@racai.ro

Lotfi A. Zadeh

Director,
Berkeley Initiative in Soft Computing (BISC)
Computer Science Division
University of California Berkeley,
Berkeley, CA 94720-1776
U.S.A.
zadeh@eecs.berkeley.edu

DATA FOR SUBSCRIBERS

Supplier: Cercetare Dezvoltare Agora Srl (Research & Development Agora Ltd.)

Fiscal code: 24747462

Headquarter: Oradea, Piata Tineretului Nr.8, Bihor, Romania, Zip code 410526

Bank: BANCA COMERCIALA FERROVIARA S.A. ORADEA

Bank address: P-ta Unirii Nr. 8, Oradea, Bihor, România

IBAN Account for EURO: RO50BFER248000014038EU01

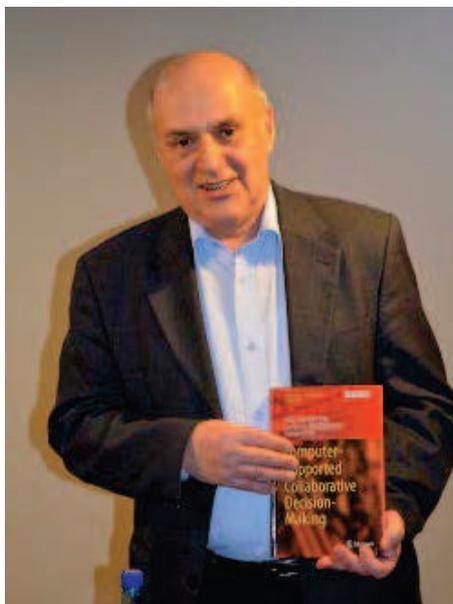
SWIFT CODE (eq.BIC): BFER

Contents

Contributions of Florin Gheorghe Filip in Information Science and Technology I. Dzitac, M.J. Manolescu	449
Obfuscation-based Malware Update: A comparison of Manual and Automated Methods C. Barriá, D. Cordero, C. Cubillos, M. Palma, D. Cabrera	461
Asymptotically Unbiased Estimation of A Nonsymmetric Dependence Measure Applied to Sensor Data Analytics and Financial Time Series A. Caçaron, R. Andonie, Y. Chueh	475
The Logistic Regression from the Viewpoint of the Factor Space Theory Q.F. Cheng, T.T. Wang, S.C. Guo, D.Y. Zhang, K. Jing, L. Feng, Z.F. Zhao, P.Z. Wang	492
An Approach for Construction of Augmented Reality Systems using Natural Markers and Mobile Sensors in Industrial Fields D. Gomes Jr., P. Reis, A. Paiva, A. Silva, G. Braz Jr., M. Gattass, A. Araújo	507
PID and Fuzzy-PID Control Model for Quadcopter Attitude with Disturbance Parameter E. Kuantama, T. Vesselenyi, S. Dzitac, R. Tarca	519
Automatic Generation Control by Hybrid Invasive Weed Optimization and Pattern Search Tuned 2-DOF PID Controller N. Manoharan, S.S. Dash, K.S. Rajesh, S. Panda	533
A Solution for Interoperability in Crisis Management F.J. Pérez, M. Zambrano, M. Esteve, C. Palau	550
Coverage Hole Recovery Algorithm Based on Molecule Model in Heterogeneous WSNs X.L. Song, Y.Z. Gong, D.H. Jin, Q.Y. Li, H.C. Jing	562
High-speed Train Control System Big Data Analysis Based on Fuzzy RDF Model and Uncertain Reasoning D. Zhang	577
Author index	592

Contributions of Florin Gheorghe Filip in Information Science and Technology

I. Dzitac, M.J. Manolescu



Acad. Florin Gheorghe Filip at 70 years
(Born on July 25, 1947, Bucharest, Romania)

Ioan Dzitac^{1,2} ; **Misu-Jan Manolescu**^{2,*}

1. Aurel Vlaicu University of Arad
310330 Arad, Elena Dragoi, 2, Romania

ioan.dzitac@uav.ro, mmj@univagora.ro

*Corresponding author: mmj@univagora.ro

2. Agora University of Oradea

410526 Oradea, P-ta Tineretului 8, Romania,

Abstract: Romanian scientist Florin Gheorghe Filip was born in 1947 and this year he turns 70. F.G. Filip is an engineer and Ph.D. in control engineering and computer science. Still very young, he became corresponding member of the Romanian Academy in 1991 (when he was only 44 years old), and, at 52 years old (1999), become full member in the highest learned society of Romania. From 1970 to 2000, he worked at the National R&D Institute in Informatics Bucharest (ICI). For 10 years, during 2000-2010, he was Vice President of the Romanian Academy. In 2010, he was elected President of the "Information Science and Technology" section of Romanian Academy. At present, he is the director of the Romanian Academy Library. His fields of scientific interest have been: decision support systems (DSS), large-scale systems control and optimization, technology management and foresight and IT application to cultural domain . He has authored/co-authored over 300 technical papers, 13 monographs, and edited/co-edited 24 contributed volumes.

Keywords: decision support systems (DSS), hierarchical large-scale systems, managerial activity, recognition, scientific publications.

1 Introduction

"To whom it may concern ...

... Dr. Filip has shown himself to be a very innovative and productive researcher whose papers are equivalent in quality and cutting-edge findings to those from the best research organizations of the world. Witness his publications in *Automatica*, *Computers in Industry* and other leading journals in the field. The establishment of Dr. Filip's stature as an internationally recognized researcher in his field and the acceptance of his work are shown by the large number of international conferences where he has served as session chairman and/or program committee member and many invited seminars he has presented in other countries ..." (T.J. Williams).

This is what Theodore J. Williams, Professor of Engineering and Director, Purdue Lab. for Applied Industrial Control, Purdue University, wrote about F.G. Filip in August 1994.

In the sequel we will highlight several aspects of Acad. Filip's professional life.

Similar works within this paper has been published in 2007, with occasion of 60 years from birth of Acad. Filip, in *The Computer Science Journal of Moldova* [1] and in the *International Journal of Computers Communications & Control* [8].

2 A short biographical note

Florin Gheorghe Filip was born on 25 of July 1947 in Bucharest, Romania. He graduated in Control Engineering from Politehnica University Bucharest (PUB) in 1970 and received Ph.D. from the same university in 1982.

In 1974, he was a guest researcher in Swedish universities (Chalmers TH in Gothenburg, Teknikum Uppsala, University of Lund, KTH, Stockholm). In 1993 and 1996 he was a visitor of German Fraunhofer institutes: FIRST Berlin and IITB Karlsruhe, respectively. In 1983, Prof. Filip was invited to deliver a 10-day course on hierarchical control systems at Institute of Automation of the Chinese Academy of Sciences in Shenyang, China. From 2006 to 2011, he spent one month each year as a *chercheur-enseignant* at École Centrale de Lille (France).

He was elected as a corresponding member of the Romanian Academy (RA) in 1991 and became full member of the RA in 1999. During 2000-2010, he was Vice-president of the Academy (elected in 2000, and re-elected in 2002 and 2006). In 2010 was elected as a president of the 14th Section "Information Science and Technology" of the RA (re-elected in 2015). Also he was elected as a honorary member of Academy of Sciences of Moldova and Romanian Academy of Technical Sciences in 2007.

Prof. Filip is one of the founding members of the International Academy for IT and Quantitative Management (IAITQM), founded in 2012 at University of Nebraska in Omaha, USA. He was President of IFAC TC 5.4 SML - Large Scale Systems (2002-2008) and a member of two other IFAC technical committees: TC 4.5 (HMS), TC 5.2 (MMC) and one IFIP WG 5.12 (Enterprise Integration).

Prof Filip was the managing director of ICI- National Institute for R&D in Informatics (1991-1997), and the president of Scientific Council of ICI (1995-2003). Presently, he is the director of the Romanian Academy Library.

F.G. Filip gave master courses on computer applications, expert systems, computer supported decision-making, and enterprise engineering at PUB and other Romanian universities such as Agora University Oradea, University of Bucharest, Lucian Blaga University of Sibiu, Valahia University of Targoviste, Academy of Economic Studies (ASE) Bucharest.

He has supervised doctoral studies in PUB and Institute for Artificial Intelligence (ICIA) of the Romanian Academy. Four of his former PhD students (A. Alexandru, M. Cioca, L. Duță,

and C.B Zamfirescu) are at present professors in various universities and supervise their PhD students, and C. Căndeia is the leader of ROPARDO, a successful software company.

F. G. Filip was an adjunct professor at the University of the Chinese Academy of Sciences (2013-2016).

He delivered invited lectures and seminars at the invitation of universities and research centers in England (City University, London, 1985), Austria (Insbruck University, 1996), Brazil (Espírito Santo University, Vitoria, 1996), Czechoslovakia (CAS UTIA, 1983), China (University of CAS School of Management and University of Finance and Economics, Chengdu, both in 2013; CAS 2016), France (LAAS Toulouse, 1994; EC Lille, 2008), Poland (Warsaw Polytechnic University, 1980), Sweden (Chalmers TH, Göteborg, 1974 ; Uppsala University, 2011), Germany (TH Ilmenau, 1983; ZKI Berlin, 1991; Fraunhofer FIRST, 1995).

His main personal scientific interests include: hierarchical optimization and control of large-scale systems, decision support systems (DSS), applications of IT in the cultural sector, technology management and foresight.

From the very beginning of his carrier he was interested to find effective solutions to real-life problems.

In 1972, as a young engineer, he was appointed to implement a teleprocessing information system in the largest steelworks in Romania, in Galați city (250 km away from Bucharest). He noticed the very high length of the lines which were to connect the central computer to the terminals placed in various production sections and decided to devise an algorithm based on graph theory to minimize the total length and, consequently, the costs.

While as a guest researcher in the Royal Technical University (KTH) Stockholm, in 1974, devised an efficient heuristic algorithm. Then, he approached many practical process and production control problems he identified in various industries (mainly refineries, chemical plants, water systems, discrete part manufacturing), tried to find adequate scientific methods to solve them and then came back to applications with a view to deploying the corresponding computerized devised algorithms.

Prof. Filip was an IPC member of more than 50 international conferences held in Europe, USA, South America, Asia and Africa. Since 2008 he has been the IPC chair of International Conference on Computers Communications and Control. Also chaired the IPC of several other conferences held in Chile (2008), France (2010), Romania (2006, 2007, 2008, 2009, 2010, 2012, 2014, 2016), Portugal (2010), Germany (2015). He is one of the honorary chairs of the ITQM (Information Technology and Quantitative Management). He was invited to deliver plenary talk at scientific conferences held in England (LISS2013, Reading), Chile (IFAC MCPL2004, University of the Americas, Santiago, IFAC Workshop on Logistics, Santiago, IEEE CESA2012, Santiago), China (CESA2005, University Tzingshua Beijing, ITQM2013, Suzhou), France (CSCS2013, Villeneuve d'Asq), Portugal (IFAC MCPL2010, Coimbra), Spain (IFAC BASYS2010, Valencia; CIO2012, Vigo), Tunisia (CESA1997, Hammamet).

2.1 Several professional highlights

1. Creating of one of the first Romanian CAD (Computer Aided Design) packages, containing original optimization algorithms in graphs called OPTCONF - Designing Equipment Configurations (1974);

2. Designing the first Romanian experimental Decision Support Systems for a steady and discreet production management, with a demonstration of their operation in teleprocessing mode at the World Cybernetics Conference in Bucharest (1976) (on the FELIX256 computer placed in ICI connected to the IBM VT 52 terminal and CII MITRA minicomputer, located at the Sala Palatului (Palace Hall));



D. Tufis, I. Dzitac, L.A. Zadeh M.J. Manolescu, F.G. Filip (ICCCC2008, Oradea)

3. Coordination of teams who built DISPATCHER DSS family (containing original optimization algorithms for discrete time systems with hard constraints, sparse matrices, and slowly variable in time parameters) implemented in refineries, chemical plants and retention and distribution water systems (1980-2000).

4. Refining, adapting, and using control theory methods to model the interacting entities of the cultural sector and creative industries (since 1996).

2.2 Recognitions and awards

Prof. Filip is a Honorary Magister of Dunărea de Jos University of Galați (since 2006), a honorary member of the Agora University Senate (since 2007) and has received Doctor Honoris Causa title from Lucian Blaga University of Sibiu (November 2000), Valahia University, Târgoviște (2007), Ovidius University, Constanta (2007), Ecole Centrale de Lille, France (2007), Politehnica University of Timișoara (2009), Agora University of Oradea, Academy of Economic Studies of Bucharest (2014), University Petrol-Gaze of Ploiești (2017), and University of Pitești (2017).

Prof. Filip received the National Orders "Serviciu Credincios" (Contentious Service), at rank of "Great Cross" and "Steaua Romaniei" (Star of Romania) at rank of "Knight" from the presidents of Romania in 2000 and 2017, respectively, and "Marin Drimov" medal from the Bulgarian Academy of Sciences. He is the recipient of several prizes such as: COPY RO Prize for the best book in IT for the book "Computer aided decision-making" (in 2002), the Grigore Moisil Prize, and Medal "Man of the year 1999" from the Association for Economic Informatics, "W. Scott Jr. in IT" Award from the IAITQM for "remarkable record in the theory, algorithm, education and applications of information technology" (2013), "Stefan Odobleja" Award in Information Science and Technology from the Academy of Scientists for the book *Optimization dans les sciences de l'ingenieur* (2015).



"Walter Scott Jr. in IT Award" for F.G. Filip by the IAITQM at ITQM2013, Suzhou, China (Anonymous Sponsor), Cheng-Few Lee, Cheng Siwei, Florin Gheorghe Filip, Daniel Berg, Yong Shi

3 Publications

3.1 Published books (monographs and volumes)

F.G. Filip is author/co-author of 13 monographs, and edited/co-edited 24 contributed volumes, such as: [2], [17], [18], [20], [34], [35], [43], etc.

The summary of the monograph *Computer-Supported Collaborative Decision-Making* [34] is presented at Springer web page [45]:

Addresses specific concepts, technologies, and systems for collaborative activities with particular emphasis on decision-making;

Balanced presentation of well consolidated and modern methodologies, together with pacing information and communications;

Includes various industrial, financial and "culture economy" applications.

"This is a book about how management and control decisions are made by persons who collaborate and possibly use the support of an information system. The decision is the result of human conscious activities aiming at choosing a course of action for attaining a certain objective (or a set of objectives). The act of collaboration implies that several entities who work together and share responsibilities to jointly plan, implement and evaluate a program of activities to achieve the common goals.

The book is intended to present a balanced view of the domain to include both well-established concepts and a selection of new results in the domains of methods and key technologies. It is meant to answer several questions, such as:

- a) How are evolving the business models towards the ever more collaborative schemes?;
- b) What is the role of the decision-maker in the new context?;
- c) What are the basic attributes and trends in the domain of decision-supporting information systems?;

- d) Which are the basic methods to aggregate the individual preferences?;
- e) What is the impact of modern information and communication technologies on the design and usage of decision support systems for groups of people?"

3.2 Most cited papers

Florin Gheorghe Filip has authored/co-authored over 300 scientific papers. For example, some selected papers authored by F.G. Filip are [9] - [16] or co-authored with other international scientists [3]- [7], [19], [21]- [25], [27] - [38], [40], [41], etc.

Only in Web of Science - Core Collection, the papers authored/co-authored by F.G. Filip has been cited over 200 times.

The most cited paper authored by F.G. Filip is *Decision support and control for large-scale complex systems* [15]. This paper has been cited in a lot important international journals, indexed in Web of Science, such as: *Robotics and Autonomous Systems*, *Information Sciences*, *International Journal of Systems Science*, *Technological and Economic Development of Economy*, *IEEE Transactions on Industrial Electronics*, *Studies in Informatics and Control*, *Romanian Journal of Information Science and Technology*, *Quality and Reliability Engineering International*, *Biometric and Intelligent Decision Making Support*, *International Journal of Computers Communications & Control*, *Simulation Modelling Practice and Theory*, and others.

Many citations in Web of Science has also other papers such as: *A Decision-Making Perspective for Designing and Building Information Systems* [16], *Job Shop Scheduling Optimization in Real-Time Production Control* [29], *DSS in Numbers* [33], *Multilevel optimization algorithms in computer-aided production control in process industry* [23], etc.

4 Science and research management

1. Management of ICI (National Institute of Informatics) during 1991-1997, as a director-general. At the time, ICI obtained about 50% of the IT research projects funded by the European Commission and it was placed (in 1997) first among the most famous 10 research institutes in Romania in the Report of the Coopers & Lybrand Company Consortium, submitted to the European Commission. In 1992, ICI team introduced Internet into the country through the RNC network 1992;

2. He was part of the first team of the five evaluators from Romania for the project proposals for the EU research program called EC-PECO (European Commission - Pays de l'Europe Centrale et Orientale) in 1992;

3. He introduced the Romanian research funding system through competition in the field of "Informatics" (1992) and co-founder and deputy chair (1996-2003) of the Council for the Romanian Academy Grants;

4. He was the Secretary General of the Romanian National Advisory Council for Research-Development, and Innovation (1997-2010);

6. Member in the "IST Prize" jury, organized by EUROCASE (Federation of academies of Applied and Engineering Sciences in Europe) for 10 years (1997-2007);

7. The first Romanian scientist in the IT Advisory Group (ISTAG) of the European Commission (2001), and NATO Science for Peace subcommittee for "Computer Networks" (2000).

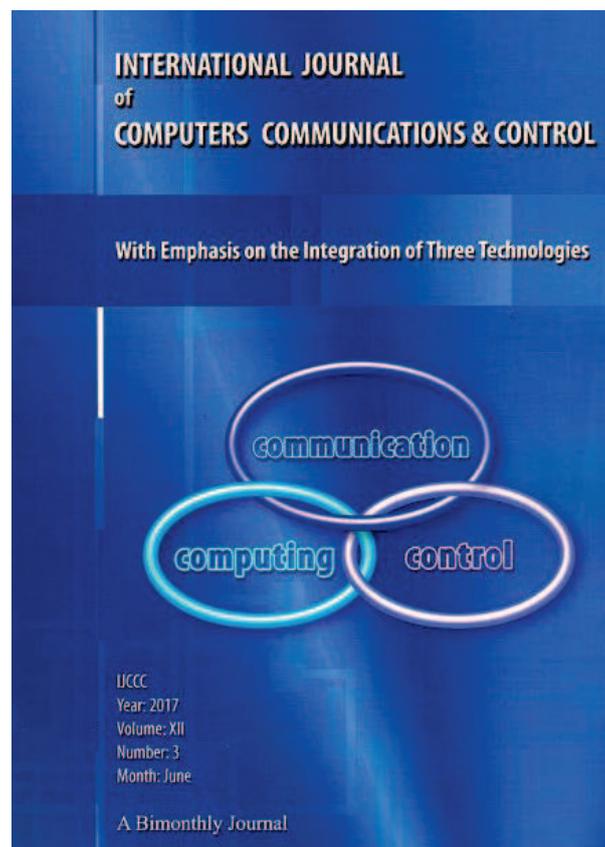
5 Editorial activities

Acad. Filip is a member in the editorial staff of several scientific journals, such as:

1. Systems Analysis, Modeling and Simulation (Taylor and Francis, 1993-2004);

2. International J. of Critical Infrastructures (Inderscience Publishers, since 2004);
3. Computer Journal of Moldova (Chisinau, since 1993);
4. Information Technologies and Control (Sofia, since 1998);
5. Romanian Journal of Information Science and Technology (Romanian Academy);
6. Studies in Informatics and Control (ICI, founder and chief-editor, from 1989);
7. International Journal of Computers Communications & Control (Agora Univ., chief-editor from 2006);
8. Control Engineering and Applied Informatics (SRAIT);
9. Romanian Journal of Informatics and Automatics (ICI, founder);
10. Romanian Journal of Automatics;
11. Technological and Economic Development of Economy (Taylor & Francis);
12. International Journal of Information Technology & Decision Making (Science Direct);
13. Advances in Electrical and Computer Engineering (Stefan Cel Mare Univ.);
14. Financial Innovation (Springer, since 2016);
15. Journal of System and Management Sciences (Beijing Jiatong Univ., China).

F.G.Filip is the founder and Editor-in-Chief of Studies in Informatics and Control (1991), and co-founder and Editor-in-Chief of International Journal of Computers Communications & Control (2006).



International Journal of Computers Communications & Control
(Founded in 2006 at Agora University by I. Dzitac, F.G. Filip, M.J. Manolescu)

6 Conclusion

Florin Gheorghe Filip authored/co-authored over 300 papers published in international journals (IFAC J. Automatica, IFAC J. Control Engineering Practice, Annual Reviews in Control, Computers in Industry, Large Scale Systems etc.) and contributed volumes released by international publishing houses (Pergamon Press, North Holland, Elsevier, Kluwer, Chapman & Hall and so on). He is also the author/co-author of 13 monographs (published by Editura Tehnică - Bucharest, Hermes-Lavoisier - Paris, J. Wiley & Sons London, Springer) and editor/co-editor of 24 volumes of contributions (published by Editing House of the Romanian Academy, Pergamon Press, Elsevier, Institute of Physics Melville - USA, IEEE Computer Society - Los Alamitos - USA). His main scientific contributions include: hierarchical optimization and control of large-scale systems, decision support systems (DSS), applications of IT in the cultural sector, technology management and foresight. In our bibliographic list are presented 42 selected works (papers, monographs, volumes) authored/co-authored by F.G. Filip [2-7], [9-44].

Acad. Filip have a good international impact and many collaborators, from countries such as: *Brazil* (L.F. Autran Monteiro Gomes), *Bulgaria* (K. Boyanov, B. Sendov), *Chile* (F. Cordova, C. Lagos, G. Lefranc), *China* (G. Kou, Y. Peng, C. Siwei, Y. Shi, M. Tang), *Czech Republic* (K. Bakule, J. Zavorka), *Finland* (K. Leiviskä), *France* (J. Bernussou, P. Borne, A. Dolgui, A. El Kamel, H. Panetto, A. Titli, P. Zarate), *Germany* (R. Popescu-Zeletin, K. Reinisch, A. Sydow), *Greece* (P. Groumpos, G. Metakides, P. Spirakis), *Hungary* (G. Kovacs, L. Monostori), *Lithuania* (G. Dzemyda, A. Kaklauskas, Z. Turskis, E.K. Zavadskas), *Mexic* (A. Molina), *Poland* (W. Findeisen, M.A. Brdys, K. Malinowski), *Portugal* (A.D. Correia, L.M. Camarinha Matos), *Republic of Moldova* (G. Căpățână, I. Cojocaru, S. Cojocaru, G. Duca, C. Gaidric), *Romania* (B. Bărbat, V. Balas, M. Bizoi, G. Bologna, I. Buciu, C. Ciurea, H. Dragomirescu, M. Guran, D. Donciulescu, I. Dumitrache, I. Dzitac, E.A. Iancu, A. Ionita, I. Ivan, R. Lile, A. Manolescu, M.J. Manolescu, I. Moisil, G. Neagu, S. Nicoară, S.I. Nițchi, H. Oros, N. Paraschiv, D. Popescu, R.E. Precup, O. Proștean, D. Ștefănoiu, A.M. Suduc, H.N. Teodorescu, D. Tufiș, C.B. Zamfirescu), *Russia* (F. Aleskerov), *Spain* (E. Herrera -Viedma, A. Ortiz), *Sweden* (A. Hedin), *UK* (P.D. Roberts), and from *USA* (R. Andonie, D. Berg, S.M. Gupta, S.Y. Nof, D.J. Power, G. Tecuci, T.J. Williams, A.B. Whinston, L.A. Zadeh).

We cannot come to an end without mentioning the man Florin Gheorghe Filip. All of us that know him, if asked which are the first three words coming in our minds we shall not hesitate to tell: intuitive, generous and catalyst. He is able to intuit a young researcher, a valuable research team or an academic institution with prospects. Full of modesty, he does not know envy and is always eager to support those needing his advices, with a generosity specific to great spirits. His behavior is always of a catalyst, precipitating many beneficial reactions, support without waiting for a reward.

Acknowledgment

We are honoured and we grateful to Prof. Filip for his contributions, in his capacity of Editor-in-Chief, of our *International Journal of Computers Communications & Control* (co-founded in 2006 with I. Dzitac and M.J. Manolescu at Agora University in Oradea, Romania).

Bibliography

- [1] Bărbat B.E. (2007), Filip the Catalyst, *The Computer Science Journal of Moldova*, 15(2), 115-119, 2007.
- [2] Borne P., Popescu D., Filip F.G., Ștefanioiu D. (2012), *Optimisation en sciences de l'ingénieur*, Hermes-Lavoisier, Paris, 2012 (in French).
- [3] Brândaș C., Panzaru C., Filip F. G. (2016), Data driven decision support systems: an application case in labour market analysis, *Romanian Journal of Information Science and Technology*, 19(1-2), 65-77, 2016.
- [4] Căndea C., Filip F.G. (2016), Towards intelligent collaborative decision support platforms, *Studies in Informatics and Control*, 25(2), 143-152, 2016.
- [5] Duță L., Filip F.G., Henrioud J.M.(2003), A method for dealing with multi-objective optimization problem of disassembly processes, *Assembly and Task Planning, 2003. Proceedings of the IEEE International Symposium*, 163-168, 2003.
- [6] Duță L; Filip F.G., (2008), Control and Decision-making Process in Disassembling Used Electronic Products, *Studies in Informatics and Control*, 17(1), 17-26, 2008.
- [7] Duță L., Filip F.G., Popescu C. (2008), Evolutionary programming in disassembly decision making *International Journal of Computers Communications & Control*, 3(Suppl.), 282-286, 2008.
- [8] Dzitac I., Manolescu M.J., Oros H., Văleanu E. (2007); 60 Years from Birth of Academician F.G. Filip, *International Journal of Computers Communications & Control*, 2(3), 209-216 , 2007.
- [9] Filip F.G. (1989), Creativity and decision support systems, *Studies and Researches in Computer and Informatics*, 1(1), 41-49, 1989.
- [10] Filip F.G. (1991), System analysis and expert systems techniques for operative decision making, *Systems Analysis Modeling Simulation*, Gordon and Breach Science Publishers, 8(3), 213-219, 1991.
- [11] Filip F.G. (1994), Evolutions in systems analysis, modelling and simulation in Computer based industrial automation, *Systems Analysis, Modelling and Simulation*, 15, 135-149, 1994.
- [12] Filip F.G. (1995), IT culture dissemination in Romania: experiments and achievements, In A. Inzelt and R. Coenen (Eds.), *Knowledge Technology Transfer and Forecasting*, Kluwer Academic, Dordrecht, 149-160, 1995.
- [13] Filip F.G. (1995), Towards more humanized DSS, In: L.M. Camarinha Matos and H. Afsarmanesh (Eds.), *Balanced Automation Systems: Architectures and Design Methods*, Chapman & Hall, London, 230-240, 1995.
- [14] Filip F.G. (1998), Optimization models with sparse matrices and relatively constant parameters, *Systems Analysis Modelling and Simulation*, 33, 407-430, 1998.
- [15] Filip F.G. (2008), Decision support and control for large-scale complex systems, *Annual Reviews in Control*, 32(1), 61-70, 2008.

- [16] Filip F.G. (2012), A Decision-Making Perspective for Designing and Building Information Systems, *International Journal of Computers Communications & Control*, 7(2), 264-272, 2012.
- [17] Filip F.G. (2002), *Decizie asistată de calculator decizii, decidenți, metode și instrumente de bază*, Editura Tehnică, București, 2002 (in Romanian).
- [18] Filip. F.G. (2007), *Sisteme suport pentru decizii*, Editura Tehnică, București, 2007 (in Romanian).
- [19] Filip F.G., Alexandru A., Socol I. (1997), Technology management and international cooperation: several success stories, *J. Human Systems Management*, 16, 223-229, 1997.
- [20] Filip F.G., Bărbat B. (1999), *Informatică industrială: Noi paradigme și aplicații*, Editura Tehnică, București, 1999 (in Romanian).
- [21] Filip F.G., Ciurea C., Dragomirescu H., Ivan I. (2015), Cultural heritage and modern information and communication technologies, *Technological and Economic Development of Economy*, 21(3), 441-459, 2015.
- [22] Filip F.G., Donciulescu D. A. , Filip C.I. (2001), A cybernetic model of computerization of the cultural heritage, *Computer Science Journal of Moldova*, 9(2), 101 -112, 2001.
- [23] Filip F.G. , Donciulescu D.A., Gașpar R., Muratcea M., Orășanu L. (1985), Multilevel optimization algorithms in computer-aided production control in process industry, *Computers in Industry*, 6(1), 47-57, 1985.
- [24] Filip F.G., Donciulescu D.A. (1983), On on-line direct dynamic coordination method in process industry, *IFAC J. Automatica*, Pergamon Press, Oxford, 19(1), 317 - 320, 1983.
- [25] Filip F.G., Dragomirescu H., Predescu R., Ilie R. (2004), IT tools for foresight studies, *Studies in Informatics and Control*, 13(3), 161-168, 2004.
- [26] Filip F.G., Dumitrache I., Iliescu S.S. (Eds.) (2002), *Large Scale Systems: Theory and Applications*, Elsevier Science Ltd. (Pergamon), Oxford, 2002.
- [27] Filip F.G., Herrera-Viedma E. (2014), Big data in the European Union, *The Bridge*, 44(4), 33-37, 2014.
- [28] Filip F.G., Leiviskä K. (2009), Large-scale complex systems, In: Nof S.Y. (Ed.), *Springer Handbook of Automation*, Springer, Berlin Heidelberg, 619-638, 2009.
- [29] Filip F.G., Neagu G., Donciulescu D. (1983), Job Shop scheduling optimisation in real-time production control, *Computers in Industry*, 4(4), 395-403, 1983.
- [30] Filip F.G., Neagu G., Donciulescu D. (1996), DSSfM: from technology to decision making, *Proceedings 14th IFAC World Congress*, San Francisco, Vol B, 367-372, 199
- [31] Filip F.G., Neagu G. (1993), CIM in continuous and discrete manufacturing: object-oriented generic modeling, *IFAC J. Control Eng. Practice*, 1(5), 815-825, 1993.
- [32] Filip F.G., Popescu D., Mateescu M. (2008), Optimal decisions for complex systems-software packages, *Mathematics and Computers in Simulation*, 76(5-6), 422-429, 2008.
- [33] Filip F.G., Suduc A.M., Bizoi M. (2014), DSS in Numbers, *Technological and Economic Development of Economy*, 20(1), 154-164, 2014.

- [34] Filip F.G., Zamfirescu C.B., Ciurea C. (2016), *Computer-Supported Collaborative Decision-Making*, Springer, 2016.
- [35] Guran M., Filip F.G. (1986), *Sisteme ierarhizate și în timp real cu prelucrarea distribuită a datelor*, Editura Tehnică, București, 1986 (in Romanian).
- [36] Guran M., Filip F.G., Donciulescu D., Orășanu L. (1985), Hierarchical optimization in computer-aided dispatcher systems in the process industry, *Large scale systems*, 8(2), 157-167, 1985.
- [37] Nicoară E.S., Filip F.G., Paraschiv N. (2011), Simulation-based Optimization Using Genetic Algorithms for Multi-objective Flexible JSSP, *Studies in Informatics and Control*, 20(4), 333-344, 2011.
- [38] Nof S.Y., Morel G., Monostori L., Molina A., Filip F.G. (2006), From plant and logistics control to multi-enterprise collaboration, *Annual Reviews in Control*, Elsevier, 30(1), 55-68, 2006.
- [39] Pascu C., Filip F.G. (Eds.) (2005); *Visions on Information Society Future in an Enlarged Europe*, Editura Academiei, București, 2005.
- [40] Seok H., Nof S.Y., Filip F.G. (2012), Sustainability decision support system based on collaborative control theory *Annual Reviews in Control*, 36(1), 85-100, 2012.
- [41] Reșteanu C., Filip F.G., Ionescu C., Somodi M. (1996), On optimal choice problem solving, In: *Systems, Man, and Cybernetics, 1996, IEEE International Conference*, 3, 1864-1869, 1996.
- [42] Ștefanoiu D., Borne P., Popescu D., Filip F.G., El Kamel A. (2014), *Optimization in Engineering Sciences: Approximate and Metaheuristic Methods*, John Wiley & Sons, London, 2014.
- [43] Zadeh L.A., Tufiș D., Filip F.G., Dzitac I. (Eds.) (2008), *From Natural Language to Soft Computing: New Paradigms in Artificial Intelligence*, Editing House of the Romanian Academy, 2008.
- [44] Zamfirescu C.B., Filip F.G. (2010), Swarming Models for Facilitating Collaborative Decisions *International Journal of Computers Communications & Control*, 5(1), 125-137, 2010.
- [45] <http://www.springer.com/gp/book/9783319472195>

Obfuscation-based Malware Update: A comparison of Manual and Automated Methods

C. Barría, D. Cordero, C. Cubillos, M. Palma, D. Cabrera

Cristian Barría, Claudio Cubillos*

Pontificia Universidad Católica de Valparaíso
Valparaíso, Chile
cristian.barría@udp.cl

*Corresponding author:claudio.cubillos@pucv.cl

David Cordero

Universidad Andrés Bello
Santiago, Chile
d.cordero.v@gmail.com

Miguel Palma

Universidad Tecnológica de Chile
Santiago, Chile
rmiguel.palma06@inacapmail.cl

Daniel Cabrera

Universidad de Valparaíso
Valparaíso, Chile
daniel.cabrera@uv.cl

Abstract: This research presents a proposal of malware classification and its update based on capacity and obfuscation. This article is an extension of [4]^a, and describes the procedure for malware updating, that is, to take obsolete malware that is already detectable by antiviruses, update it through obfuscation techniques and thus making it undetectable again. As the updating of malware is generally performed manually, an automatic solution is presented together with a comparison from the standpoint of cost and processing time. The automated method proved to be more reliable, fast and less intensive in the use of resources, specially in terms of antivirus analysis and malware functionality checking times.

Keywords: Security, Malware, obfuscation techniques, cyberspace, antivirus.

^aReprinted (partial) and extended, with permission based on License Number 3954200696590 [2016] © IEEE, from "Computers Communications and Control (ICCCC), 2016 6th International Conference on".

1 Introduction

Malware is a malicious program that provides access to a system without user's authorization and executes undesirable actions. The terms malware and virus are usually used as the same concept, although they are, in fact, different [4] [19]. Initially, the use of malware was strictly associated with the investigation and protection of software developers for intellectual property (using cryptography), but throughout time its objective changed, and therefore, its importance. Currently, the aim of malware is mainly related to profit- making [8] [15].

The evolution of malware began in 1949 when Von Neumann established the idea of stored program and introduced the Automata Theory, which is the possibility of developing small replicates of a program that were capable of taking control of other programs with similar structures [13]. While the concept has many applications in science, it is quite easy to apply it to

viruses, considering that, at that time, the two concepts (i.e., virus and malware) were thought as equivalent.

In 1971, Bob Thomas created the first malicious code, named Creeper, that was capable of infecting IBM 360 machines on the ARPANET network by delivering an on-screen message that said, "I'm the Creeper, catch me if you can!" To find and eliminate the Creeper another code, called Reaper, was created. From this, it originated the current antivirus [3].

Primarily, malware affects the Critical Information Infrastructure (CII), which is of prime importance to both, the attackers and the defenders. It is important to highlight that a successful virus leaves the CII defenseless. This is why for both, the defenders and the offenders, there is a necessity of implementing a Security Information Management System (SIMI) with the objective of preserving the reliability, integrity and availability of the information in order to act upon these threats [18].

In the context of this research, obfuscation will be defined as the execution of transformations in a malware code (source or binary) that changes its appearance through the realization of a series of steps. This maintains the malware functionality and allows it to take control of the systems [6]. It is fundamental to continue carrying out research regarding obfuscation, in order to contribute to its analysis from a malware updating perspective. As the malware updating process is generally performed manually, an automatic solution is to be introduced so to compare both in terms of time and costs.

This research, an extended version of a previous work [4], provides a proposal of malware classification based on obfuscation. The novelty of the present work relies on; first, providing a summary of the obfuscation-based malware classification tackled in the previous work [4]; second, formalizing the malware update process and making an empirical demonstration on how a malware, that is detectable by antiviruses, could be reused by updating it and making it undetectable through obfuscation; third, comparing the manual and the automated methods for malware updating through an example.

In section 2, this paper presents a summary of malware classification, while in section 3 the use of obfuscation through a crypter for updating the detected malware is explained. In section 4, a chart of the malware updating and its classification is proposed. In section 5, the proposed malware updating process is formalized and explained through an example. In section 6, a software prototype of an automated version of the obfuscation process is presented. Section 7, provides an evaluation and comparison of the manual and automated solutions. Finally, in section 8, the conclusions of this research are presented and future work.

2 Related work

In this section, a brief overview of the different malware classification is provided. As it is known, there are different malware classifications stemming from different perspectives. For example, classifications regarding malware infection mechanisms, and others regarding means of malware spread. Although these are used in different investigations, there is no given classification from the perspective of obfuscation, as a common vision has not been agreed yet [17].

As it is shown in Figure 1, a structure that facilitates the malware classification (species) could be established according to the class, type and generation. This classification will enable to succinctly present a part of the essential information about the malware non-detectable capacities.

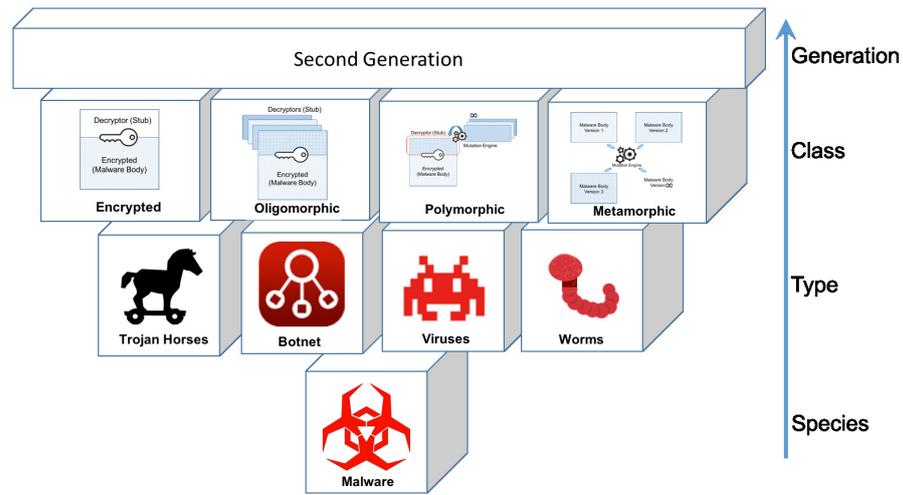


Figure 1: Malware classification

3 Malware updating based on obfuscation

According to the malware classification provided in previous investigations, first-generation malware were not considered part of the inverted pyramid shown in Figure 1, as they did not present any modifications in their code and they are commonly denominated No-Stealth [12] because of their obsolescence.

Although No-Stealth malware are detected by an antivirus, currently they are capable of being transformed into second-generation malware using crypters. Crypters are easy to obtain from specialized information security websites, and these could apply obfuscation techniques to any type of file. However, in this research, crypters will be applied to malware, yet not alter their functionality in any of their states [6] [9] [20].

3.1 Update

Malware updating requires obfuscation to give rise a second generation. This can be performed by using one of the following two methods: a) The manual method. In this, the operation is performed by a person and is applied to the Encryptor and Oligomorphic malware types. b) The automated method. This refers to a malware that can automatically undergo the process of obfuscation without requiring a human operator and is applied to Polymorphic and Metamorphic malware [14], as it is shown in Table 1.

Table 1: Obfuscation update

MALWARE	Manual	Automated
Encryption	X	
Oligomorphic	X	
Polymorphic		X
Metamorphic		X

3.2 Procedures

Among the procedures that are highlighted in various sources of information related to computer security, the following can be identified: AvFucker, DSplit, RIT, Hexing and XOR. These aim at generating a binary of the original code, producing a modification of a relevant hexadecimal value to evade the antivirus. Figure 2 allows us to perform an updating to the malware through the obfuscation process, which will depend on malware status, method, technique and procedure employed.

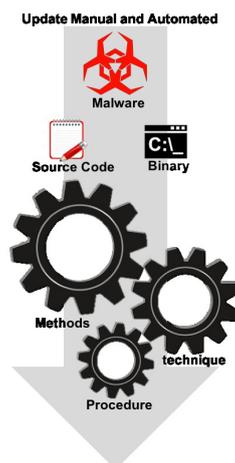


Figure 2: Malware obfuscation update.

In the case of an antivirus detection, a decryptor or stub is resorted to apply the obfuscation method, in order to re-apply the crypter to malware so to achieve its renovation and, thereby its ability of evasion. The report provided by the antivirus detection is directly given to the stub, if this gets detected it directly compromises the encrypted malware. This is put into practice through the following tests:

- Malware construction (No-Stealth)

The construction of a malicious code has the objective of obtaining documents from a third party, including certain directories and various extensions from a computer. A part of the code is presented (for confidentiality reasons) in Figure 3.

```

user = getpass.getuser()
a= r'C:/Users/'
b= user+'/'
c= r"Desktop"
d= r"Documents"
e= r"Downloads"
ficheros=os.listdir(a+b+c)
mvdokumentos=os.listdir(a+b+d)
download=os.listdir(a+b+e)
included_extenstions = ['docx', 'doc', 'pdf']
file_names = [fn for fn in ficheros if any([fn.endswith(ext) for ext in included_extenstions])]

```

Figure 3: No-stealth malware code portion

- Antivirus tests

The analysis of a malicious code is performed by a specific antivirus (NOD 32) signature, which has the ability to incorporate other antivirus engines. As it is shown in Figure 4, this code is detected and reported as an element of risk.



Figure 4: Antivirus test

- Obfuscation application

After the malicious code has been detected, obfuscation techniques are applied in order to subject the malware to a crypter (encryption / decryption), as shown in Figure 5. Once the process has been completed, a detectable code may become undetectable.



Figure 5: Crypter application

- Screening

Once the malicious code has undergone obfuscation, the malware is subjected to the antivirus (NOD 32) again, under the same conditions described above (same house and built motors). As it is shown in Figure 6, the antivirus always generates a reporting message in case the malware is either detected or not. In case the malware goes undetected, the antivirus logs does not consider this file a threat, as shown in Figure 7, and therefore, the malware may fulfill its malicious intent.



Figure 6: Antivirus test.

As shown in Figure 8, it is evident that a first-generation malware could be subjected to obfuscation techniques. In order to do so, a crypter tool is applied to transform this first-generation malware into a second-generation, and thereby, rising its hierarchical criteria.

4 Obfuscation process and malware classification proposal

In accordance with the malware classification that was previously described, a first-generation malware is not considered because it has not undergone the obfuscation techniques yet. However, it is possible to prove that despite the obsolescence of this No-Stealth malware, this is still present in the cyberspace and need to be considered. Furthermore, this No-Stealth malware could be transformed into a threat only by applying an obfuscation procedure, resulting in this becoming undetectable by an antivirus. Taking into account the points previously exposed, a flow chart is proposed for the updating of malware, as shown in Figure 9.

Anál. bot. dcho No hay amenazas

Figure 7: Antivirus logs

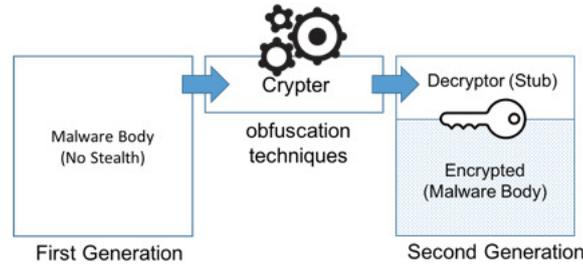


Figure 8: Malware transformation

Thus, in the first inverted pyramid previously described the No Stealth malware should be taken into consideration because this could be related to any other malware either belonging to a class (virus, worms, botnet, and Trojan horses), or belonging to the first generation. As it is shown in Figure 10, a revised inverted pyramid is proposed in order to include a complete structure that classifies malware.

5 Current malware updating process

The process of detecting malware through the recognition of their signatures, that these are later saved into the antivirus databases. However, through an updating process, these malware could be transformed and become undetectable without losing its functionality as it is described by Barría et al [4].

When considering the diagram in Figure 11 [11], it could be observed that the updating process of a malware starts in its binary state, and then a crypter is required in order to apply the method of encryption. After this, the crypter output must undergo the antivirus detection (signature-based analysis). In case the encrypted malware goes undetected, it means the updating process has been successful. In the specific case of this work, the obfuscation process employs the Dead-Code Insertion technique [1]. For all aforementioned, the modders defines this as the AvFucker [11].

This process involves a series of procedures in which the malware is encrypted by a Crypter

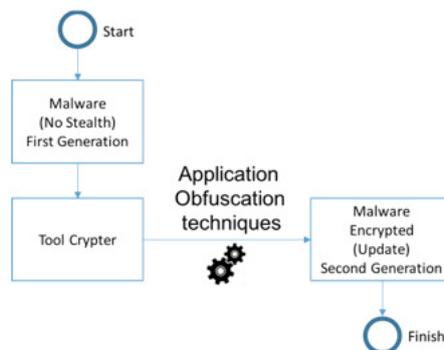


Figure 9: Proposed obfuscation procedure.

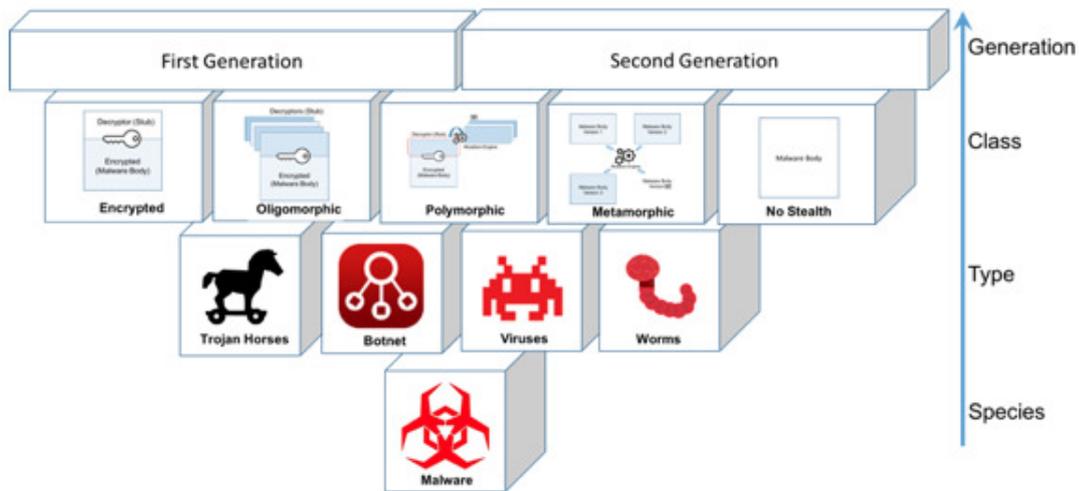


Figure 10: Proposed malware classification

and is then reviewed by one or more anti-malware systems. If it is detected, an insertion of a dead code, based on the process of obfuscation, followed by a re-encryption of the resulting code with a Crypter is performed.

Finally, the test is performed against the anti-malware or systems, and the cycle is repeated until a program is obtained with a signature that is no longer detected by these systems. A graphical explanation of this process is illustrated below in Figure 11 [10]. When the crypter output is contrasted through detection techniques based on signatures [2] and is detected as malware, the procedure is applied as a whole. This means that the stub, the key and the enciphered malware are included, although only the stub area is detected. This procedure consists of the malware copy creation (crypter output), whose quantity is given by the file size, according to the following equation:

$$Nc = (St + K + Mc)$$

where:

Nc= Copy number.

St= Stub bytes size.

K = Key bytes size.

Mc= Enciphered malware bytes size.

Each one of these copies, is created for the Dead Code Insertion (00, 90), from byte to byte, from the initial to the final offset, correlatively [7], as observed in Figure 12.

Once all copies are made, they are analyzed by the antivirus. In this way, only files detected as non-malicious are obtained, and the rest are removed when matching with the antivirus registered signature. A group of files is then obtained that is able to avoid the antivirus, but its hexadecimal code has been modified, so it is necessary to verify that the functionality is not affected.

What was previously presented is not viable if a file of 1 Mb output is used as an example:

$$Nc = (St + K + Mc) = 1.048.576 \text{ copies}$$

$$Nc^2 = \text{Space for the copies on the hard disk}$$

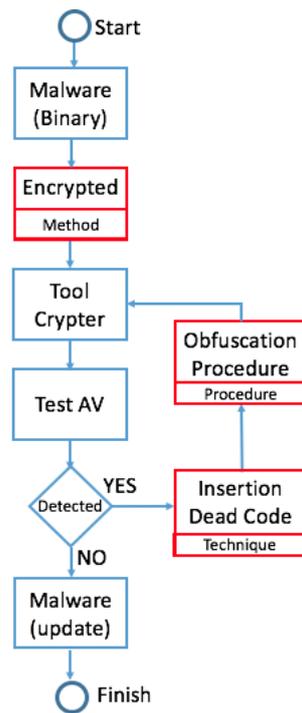


Figure 11: Diagram of the malware update process.

$$Nc^2 = (1.048.576)^2 \text{ bytes} = 1 \text{ Terabyte}$$

The previous analysis implies the following problems: a) capacity, since the number of files created could be higher than the number of information terabytes directly affecting the hardware capacity, which also affects equipment investment; b) test, as a function of the time in which the antivirus performs the analysis on each of the copies; c) functionality, a revision that must be made on an individual basis to the group of undetected files by the antivirus and then prove that the malware functionality is not affected.

This is how the procedure seeks to be optimized by increasing the Dead Code Insertion. If the 1 Mb output file is used, it is possible to create copies every 1000 bytes (offset with 0.1 percent of the total file size), and the proper analysis is made, considering only the portion of 1000 bytes that is undetected. Copies are then created every 100 bytes over it, and this procedure is repeated until reaching the portion of 1 byte [5], considering that the ranges are not necessarily consecutive. Thus, hardware use increases, and the number of created files decreases, so the antivirus testing time also decreases.

Nevertheless, the problem of the malware operation test persists because is needed a controlled environment that allows typical functions of this type of malicious code to be proved, such as persistence, connectivity and spreading. This takes significant time and a great amount of resources for the modder [11]. Thus, a tool called an "annotator" is used [16], which permits the functions described above to be verified by replacing input malware on the crypter. In addition, another characteristic is its tiny size, further allowing optimization of the creation of copies and reducing the antivirus analysis and detection time.

Currently, the iteration of this model is carried out with the support of various tools, including crypters, hex editors, and anti-malware engine systems, among others. It also usually uses a tool to generate a number n copies of the original file, modifying each copy in a different range of offsets and replacing those locations with a dead code. While the use of such tools represents

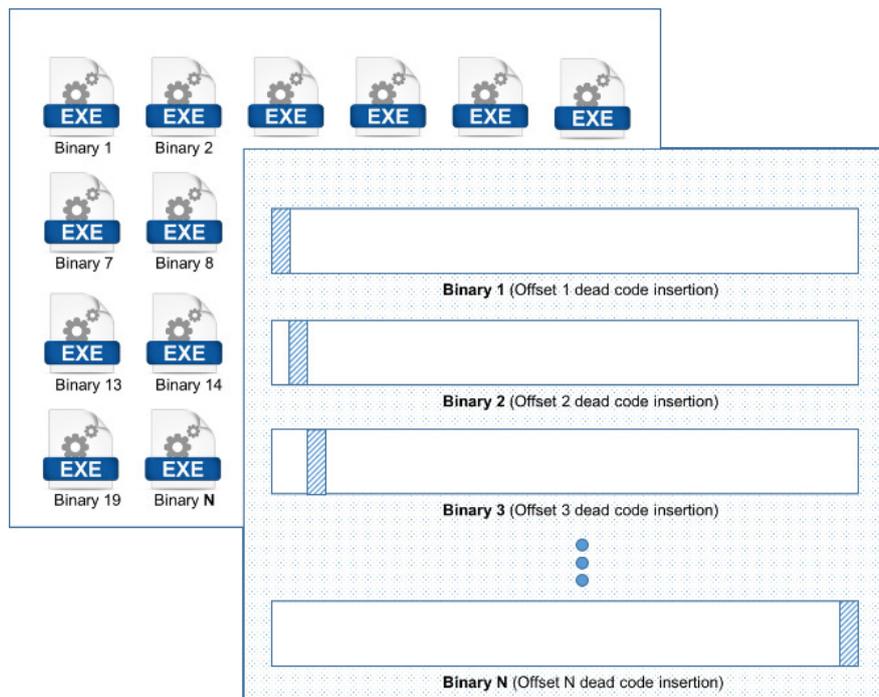


Figure 12: Malware copy generation

a significant reduction in the time required compared to the performance of the same process manually (e.g., editing each range of offsets using an editor), this part of the process adds more time to the total iteration, increasing the final cost of the process. Having identified this difficulty, the ability to automate this iteration emerges as a natural choice in regard to reducing time and monetary costs.

6 Software prototype validation

An algorithm was programmed in Python version 2, with the addition of PyWinAuto library to automate the process under study, exposed in Figure 13.

The automatic process and its algorithm provides the advantage of reducing the execution time of each task, while leaving the structure of the process itself intact, thus allowing an objective performance comparison between the original manual process and the automated one.

Avoiding human factors on the execution of each task reduces the risk of certain kind of errors, mostly related to distraction, fatigue, among others. This can be more critical in certain tasks, compared to others.

For example, selecting the malware sample and using the Crypter tool on it (the first task), can be considered a fairly simple and straightforward task. But replicating the file an undetermined amount of times, with different ranges, can eventually involve the repetition of an action hundreds of times, and each one requiring the exact selection of specific ranges, and offsets which is error prone. In the automated method, the algorithm itself will not suffer from fatigue after 5 hours of uninterrupted work, for example, while humans surely will, being a more reliable solution.

The general process involves five phases: malware encryption, offsets identification, obfuscation through dead code insertion, anti-malware analysis, and functionality analysis. The encryption phase, as its name says, involves the encryption of the malware code through the use of

a crypter software. The offset identification regards the activity of splitting the file for making an exhaustive search looking for the point (offset) in which changing that bit provides a change in the overall malware signature recognized by anti-malware software, thus turning it in undetectable. The anti-malware analysis refers to antivirus check to be performed to the file sample with the modified offset in order to check if it is detected with the virus. Please consider that this step does not involve only a single antivirus software but a hole set of them, four or five different at least in order to ensure undetectability. The phase of functionality analysis, comes after the anti malware analysis for each of the files with their diverse offsets and pursues to check that the dead code insertion has not altered or braked the functionality to the malware within the file. The algorithm automatically goes through each iteration, changes the values for each range and advances until the end of the whole process. The only human interaction required is starting the script, since the automated process replaces every task originally executed by human resources.

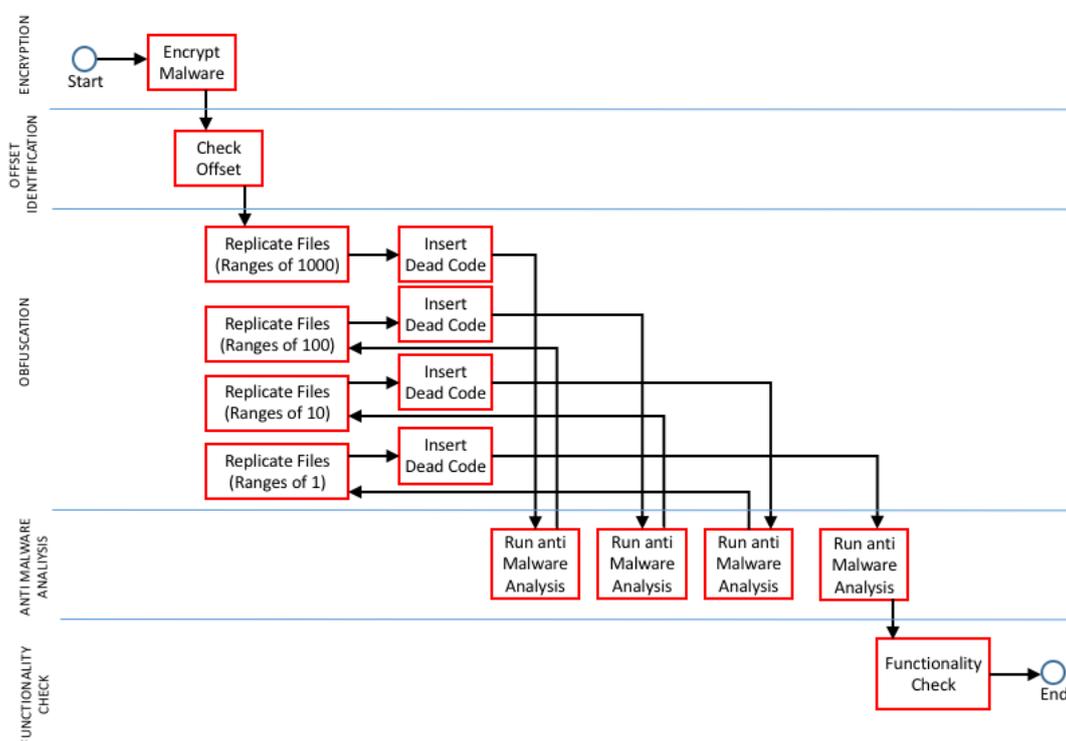


Figure 13: Malware update automated process

7 Experimentation

Once the prototype was prepared, we compared the time required to complete an instance of the process, from beginning to end, between the automated case and the original case then, it was used simulation software, to evaluate the time required in 50 instances.

To standardize the experimental conditions in both cases, the same tools and base file containing the signature of the malware were used, as detailed in the following list:

- Base File
- Crypter tool
- Hex editing tool and File Replication "Offset Locator"

- Anti-malware Engine "NOD32"

Both processes were executed on the same computer with the same operating system (Windows 7). Finally, the economic dimension of each process was evaluated, and each hour of work was assigned the value of CLP 1100.

7.1 Experimental evaluation

The manual process execution required 420 minutes (7 hours), while the prototype software process was completed in 5 minutes. In both cases, it was possible to obtain a series of files that successfully evaded the anti-malware detection engine.

Considering the value assigned to each hour of work, we observe the following, in Table 2.

Table 2: Time and cost comparative table evaluated experimentally.

	Manual Process	Automated Process
Time (hours)	7	0,083
Cost (CLP)	7.700	1.100

7.2 Simulated evaluation

Fifty process execution cycles in the simulation software generated the following results, in Table 3:

Table 3: Time and cost comparative table evaluated by simulation

	Manual Process	Automated Process
Time (hours)	350	4,2
Cost (CLP)	385.000	5.500

Due to the cost being calculated per hour, the automated process value was still considered as a full worked hour, regardless of having completed the entire set of tasks in much less time than that.

It is also important to note, that the human executed tasks may differ in terms of time needed, because of the possibility of mistakes during the execution of any action, thus causing an increase of the time needed and, in consequence, an increase in its cost. During this experimental evaluation, however, there were neither difficulties nor failures during the execution of the entire process, so the values provided are as close as possible to an ideal execution.

Although these values already provide clear evidence of the cost reductions obtained thanks to this automation, this difference may be even higher, because the human resources simulated on the manual process were considered with ideal conditions, such as 100 percent correct execution of tasks (no possibility of mistakes) and continuous, uninterrupted work hours during the whole process. As can be expected, actual human resources will not present such performance, meaning a higher cost for the manual process.

7.3 Cost and time comparison

Figure 14 and Figure 15 present the same data, but from a different perspective, which helps to easily perceive the cost reduction obtained with this automation.

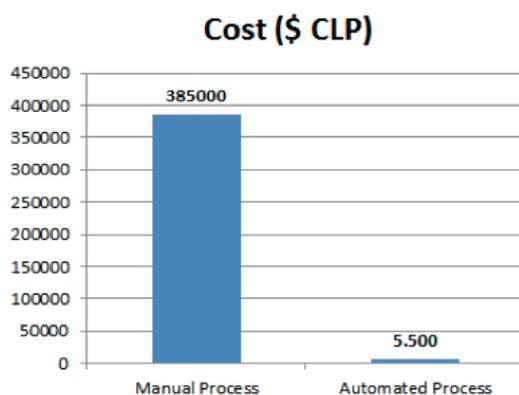


Figure 14: Cost comparison between manual process and automated process under simulated evaluation

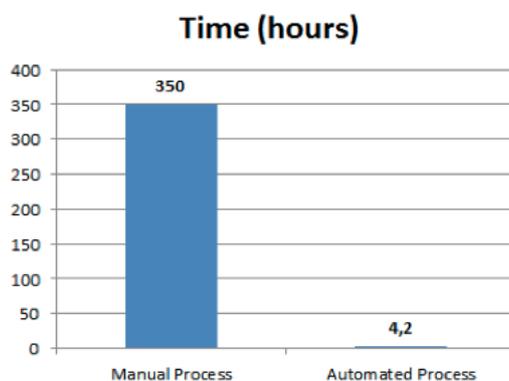


Figure 15: Time comparison between manual process and automated process under simulated evaluation

The numbers speak by themselves, as can be easily noticed that the same result can be achieved with way less time and money, thus exposing the advantages of automating the process using the algorithm. As mentioned before, the structure of the whole process was left untouched, meaning that this comparison is not affected by neither adding, removing nor modifying tasks. Each task remains the same, but its execution is improved via automation.

8 Conclusions and future work

Previous researches have provided a malware classifications that has been generated from multiple perspectives; however, there is no classification generated on the basis of the obfuscation perspective. Consequently, the first contribution of the present work is to provide a classification that takes into account the obfuscation procedure. This revised classification allows us to hierarchically structure the malware based on their capacities. Furthermore, considering that these malware are still present in the cyberspace paves the way to create countermeasures to preserve the security of the information systems.

Because of the rapid evolution of second-generation malware, those that are considered to be part of the first-generation have been described as obsolete, and therefore unimportant. However, this research has proved that through the application of methods, techniques, procedures and tools of obfuscation it is possible that No-Sealth malware could be updated to the point of being

transformed into second-generation malware, and therefore, go undetectable by the antivirus systems. Similarly, the present work incorporates this updating procedure through a flow chart that considers this classification proposal of the revised inverted pyramid previously presented.

This work also explains that this updating procedure could be put into practice from a manual method to an automated method. It is evident that in the case of the automated method, the resources of time and costs are more efficiently employed. This represents a relevant challenge for the malware developers, who day-to-day have to deal with between those who protect and those who evade the system of data protection.

Future work may consider the improvement of this malware updating process in order to employ new techniques with the aim of searching matched patterns and making this process more efficient.

Acknowledgment

The present work has been partially supported by Pontificia Universidad Católica de Valparaíso through scholarship INF-PUCV-2016.

Bibliography

- [1] Balakrishnan A., Schulze C. (2005); Code obfuscation literature surveyt, *CS701 Construction of compilers*, vol. 19, 2005.
- [2] Bazrafshan Z., Hashemi H., Fard S. M. H., Hamzeh A. (2013), Survey on heuristic malware detection techniquet, *Information and Knowledge Technology (IKT), 2013 5th Conference on*, 113-120, 2013. doi: 10.1109/IKT.2013.6620049
- [3] Balakrishnan A., Schulze C. (2010), Code obfuscation literature survey, *CS701 Construction of Compilers*, URL <http://pages.cs.wisc.edu/~arinib/writeup.pdf>, 19, 1-10, 2005.
- [4] Barria C., Cordero D., Cubillos C., Palma M. (2016), Proposed classification of malware, based on obfuscation, *2016 6th International Conference on Computers Communications and Control (ICCCC)*, IEEE Xplore 2016, ISBN: 978-1-5090-1735-5, 37-44, 2016. <http://dx.doi.org/10.1109/ICCCC.2016.7496735>
- [5] Barria C., Cordero D., Cubillos C., Osses R, Obfuscation procedure based in dead code insertion into cryptert, *2016 6th International Conference on Computers Communications and Control*, IEEE Xplore 2016, ISBN: 978-1-5090-1735-5, 23 - 29, 2016. <http://dx.doi.org/10.1109/ICCCC.2016.7496733>
- [6] Egele M., Scholte T., Kirda E., Kruegel C(2008), A Survey on Automated Dynamic Malware-analysis Techniques and Tools, *ACM Comput.Surv.*, 44(2), 1-6, 2008. <http://dx.doi.org/10.1145/2089125.2089126>
- [7] Khurram M., Syed Noor-ul-Hassan S., Zikria Y. B., Nassar I.(2010), Evading Virus Detection Using Code Obfuscation, *Future Generation Information Technology: Second International Conference, FGIT 2010*, 394-401, 2010. <http://dx.doi.org/10.1007/978-3-642-17569-5-39>
- [8] Konstantinou E., Wolthusen S. (2008), *Metamorphic virus: Analysis and detection*, Technical report, Royal Holloway University of London, vol. 15, 2008.
- [9] Kolter J., Maloof M. (2006), Learning to Detect and Classify Malicious Executables in the Wild, *Journal of Machine Learning Research*, 7(7), 2721-2744, 2006.

-
- [10] Kumar A., Shrivastava V. (2013), BASIC: Brief Analytical Survey on Metamorphic Code, *International Journal of Advanced Research in Computer and Communication Engineering*, 2(9), 1-5, 2013.
- [11] Kumar B., Prajapati A. (2013), Modelling and Simulation: Cyber War, *Procedia Technology*, 10, 987-997, 2013. <http://dx.doi.org/10.1016/j.protcy.2013.12.447>
- [12] Livingston W. (2007), COTS: Commercial Off-The-Shell for Custom Off-The-Shelf, CrossTalk , www.stsc.hill.af.mil, 31-31, 2007
- [13] Neumann J. (1996), *Theory of self-reproducing automata*, University of Illinois Press, Edited and completed by A. Burks, 1996.
- [14] [Online] ISO, 9241-11:1998, Ergonomic requirements for office work with visual display terminals (VDTs) - Part 11: Guidance on usability, March 1998.
- [15] [Online] Available: www.trendmicro.com/cloud-content/us/pdfs/security-intelligence/reports/rpt-cashing-in-on-digital-information.pdf, TrendMicro, *Roundup, 2013 Annual Security*, 2013.
- [16] [Online]. Available: www.securitybydefault.com/2013/09/crypters-localizando-firmas-de-los.html. A. Pasamar, *CRYPTERS: Localizando firmas de los antivirus*, September 2013. [Last Access: September 28 2016].
- [17] Rad B., Masrom M., Ibrahim S. (2012), Camouflage in malware: from encryption to metamorphism, *International Journal of Computer Science and Network Security*, 12, 74-83, 2012.
- [18] Vinod P., Jaipur R., Laxmi R., Gaur M. (2009), Survey on malware detection methods, *Proceedings of the 3rd Hackers? Workshop on Computer and Internet Security*, 74-79, 2009.
- [19] You I., Yim K. (2010), Malware obfuscation techniques: A brief survey, *Proceedings of the 2010 International Conference on Broadband, Wireless Computing, Communication and Applications*, 297-300, 2010. <http://dx.doi.org/10.1109/BWCCA.2010.85>
- [20] Zhang Q., Reeves D. (2007), Metaaware: Identifying metamorphic malware, *Computer Security Applications Conference, ACSAC 2007*, 411-420, 2007.

Asymptotically Unbiased Estimation of A Nonsymmetric Dependence Measure Applied to Sensor Data Analytics and Financial Time Series

A. Cațaron, R. Andonie, Y. Chueh

Angel Cațaron

Department of Electronics and Computers
Transilvania University of Brașov, Romania
cataron@unitbv.ro

Razvan Andonie

1. Department of Computer Science
Central Washington University, USA
2. Department of Electronics and Computers
Transilvania University of Brașov, Romania
*Corresponding author: andonie@cwu.edu

Yvonne Chueh

Department of Mathematics
Central Washington University, USA
chueh@cwu.edu

Abstract: A fundamental concept frequently applied to statistical machine learning is the detection of dependencies between unknown random variables found from data samples. In previous work, we have introduced a nonparametric unilateral dependence measure based on Onicescu's information energy and a kNN method for estimating this measure from an available sample set of discrete or continuous variables. This paper provides the formal proofs which show that the estimator is asymptotically unbiased and has asymptotic zero variance when the sample size increases. It implies that the estimator has good statistical qualities. We investigate the performance of the estimator for data analysis applications in sensor data analysis and financial time series.

Keywords: machine learning, sensor data analytics, financial time series, statistical inference, information energy, nonsymmetric dependence measure, big data analytics.

1 Introduction

Statistical machine learning is based on the strong assumption that we use a *representative* training set of samples to infer a model. In this case, we select a random sample of the population, perform a statistical analysis on this sample, and use these results as an estimation to the desired statistical characteristics of the population as a whole. The accuracy of the estimation depends on the representativeness of the data sample. We gauge the representativeness of a sample by how well its statistical characteristics reflect the probabilistic characteristics of the entire population. Many standard techniques may be used to select a representative sample set [16]. However, if we do not use expert knowledge, selecting the most representative training set from a given dataset was proved to be computationally difficult (NP-hard) [10].

The problem is actually more complex, since in most applications the complete dataset is either unknown or too large to be analyzed. Therefore, we have to rely on a more or less representative training set. For example, a common statistical machine learning problem is to estimate information theory measures (such as entropy) from available training sets. This can be

reduced to the construction of an estimate of a density function from the observed data [21]. We refer here only to nonparametric estimation, where less rigid assumptions will be made about the distribution of the observed data. Although it will be assumed that the distribution has the probability density f , the data will be allowed to speak for themselves in determining the estimate of f more than would be the case if f were constrained to fall in a given parametric family.

The estimation of information theory measures has an important application area - the detection of dependency relationships between unknown random variables represented by data samples. There are two information theory strategies one can adopt when studying the relationship between two random variables: the first is to measure their interdependence thought as a mutual attribute and the second is to measure how much one system depends on the other. In the first case we have symmetric (bilateral) measures of dependence, whereas in the second one we have nonsymmetric (unilateral) measures. The literature review summarizes these two strategies as follows.

Strategy I. Several symmetric dependence measure were proposed (see [20]). Among them, the Shannon entropy based mutual information (MI), $MI(X, Y) = H(X) + H(Y) - H(X, Y) = MI(Y, X)$, which measures the information interdependence between two random variables X and Y . Estimating MI is known to be a non-trivial task [3]. Naïve estimations (which attempt to construct a histogram where every point is the center of a sampling interval) are plagued with both systematic (bias) and statistical errors. Other more sophisticated estimation methods were proposed [19], [23], [14]: kernel density estimators, B-spline estimators, k -th nearest neighbor (kNN) estimators, and wavelet density estimators. An ideal estimator does not exist, instead the choice of the estimator depends on the structure of data to be analyzed. It is not possible to design an estimator that minimizes both the bias and the variance to arbitrarily small values. The existing studies have shown that there is always a delicate trade off between the two types of errors [3].

Strategy II. Several continuous entropy-like nonsymmetric dependence measures have been proposed [15]. But information measures can also refer to certainty (not only to uncertainty, like Shannon's entropy), and probability can be considered as a measure of certainty. For instance, Onicescu's information energy (IE) was interpreted by several authors as a measure of expected commonness, a measure of average certainty, or as a measure of concentration. The second strategy can be illustrated by a unilateral dependence measure defined for discrete random variables by Andonie *et al.* [2]: $o(X, Y) = IE(X|Y) - IE(X)$, where $IE(X) = \sum_{k=1}^n p_k^2$ is Onicescu's information energy of a discrete random variable X with probabilities p_k , and $IE(X|Y)$ is the conditional information energy between two variables, as defined in [18].

For a continuous random variable X with probability density function $f(x)$, the IE is [11, 18]:

$$IE(X) = E(f(X)) = \int_{-\infty}^{+\infty} f(x)f(x)dx = \int_{-\infty}^{+\infty} f^2(x)dx \quad (1)$$

where X is a random variable with probability density function $f(x)$. In other words, $IE(X)$ is the expectation of the values of a density function f . For two continuous random variables X and Y , with their joint probability density function $f(x, y)$, we introduced the continuous version of the $o(X, Y)$ measure in [1, 8]:

$$o(X, Y) = IE(X|Y) - IE(X) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x, y)f(x|y)dy dx - \int_{-\infty}^{+\infty} f^2(x)dx \quad (2)$$

The discrete and the continuous $o(X, Y)$ measure the "additional" average certainty (or information) of X occurring under the condition that Y has already or simultaneously occurred

(or is certain) “over” the average certainty of X when the certainty (or information) of Y is not available. Thus, $o(X, Y)$ can be regarded as an indicator of the unilateral dependence of X upon Y . As a unilateral measure, $o(X, Y)$ gives a different insight than the MI into the dependencies between pairs of random variables.

The $o(X, Y)$ measure can easily be computed if the data sample is extracted from known parametric probability distributions. When the underlying distribution of data sample is unknown, the $o(X, Y)$ has to be estimated. More formally, we have to estimate $o(X, Y)$ from a random sample x_1, x_2, \dots, x_n of n d -dimensional observations from a distribution with the unknown probability density $f(x)$. This problem is even more difficult if the number of available points is small.

Contributions. In previous work [4, 5], we introduced a kNN method for estimating IE and $o(X, Y)$ from data samples of discrete or continuous variables. In the present paper, we give a more detailed description of this method. For the first time, we provide the formal proofs showing that the estimator is asymptotically unbiased and has asymptotic zero variance when the sample size increases. This implies that the estimator has good statistical qualities. The potential application of our results in statistical machine learning for multivariate data drives the motivation for the present work and we refer to two real world applications. The first one is a study of the unilateral interactions between temperature sensor data in a refrigerator room. The second one is an analysis of the dynamics of interaction between assets and liabilities evidenced by the historical event that matches the time series formed by the unilateral dependence measures observed from the financial time series reported by Kodak and Apple. This evolution is captured by our unilateral dependence measure over time.

The rest of the paper is structured as follows. In Section 2 we summarize our previous work, we review the kNN estimation method for the continuous $o(X, Y)$ measure, and we derive our IE estimator in a d -dimensional space. In Section 3 we analyze and prove the asymptotic unbiased behavior of this IE estimator. In Section 4 we derive the $o(X, Y)$ estimator for the purpose of measuring the unilateral dependence relation between a pair of random variables. Section 5 discusses applications and Section 6 contains our final remarks.

2 Background

Fitting multi-dimensional data to joint density functions is challenging. We aim to summarize the kNN estimation technique which can be used for this problem. We will refer to our previous results on IE and $o(X, Y)$ kNN estimation.

2.1 kNN estimation of the unilateral dependence measure in an unidimensional space

Our goal is to approximate IE and $o(X, Y)$ from the available dataset, for random variables X and Y with unknown densities, using the kNN method. The kNN estimators represent an attempt to adapt the amount of smoothing to the “local” density of data. The degree of smoothing is controlled by an integer k , chosen to be considerably smaller than the sample size. We define the distance $R_{i,k,n}$ between two points on the line as $|x_i - x_k|$ out of n available points, and for each x_i we define the distances $R_{i,1,n} \leq R_{i,2,n} \leq \dots \leq R_{i,n,n}$, arranged in ascending order, from x_i to the points of the sample.

The kNN density estimate $f(x_i)$ is defined by [21]:

$$\hat{f}(x_i) = \frac{k}{2nR_{i,k,n}}$$

The kNN was used for non-parametrical estimate of the entropy based on the k -th nearest neighbor distance between n points in a sample, where k is a fixed parameter and $k \leq n - 1$. Based on the first nearest neighbor distances, Leonenko *et al.* [13] introduced an asymptotic unbiased and consistent estimator H_n of the entropy $H(f)$ in a multidimensional space. When the sample points are very close one to each other, small fluctuations in their distances produce high fluctuations of H_n . In order to overcome this problem, Singh *et al.* [22] defined an entropy estimator based on the k -th nearest neighbor distances. A kNN estimator of the Kullback-Leibler divergence was introduced by Wang *et al.* [24]. Faivishevsky *et al.* used a mean of several kNN estimators, corresponding to different values of k , to increase the smoothness of the estimation [9].

The kNN method works well if the value of k is optimally chosen, and it outperforms histogram methods [23]. However, there is no model selection method for determining the number of nearest neighbors k , and this is a known limitation.

2.2 kNN estimation of the IE and the $o(X, Y)$ measure

In [5], we introduced a kNN IE estimator, and we stated in [4], without mathematical proofs, that this estimator is consistent. By definition [12], a statistic whose mathematical expectation is equal to a parameter is called *unbiased* estimator of that parameter. Otherwise, the statistic is said to be *biased*. A statistical estimator that converges asymptotically to a parameter is called *consistent* estimator of that parameter. In the following, we will review this result, introducing also the basic mathematical notations used in Section 3.

The IE is the average of $f(x)$, therefore we have to estimate $f(x)$. The n observations from our samples have the same probability $\frac{1}{n}$. A convenient estimator of the IE is:

$$\hat{IE}_k^{(n)}(X) = \frac{1}{n} \sum_{i=1}^n \hat{f}(x_i). \quad (3)$$

The mass probability at value x_i determined by its k nearest surrounding points and standardized by the sphere volume $V_k(x_i)$ these k nearest observations occupy measures the density of probability at value x_i . The k -th nearest neighbor estimate is defined by [4], [21]:

$$\hat{f}(x_i) = \frac{\frac{k}{n}}{V_k(x_i)}, i = 1 \dots n \quad (4)$$

where $V_k(x_i)$ is the volume of the d -dimensional sphere centered in x_i , with the radius $R_{i,k,n}$ measuring the Euclidean distance from x_i to the k -th nearest data point. The estimate (4) can be re-written as:

$$\hat{f}(x_i) = \frac{k}{nV_d R_{i,k,n}^d}, i = 1 \dots n. \quad (5)$$

given $V_k(x_i) = V_d R_{i,k,n}^d$ where V_d is the radius of the unit sphere in d -dimensions, that is $V_1 = 2$, $V_2 = \pi$, $V_3 = \frac{4\pi}{3}$, etc.

By substituting $\hat{f}(x_i)$ in (3), we finally obtain the following IE approximation:

$$\hat{IE}_k^{(n)}(f) = \frac{1}{n} \sum_{i=1}^n \frac{k}{nV_d R_{i,k,n}^d}. \quad (6)$$

In [6], we showed how to estimate the $o(X, Y)$ dependence measure from an available sample set of discrete or continuous variables and we experimentally compared the results of the kNN estimator with the naïve histogram estimator.

3 Asymptotic behavior of the IE approximator

Consistency of an estimator means that as the sample size gets large the estimate gets closer and closer to the true value of the parameter. Unbiasedness is a statement about the expected value of the sampling distribution of the estimator. The ideal situation, of course, is to have an unbiased consistent estimator. This may be very difficult to achieve. Yet unbiasedness is not essential, and just a little bias is permitted as long as the estimator is consistent. Therefore, an asymptotically unbiased consistent estimator may be acceptable.

In [4], we stated (without proofs) that the IE approximator is asymptotically unbiased and consistent. In this section, we provide the formal proofs for this statement.

Lemma 1. *Let us consider the identically distributed random variables $T_1^{(n)}, T_2^{(n)}, \dots, T_n^{(n)}$ defined by*

$$T_i^{(n)} = \frac{k}{nV_d R_{i,k,n}^d}. \tag{7}$$

For a given x , the random variable T_x has the following probability density function:

$$h_{T_x}(y) = \frac{f(x) \left[\frac{k}{y} f(x) \right]^k}{y^2(k-1)!} e^{-\frac{k}{y} f(x)}, 0 < y < \infty.$$

Proof: See Appendix 6. □

Theorem 2 (Asymptotically Unbiasedness). *The information energy estimator $\hat{IE}_k^{(n)}(f)$ is asymptotically unbiased.*

Proof:

We will find the asymptotic value of the average:

$$\lim_{n \rightarrow \infty} E \left[\hat{IE}_k^{(n)}(f) \right].$$

From Lemma 1, for a given x , the random variable T_x has the pdf:

$$h_{T_x}(y) = \frac{f(x) \left[\frac{k}{y} f(x) \right]^k}{y^2(k-1)!} e^{-\frac{k}{y} f(x)}, 0 < y < \infty.$$

The conditional expectation of $T_1^{(n)} | X_1 = x$, given the fixed center $X_1 = x$ is:

$$\begin{aligned} \lim_{n \rightarrow \infty} E \left[T_1^{(n)} | X_1 = x \right] &= \int_0^\infty y h(y) dy \\ &= \int_0^\infty \frac{f(x) (k f(x))^k}{(k-1)!} \cdot y^{-1-k} e^{-\frac{k f(x)}{y}} dy \\ &= \frac{f(x) (k f(x))^k}{(k-1)!} \int_0^\infty y^{-1-k} e^{-\frac{k f(x)}{y}} dy. \end{aligned}$$

We make the following change of variable: $u = \frac{k f(x)}{y}$, with $y \in (0, +\infty)$. The corresponding $u \in (+\infty, 0)$ is decreasing. We have:

$$du = -\frac{1}{y^2} k f(x) dy$$

and

$$y^{-1-k} = \left(\frac{u}{kf(x)} \right)^{k+1}.$$

We obtain:

$$\int_0^\infty yh(y)dy = \frac{f(x)(kf(x))^k}{(k-1)!} \int_\infty^0 -u^{k+1}e^{-u}du \cdot \frac{1}{(kf(x))^{k+1}} \left(\frac{kf(x)}{u} \right)^2 \cdot \frac{1}{kf(x)}.$$

The negative sign before u^{k+1} allows us to change the limits of the integral:

$$\begin{aligned} \int_0^\infty yh(y)dy &= \frac{f(x)(kf(x))^k}{(k-1)!} \int_0^\infty u^{k-1}e^{-u}du \cdot \frac{1}{(kf(x))^k} \\ &= \frac{f(x)}{(k-1)!} \int_0^\infty u^{k-1}e^{-u}du = \frac{f(x)}{(k-1)!} \Gamma(k) = f(x). \end{aligned}$$

We take the complete expectation of the above, then we have:

$$\lim_{n \rightarrow \infty} E \left[E \left[T_1^{(n)} | X_1 = x \right] \right] = \lim_{n \rightarrow \infty} E \left[T_1^{(n)} \right] = E[f(x)].$$

Let us define $\hat{G}^{(n)}(f) = \frac{1}{n} \sum_{i=1}^n T_i^{(n)}$, the unbiased estimator of $f(x)$ (like in [22]). We obtain:

$$E \left[\hat{G}^{(n)}(f) \right] = \frac{1}{n} \sum_{i=1}^n E \left[T_i^{(n)} \right] = \frac{1}{n} [E[f(x)] + \dots + E[f(x)]] = E[f(x)].$$

$$\lim_{n \rightarrow \infty} E \left[\hat{G}^{(n)}(f) \right] = E[f(x)] = \int f(x) \cdot f(x) dx = IE(f).$$

This proves that our IE estimator is asymptotically unbiased. □

Next, we aim to prove that it has asymptotic zero variance.

Lemma 3. *Given the identically distributed random variables $T_1^{(n)}, T_2^{(n)}, \dots, T_n^{(n)}$ defined by (7), we have:*

$$\lim_{n \rightarrow \infty} \frac{1}{n} E \left[T_1^{(n)2} \right] - \lim_{n \rightarrow \infty} \frac{1}{n} E \left[T_1^{(n)} \right]^2 = 0.$$

Proof: See Appendix 6. □

Lemma 4. *Given the identically distributed random variables $T_1^{(n)}, T_2^{(n)}, \dots, T_n^{(n)}$ defined by (7), the following relation is true:*

$$\lim_{n \rightarrow \infty} \frac{n(n-1)}{n^2} \left(E \left[T_1^{(n)} T_2^{(n)} \right] - E \left[T_1^{(n)} \right] E \left[T_2^{(n)} \right] \right) = 0.$$

Proof: See Appendix 6. □

Theorem 5 (Asymptotic Zero Variance).

$$\lim_{n \rightarrow \infty} Var \left[\hat{G}_k^{(n)}(f) \right] = 0.$$

Proof: $T_1^{(n)}, T_2^{(n)}, \dots, T_n^{(n)}$ are identically distributed.

$$\begin{aligned} \text{Var} \left[\hat{G}_k^{(n)}(f) \right] &= \frac{1}{n^2} \left[\sum_{i=1}^n \text{Var} \left[T_i^{(n)} \right] + \sum_{\substack{i \neq j \\ i < j}} 2\text{Cov} \left(T_i^{(n)}, T_j^{(n)} \right) \right] \\ &= \frac{1}{n} \text{Var} \left[T_1^{(n)} + \frac{n(n-1)}{n^2} \right] \text{Cov} \left(T_1^{(n)}, T_2^{(n)} \right). \end{aligned}$$

$$\begin{aligned} \lim_{n \rightarrow \infty} \text{Var} \left[\hat{G}_k^{(n)}(f) \right] &= \lim_{n \rightarrow \infty} \frac{1}{n} E \left[T_1^{(n)^2} \right] - \lim_{n \rightarrow \infty} \frac{1}{n} E \left[T_1^{(n)} \right]^2 \\ &\quad + \lim_{n \rightarrow \infty} \frac{n(n-1)}{n^2} \left(E \left[T_1^{(n)} T_2^{(n)} \right] - E \left[T_1^{(n)} \right] E \left[T_2^{(n)} \right] \right). \end{aligned}$$

From Lemma 3 and Lemma 4, we obtain:

$$\lim_{n \rightarrow \infty} \text{Var} \left[\hat{G}_k^{(n)}(f) \right] = 0.$$

□

Our main result is synthesized in:

Theorem 6 (Consistency). *The information energy estimator $\hat{IE}_k^{(n)}(f)$ is consistent.*

Proof: This results from Theorems 2 and 5, and the following property: An asymptotically unbiased estimator with asymptotic zero variance is consistent [17]. □

4 The kNN $o(X, Y)$ estimator

Our goal is to infer $o(X, Y)$ from the random samples x_1, x_2, \dots, x_n . We will use the results from Section 2.2 to deduct the kNN estimator of $o(X, Y)$.

First, we substitute $\widehat{IE}_k^{(n)}(X)$ from eq. (6) in eq. (2):

$$\hat{o}(X, Y) = \widehat{IE}_k^{(n)}(X|Y) - \widehat{IE}_k^{(n)}(X) \tag{8}$$

where:

$$\widehat{IE}_k^{(n)}(X|Y) = \sum_{j=1}^m \hat{f}(y_j) \widehat{IE}_k^{(n)}(X|y_j) \tag{9}$$

and

$$\widehat{IE}_k^{(n)}(X) = \frac{1}{n} \sum_{i=1}^n \frac{k_1}{nV_{d_1(X)}R_i^{d_1}}, \tag{10}$$

is an adaptation of eq. (6).

We can write:

$$\widehat{IE}_k^{(n)}(X|y_j) = \frac{1}{n} \sum_{i=1}^n \hat{f}(x_i|y_j) = \frac{1}{n} \sum_{i=1}^n \frac{\hat{f}(x_i, y_j)}{\hat{f}(y_j)} = \frac{1}{n} \sum_{i=1}^n \frac{\hat{f}(x_i, y_j)}{\hat{f}(y_j)}, \quad (11)$$

where

$$\hat{f}(x_i|y_j) = \frac{\hat{f}(x_i, y_j)}{\hat{f}(y_j)}. \quad (12)$$

From (9) and (11) we can write:

$$\widehat{IE}_k^{(n)}(X|Y) = \sum_{j=1}^m \hat{f}(y_j) \frac{1}{n} \sum_{i=1}^n \frac{\hat{f}(x_i, y_j)}{\hat{f}(y_j)} = \frac{1}{n} \sum_{j=1}^m \sum_{i=1}^n \hat{f}(x_i, y_j).$$

The estimate of the joint probability density can be written as:

$$\hat{f}(x_i, y_j) = \frac{k_2}{pV_{d_2(X,Y)}R_{i,j}^{d_2}},$$

where p is the number of pairs (x_i, y_j) , then re-write the eq. (9) as:

$$\widehat{IE}_k^{(n)}(X|Y) = \frac{k_2}{npV_{d_2(X,Y)}} \sum_{j=1}^m \sum_{i=1}^n \frac{1}{R_{i,j}^{d_2}}.$$

R_i is the Euclidean distance between the reference point x_i and its k_1^{th} nearest neighbor, when the points are drawn from the one-dimensional probability distribution $f(x)$: $R_i = \|x_i - x_{i,k_1}\|$. Similarly, R_j is the Euclidean distance between the reference point y_j and its k_1^{th} nearest neighbor, when the points are drawn from the one-dimensional probability distribution $f(Y)$: $R_j = \|y_j - y_{j,k_1}\|$. Then, R_{ij} is the Euclidean distance between the reference point (x_i, y_j) and its k_2^{th} nearest neighbor, when the points are drawn from the joint probability distribution $f(X, Y)$: $R_{ij} = \sqrt{(x_{ij} - x_{ij,k_2})^2 + (y_{ij} - y_{ij,k_2})^2}$.

The estimate of $o(X, Y)$ is:

$$\hat{o}(X, Y) = \frac{k_2}{npV_{d_2(X,Y)}} \sum_{j=1}^m \sum_{i=1}^n \frac{1}{R_{i,j}^{d_2}} - \frac{k_1}{n^2V_{d_1(X)}} \sum_{i=1}^n \frac{1}{R_i^{d_1}}. \quad (13)$$

Although we do not have a general method to set the nearest neighbor parameter, Silverman [21] suggested that k should be proportional to

$$n^{4/(d+4)}. \quad (14)$$

In our case, the optimal values of k_1 and k_2 may not be equal, because the these two parameters refer to different samples. The R code for computing the estimate of $o(X, Y)$ is publicly available on GitHub¹.

5 Applications

There are tremendous opportunities for applications using $o(X, Y)$ estimation to analyze potential dependencies between data represented by samples. We illustrate with an example with numerical implementations and two applications using real-world data, all extracted from our previous work and summarized here.

¹<https://github.com/caçaron/information-energy>

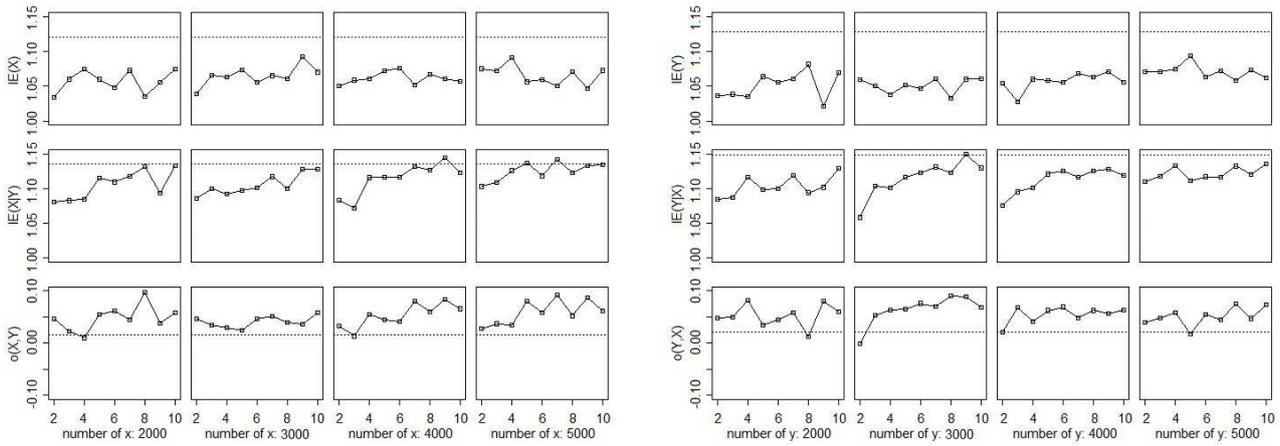


Figure 1: The kNN estimated values, where k was determined with formula (14). The theoretical values $IE(X) = 1.12$, $IE(X|Y) = 1.1351$, $o(X, Y) = 0.0151$, $IE(Y) = 1.128$, $IE(Y|X) = 1.14787$, $o(Y, X) = 0.01987$ are marked with dashed lines. The estimates converge towards their theoretical values for an increasing number of samples, under various k values (the horizontal axis): 2, 4, 6, 8, 10.

5.1 A specified joint probability distribution

Let us illustrate with the following joint probability density function from [6]:

$$f_{X,Y}(x, y) = \frac{6}{5} (x + y^2), x \in [0, 1], y \in [0, 1], \tag{15}$$

which has the marginal probability density functions

$$f_X(x) = \int_0^1 \frac{6}{5} (x + y^2) dy = \frac{6}{5} \left(x + \frac{1}{3} \right) \tag{16}$$

and

$$f_Y(y) = \int_0^1 \frac{6}{5} (x + y^2) dx = \frac{6}{5} \left(\frac{1}{2} + y^2 \right). \tag{17}$$

The conditional probability density function is:

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x, y)}{f_Y(y)} = \frac{x + y^2}{\frac{1}{2} + y^2},$$

and the theoretical value of $o(X, Y)$ is:

$$o(X, Y) = IE(X|Y) - IE(X)$$

$$IE(X) = \int_0^1 f^2(x) dx = 1.12$$

$$IE(X|Y) = \int_0^1 \int_0^1 f_{X,Y}(x, y) f_{X|Y}(x|y) dx dy = 1.1351$$

and

$$o(X, Y) = 1.1351 - 1.12 = 0.0151.$$

Similarly, we obtain $IE(Y) = 1.128$, $IE(Y|X) = 1.14787$, and $o(Y, X) = 0.01987$. We observe that $o(X, Y) < o(Y, X)$. The theoretical values of $IE(X)$, $IE(Y)$, $IE(X|Y)$, $IE(Y|X)$, $o(X, Y)$, and $o(Y, X)$ are represented in Fig. 1 with dashed lines. Their kNN estimates are represented with continuous lines, for an increasing number of data samples.

5.2 Sensor data

When studying the interaction between two random variable, why is a unilateral dependence measure useful, and why do we not simply use the well-known MI? Let us consider two sets of experiments, characterized respectively by random variables X and Y . The experiments run simultaneously and interact probabilistically. Our question is which variable influences probabilistically more the other one. Thus, X can be viewed as $X|Y$ and Y can be viewed as $Y|X$. While the correlation quantifies linear dependence and MI describes the degree of interdependence between two random variables, $o(X, Y)$ helps us understand which random variable, X or Y , has a higher influence on the other one. If both X and $X|Y$ are available, we can estimate $IE(X|Y)$ as well as $o(X, Y)$, and similarly for Y and $Y|X$.

When the data is acquired from real world experiments or from simulators, we need to store series of values for X and Y featuring the two phenomena running independently, as well as measurements of values generated by the two phenomena running simultaneously, in order to capture $X|Y$ and $Y|X$. Moreover, the precision of the $IE(X|Y)$ and $o(X, Y)$ estimators increase when more values of $X|Y$ are available for each value of Y .

In our application, which can be found in [6], we study the evolution of the temperature measured at the surface of two packages, X and Y , introduced in a refrigerated room. When one package at the outside temperature is placed in the room, it is getting cold under the influence of the air conditioning devices. The situation changes when a second package is placed in the room because it will change the ambient conditions and the temperature may decrease slower. Emulating the real world circumstances, we are able to measure the temperatures with various experimental setups. When we only have one package in the room, we measure the time series of temperatures as samples of an unconditional random variables. When we have both packages in the room, we measure simultaneously the time series of temperatures as conditional random values.

According to these experiments, we can assess if package X has a stronger influence on package Y , or vice versa. This might lead to the decision to remove the sensor which is more influenced by the other.

5.3 Financial time series

Studying the relationship between stocks is a very interesting business case for finance. This is just an example which can be extended to all kind of financial parameters of companies. In [7] we have presented an application using the public available data offered by Kodak and Apple for the period 1999–2014. We analyzed the unilateral dependence between the Kodak and Apple series in order to understand how they influence one to each other. We aim to describe in the following our general approach for detecting unilateral dependencies between time series variables, without specifically referring to the Kodak-Apple case.

The information energy and the unilateral dependence measure both are the result of two random variables. The estimation becomes more precise when the number of observations is high. When we analyze time series, we consider the series as a set of n observations $x_i, i = 1 \dots n$ of the random variable X . Our goal is to find synchronous relations between two time series which have been considered as two samples of two random variables X and Y . A point x_i from the first series was paired with a point y_i from the second one producing a joint observation, with

$i = 1, \dots, n$. Therefore, from two time series with n points each we obtain n pairs (x_i, y_i) . The unilateral dependence $o(X, Y)$ between X and Y can be estimated with formula (13) based on X and (X, Y) , while the unilateral dependence $o(Y, X)$ can be estimated with the same formula based on Y and (Y, X) . We note that the set of pairs (X, Y) is equal with the set of pairs (Y, X) since for each observation x_i exactly one observation y_i is used.

It is interesting to study the evolution of the unilateral dependence in time. The values $o(X, Y)$ and $o(Y, X)$ can be estimated at the moment of time t_m using the history of the m observations of x_i and y_i between the initial moments t_0 and t_m . A new set of two observations x_{m+1} and y_{m+1} allow us to re-estimate $o(X, Y)$ and $o(Y, X)$.

Our model captures not only the unilateral dependencies between two simultaneous time series, but also how these dependencies evolve in time, which can be valuable and insightful for forecast or investment in financial research or applications.

6 Conclusion

We introduced a non-parametric asymptotically unbiased and consistent estimator of the IE, and the unilateral dependency measure $o(X, Y)$ between random variables X and Y . We estimated $o(X, Y)$ from available data samples using the kNN technique. In our applications, we showed how the nonsymmetric dependence measure can provide information that cannot be expressed by a symmetric measure. The examples illustrated in this paper are all one-dimensional random variables, for the purpose of simplicity. The kNN estimation can be also applied on multi-dimensional variables. For instance, the $o(X, Y)$ measure can be used to analyze the unilateral dependence between bi-variate time-series data acquired from various fields.

Because of the mathematical properties proved here, the $o(X, Y)$ estimator works the best for large data sets, so it is suitable for big data analytics. Our method can be applied to both continuous and discrete variable spaces, meaning that we can use it both in classification and regression problems.

Bibliography

- [1] Andonie R., Cațaron A. (2004), An informational energy LVQ approach for feature ranking, *European Symposium on Artificial Neural Networks 2004, pages In d-side publications*, 471–476, 2004.
- [2] Andonie R. (1986), Interacting systems and informational energy, *Foundation of Control Engineering*, 11, 53–59, 1986.
- [3] Bonachela J.A., Hinrichsen H., Miguel A. Munoz M.A. (2008), Entropy estimates of small data sets, *MATH.THEOR.*, 41(20), 1-20, 2008.
- [4] Cațaron A., Andonie R., Chueh Y. (2013), Asymptotically unbiased estimator of the informational energy with kNN, *International Journal of Computers Communications & Control*, 8(5), 689–698, 2013.
- [5] Cațaron A., Andonie R. (2012), How to infer the informational energy from small datasets, *Optimization of Electrical and Electronic Equipment (OPTIM), 2012 13th International Conference on*, 1065–1070, 2012.
- [6] Cațaron A., Andonie R., Chueh Y. (2014), kNN estimation of the unilateral dependency measure between random variables, *2014 IEEE Symposium on Computational Intelligence and Data Mining, (CIDM 2014)*, Orlando, FL, USA, 471–478, 2014.

-
- [7] Caçaron A., Andonie R., Chueh Y. (2015), Financial data analysis using the informational energy unilateral dependency measure, *Proceedings of the International Joint Conference on Neural Networks, (IJCNN 2015)*, Killarney, Ireland, 1-8, 2015.
- [8] Chueh Y., Caçaron A., Andonie R. (2016), Mortality rate modeling of joint lives and survivor insurance contracts tested by a novel unilateral dependence measure, *2016 IEEE Symposium Series on Computational Intelligence, SSCI 2016*, Athens, Greece, December 6-9, 2016, 1–8, 2016.
- [9] Faivishevsky L., Goldberger J. (2008), ICA based on a smooth estimation of the differential entropy, *NIPS*, 1-8, 2008.
- [10] Gamez J.E., Modave F., Kosheleva O. (2008), Selecting the most representative sample is NP-hard: Need for expert (fuzzy) knowledge, *Fuzzy Systems, 2008. FUZZ-IEEE 2008. (IEEE World Congress on Computational Intelligence). IEEE International Conference on*, 1069–1074, 2008.
- [11] Guiaşu S. (1977), *Information theory with applications*, McGraw Hill, New York, 1977.
- [12] Hogg R.V., McKean J., Allen T. Craig A.T. (2006), *Introduction To Mathematical Statistics, 6/E*, Pearson Education, 2006.
- [13] Kozachenko L. F., Leonenko N. N. (1987), Sample estimate of the entropy of a random vector, *Probl. Peredachi Inf.*, 23(2), 9–16, 1987.
- [14] Kraskov A., Stögbauer H., Grassberger P. (2004), Estimating mutual information, *Phys. Rev. E*, 69, 1–16, 2004.
- [15] Li H. (2015), On nonsymmetric nonparametric measures of dependence, arXiv:1502.03850, 2015.
- [16] Lohr H. (1999), *Sampling: Design and Analysis*, Duxbury Press, 1999.
- [17] Miller M., Miller M. (2003), *John E. Freund's mathematical statistics with applications*, Pearson/Prentice Hall, Upper Saddle River, New Jersey, 7th edition, 2003.
- [18] Onicescu O. (1966), Theorie de l'information. Energie informationelle, *C. R. Acad. Sci. Paris, Ser. A-B*, 263, 841–842, 1966.
- [19] Paninski L. (2003), Estimation of entropy and mutual information, *Neural Comput.*, 15, 1191–1253, 2003.
- [20] Schweizer B., Wolff E. F. (1981), On nonparametric measures of dependence for random variables, *Ann. Statist.*, 9:879–885, 1981.
- [21] Silverman B.W. (1986), *Density Estimation for Statistics and Data Analysis (Chapman & Hall/CRC Monographs on Statistics & Applied Probability)*, Chapman and Hall/CRC, 1986.
- [22] Singh H., Misra N., Hnizdo V., Fedorowicz A., Demchuk E. (2003), Nearest neighbor estimates of entropy, *American Journal of Mathematical and Management Sciences*, 23, 301–321, 2003.
- [23] Walters-Williams J., Li Y. (2009), Estimation of mutual information: A survey, *Proceedings of the 4th International Conference on Rough Sets and Knowledge Technology*, Springer-Verlag, Berlin, Heidelberg, 389–396, 2009.

- [24] Wang Q., Kulkarni S. R., Verdu S. (2006), A nearest-neighbor approach to estimating divergence between continuous random vectors, *Proc. of the IEEE International Symposium on Information Theory*, Seattle, WA, 242-246, 2006.

Appendix

Proof of Lemma 1

The random variables $T_1^{(n)}, T_2^{(n)}, \dots, T_n^{(n)}$ are identically distributed, therefore:

$$E \left[\hat{I}E_k^{(n)}(f) \right] = E \left[T_1^{(n)} \right].$$

Let us denote $\rho_{r,n} = \left(\frac{k}{nV_d r} \right)^{\frac{1}{d}}$, with $\rho_{r,n} \rightarrow 0$ when $n \rightarrow \infty$. For a real number r we have:

$$P \left[T_1^{(n)} > r | X_1 = x \right] = P \left[\rho_{r,n} > R_{1,k,n} | X_1 = x \right], \quad (18)$$

because:

$$T_1^{(n)} > r \Leftrightarrow \frac{k}{nV_d R_{1,k,n}^d} > r \Rightarrow \frac{k}{nV_d r} > R_{1,k,n}^d \Rightarrow \left(\frac{k}{nV_d r} \right)^{\frac{1}{d}} > R_{1,k,n}$$

We write the probability $P \left[T_1^{(n)} > r | X_1 = x \right]$ as a binomial distribution, and approximate it with the Poisson probability function.

The binomial distribution is given by $f(k; n, p) = \binom{n}{k} p^k (1-p)^{n-k}$, and the binomial distribution of the cumulative distribution function is $P(X \leq x) = \sum_{x_i \leq x} P(X = x_i) = \sum_{x_i \leq x} f(x)$. The probability expressed by the Poisson formula is $P(x) \approx \frac{e^{-np} (np)^x}{x!}$, where p is the successful probability out of n samples.

Using the Poisson approximation for the binomial distribution by letting the sample size $n \rightarrow \infty$ we get:

$$\begin{aligned} P[\rho_{r,n} > R_{1,k,n} | X_1 = x] &= 1 - P[R_{1,k,n} \geq \rho_{r,n}] \\ &= 1 - \sum_{i=1}^k \binom{n-1}{i} [P(S_{\rho_{r,n}})(1 - P(S_{\rho_{r,n}}))]^{n-i-1} = P[T_x > r]. \end{aligned}$$

We note that:

$$\begin{aligned} P(S_{\rho_{r,n}}) &= \int_{S_{\rho_{r,n}}} f(t) dt, \\ \rho_{r,n} &= \left(\frac{k}{nV_d r} \right)^{\frac{1}{d}}, \\ V_\rho &= V_d \rho^d = V_d \frac{k}{nV_d r} = \frac{k}{nr} \Rightarrow n = \frac{k}{rV_\rho}. \end{aligned}$$

Let $n \rightarrow \infty$, then:

$$\lim_{n \rightarrow \infty} nP(S_{\rho r, n}) = \frac{k}{r} \lim_{n \rightarrow \infty} \frac{P(S_\rho)}{V_\rho} = \frac{k}{r} f(x).$$

Thus, from the Poisson approximation we obtain:

$$P[T_x > r] = 1 - \sum_{i=0}^k \frac{\left(\frac{k}{r} f(x)\right)^i}{i!} e^{-\frac{k}{r} f(x)}.$$

To calculate the pdf for random variable T_x , we differentiate its cumulative density function with respect to r :

$$\begin{aligned} \frac{d}{dr} P[T_x \leq r] &= \frac{d}{dr} [1 - P[T_x > r]] = \frac{d}{dr} \left[1 - \left(1 - \sum_{i=0}^k \frac{\left[\frac{k}{r} f(x)\right]^i}{i!} e^{-\frac{k}{r} f(x)} \right) \right] \\ &= \frac{d}{dr} \sum_{i=0}^k \frac{\left[\frac{k}{r} f(x)\right]^i}{i!} e^{-\frac{k}{r} f(x)} = \frac{d}{dr} \left[e^{-\frac{k}{r} f(x)} + \sum_{i=1}^k k \frac{\left[\frac{k}{r} f(x)\right]^i}{i!} e^{-\frac{k}{r} f(x)} \right] \\ &= \frac{k f(x)}{r^2} e^{-\frac{k}{r} f(x)} + \sum_{i=1}^k \frac{\left[\frac{k}{r} f(x)\right]^{i-1}}{(i-1)!} \left(-\frac{k f(x)}{r^2} \right) e^{-\frac{k}{r} f(x)} + \sum_{i=1}^k \frac{\left[\frac{k}{r} f(x)\right]^i}{i!} \cdot \frac{k f(x)}{r^2} e^{-\frac{k}{r} f(x)} \\ &= \frac{k f(x)}{r^2} e^{-\frac{k}{r} f(x)} - \sum_{i=1}^k \frac{k f(x)}{r^2} \cdot \frac{\left[\frac{k}{r} f(x)\right]^{i-1}}{(i-1)!} e^{-\frac{k}{r} f(x)} + \sum_{i=1}^k \frac{k f(x)}{r^2} \cdot \frac{\left[\frac{k}{r} f(x)\right]^i}{i!} e^{-\frac{k}{r} f(x)} \\ &= \frac{k f(x)}{r^2} e^{-\frac{k}{r} f(x)} - \frac{k f(x)}{r^2} e^{-\frac{k}{r} f(x)} + \frac{k f(x)}{r^2} \cdot \frac{\left[\frac{k}{r} f(x)\right]^k}{k!} e^{-\frac{k}{r} f(x)} = \frac{e^{-\frac{k}{r} f(x)} f(x) \left[\frac{k}{r} f(x)\right]^k}{r^2 (k-1)!}. \end{aligned}$$

Then, for a given x , the random variable T_x has the pdf:

$$h_{T_x}(y) = \frac{f(x) \left[\frac{k}{y} f(x)\right]^k}{y^2 (k-1)!} e^{-\frac{k}{y} f(x)}, 0 < y < \infty.$$

Proof of Lemma 3

We have:

$$\begin{aligned} \lim_{n \rightarrow \infty} E \left[T_1^{(n)^2} \mid X_1 = x \right] &= \int_0^\infty y^2 h(y) dy \\ &= \int_0^\infty \frac{f(x) (k f(x))^k}{(k-1)!} y^{-k} e^{-\frac{k f(x)}{y}} dy \\ &= \frac{f(x) (k f(x))^k}{(k-1)!} \int_0^\infty y^{-k} e^{-\frac{k f(x)}{y}} dy. \end{aligned}$$

We make the following substitution: $u = \frac{k f(x)}{y}$, with $y \in (0, +\infty)$, $u \in (+\infty, 0)$ decreasing, and $y = \frac{k f(x)}{u}$. Then $du = -\frac{1}{y^2} k f(x) dy$ and $y^{-k} = \left(\frac{u}{k f(x)}\right)^k$.

We obtain:

$$\begin{aligned} \lim_{n \rightarrow \infty} E \left[\left(T_1^{(n)} \right)^2 \mid X_1 = x \right] &= \frac{f(x) (kf(x))^k}{(k-1)!} \int_{\infty}^0 - \left(\frac{u}{kf(x)} \right)^k e^{-\frac{kf(x)}{y} y^2} \frac{1}{kf(x)} du \\ &= \frac{f(x) (kf(x))^k}{(k-1)!} \int_{\infty}^0 -u^{k-2} e^{-u} \left(\frac{1}{kf(x)} \right)^{k-1} du \\ &= \frac{f(x) (kf(x))^k}{(k-1)!} \int_0^{\infty} -u^{k-2} e^{-u} du \\ &= \frac{kf(x)^2}{(k-1)!} \Gamma(k-1) = \frac{k}{k-1} f(x)^2. \end{aligned}$$

$$\begin{aligned} \lim_{n \rightarrow \infty} E \left[\left(T_1^{(n)} \right)^2 \right] &= \lim_{n \rightarrow \infty} E \left[E \left[\left(T_1^{(n)} \right)^2 \mid X_1 = x \right] \right] \\ &= \lim_{n \rightarrow \infty} E \left[\frac{k}{k-1} f(x)^2 \right] \\ &= \frac{k}{k-1} E [f(x)^2] \\ &= \frac{k}{k-1} \left[Var [f(x)] + (E [f(x)])^2 \right] \\ &= \frac{k}{k-1} \left[Var [f(x)] + IE(f)^2 \right]. \end{aligned}$$

$$\begin{aligned} \lim_{n \rightarrow \infty} Var \left[T_1^{(n)} \right] &= \lim_{n \rightarrow \infty} E \left[\left(T_1^{(n)} \right)^2 \right] - \lim_{n \rightarrow \infty} \left(E \left[T_1^{(n)} \right] \right)^2 \\ &= \frac{k}{k-1} Var [f(x)] + \frac{k}{k-1} IE(f)^2 - IE(f)^2 \\ &= \frac{k}{k-1} Var [f(x)] + \frac{1}{k-1} IE(f)^2. \end{aligned}$$

$$\lim_{n \rightarrow \infty} \frac{1}{n} Var \left[T_1^{(n)} \right] = \lim_{n \rightarrow \infty} \frac{1}{n} \left[\frac{k}{k-1} Var [f(x)] + \frac{1}{k-1} IE(f)^2 \right] = 0.$$

Proof of Lemma 4

The relation above can be proved if the limiting covariance between $T_1^{(n)}$ and $T_2^{(n)}$ is zero. This can be achieved by showing limiting probability distributions of $T_1^{(n)}$ and $T_2^{(n)}$ are independent.

Given the real number $r, s > 0$:

$$\begin{aligned} &P \left[T_1^{(n)} > r, T_1^{(n)} > s \mid X_1 = x, X_2 = y \right] \\ &= P \left[R_{1,k,n} < \rho_{r,n}, R_{2,k,n} < \rho_{s,n} \mid X_1 = x, X_2 = y \right] \\ &= P \left[\text{at least } k \text{ of } X_3, X_4, \dots, X_n \in S_{\rho_{r,n};x} \text{ and} \right. \\ &\quad \left. \text{at least } k \text{ of } X_3, X_4, \dots, X_n \in S_{\rho_{s,n};y} \right]. \end{aligned}$$

Note: For $x \neq y$, $\rho_{r,n} \rightarrow 0$ and $\rho_{s,n} \rightarrow 0$ as $n \rightarrow +\infty$. For large enough n , we have $S_{\rho_{r,n};x} \cap S_{\rho_{s,n};y} = \Phi$.

We obtain:

$$\begin{aligned} &P \left[T_1^{(n)} > r, T_1^{(n)} > s \mid X_1 = x, X_2 = y \right] \\ &= \sum_{\substack{k \leq i, j \\ i+j \leq n-2}} \frac{(n-2)!}{i!j!(n-2-i-j)!} [P(S_{\rho_{r,n};x})]^i [P(S_{\rho_{s,n};y})]^j \cdot \\ &\quad [1 - P(S_{\rho_{r,n};x}) - P(S_{\rho_{s,n};y})]^{n-2-i-j} \end{aligned}$$

where $P(S_{\rho_{r,n};x}) = \int_{S_{\rho_{r,n};x}} f(t) dt$.

Recall:

$$\lim_{n \rightarrow \infty} nP(S_{\rho_{r,n};x}) = \frac{k}{r} f(x) = \lim_{n \rightarrow \infty} \frac{k}{r} \cdot \frac{P(S_{\rho_{r,n};x})}{V_{\rho}}$$

Then, we have:

$$\begin{aligned} &P \left[T_1^{(n)} > r, T_1^{(n)} > s \mid X_1 = x, X_2 = y \right] \\ &= \sum_{\substack{k \leq i, j \\ i+j \leq n-2}} \frac{(n-2)!}{i!j!(n-2-i-j)!} \left(\frac{k}{nr} \right)^i \left(\frac{k}{ns} \right)^j \cdot \\ &\quad \left(\frac{P(S_{\rho_{r,n};x})}{V_{\rho_{r,n}}} \right)^i \left(\frac{P(S_{\rho_{s,n};y})}{V_{\rho_{s,n}}} \right)^j \cdot \\ &\quad \left[1 - \frac{1}{n} \left(\frac{k}{r} \cdot \frac{P(S_{\rho_{r,n};x})}{V_{\rho_{r,n}}} + \frac{k}{s} \cdot \frac{P(S_{\rho_{s,n};y})}{V_{\rho_{s,n}}} \right) \right]^{n-2-i-j} \end{aligned}$$

Since:

$$\begin{aligned} &\frac{(n-2)!}{(n-2-i-j)!n^i n^j} \rightarrow 1 \text{ as } n \rightarrow \infty \\ &\left[1 - \frac{1}{n} \left(\frac{k}{r} \cdot \frac{P(S_{\rho_r})}{V_{\rho_r}} + \frac{k}{s} \cdot \frac{P(S_{\rho_s})}{V_{\rho_s}} \right) \right]^{n-2-i-j} \rightarrow e^{-\left(\frac{k}{r} f(x) + \frac{k}{s} f(y)\right)} \text{ as } n \rightarrow \infty \\ &\lim_{n \rightarrow \infty} \left(1 - \frac{1}{n} \right)^n = e^{-1}. \end{aligned}$$

we obtain:

$$\begin{aligned}
 & \lim_{n \rightarrow \infty} P \left[T_1^{(n)} > r, T_2^{(n)} > s \mid X_1 = x, X_2 = y \right] \\
 &= \sum_{i=k}^{n-2-k} \sum_{j=k}^{n-2-k} \frac{k^i}{i!r^i} [f(x)]^i \cdot \frac{k^j}{j!s^j} [f(y)]^j \cdot e^{-\left(\frac{k}{r}f(x) + \frac{k}{s}f(y)\right)} \\
 &= \left[\sum_{i=k}^{n-2-k} \frac{k^i}{i!r^i} [f(x)]^i e^{-\frac{k}{r}f(x)} \right] \cdot \left[\sum_{j=k}^{n-2-k} \frac{k^j}{j!s^j} [f(y)]^j e^{-\frac{k}{s}f(y)} \right] \\
 &= P[T_x > r] \cdot P[T_y > s],
 \end{aligned}$$

where for given z , the random variable T_z has the *pdf*:

$$h_{T_x} = \frac{f(z) \left(\frac{k}{y}f(z)\right)^k}{(k-1)!y^2} e^{-\frac{kf(z)}{y}} = \frac{k^k f(z)^{k+1}}{(k-1)!y^{k+2}} e^{-\frac{kf(z)}{y}}.$$

Therefore, by the theorem of independent variables:

$$\begin{aligned}
 & \lim_{n \rightarrow \infty} E[T_1^{(n)} T_2^{(n)}] = \lim_{n \rightarrow \infty} \left[E[T_1^{(n)}] \cdot E[T_2^{(n)}] \right] \\
 & \Rightarrow \lim_{n \rightarrow \infty} Cov[T_1^{(n)}, T_2^{(n)}] = 0.
 \end{aligned}$$

The Logistic Regression from the Viewpoint of the Factor Space Theory

Q.F. Cheng, T.T. Wang, S.C. Guo, D.Y. Zhang, K. Jing, L. Feng, Z.F. Zhao, P.Z. Wang

Qifeng Cheng

Research and Development Center,
China Academy of Launch Vehicle Technology,
Fengtai District, Beijing, 100076, P. R. China

Tiantian Wang

School of Science
Liaoning Technical University,
Fuxin, 123000, P. R. China

Sicong Guo, Peizhuang Wang

Institute of Intelligent Engineering and Mathematics
Liaoning Technical University,
Fuxin, 123000, P. R. China

Dayi Zhang*, Kai Jing, Liang Feng, Zhifeng Zhao

General Hospital of Coal Mining Industry Group Fuxin,
Fuxin, 123000, P. R. China

*Corresponding author:18242875460@163.com

Abstract: Logistic regression plays an important role in machine learning. People excitingly use it in conceptual matching yet with some details to be understood further. This paper aims to present a reasonable statement on logistic regression based on fuzzy sets and the factor space theory. An example about breast cancer diagnosis is displayed to show how the factor space theory can be incorporated into the understanding and use of logistic regression.

Keywords: logistic regression, factor space theory, fuzzy sets, logistic membership function

1 Introduction

In 1965, Zadeh put forward the concept of fuzzy subset which made ordinary subset generalize that the range of grade of membership has been relaxed from binary variables $\{0,1\}$ to continuous variables $[0,1]$ [23]. Today, we are facing the tide of big data, the essential meaning of fuzzy set is shined by its original definition still. Along the way, the factor space theory has been established to provide a deeper mathematical foundation for artificial intelligence [13–15]. The factor space theory builds a bridge connecting certainty and uncertainty and the two sides can be transferred to each other by changing the dimension of related factor space [16].

The falling shadow theory [17] based on factor space was developed to compare fuzziness with randomness. No matter randomness or fuzziness, they are both caused by a lack of factors. Randomness is a kind of uncertainty, which is caused by the lack of conditional factors for predicating. Randomness breaks the law of causality, while probability theory replaces it by a generalized causality law: even though insufficient conditions can not determine whether an event will occur or not, they determine the occurrence of an event with a certain probability. Fuzziness is another kind of uncertainty, which is caused by the lack of identifying factors for recognition. Fuzziness breaks the law of excluded middle, while fuzzy theory replaces it by a generalized law

of excluded middle: even though insufficient identifying factors can not determine if an object conforms a concept, they identify an object to a concept with a certain membership degree. An important relationship between fuzziness and randomness is emerged from the similarity: the fuzziness phenomenon on the ground, the universe U , can be described as a randomness phenomenon on the sky, the power of U [22]. Subjective reliability is the non-additive measure since they are the fallen of probability from sky to ground, this is the core idea of the falling shadow theory.

Logistic regression analysis (LRA) is a simple but effective method which is widely used in classification, extending the techniques of multiple regression analysis to research situations in which the outcome variable is categorical [3]. In stratification research, demographic research and social medicine, the use of logistic regression is routine [8]. Some related works have been done on logistic regression. A forward stepwise logistic regression analysis is conducted and the results show that the peer ratings of collaboration can predict you belong to Learning Problem group or Not Learning Problem group [10]. Split-sample, cross-validation and bootstrapping methods for estimation of internal validity of a logistic regression model are compared, which ends up with the recommendation of bootstrapping [12]. The number of events per variable (EPV) in LRA has influences on the validity of the model, and it turns out that low EPV plays a major role [9]. Yet these works only focus on the strategic level of logistic regression, deep understanding hidden behind this common model should be mined.

In artificial intelligence, the regressed curves can be regarded as membership functions. The selection of the types of regressed curves decides the quality of curve fitting partially. Logistic regression is fitted by the logistic membership function. In this article, we will expand the discussion about logistic regression based on factor space theory and fuzzy sets which can present a relatively different state from before.

The paper is proceeded as follows. We introduce the types of membership functions in Section 2. The logistic membership function is put forward in Section 3. In Section 4, logistic regression based on the factor space theory is discussed further and an algorithm is proposed accordingly. In Section 5, an example is used to identify the effectiveness of the proposed algorithm. Finally, a brief conclusion is made in Section 6.

2 Type of membership functions

Definition 1. A fuzzy subset A defined on universe of discussion U is a mapping $\mu_A: U \rightarrow [0,1]$, $\mu_A(u)$ is called the membership degree of u with respect to A [23].

From Definition 1, two fundamental meanings are revealed: firstly, a fuzzy subset stands for the extension of a fuzzy concept, it is a milestone of intelligence mathematics; secondly, a fuzzy subset builds a bridge to step over the gap between quantitative and qualitative phenomena, it was the main bottleneck of information revolution. While the falling shadow theory can provide a deeper mathematical foundation for the revolutionary change [5, 6, 18]. Definitions 2, 3 and 4 are listed to show how the falling shadow of a random subset is generated.

Definition 2. Let (U, \mathcal{B}) be a measurable space. $(\mathcal{P}(U), \underline{\mathcal{B}})$ is called a super-measurable space defined on (U, \mathcal{B}) if $(\mathcal{P}(U), \underline{\mathcal{B}})$ is a measurable space.

Definition 3. Given a probabilistic field (Ω, \mathcal{F}, p) and a super-measurable space $(\mathcal{P}(U), \underline{\mathcal{B}})$ on (U, \mathcal{B}) . A mapping $\xi: \Omega \rightarrow \mathcal{P}(U)$ is called a random set if $\xi^{-1}(\mathcal{A}) = \{\omega \in \Omega \mid \xi(\omega) \in \mathcal{A}\} \in \mathcal{F}$ whenever $\mathcal{A} \in \underline{\mathcal{B}}$.

A probability distribution can be induced on $\underline{\mathcal{B}}$ from p on \mathcal{F} through the mapping ξ ,

$$P(\mathcal{A}) = P(\xi^{-1}(\mathcal{A})), \mathcal{A} \in \underline{\mathcal{B}},$$

which makes $(\mathcal{P}(U), \underline{\mathcal{B}})$ a super probabilistic field [19].

Definition 4. Given a super probabilistic field $(\mathcal{P}(U), \underline{\mathcal{B}})$ and a random set $\xi : \Omega \rightarrow \mathcal{P}(U)$, we call a fuzzy subset A_ξ on U the falling shadow of ξ if $\mu_{A_\xi}(u) = P(\omega \mid u \in \xi(\omega))$ for all $u \in U$.

It is revealed in Definition 4 that μ_{A_ξ} can be viewed as the covering function of a random set ξ on U . The random set is called clouds, and the fuzzy set is called the falling shadow of the clouds. The thicker the cloud, the higher the darkness of the shadow of the cloud [19]. Meanwhile, according to Definition 1, μ_{A_ξ} is also the membership function with respect to fuzzy subset A_ξ .

Definition 5. A membership function μ_{A_ξ} is also called the possibility distribution of a concept A_ξ on U .

Possibility varies from probability. According to Definition 5, possibility stands for the covering chance of ξ to u and it does not hold exclusiveness; while probability holds the exclusiveness and stands for the chance of monopolization [20].

Theorem 1. Let $U = (-\infty, +\infty)$ be the one dimensional state space. Given a random interval r with falling shadow μ_A , let ζ be the left extreme point of random interval r , then ζ is a random variable defined on $(U, \underline{\mathcal{B}})$. And the possibility distribution of the concept A is the same as the distribution function $F(u)$ of $\zeta : \mu_A(u) = F(u) = P(\zeta \leq u)$.

Proof: Given probabilistic field (Ω, \mathcal{F}, p) , measurable space $(U, \underline{\mathcal{B}})$ and super-measurable space $(\mathcal{P}(U), \underline{\mathcal{B}})$.

Since r is a random set, then

$$r^{-1}(\mathcal{A}) = \{\omega \in \Omega \mid r(\omega) \in \mathcal{A}\} \in \mathcal{F}$$

whenever $\mathcal{A} \in \underline{\mathcal{B}}$.

Because ζ is the left extreme point of r , then for each $\omega \in \Omega$, there exists $\beta \in \underline{\mathcal{B}}$, which makes

$$r(\omega) \in \mathcal{A} \implies \zeta(\omega) \in \beta.$$

Since $\zeta(\omega)$ is a real number on the interval $(-\infty, +\infty)$, according to the definition of random variable [21], it is obvious that ζ is a random variable. And

$$\mu_A(u) = P\{\omega \mid u \in r(\omega)\} = P\{\omega \mid \zeta(\omega) \leq u\} = P(\zeta \leq u).$$

□

From Theorem 1, we can distinguish possibility distribution from probability distribution and combine the two terminologies by means of density function and distributed function respectively.

The membership function μ_A determines the extent that u belongs to a concept A . For one concept, different types of membership functions exhibit different membership degree for a certain point u , making it necessary to clarify which type of membership functions should be chosen accordingly.

Definition 6. A is a fuzzy subset defined on universe of U whose membership function is μ_A . If $\mu_A(x) > \min\{\mu_A(a), \mu_A(b)\}$ for any $a < x < b$, then A is a convex fuzzy subset [4].

A convex fuzzy set divides A into five parts with four points l^-, l^+, u^-, u^+ ($l^- \leq l^+ \leq u^- \leq u^+$):

$$\mathbf{k} = (-\infty, l^-], \mathbf{l} = (l^-, l^+], \mathbf{t} = (l^+, u^-], \mathbf{u} = (u^-, u^+], \mathbf{v} = (u^+, +\infty).$$

$\mu_A \equiv 0$ when $x \in \mathbf{k}$ or $x \in \mathbf{v}$, $\mu_A \equiv 1$ when $x \in \mathbf{t}$. $(l^-, \mu_A(l^-))$ and $(u^+, \mu_A(u^+))$ are the inflection points.

Definition 7. We call $|\mathbf{l}|$ and $|\mathbf{u}|$ the lower and upper interim length respectively.

Even though the shape of membership functions is variable with countless changes, the essential variations focus on the two interim periods. The length of interim reflects the degree of fuzziness with respect to a membership function, the narrower the length of interim, the more precise the representation of a concept. For ease of simplicity, we only discuss the membership function on the left fuzzy interval \mathbf{l} formed by the distribution of left extreme point ζ on the interval. There are three common types of probability density functions of ζ : uniform type, cosine type and normal type. Due to space limitation, only uniform distribution of ζ is displayed here to show how the membership function is generated.

For example, ζ is uniformly distributed on fuzzy segment $(l^-, l^+]$, the probability density function of ζ is

$$f_1(x) = \frac{1}{l^+ - l^-}, \quad x \in (l^-, l^+].$$

According to Theorem 1, the membership curve of the concept on this interval is

$$\mu_1(x) = P_1(\zeta \leq x) = \frac{x - l^-}{l^+ - l^-}. \tag{1}$$

It is a straight line ranging from 0 to 1 on fuzzy segment $(l^-, l^+]$.

Curve fitting is the foundation of optimization, the quality of fitting depends on whether the curve chosen is appropriate or not. In this article, we introduce another type of membership function which is frequently used in classification but lacking deeper understanding, as well as expanding the discussion on it from the viewpoint of factor space.

3 Logistic membership function

Let γ be the random variable defined on U , indicating attribute x . And let y be the indication variable of concept α with the extension A defined on U , which takes value 1 when $u \in A$ and 0 else. Denote $P_x = P\{y = 1 | \gamma = x\}$, which is the possibility distribution of the concept A with respect to the variable x . To estimate the possibility, use the *maximal likelihood principle*. Consider a series of sampling points $(x_1, y_1), \dots, (x_m, y_m)$, the likelihood function is as follows:

$$\underline{L} = \prod_{i=1}^m P_{x_i}^{y_i} (1 - P_{x_i})^{1-y_i}. \tag{2}$$

It is not easy to calculate the derivative, put logarithm on it and get the new likelihood

function:

$$\begin{aligned}
L &= \ln \prod_{i=1}^m P_{x_i}^{y_i} (1 - P_{x_i})^{1-y_i} \\
&= \sum_{i=1}^m (y_i \ln P_{x_i} + (1 - y_i) \ln(1 - P_{x_i})) \\
&= \sum_{i=1}^m (y_i (\ln P_{x_i} - \ln(1 - P_{x_i}))) + \sum_{i=1}^m \ln(1 - P_{x_i}) \\
&= \sum_{i=1}^m (y_i \ln \frac{P_{x_i}}{1 - P_{x_i}}) + \sum_{i=1}^m \ln(1 - P_{x_i}).
\end{aligned} \tag{3}$$

Since $\ln \frac{P_x}{1-P_x}$ varies on the interval $[0, +\infty)$, also being related to the attribute x , we define $\ln \frac{P_x}{1-P_x} = ax + b$, then $P_x = \frac{1}{1+e^{-(ax+b)}}$. And Equation (3) can be transferred into

$$L = \sum_{i=1}^m (y_i(ax_i + b)) - \sum_{i=1}^m \ln(1 + e^{ax_i+b}). \tag{4}$$

Maximize L so that a and b can be achieved. This is an optimization problem and some strategies solving this type of problems such as *gradient descent* [?] can be applied.

Definition 8. The logistic regression function is defined as:

$$\phi(\mathbf{x}; \mathbf{a}, b) = \frac{1}{1 + e^{-(\mathbf{a}^\top \mathbf{x} + b)}}$$

$\mathbf{x}, \mathbf{a} \in R^n$, $\mathbf{a}^\top \mathbf{x} = a_1x_1 + \dots + a_nx_n$. When $n = 1$, $\phi(x; a, b) = \frac{1}{1+e^{-(ax+b)}}$; $\phi(x; 1, 0) = \frac{1}{1+e^{-x}}$.

It is obvious that this method of estimating a and b is kind of complex, while simplicity is what we pursue eventually. Another method which can transfer the problem into a linear problem will be introduced in the next section.

Since $P_A(x) = \frac{1}{1+e^{-(ax+b)}}$ is central symmetric with respect to the point $(-\frac{b}{a}, \frac{1}{2})$, we can get

$$P_{\neg A}(x) = P_A(-\frac{2b}{a} - x) = \frac{1}{1 + e^{-a(-\frac{2b}{a} - x) + b}} = \frac{1}{1 + e^{ax+b}}.$$

And it is easy to know

$$P_A(x) + P_{\neg A}(x) = \frac{1}{1 + e^{-(ax+b)}} + \frac{1}{1 + e^{ax+b}} = 1.$$

With this property, the logistic regression function $P(x) = \frac{1}{1+e^{-(ax+b)}}$ can be seen as the logistic membership function. And this property is available among another three types of membership functions referred in Section 2, the details of which are omitted here.

Definition 9. We call $P(x) = \frac{1}{1+e^{-a(x-x)}}$ the logistic interim function.

3.1 Logistic regression in risk attributable factor space

The factor space theory paves the way for logistic regression since some concepts find their footholds in the factor space, which contributes to the better understanding of logistic regression.

Definition 10. A factor space defined on universe of discussion U is a family of set $\Psi = (\{X(f)\}_{f \in F}; U)$ satisfying:

- (1) $F = (F, \vee, \wedge, ^c, \mathbf{1}, \mathbf{0})$ is a complete Boolean algebra;
- (2) $X(\mathbf{0}) = \{\emptyset\}$;
- (3) For any $T \subseteq F$, if $\{f|f \in T\}$ are irreducible (i.e., $s \neq t \Rightarrow s \wedge t = \mathbf{0}$ ($s, t \in T$)), then

$$X(\{f|f \in T\}) = \prod_{f \in T} X(f)$$

(\prod stands for Cartesian product)

- (4) $\forall f \in F$, there is a mapping with same symbol $f : U \rightarrow X(f)$.

F is called the set of factors, $f \in F$ is called a factor on U . $X(f)$ is called the state space of factor f [22].

Denote $I_j = X(f_j)$ and $I = I_1 \times \dots \times I_n$.

Definition 11. Given an attribute space $O = \{o_{i_1 \dots i_n} | i_j \in I_j, j = 1, \dots, n\}$ in factor space $\Psi = (\{X(f)\}_{f \in F = \{f_1, \dots, f_n\}}; U)$. For each $i_j \in I_j$,

$$o_{i_1 \dots i_n}$$

is called a granule of $X(F)$ [5]. Denote that $q_{i_1 \dots i_n} = P\{u | F(u) = o_{i_1 \dots i_n}\}$,

$$\mathbf{P} = \{q_{i_1 \dots i_n} | i_j \in I_j, j = 1, \dots, n\}$$

is called the probability distribution of attributes.

Assumption 1. $X(f_1), X(f_2) \dots, X(f_n)$ are all partial ordered sets, it means for one attribute f_i whose range of values is $\{i_{j_1}, i_{j_2}, \dots, i_{j_r}\}$, $i_{j_1} \leq i_{j_2} \leq \dots \leq i_{j_r}$ is always correct. It should be clear that if there exists two granules $o_{1 \dots 1 \dots 1}$ and $o_{1 \dots 3 \dots 1}$, granule $o_{1 \dots 2 \dots 1}$ is anyhow certain to appear. Then we naturally suppose that \mathbf{P} is convex.

A convex probability distribution of attributes \mathbf{P} is called the background distribution with respect to universe U [7].

Definition 12. A factor space $\tilde{\Psi} = (\{X(f)\}_{f \in \tilde{F}}; U)$ is called a risk attribute factor space if $\tilde{F} = \{f_1, \dots, f_n; f_{n+1}\}$, where f_1, \dots, f_n stand for attribute factors and f_{n+1} stands for a risk factor with binary attribute space $X(f_{n+1}) = \{1, 0\}$.

Given a group of sampling points on $I \times \{1, 0\}$:

$$S = \{(x_{1i}, \dots, x_{ni}; y_i)\}_{(i=1, \dots, m)}$$

For $(i_1, \dots, i_n; i_{n+1}) \in I \times \{1, 0\}$, denote

$$q_{i_1 \dots i_n i_{n+1}} = \frac{|\{t \mid x_{1t} = i_1, \dots, x_{nt} = i_n; y_t = i_{n+1}\}|}{m} \tag{5}$$

It is obvious that $\sum_{i_1} \dots \sum_{i_n} \sum_{i_{n+1}} q_{i_1 \dots i_n i_{n+1}} = 1$.

Definition 13. $\tilde{Q} = \{i_1, \dots, i_n; i_{n+1} \in I \times \{1, 0\} \mid q_{i_1 \dots i_n i_{n+1}} > 0\}$ is the support set of $Q = \{q_{i_1 \dots i_n i_{n+1}}\}_{(i_1, \dots, i_n; i_{n+1}) \in I \times \{1, 0\}}$.

In the risk attribute factor space $\tilde{\Psi}$, logistic regression is mainly based on the support set \tilde{Q} . While the condition that $q_{i_1 \dots i_n i_{n+1}} = 0$ still exists and transformation used to handle it will be discussed in the final part of this section. Also, the definition of the logistic membership function fitted out through logistic regression get updated in $\tilde{\Psi}$.

Definition 14. The membership function of illness α :

$$P(X) = \frac{1}{1 + e^{-(\theta_0 + \theta_1 x_1 + \dots + \theta_n x_n)}} = \frac{1}{1 + e^{-\theta^\top X}},$$

where $\theta = (\theta_0, \theta_1, \dots, \theta_n)^\top$, $X = (1, x_1, \dots, x_n)^\top$. The parameters $\theta_1, \dots, \theta_n$ are called risk attribute coefficients of factors.

The bigger the coefficient θ_j , the more important the factor f_j on risk-increasing.

Definition 15. Let α be a concept with extension $A \subseteq U$ and intension $O \subseteq X(F)$, we call conditional probability $P\{u \in A | F(u) = o_{i_1 \dots i_n}\}$ the possibility of concept α under $o_{i_1 \dots i_n}$.

Remark 2. The difference between possibility and probability is clear again. Since possibility is equal to the membership degree of the factorial configuration with respect to concept α , the logistic membership function of illness α is

$$\begin{aligned} p_{i_1 \dots i_n} &= P\{u \in A | F(u) = o_{i_1 \dots i_n}\} \\ &= \frac{|\{t \mid x_{1t} = i_1, \dots, x_{nt} = i_n; y_t = 1\}|}{|\{t \mid x_{1t} = i_1, \dots, x_{nt} = i_n\}|}. \end{aligned} \quad (6)$$

Also, the probability corresponding to possibility $p_{i_1 \dots i_n}$ is

$$\begin{aligned} q_{i_1 \dots i_n, i_{n+1}=1} &= P\{u \in A | F(u) = o_{i_1 \dots i_n}\} p\{F(u) = o_{i_1 \dots i_n}\} \\ &= \frac{|\{t \mid x_{1t} = i_1, \dots, x_{nt} = i_n; y_t = 1\}|}{m}. \end{aligned} \quad (7)$$

It is obvious that

$$\sum p_{i_1 \dots i_n} \neq \sum q_{i_1 \dots i_n, i_{n+1}=1}. \quad (8)$$

Then the approach to estimate θ through logistic regression is developed as Algorithm 1, and the membership degree for the unknown samples can be calculated.

Algorithm 1 Logistic Regression Algorithm

- 1: Given sample points $S = \{(x_{1i}, \dots, x_{ni}; y_i)\} (i = 1, \dots, m)$.
- 2: For $|\{t \mid x_{1t} = i_1, \dots, x_{nt} = i_n\}| > 0 (i_1 \in I_1, \dots, i_n \in I_n)$:
- 3: **if** $|\{t \mid y_t = 1\}| > 0$ and $|\{t \mid y_t = 0\}| > 0$, **then**
- 4: calculate $P_{i_1 \dots i_n}$
- 5: let

$$y_{i_1 \dots i_n} = \ln \frac{P_{i_1 \dots i_n}}{1 - P_{i_1 \dots i_n}}$$

6: **else**

7:

$$y_{i_1 \dots i_n} = \ln \frac{|\{t \mid y_t = 1\}| + 0.5}{|\{t \mid y_t = 0\}| + 0.5} [1]$$

8: **end if**

- 9: Set $y_{i_1 \dots i_n} = \theta_0 + \theta_1 i_1 + \dots + \theta_n i_n$, do linear regression to get the coefficients $\theta_0, \theta_1, \dots, \theta_n$.
-

4 Example

To show how the factor space theory is imbedded into logistic regression, in this section, an example calculating the membership degree of breast cancer using Algorithm 1 is considered. The data extracted from [2] are shown in Table 1 and Table 2. There are 1896 samples in total which form the universe of discussion U . Four attribute factors f_1, f_2, f_3, f_4 are considered and each factor is a categorical variable whose set of value is $I_1 = \{0, 1, 2\}, I_2 = \{0, 1, 2, 3\}, I_3 = \{0, 1\}, I_4 = \{0, 1, 2\}$ respectively. The meaning of these coded numbers are shown in Table 4. Then the factor space $\Psi = (\{X(f)\}_{(f \in F=f_1, f_2, f_3, f_4)}; U)$ involving 72 different granules $o_{i_1 \dots i_4}$ of $X(F)$ can be established. The risk attribute factor space $\tilde{\Psi} = (\{X(f)\}_{(f \in \tilde{F}=f_1, f_2, f_3, f_4; f_5)}; U)$ is also confirmed, where $X(f_5) = \{1, 0\}$.

The data must be processed before using Algorithm 1. For the background distribution P , there are some granules whose $q_{i_1 \dots i_4} = 0$ which means no samples are included when $f_1 = i_1, \dots, f_4 = i_4$, then those granules should be deleted. For some granules, the samples are very small and there is little meaning to consider them, so the granules whose total samples are less than 5 are deleted. Hence 33 granules remain and the samples decrease to 1837. Algorithm 1 is applied to the 1837 samples and we can get the membership function of breast cancer as follows:

$$P(X) = \frac{1}{1 + e^{-(0.19x_1 + 0.24x_2 + 0.36x_3 + 0.80x_4 - 0.75)}}. \quad (9)$$

Equation (9) represents a hyperplane in five-dimensional space. Although the interim still exists, we ignore it on account of the complexity. The estimated membership degree for each granule can be calculated. Table 5 depicts the real possibility $P_{i_1 \dots i_4}$, the calculated possibility $P(X_i)$ and probability $q_{i_1 \dots i_4, i_5=1}$. Then Equation (8) can be verified easily:

$$\sum P_{i_1 \dots i_4} = 18.5227, \quad \sum q_{i_1 \dots i_4, i_5=1} = 0.4899.$$

In Table 5, the rows that the difference between $P_{i_1 \dots i_4}$ and $P(X_i)$ are over 0.2 are marked in red, only four granules 13th, 16th, 21st and 32nd are included. Since membership degree is estimated through linear regression, it is necessary to compare $y_{i_1 \dots i_4}$ and $\ln \frac{P(X_i)}{1-P(X_i)}$, Figure 1 shows the differences of their values. The circles represent the real differences between $y_{i_1 \dots i_4}$ and $\ln \frac{P(X_i)}{1-P(X_i)}$. Since *least-square estimation* is applied here, it is necessary to do parameter test. The vertical line segments represent the 95% confidence intervals of random residuals. It is obvious that the confidence intervals of the 16th, 21st and 32nd cases don't cover 0, yet not far away from 0. Errors from different angles indicate that logistic regression is applicable in this example.

From Equation (9), we know that θ_i ($i = 1, 2, 3, 4$) is 0.19, 0.24, 0.36, 0.80 accordingly. θ_i ($i = 1, 2, 3, 4$) are all positive numbers which means the risk of having breast cancer will grow when the coded number of factor i gets larger. This is in accordance with the existing knowledge: the risk of having breast cancer will become larger if girls get their first period at a younger age; nulliparous women and those who give the first birth when they are older will have higher risks of having breast cancer; doing previous breast biopsies¹ means you are suspected of having breast cancer, the more you do, the larger the chance of being suspected will be; the risk of having breast cancer will grow if the number of people with breast cancer in near relations goes up. This conclusion verifies the efficiency of the logistic regression from another perspective.

¹A biopsy is a medical test commonly performed by a surgeon, interventional radiologist, or an interventional cardiologist involving extraction of sample cells or tissues for examination to determine the presence or extent of a disease.

5 Conclusions

The possibility in logistic regression is equal to the membership degree in factor space. Connection between the two sides is established through the membership function. The paper shows that the factor space theory gives logistic regression a relatively different state. Meanwhile, logistic regression forms another foothold of the factor space theory in the big data era. This is of great meaning since it gives us the reason and motivation to explore the factor space theory deeply.

Acknowledgements

This work is supported in part by Natural Science Foundation of Liaoning Province, China [2015020570] and National Natural Science Foundation (NNSF) of China under Grant [61304090].

Bibliography

- [1] Bao Y.K (2011); *Data Analysis*, Tsinghua University Press, 2011.
- [2] Bruzzi B.P, Sylvan B.G, David P.B, Louise A.B, Catherine S. (1985); Estimating the Population Attributable Risk for Multiple Risk Factors Using Case-Control Data, *American Journal of Epidemiology*, 122(5), 904-914, 1985.
- [3] Dayton C.M. (1992); Logistic Regression Analysis, *University of Maryland*, 1992.
- [4] Guo S.C, Chen G. (2001); *Method of Soft Computing in Information Science*, Northeastern University Press, 2001.
- [5] Li H.X, Wang P.Z. (1994); *A Mathematical Theory on Knowledge Representation*, Tianjin Scientific and Technical Press, 1994.
- [6] Liu Z.L, Liu Y.C. (1992); *Theory of Factorial Neural Networks*, Beijing Normal University Press, 1992.
- [7] Liu H.T, Wang P.Z; *Background distribution and fuzzy background relationship (to be published)*.
- [8] Mood C. (2010); Logistic Regression: Why We Cannot Do What We Think We Can Do, and What We Can Do About It, *European Sociological Review*, 26(1), 67-82, 2010.
- [9] Peduzzi P., Concato J., Kemper K. (1996); A simulation study of the number of events per variable in logistic regression analysis, *Journal of clinical epidemiology*, 49(12), 1373-1379, 1996.
- [10] Prette Z. A. P. D, Prette A.D, Oliveira L.A.D, Gresham F.M, Vance M.J. (2012); Role of social performance in predicting learning problems: Prediction of risk using logistic regression analysis, *School Psychology International*, 33(6), 615-630, 2012.
- [11] Russell S.J, Norvig P. (2011); *Artificial Intelligence A modern Approach (Third Edition)*, Tsinghua University Press, 2011.
- [12] Steyerberga E.W, Jrb F. E. H, Borsbooma G. J. J. M. (2001); Internal validation of predictive models: Efficiency of some procedures for logistic regression analysis, *Journal of Clinical Epidemiology*, 54, 774-781, 2001.

-
- [13] Wang P.Z. (2013); Factor space and factor data bases, *Journal of Liaoning Engineering and Technology University*, 32(10), 1-8, 2013.
- [14] Wang P.Z, Guo S.C, Bao Y.K, Liu H.T. (2014); Causality analysis in factor spaces, *Journal of Liaoning Engineering and Technology University*, 33(7), 1-6, 2014.
- [15] Wang P.Z. (2015); Factor space and data science, *Journal of Liaoning Engineering and Technology University*, 34(2), 273-280, 2015.
- [16] Wang P.Z. (1981); In: *Hao BL et al (eds) Advance of statistical physics.Science*, Beijing, 1981.
- [17] Wang P.Z. (1985); *Fuzzy Sets and Falling Shadow of Random Sets*, Beijing Normal University Press, 1985.
- [18] Wang P.Z, Sugeno M. (1982); The factors field and background structure for fuzzy subsets, *Fuzzy Mathematics*, (2), 45-54, 1982.
- [19] Wang P.Z, Zhang H.M, Ma X.W, Xu W. (1991); Fuzzy set-operations represented by falling shadow theory, in: *Fuzzy Engineering toward Human Friendly Systems, Proceedings of the International Fuzzy Engineering Symposium '91, Yokohama, Japan*, 1, 82-90, 1991.
- [20] Wang P.Z, Liu H.C, Zhang X.H. (1993); Win-win strategy for probability and fuzzy mathematics, *The Journal of Fuzzy Mathematics*, 1(1), 223-231, 1993.
- [21] Wang P.Z. (2013); *Fuzzy Mathematics and Optimization*, Beijing Normal University Press, 2013.
- [22] Wang P.Z, Liu Z.L, Shi Y, Guo S.C. (2014); Factor Space, the Theoretical Base of Data Science, *Annals of Data Science*, 1(2), 233-251, 2014.
- [23] Zadeh L.A. (1965); Fuzzy Subsets, *Information and Control*, 8, 338-353, 1965.

Appendix

Table 1: Detailed distribution of cases and their matched controls in all strata defined by cross-classifying the four risk factors

Risk factor category						
Age at menarche	Age		Mothers		No. of cases	No. of controls
	at first live birth	Previous breast biopsies	plus sisters with breast cancer			
0	0	0	0		14	27
0	0	0	1		3	1
0	0	0	2		0	0
0	0	1	0		3	1
0	0	1	1		1	0
0	0	1	2		0	0
0	1	0	0		54	85
0	1	0	1		20	12
0	1	0	2		1	1
0	1	1	0		5	5
0	1	1	1		2	0
0	1	1	2		0	0
0	2	0	0		81	100
0	2	0	1		18	20
0	2	0	2		3	0
0	2	1	0		7	12
0	2	1	1		4	2
0	2	1	2		0	0
0	3	0	0		27	14
0	3	0	1		12	7
0	3	0	2		1	0
0	3	1	0		0	2
0	3	1	1		1	0
0	3	1	2		1	0
1	0	0	0		27	56
1	0	0	1		8	7
1	0	0	2		1	0
1	0	1	0		1	4
1	0	1	1		0	0
1	0	1	2		0	0
1	1	0	0		112	173
1	1	0	1		27	12

Table 2: Table 1 - Continued

Risk factor category					
Age at menarche	Age at first live birth	Previous breast biopsies	Mothers plus sisters with breast cancer	No. of cases	No. of controls
1	1	0	2	4	0
1	1	1	0	14	4
1	1	1	1	1	2
1	1	1	2	0	0
1	2	0	0	187	174
1	2	0	1	41	20
1	2	0	2	10	1
1	2	1	0	11	10
1	2	1	1	5	0
1	2	1	2	0	1
1	3	0	0	41	47
1	3	0	1	15	5
1	3	0	2	4	0
1	3	1	0	4	5
1	3	1	1	1	0
1	3	1	2	1	0
2	0	0	0	9	15
2	0	0	1	3	2
2	0	0	2	2	0
2	0	1	0	1	1
2	0	1	1	0	0
2	0	1	2	0	0
2	1	0	0	43	44
2	1	0	1	14	5
2	1	0	2	1	0
2	1	1	0	3	2
2	1	1	1	2	0
2	1	1	2	0	0
2	2	0	0	53	52
2	2	0	1	9	8
2	2	0	2	2	0

Table 3: Table 2 - Continued

Risk factor category					
Age at menarche	Age	Previous breast biopsies	Mothers	No. of cases	No. of controls
	at first live birth		plus sisters with breast cancer		
2	2	1	0	3	1
2	2	1	1	2	1
2	2	1	2	0	0
2	3	0	0	17	4
2	3	0	1	4	3
2	3	0	2	1	0
2	3	1	0	3	0
2	3	1	1	3	0
2	3	1	2	0	0

Table 4: Levels of the risk factors

Risk factor	Range	Coding
Age at menarche	< 12	2
	12-13	1
	≥ 14	0
Age at first live birth	< 20	0
	20-24	1
	25-29(or nulliparous)	2
	≥ 30	3
No. of previous breast biopsies	0 or 1	0
	≥ 1	1
No. of mothers plus sisters with breast cancer	0	0
	1	1
	≥ 2	2

Table 5: $P_{i_1 \dots i_4}$, $P(X_i)$ and $q_{i_1 \dots i_4, i_5=1}$ for 33 granules

case number	Risk factor category				$P_{i_1 \dots i_4}$	$P(X_i)$	$q_{i_1 \dots i_4, i_5=1}$
	Age at menarche	Age at first live birth	Previous breast biopsies	Mothers plus sisters with breast cancer			
1	0	0	0	0	0.3415	0.3207	0.0076
2	0	1	0	0	0.3885	0.3739	0.0294
3	0	1	0	1	0.6250	0.5711	0.0109
4	0	1	1	0	0.5000	0.4624	0.0027
5	0	2	0	0	0.4475	0.4304	0.0441
6	0	2	0	1	0.4737	0.6275	0.0098
7	0	2	1	0	0.3684	0.5211	0.0038
8	0	2	1	1	0.6667	0.7081	0.0022
9	0	3	0	0	0.6585	0.4887	0.0147
10	0	3	0	1	0.6316	0.6806	0.0065
11	1	0	0	0	0.3253	0.3640	0.0147
12	1	0	0	1	0.5333	0.5606	0.0044
13	1	0	1	0	0.2000	0.4519	0.0005
14	1	1	0	0	0.3930	0.4200	0.0610
15	1	1	0	1	0.6923	0.6175	0.0147
16	1	1	1	0	0.7778	0.5105	0.0076
17	1	2	0	0	0.5180	0.4781	0.1018
18	1	2	0	1	0.6721	0.6713	0.0223
19	1	2	0	2	0.9091	0.8199	0.0054
20	1	2	1	0	0.5238	0.5689	0.0060
21	1	2	1	1	1.0000	0.7463	0.0027
22	1	3	0	0	0.4659	0.5368	0.0223
23	1	3	0	1	0.7500	0.7210	0.0082
24	1	3	1	0	0.4444	0.6254	0.0022
25	2	0	0	0	0.3750	0.4097	0.0049
26	2	0	0	1	0.6000	0.6074	0.0016
27	2	1	0	0	0.4943	0.4675	0.0234
28	2	1	0	1	0.7368	0.6619	0.0076
29	2	1	1	0	0.6000	0.5585	0.0016
30	2	2	0	0	0.5048	0.5263	0.0289
31	2	2	0	1	0.5294	0.7124	0.0049
32	2	3	0	0	0.8095	0.5843	0.0093
33	2	3	0	1	0.5714	0.7580	0.0022

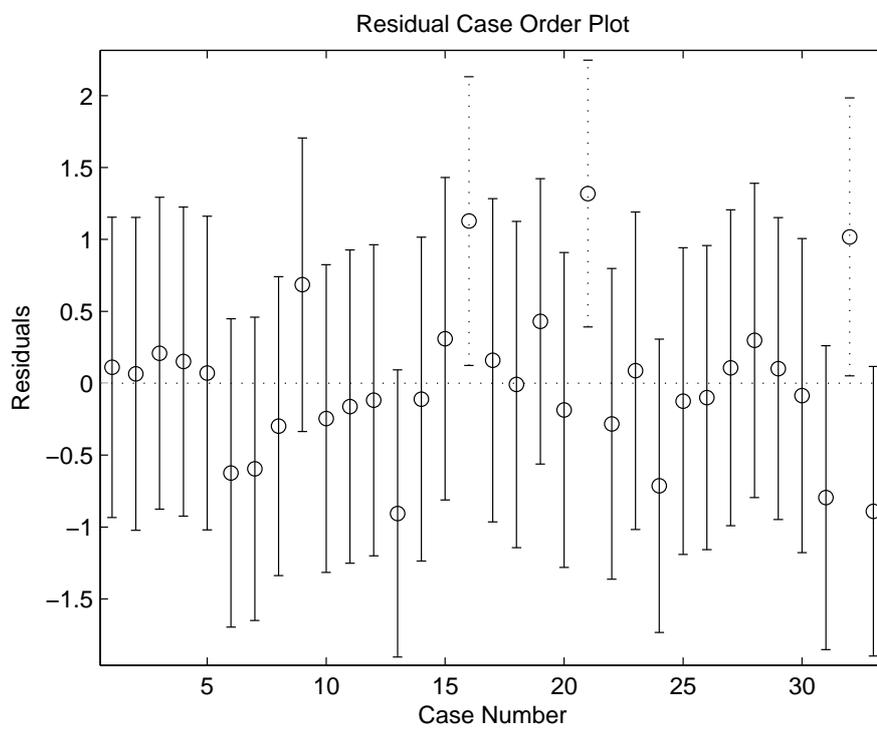


Figure 1: The differences between $y_{i_1 \dots i_4}$ and $\ln \frac{P(X_i)}{1-P(X_i)}$

An Approach for Construction of Augmented Reality Systems using Natural Markers and Mobile Sensors in Industrial Fields

D. Gomes Jr., P. Reis, A. Paiva, A. Silva, G. Braz Jr., M. Gattass, A. Araújo

Daniel Lima Gomes Júnior*

1. Federal Institute of Maranhão, São Luís, MA, Brazil
2. Applied Computing Center - Federal University of Maranhão, São Luís, MA, Brazil

*Corresponding author: daniellima@ifma.edu.br

Paulo Roberto Jansen dos Reis, Anselmo Cardoso de Paiva, Aristófanés Corrêa Silva, Geraldo Braz Junior

Applied Computing Center - Federal University of Maranhão,
São Luís, MA, Brazil

jansen@nca.ufma.br, paiva@deinf.ufma.br, ari@dee.ufma.br, geraldo.braz@ufma.br

Marcelo Gattass

Tecgraf Institute/PUC-Rio, Rio de Janeiro, RJ, Brazil
mgattass@inf.puc-rio.br

Antônio Sérgio de Araújo

Hydroelectric Company of São Francisco, Recife, PE, Brazil
asergio@chesf.gov.br

Abstract: This paper presents a methodology for the development of augmented reality (AR) visualization applications in industrial scenarios. The proposal presents the use of georeferenced natural markers detected in real time, which enables the construction of AR systems. This use of augmented visualization allows the creation of tools that can aid on-site maintenance activities for operators. AR use makes possible including information about the equipment during a specific procedure. In this work, the detection of natural markers in the scene are based on Haar-like features associated with equipment geolocalization. This approach enable the detection of equipment in multiple user's viewpoints in the industrial scenario and makes it possible the inclusion of real information about those equipment in real time as AR annotations. In this way, beyond a methodology approach, this paper presents a new way for Power System information visualization in the field that can be used in both for training and for control operations.

Keywords: Augmented Reality, Data Visualization, Natural Markers.

1 Introduction

Augmented reality (AR) is associated with the insertion of additional information or virtual objects superimposed or combined into a visualization of the real world [3]. The use of AR techniques allows to magnify user's visual experience with the addition of information to the visualized real scene. It contributes to a potential increase of user's comprehension about the physical environment where the user stands. Therefore, information or images generated by computers need to be registered to the real scene. It means that they need to be precisely and contextually positioned in the real world visualization. In addition, it allows a big variety of applications on several areas such as medicine, education, and industrial activities.

Industrial companies demand for applications that use AR technology for training and also for control and operation procedures. This technology has been used in a quite limited form,

in areas such as architecture, engineering, and civil construction [6]. The main limitation in these areas are related to high precision processing in real time. Another limiting factor is the complexity of the detection process and the object tracking in these scenes. This demands the use of strategies and algorithms of computer vision (CV) which permit a suitable time of processing for the visualization.

A strategy generally used to reduce processing time during object detection is the use of fiducial markers in the physical environment [25]. This is the insertion of objects or marks of low complexity, highly distinct of the present objects in the scene that allow to decrease the necessary computing effort for the correct spatial orientation in an AR environment. Although the use of these markers increase robustness and reduce significantly the computational cost, it has the disadvantage of requiring the inclusion of artificial marker in the real environment. Besides this, there are safety restrictions in industrial environments and incidental displacement or destruction of these markers result in the immediate inoperability of the AR system.

Using another type of AR marker, known as natural marker, the elements that already exist in the scene are detected directly without the need of fiducial markers inclusion. This approach can use the equipment of the industrial environment as markers for an AR application in an industrial plant, for example. The detection of these objects require extraction of the image features and comparison with patterns previously registered. This is considered one of the most complex tasks for the creation of AR applications. Although there are techniques which address the feature detection problem and object recognition [17] - [14], processing time is still a challenge for these systems, since in AR applications, detection needs to be done in real time.

Also there is a demand for AR applications that offer support to operators during real maintenance procedures of equipment in the field. In general, in these situations, data from equipment sensors are sent to a distant control room. The operators make decisions based on these information and in some circumstances there is the need of operators to move from the site to get access to the information. This could generate a situation in which information can be modified during the route between control room and maintenance site. This methodology proposes a solution for the creation of AR applications which enables a visualization of these information in the maintenance site, enabling a decentralization of control rooms.

This work presents a methodology which is composed by three main steps: the creation of an AR marker using a computer vision technique, followed by using sensors (GPS and compass) from mobile device to perform the correct identification of an equipment. Finally, it is presented the integration with equipment data. In our case study it is used the communication and integration with Supervisory Control and Data Acquisition Systems/Energy Management System (SCADA/EMS).

So, this methodology enables the creation and use of AR in industrial outdoor scenarios. The investigation method is based on the association of location data sensors and the detection of real scene objects that are used as natural markers. This way, we search to contribute for the use of full potential of AR application in industrial environments aiming the improvement of maintenance tasks in these scenarios.

This paper is organized as follows. The Section 2 presents and discuss several related works and in Section 3 the technical background on CV and AR. From Section 4 to Section 6 we present the methodology steps described above. In Section 7 we present the results from a case study related to substation scenario followed with conclusion and future works.

2 Related work

In [31] the concept of AR as a novel way to enhance visualization is used for educational purposes. In this work the authors highlight the need of registration, because the change of

camera viewpoint changes the position that the virtual element have in the scene. This is an obstacle to build AR mobile applications.

The use of applications in mobile devices in real time has grown rapidly and several applications for these device formats have been developed. In [12], simple techniques of image processing are used in smartphones for detection of fiducial markers to include 3D elements in real scenes in real time. This related research are enabled by the evolution of the devices and its processing capabilities in the last years. Also, according to [7], the correct position identification to include virtual information is the key for AR.

In [29] there is a strategy presented for detection of ships on the African coast which uses cascade classifiers based in Haar-like features. The work shows a detection system which generally uses the vessel's transponder for monitoring, but in sabotage cases or in transponder shutdown, the solution is to detect it with radar images and a combination of techniques.

The use of techniques such as template matching demonstrate a creation approach of AR environments based in image processing. Besides that, the adoption of mobile devices brought a great advance for AR systems and, according to [26], generally techniques similar to SURF (Speeded- Up Robust Features) [?] have been used for natural markers detection.

Object detection is a fundamental task for AR applications to be used in outdoor environments. The detection results may be used for objects recognition, tracking, and construction of environment maps. In [11] the detection is applied for the use in security robots of the substation with use of object detection algorithm with processing based in cloud computing. Another interesting strategy presented is to reduce the comparison area with costly algorithms such as SURF and SIFT [17] to optimize the processing.

Several works like [16] and [8] point to the question of data communication and architecture questions for electric power systems. Although the emergence of new devices suggest possibilities of using newer technologies in these scenarios. Our work present an approach for new ways of visualization in power system substations.

Despite the importance of new visualizations for power systems, accordingly to [20], in the last 20 years there were few proposals or new ideas in visualization approaches for electric system data. Even though some works were done like [18], [19], [23], [33] and [24]. The above mentioned analysis concludes that visualizations must have the intention of replacing textual data or numeric information and must be explored the visual patterns for a mapping, the most natural possible way, aggregating meaning of the data for the visualization. Information must be understood naturally and colors must be used, but carefully not to generate discomfort for system operators.

A more recent work [1] proposed the use of QR code for AR markers and the use of IEC 61850 communication for automation integrated with SCADA information. According to the authors, many solutions use AR applications for simulation and operator training, and there are several situations where it might be advantageous to have the capability of annotating the real process with information. Besides all important contributions, the above-mentioned solution still needs to include the fiducial markers in real equipment, differently from proposed in our work.

3 Computer vision and augmented reality

We can define AR as a system that combines real world visualization with virtual information objects, requiring real time interactivity and 3D world registration.

Tracking and registration is one of the fundamental tasks for AR systems. These systems must work in real time and, aim to get a credible scene for the user, the real and virtual cameras should be mapped in such a way that both environments precisely match. Thus AR system can take advantage of CV techniques. Because AR can rely on visual features that are naturally present on the real scene, avoiding the need for engineering the environment.

CV aims to obtain geometrical, topological or physical information from an image and the objects which are present in that image. These information may allow the recognition of patterns, object classification, robot movement, among other possibilities. Furthermore, the digital images carry with them information such as colors, light intensity that permit an image analysis through image processing.

To supply the AR need for tracking, fiducial markers are commonly used. A fiducial marker is an easily detection used as a point of reference to an object targeted for tracking. An options is to use natural information of the scene (natural markers). Thus, computer vision techniques can be used with additional information such as georeferencing data, gyroscope data, thermal sensors, among others sensors.

These natural markers generally are specific for each type of application and the object type to be identified. For processing the inclusion of 3D virtual elements, there is a need to align the object coordinate system to the world coordinate system for a proper virtual information placement in the real scene.

4 Natural 3D marker for detection of objects using Haar-like features

To use natural markers, we need to detect the objects that are in the scene image. This depends on the correct selection of the representation model. There are several techniques to detect objects in scene and each different kind of object changes the detection approach. In some situations it is necessary to modify the whole training, even using the same detection technique.

The proposed methodology uses a technique proposed in [32] and extended by [15]. This is a a robust framework for the construction of fast object detectors, using machine learning, that reach high detection rates. The version used includes a rectangular feature rotation, enriching the detection algorithm and keeping calculation efficiency.

Combining the aspects of the works mentioned above, the proposed 3D marker uses cascade classifier based on Haar features. The process is based on two main stages of training. Firstly, the classifier is trained with images containing the object to be detected and these images are called positive examples. During processing these images are resized into a commom size defined empirically. The second stage is responsible for training the negative examples, which are arbitrary images that do not contain the object trained in first step. Our approach have used several images from real industrial environment in both stages.

This is because there is the requirement to detect an object regardless of where the user stands, the proposed work used images from every side of the equipment for the object detector's training. Figure 1 shows that each image matches an equipment side that is used in training process. The overall process of training and detection is presented in Figure 2.

The equipment detection process is performed for each frame acquired from the camera. It is initially performed necessary adjustments in the image related to camera position configuration and operational system, then it executes related procedures for feature detection in image with cascade classifier.

This work have used transformer initially as a case study for detection, since this one is one of the most important equipment inside the substation.

As result of the training, a three dimensional natural marker is created, in which object detection is performed independently of user's viewpoint. Using this approach, it is possible to build AR environments with use of 3D natural markers, without the need to include fiducial markers.

The advantages of using natural markers can be seen in Table 1. It is worth mentioning that

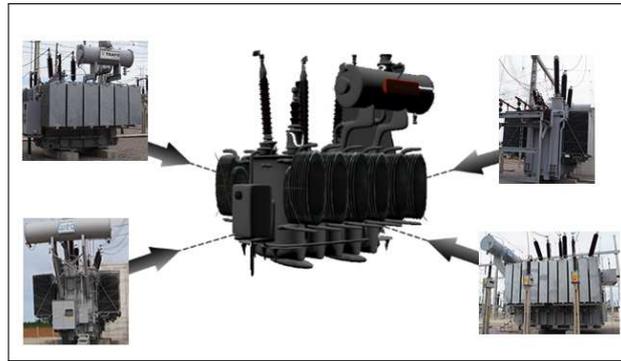


Figure 1: Image acquisition from different sides of object for training the cascade classifier

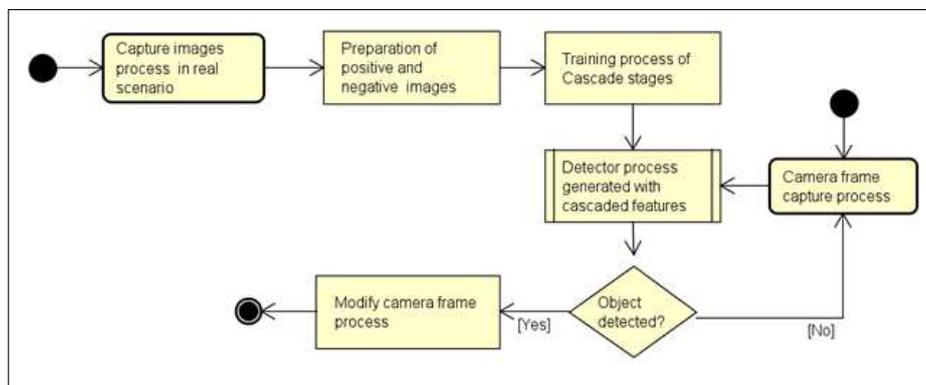


Figure 2: Processes from training to detection of objects in camera

the natural marker corresponds to an object that already exists on scene without modification of it.

Table 1: Comparison of AR markers

Technical feature	Fiducial marker	QR Code	Natural marker
Training independence	No	Yes	No
Model storing	Local	Internet	Local or Internet
Environment remains unchanged	No	No	Yes
Multiple viewpoint detection with a single loaded marker	No	No	Yes

The first step for the creation of the natural 3D marker is the detection. After the correct detection of the object for each camera frame of camera in real time, the next steps consist in associating this information with predefined information about the objects existing in scenario and the use of sensors to be combined with result of detection.

5 Sensors for automatic identification of equipment associated to detection

After the equipment detection we need to identify it. This is a required task to allow the query of the equipment information from automation and control systems. The methodology

proposed herein suggests the integrated use of more than one sensor type. The sensors are used in identifying the natural 3D markers.

Most devices, such as smartphones and digital cameras, have auxiliary sensors for geolocation and orientation. For applications that do not need big precision, only using GPS sensor is possible to solve location problem that has been used mainly in situations with map localizations.

In outdoor applications, such as detection of an equipment inside an industrial area, the use of GPS data does not only constitute a reasonable solution. Due to the error inherent to GPS localization (up to 7.8 meters with 95% of confidence). But, this error can present variation depending on atmospheric effects, reception quality and sky blockage [10]. There are some frameworks for construction of AR environment that use just GPS data as main source of information. These frameworks add virtual information to the real world but this GPS error variation causes annotation out of place in real application.

Besides GPS, this methodology also uses compass or magnetometer sensor. This sensor works like a pointer to the north pole. After the transformer detection, the algorithm considers three main variables: the operator location at the time of the image acquisition from mobile device, the transformer's location present in the substation and the compass orientation sensor for field of view (FOV) calculation from the user point of view, as shown in Figure 3a. The white circles in Figure 3b are user position or equipment inside the substation.

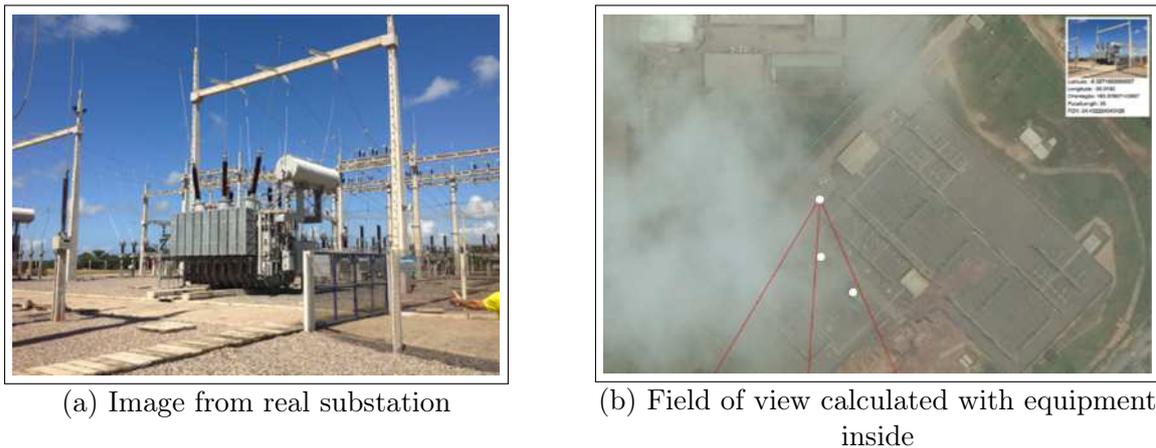


Figure 3: The proposed method for equipment identification from GPS integrated information, compass, user's location, transformers location and image obtained from camera.

The FOV is computed from intrinsic parameters of the camera, as described in Equation 1. The α angle is half of FOV and used to identify which equipment are inside the obtained image.

$$\alpha = \arctan \frac{s}{2f} \quad \begin{array}{l} s = \text{Sensor dimension} \\ f = \text{Focal length} \end{array} \quad (1)$$

Besides that, we use camera information such as sensor size and focal distance to compute the FOV to verify objects inserted inside the camera frame.

By obtaining this FOV, the objects position is analyzed in relation to the central line of FOV and the boundary lines, right and left, in relation to user's viewpoint. With this information it is possible to identify which equipment were detected in the image. With this intention the distance is calculated to the central line and the object is detected at the right side or at the left of the observer (Figure 4). This approach still allows to identify more than one detected object on the same image.

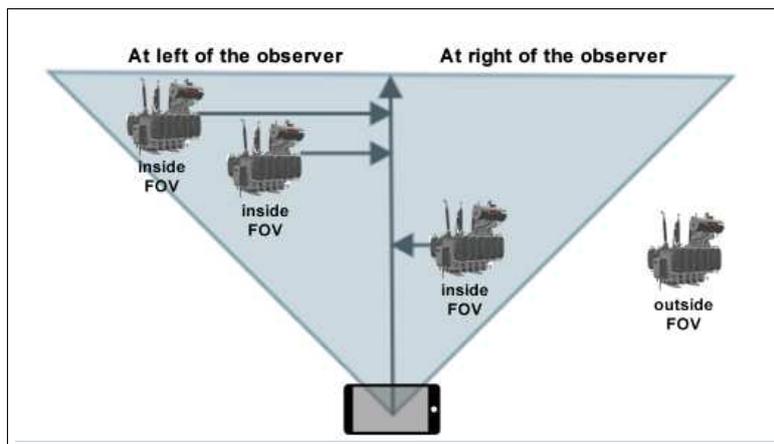


Figure 4: Identification of objects inside the observer viewpoint.

Due to degree of reliability of GPS data in low cost sensors, this proposal can be used on outdoor environments. In indoor scenarios the GPS error increases and the application behavior depends on these data. The strategy of joining 3D natural markers associated with georeferenced data and compass minimizes the error compared to use only GPS data. This strategy works as an auxiliary information layer, containing detection data of the desired object, inside the AR system.

6 Communication and integration to SCADA/EMS

After detection and identification of the equipment inside the AR scenario, a data request is performed about the equipment to SCADA/EMS communication system based on identified equipment tag. Each equipment has a unique tag inside the Open System for Energy Management (SAGE) and through this tag the associated information is obtained.

With this methodology is possible to build AR systems, including real information which could help the system operators in control process and problem identification inside the energy substation, on the field.

Besides that, critical information may be specified, showing just alert states based on defined alarms in the system. Decreasing the data quantity observed by operator.

This integration architecture is based in AGITOServer [27], in which control systems are accessed through interfaces via remote calls (RPC) provided by SAGE and Operator Training Simulator (OTS). Integration with database provide access and control to modify data via TCP/IP protocol available in socket format. After that, it is possible to send and receive messages with data of the power system in JSON format (Figure 5).

This architecture allows perform queries to equipment data, such as equipment electric parameters, operation state (open or closed switch), among others. Through message exchange model, this solution allows to obtain the desired information request about the equipment and add it to the visualization.

Furthermore, the OTS training environment is similar to real environment, only with simulated data of operation state of the systems and chosen obviously in this way for safety reasons of the electric system and enabling consultation application tests to the data.

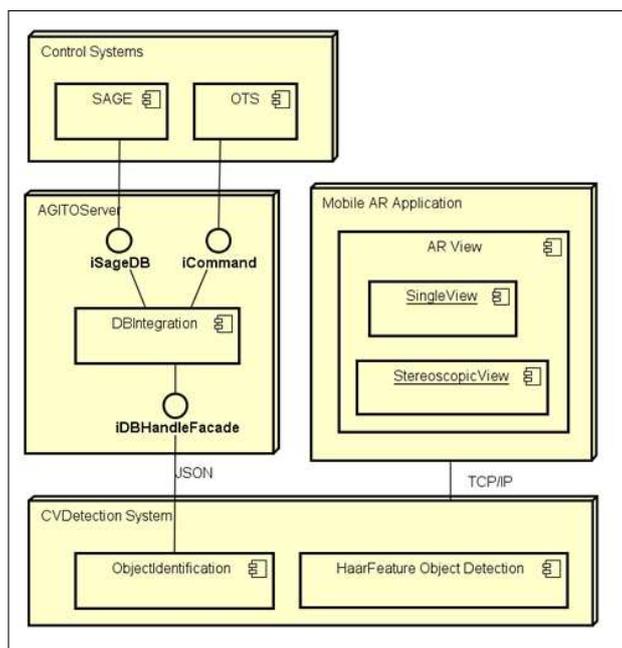


Figure 5: Architecture of acquisition of integrated data to the SAGE/OTS

7 Results and case study: AR visualization in energy substation

The evaluation of the proposed methodology was applied using mobile devices, which allow rendering different visualization formats. It might enable the construction of two augmented reality versions of application: a tablet view version with a single display and a head-mounted display (HMD) version using smartphone integrated to Google Carboard SDK [9].

This possibility of including a mobile tool for operators enables that the user can perform information acquisition in the operation field, with no need of returning to operation center to access a given system which contains desired information. By using the auxiliary use of AR it is possible to add information about the task being executed during operation.

Images used in this study were obtained from real operating environments of CHESF. More specifically at CHESF substations SUAPE II and SUAPE III in Recife-PE, and at CHESF substation Extremoz in Natal-RN, all located in Brazil.

The approach used cascade classifier based in Haar-like features for transformer detection. In this approach the classifier training is performed with negative and positive images set and detection quality depends on the amount of training images.

The obtained results initially used a base of positive images (transformer pictures) and negative images, which include basically areas around the substation itself do not containing the transformer. The results can be seen in Table 2 using amount of $p = 419$ (positive images) and $n = 270$ (negative images) during training and also, for the detection test, a total of 40 pictures of equipment were used.

For elimination of false-positives, the Haar cascade method demands an image base as large as possible, and this statement is valid both for positive images as for negative ones.

These tests show that as we increase the quantity of stages, the algorithm minimizes incorrect detection features. But, to be able to increase equipment detection in these conditions is necessary to increase the training base.

Applications have been developed for devices with android operational system and they have been installed in tablet version (single view) as shown in Figure 6 and HMD combined with

Table 2: Equipment detection with Haar Cascade classifier

Cascade Stages	False positives (FP)	Total features	Detection rate (%)
14	95	222	78.05%
15	73	181	78.05%
16	54	131	80.49%
17	36	120	80.49%
18	22	99	82.93%
19	13	87	82.93%
20	11	73	80.49%
21	10	66	82.93%
22	1	46	70.73%
23	0	33	53.66%
24	0	31	48.78%
25	0	24	43.90%
26	0	18	34.15%

smartphone (using stereoscopic view) in Figure 7a and Figure 7b.



Figure 6: AR application visualization in tablet (single view)

The work proposes real data integration after stages taken related to equipment detection system. The SAGE system is widely used by several electric power system companies in Brazil. According to [5] the SAGE system may be used in substations and power plant and it supports several hardware, including different manufacturers.

In addition to SAGE, the use of OTS has permitted an environment for training operators without the necessity of being connected to real equipment. Thus, it enables running several simulations of possible energy system scenarios.

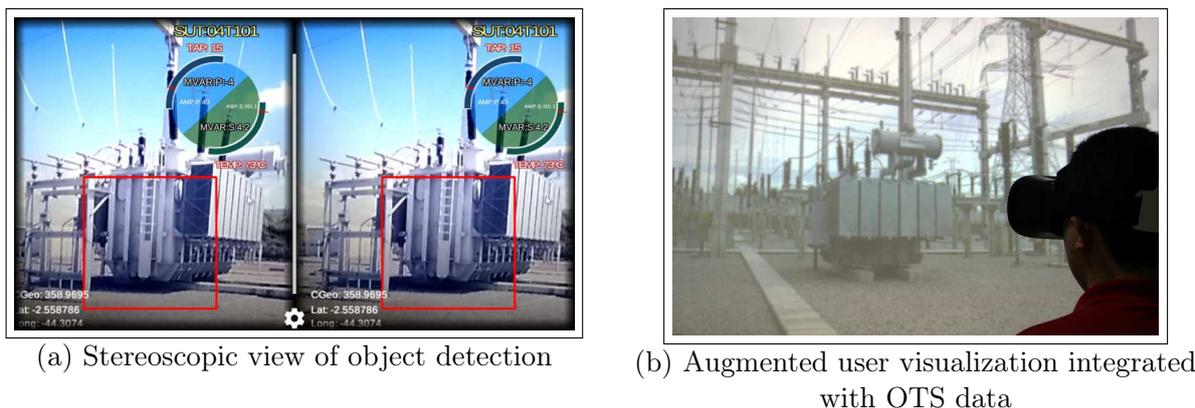


Figure 7: AR application visualization in HMD (stereoscopic view)

8 Conclusion and future work

This work presents two main contributions: the first one, and most important, it is a methodology for creation of AR systems using georeferenced natural 3D markers. It uses image processing techniques and sensor information present in mobile devices to provide identification for detected objects in outdoor environments. This contribution has applicability in others scenarios and several areas of application.

The second contribution is the application constructed itself using the natural marker and integration with mobile sensors, as a case study. The application integrated with SCADA/EMS and equipment's real information create a new visualization format in power system operational environment. This last innovation intends to improve the data visualization inside industrial environments.

This new visualization is applied in stereoscopic view display, which enables the operator to modify the real world through addition of virtual elements (information or auxiliary data of equipment state). It allows the effective creation of an innovative AR solution and the use of this technology can be used no only for training but also for operation tasks.

So, the next step will be experimenting this methodology in real industrial environments. There are many points to test about its usability, mainly on the stereoscopic visualization, due the need of reduce operator's visual discomfort infringed by the use of HMD for several hours. We also will conduce experiments to detect the pose of the object and allow a virtual analysis of it.

Bibliography

- [1] Antonijević M.; Sučić S.; Keserica H. (2016); Augmented Reality for Substation Automation by Utilizing IEC 61850 Communication, *39th International Convention on Information and Communication Technology, Electronics and Microelectronics - Distributed Computing, Visualization and Biomedical Engineering*, Opatija, Croatia, 2016.
- [2] Aukstakalnis S.; Blatner D. (1992); *Silicon Mirage - The Art and Science of Virtual Reality*, Peachpit Press, ISBN: 978-0938151821, 317, 1992.
- [3] Azuma R.T. (1997); A Survey of Augmented Reality, *Presence: Teleoperators and Virtual Environments*, 355-385, 1997.

-
- [4] Bay H.; Ess A.; Tuytelaars T.; Gool L. V. (2008); Speeded-up Robust Features (SURF), *Computer Vision and Image Understanding*, 110(3), 346-359, 2008.
- [5] CEPEL (2004); Open Energy Management System (*Sistema Aberto de Gerenciamento de Energia*), [Online] Available: <http://sage.cepel.br>
- [6] Côté S.; Barnard J.; Snyder R. (2013); Offline Spatial Panoramic Video Augmentation for Visual Communication in the AEC Industry, *13th International Conference on Construction Applications of Virtual Reality (CONVR 2013)*, 82-89, 2013.
- [7] Dai J.; Zhang L. (2011); Registered Localization Method-Based on OpenCV for Augmented Reality, *International Conference of Soft Computing and Pattern Recognition (SoCPaR)*, 492-495, 2011.
- [8] Farag H. E. Z.; El-Saadany E.F. (2012); A Novel Cooperative Protocol for Distributed Voltage Control in Active Distribution Systems, *IEEE Transactions on Power Systems*, 1645-1656, 2012.
- [9] Google (2016); Google Cardboard, [Online] Available at <http://googlecardboard.com.br>
- [10] GPS (2014); GPS Accuracy, *Official U.S. Government information about the Global Positioning System*, [Online] Available: <http://www.gps.gov/systems/gps/performance/accuracy/>
- [11] Jian Y.; Xin W.; Xue Z.; Zhenyou D. (2015); Cloud Computing and Visual Attention Based Object Detection for Power Substation Surveillance Robots, *IEEE 28th Canadian Conference on Electrical and Computer Engineering*, 337-342, 2015.
- [12] Kim J.; Jun H. (2011); Implementation of Image Processing and Augmented Reality Programs for Smart Mobile Device, *The 6th International Forum on Strategic Technology*, 1070-1073, 2011.
- [13] Li Z.; Wang J.; Sun H.; Guo Q. (2015); Transmission Contingency Analysis Based on Integrated Transmission and Distribution Power Flow in Smart Grid, *IEEE Transactions on Power Systems*, 3356-3367, 2015.
- [14] Liang M.; Min H.; Luo R.; Zhu J. (2015); Simultaneous Recognition and Modeling for Learning 3-D Object Models From Everyday Scenes, *IEEE Transactions on Cybernetics*, 45(10), 2237-2248, 2015.
- [15] Lienhart, R.; Maydt, J. (2002); An Extended Set of Haar-like Features for Rapid Object Detection, *IEEE International Conference on Image Processing (ICIP)*, 900-903, 2002.
- [16] Lin H.; Sambamoorthy S.; Shukla S.; Thorp, J.; Mili L. (2011); Power System and Communication Network Co-Simulation for Smart Grid Applications, *Innovative Smart Grid Technologies (ISGT)*, 1-6, 2011.
- [17] Lowe D.G. (2004); Distinctive Image Features from Scale-Invariant Keypoints, *International Journal of Computer Vision*, 60(2), 91-110, 2004.
- [18] Mahadev P. M.; Christie R. D. (1993); Envisioning Power System Data: Concepts and a Prototype System State Representation, *IEEE Transactions on Power Systems*, 8(3), 1084-1090, 1993.

- [19] Mahadev P. M.; Christie R. D. (1994); Envisioning power system data: vulnerability and severity representations for static security assessment, *IEEE Transactions on Power Systems*, 9(4), 1915-1920, 1994.
- [20] Mikkelsen C.; Johansson J.; Cooper M. (2012); Visualization of Power System Data on Situation Overview Displays, *16th International Conference on Information Visualisation*, 188-197, 2012.
- [21] Milgram P.; Kishino F. (1994); A Taxonomy of Mixed Reality Visual Displays, *IEICE Trans. Information and Systems*, 1321-1329, 1994.
- [22] Nigam S.; Deb K.; Khare A. (2013); Moment Invariants based Object Recognition for Different Pose and Appearances in Real Scenes, *International Conference on Informatics, Electronics & Vision (ICIEV)*, ISBN: 978-1-4799-0397-9, 1-5, 2013.
- [23] Overbye T. J.; Weber J. D. (2000); New Methods for the Visualization of Electric Power System Information, *IEEE Symposium on Information Visualization*, ISSN: 1522-404X, 1-9, 2000.
- [24] Overbye T. J.; Weber J. D. (2000); Visualization of power system data, *Proceedings of the 33rd Annual Hawaii International Conference on System Sciences*, 1-7, 2000.
- [25] Pribeanu C. (2012); Specification and Validation of a Formative Index to Evaluate the Ergonomic Quality of an AR-based Educational Platform, *International Journal of Computers Communications & Control*, 7(4), 721-732, 2012.
- [26] Prochazk D.; Stencl M.; Popelka O.; Stastny J. (2011); Mobile Augmented Reality Applications, *Proceedings of Mendel 2011: 17th International Conference on Soft Computing*, 469-476, 2011.
- [27] Ribeiro T. R.; Reis P. R. J.; Braz Jr., G.; Paiva A. C.; Silva A. C.; Maia I. M. O.; Araújo A. S. (2014); Agito: Virtual reality environment for power systems substations operators training, *International Conference on Augmented and Virtual Reality (AVR)*, 113-123, 2014.
- [28] Rublee E.; Rabaud, V.; Konolige K.; Bradski, G. (2011); ORB: an efficient alternative to SIFT or SURF, *IEEE International Conference on Computer Vision*, 2564-2571, 2011.
- [29] Schwegmann C.P.; Kleynhans W.; Salmon B.P. (2014); Ship detection in South African oceans using SAR, CFAR, and a Haar-like feature classifier, *Geoscience and Remote Sensing Symposium (IGARSS)*, 557-570, 2014.
- [30] Szeliski R. (2011); *Computer Vision: Algorithms and Applications*, Springer-Verlag London, ISBN: 978-1-84882-935-0, 812, 2011.
- [31] Vargas H.; Farias G.; Sanchez J.; Dormido S.; Esquembre F. (2013); Using Augmented Reality in Remote Laboratories, *International Journal of Computers Communications & Control*, 8(4), 622-634, 2013.
- [32] Viola P.; Jones M. J. (2001); Rapid Object Detection using a Boosted Cascade of Simple Features, *Computer Vision and Pattern Recognition (CVPR)*, 511-518, 2001.
- [33] Weber J. D.; Overbye T. J. (2000); Voltage Contours for Power System Visualization, *IEEE Transactions on Power Systems*, 15(1), 404-409, 2000.

PID and Fuzzy-PID Control Model for Quadcopter Attitude with Disturbance Parameter

E. Kuantama, T. Vesselenyi, S. Dzitac, R. Tarca

Endrowednes Kuantama*, **Tiberiu Vesselenyi**,
Simona Dzitac, **Radu Tarca**

University of Oradea

1 Universitatii St., Oradea, Romania

endrowednes@gmail.com, vtiberiu@uoradea.ro,

simona@dzitac.ro, rtarca@uoradea.ro

*Corresponding author: endrowednes@gmail.com

Abstract: This paper aims to present data analysis of quadcopter dynamic attitude on a circular trajectory, specifically by comparing the modeling results of conventional Proportional Integral Derivative (PID) and Fuzzy-PID controllers. Simulations of attitude stability with both control systems were done using Simulink toolbox from Matlab so the identification of each control system is clearly seen. Each control system algorithm related to roll and pitch angles which affects the horizontal movement on a circular trajectory is explained in detail. The outcome of each tuning variable of both control systems on the output movement is observable while the error magnitude can be compared with the reference angles. To obtain a deeper analysis, wind disturbance on each axis was added to the model, thus differences between each control system are more recognizable. According to simulation results, the Fuzzy-PID controller has relatively smaller errors than the PID controller and has a better capability to reject disturbances. The scaling factors of gain values of the two controllers also play a vital role in their design.

Keywords: Quadcopter, PID, fuzzy-PID, Simulink, attitude.

1 Introduction

A quadcopter, as its name suggests, has four symmetrical rotor-propellers which are mounted at the same height on each edge, forming a cross with identical radius from its center [9]. Each rotor propeller creates an upward thrust which is highly maneuverable by controlling the speed variation, thus allowing the quadcopter for vertical take-off, landing and even hovering. However, to achieve good stability as well as to minimize interference in every movement, the quadcopter attitude must be converted into a suitable control algorithm. This is considered to be a challenging problem because quadcopters have a complex dynamic system with high nonlinearities, strong couplings, and under actuation [13]. According to the control strategy of quadcopters, the flight control systems can be divided into 3 types, i.e., linear flight control system, nonlinear flight control system, and learning based flight control system [5]. The linear flight control system, or feedback linearization, is a very useful method to decouple and linearize quadcopter attitude models. This paper will review in detail the algorithm of linear control system for quadcopters using conventional PID (Proportional Integral Derivative) and Fuzzy-PID control. Considering the advantages of the Ziegler-Nichols rules for tuning the PID controller [22] and the feedback linearization method, we adopted this algorithm for the quadcopter attitude control and we analyzed the obtained results. The way of tuning PID controller parameters with mathematical modeling into feedback linearization for stability attitude is crucial. The way of tuning PID controller parameters with mathematical modeling using feedback linearization is crucial for attitude stability. On the other hand, Fuzzy-PID algorithms have limitless ranges to flawlessly perform when rapid changes are encountered in high dynamic environments, moreover

it does not require complex mathematical models [7]. A comparison of the PID and Fuzzy-PID tuning is presented in which both controllers have been studied based on a dynamic model of a quadcopter.

Modeling of the quadcopter control and movement has been done using Simulink Matlab. Through this modeling and control algorithm, differences between angular velocity with and without PID and FUZZY-PID control can be studied, as well as the impact of disturbances on the control system. To understand the control system of quadcopters, firstly one must comprehend the concept of quadcopter attitude based on the Newton-Euler equation. Details of the quadcopter dynamic movement using this equation have been already covered in previous researches [12], [15]. This time, Newton-Euler equation is adapted to develop a quadcopter attitude model in which the quadcopter moves in a full circle while using PID and Fuzzy-PID control systems for dynamic movement control. Another aim of this study is to investigate the behavior of the quadcopter control system when faced with wind disturbances from multiple directions.

2 Quadcopter attitude modelling

Vertical thrust is the result of combined angular velocity of four rotors which are directly connected to the propellers, thus driving the quadcopter attitude as a whole. Naturally, rotor propellers placed in opposed position will rotate in the same direction and side-by-side ones will rotate oppositely. Controlling the angular velocity of the four rotors will consequently result in a dynamic movement with force and moment transmission which can lead to lift, pitch, roll or yaw rotations moment inertia. Fig. 1. illustrates these rotations, and are explained as follows: rotation around X-axis as roll movement, rotation around Y-axis as pitch movement, and rotation around Z-axis as yaw movement.

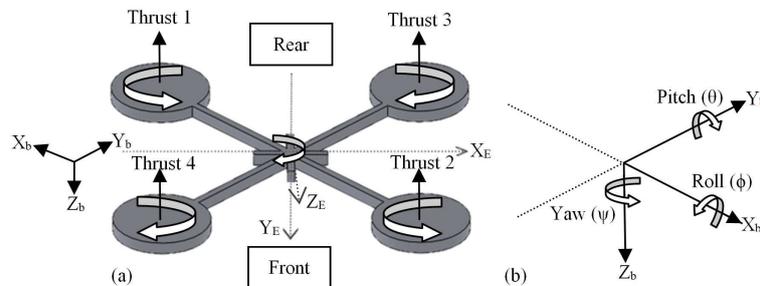


Figure 1: Quadcopter schematic: (a) The structure of a quadcopter (b) Description of Pitch, Roll, and Yaw angles

Quadcopter attitude has a total of six Degrees of Freedom (DOF), half are allotted to translational position in an Earth inertial frame and the rest to angular positions of the quadcopter's rigid body. All DOFs are covered by the linear and angular orientation vectors of the quadcopter, thus it can be seen that the quadcopter has four inputs from the rotor propellers and six state outputs of the dynamic and complex system. In order to control these, the process of quadcopter modeling can be approached by interpreting any attitude movement with two reference frames in the form of the Earth inertial frame (coordinate system) axes (X_E , Y_E , Z_E) and the body frame axes (X_b , Y_b , Z_b). From this point, we can formulate the equations and parameters needed to establish the transformation matrix based on the Newton-Euler equation model [3]. Each parameter needed for attitude modeling can be seen in Table. 1.

Table 1: Parameter for attitude modelling algorithm

Movement	Translation Positions	Angular Positions	Translation Velocity	Angular Velocity	Inertial Matrix
Roll	x	ϕ	v_x	p	I_{xx}
Pitch	y	θ	v_y	q	I_{yy}
Yaw	z	ψ	v_z	r	I_{zz}

2.1 Newton-Euler kinematics model

The Newton-Euler modeling was adapted in designing linear (L_T) and angular motion (R_T) expressions of the quadcopter which can deliver lift up forces where change of attitude occur on the orientation of the body frame axes with respect to the Earth inertial frame axes. Orientation change of body frame axes results in angle alteration on roll - $R_x(\phi)$, pitch - $R_y(\theta)$, and yaw - $R_z(\Psi)$, respectively. Based on the kinematic moving frame theory, the equations using transformation matrices for angular motion on each axis or for every movement can be written. Matrix transformation between two rectangular coordinate systems is orthogonal and can be converted to 3x3 matrices for every axis. Eq.(1). shows the rotation matrix around X-, Y-, and Z-axis and the transformation of the overall rotation movement (R_T) of the body frame can be calculated using Eq.(2).

$$R_x(\phi) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\phi & \sin\phi \\ 0 & -\sin\phi & \cos\phi \end{bmatrix}; R_y(\theta) = \begin{bmatrix} \cos\theta & 0 & -\sin\theta \\ 0 & 1 & 0 \\ \sin\theta & 0 & \cos\theta \end{bmatrix}; R_z(\psi) = \begin{bmatrix} \cos\psi & \sin\psi & 0 \\ -\sin\psi & \cos\psi & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad (1)$$

$$R_T = R_x(\phi)R_y(\theta)R_z(\psi) \quad (2)$$

The rate of change in linear motion from one position to another is related to the angular velocity and the difference in angle which will result in alteration of system output x, y, z position. This alteration can be computed using the angular rates measured by sensor of the quadcopter (p, q, r). The rotation angle in Eq.(1). is used to compute the angular velocity of the quadcopter. The angular velocity with respect to the body frame can be computed by transforming the matrix from angular position $[p \ q \ r]^T$ to Euler angle rates $[\phi' \ \theta' \ \psi']^T$ given in Eq.(4). In fact, the values of p, q and r are obtained through orientation sensor readings on the quadcopter, thus the alteration of the angles of the body frame can be computed using the inverse of Eq.(4)., as it is seen in Eq.(5).

$$\mathcal{V} = \begin{bmatrix} p \\ q \\ r \end{bmatrix} = R_x(\phi)R_y(\theta)R_z(\psi) \begin{bmatrix} 0 \\ 0 \\ \psi' \end{bmatrix} + R_x(\phi)R_y(\theta) \begin{bmatrix} 0 \\ \theta' \\ 0 \end{bmatrix} + R_x(\phi) \begin{bmatrix} \phi' \\ 0 \\ 0 \end{bmatrix} \quad (3)$$

$$\begin{bmatrix} p \\ q \\ r \end{bmatrix} = \begin{bmatrix} 1 & 0 & -\sin\theta \\ 0 & \cos\phi & \sin\phi\cos\theta \\ 0 & -\sin\phi & \cos\theta\cos\phi \end{bmatrix} \begin{bmatrix} \phi' \\ \theta' \\ \psi' \end{bmatrix} \quad (4)$$

$$\begin{bmatrix} \phi' \\ \theta' \\ \psi' \end{bmatrix} = \begin{bmatrix} 1 & \sin\phi\tan\theta & \cos\phi\tan\theta \\ 0 & \cos\phi & -\sin\phi \\ 0 & \sin\phi/\cos\theta & \cos\phi/\cos\theta \end{bmatrix} \begin{bmatrix} p \\ q \\ r \end{bmatrix} \quad (5)$$

The dynamic movement of the quadcopter, both on angular and translational basis are expressed with the Newton-Euler equation of motion as shown in Eq.(6). The angular acceleration

of the inertia ($I\omega'$), the centripetal forces ($\omega \times (I\omega)$) and the gyroscopic forces (γ) are equal to the external torque applied to the bodyframe (τ). From this equation, angular velocity vector (ω') can be determined as shown in Eq.(8).

$$\tau = I\omega' + (\omega \times (I\omega)) + \gamma \quad (6)$$

$$\omega' = I^{-1}[-\omega \times (I\omega) + \tau - \gamma] \quad (7)$$

$$\begin{bmatrix} p' \\ q' \\ r' \end{bmatrix} = I^{-1} \left(\begin{bmatrix} (I_{yy} - I_{zz})qr \\ (I_{zz} - I_{xx})pr \\ (I_{xx} - I_{yy})pq \end{bmatrix} + \begin{bmatrix} \tau_\phi \\ \tau_\theta \\ \tau_\psi \end{bmatrix} - I_m \omega_i \begin{bmatrix} q \\ -p \\ 0 \end{bmatrix} \right) \quad (8)$$

Where I_m is the Inertia matrix, ω is the angular velocity, ω' is the angular position, and ω_i is the scalar angular speed of rotor. The inertia matrix for the quadcopter is diagonal due to the symmetry of the quadcopter. Quadcopter's mass and its geometric distribution (especially inertia) are affecting the entire dynamics of the system. The symbols I_{xx} , I_{yy} , I_{zz} are the moments of inertia matrices which are the torque of the rotors.

2.2 Circular trajectory model

Analysis of attitude control was done by modeling the dynamic of the quadcopter moving on a circular trajectory. This implies that the X-Y position on the circular trajectory needs to be computed using stability control on roll and pitch motions while yaw angle is zero. The path of motion along X- and Y-axes can be seen in Fig. 2. [18].

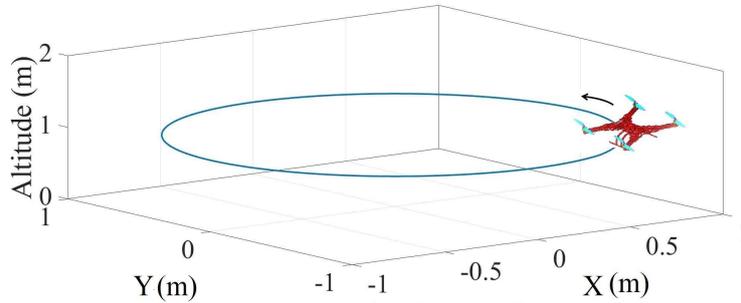


Figure 2: Quadcopter attitude trajectory model

As it is known, two rotor propellers will cause movement on each axis around the center of gravity. In this modeling, rotors placed on the same axis will generate opposite thrust values, or in other words, the increasing thrust on one side is linear with the decreasing thrust on the other side in the same axis. The rotation torques on the body frame which derives from the rotor on each axis are dependent upon two parameters, i.e., the distance between rotor and quadcopter's centroid (L) and the magnitude of thrust generated by the rotors on the same axis (X-axis or Y-axis). The torque equation on the body frame on each axis is given in Eq.(9).

$$\tau_x = L(T_{rotor4} - T_{rotor2}); \tau_y = L(T_{rotor3} - T_{rotor1}) \quad (9)$$

From the quadcopter theory and attitude parameter, basic modeling was done with only a proportional control system. Close loop feedback was designed to ensure a good functionality of the algorithm for quadcopter attitude, as well as to evaluate the output of the angular velocity on every axis. There are three subsystems of the designed model, i.e., input parameters in form

of inertia matrices I_{xx} , I_{yy} , I_{zz} and distance between rotor and centroid (L), the algorithm for quadcopter attitude and circular trajectory, and lastly the output in form of quadcopter movement on X- and Y-axis. The simulation block diagram for this model can be seen in Fig. 3. The quadcopter parameters used in this model are as follows: $L = 0.25$ m and $I_{xx} = I_{yy} = 0.005$ $kg.m^2$. Because the modeling focused only on X- and Y-axis, I_{zz} was temporarily unneeded.

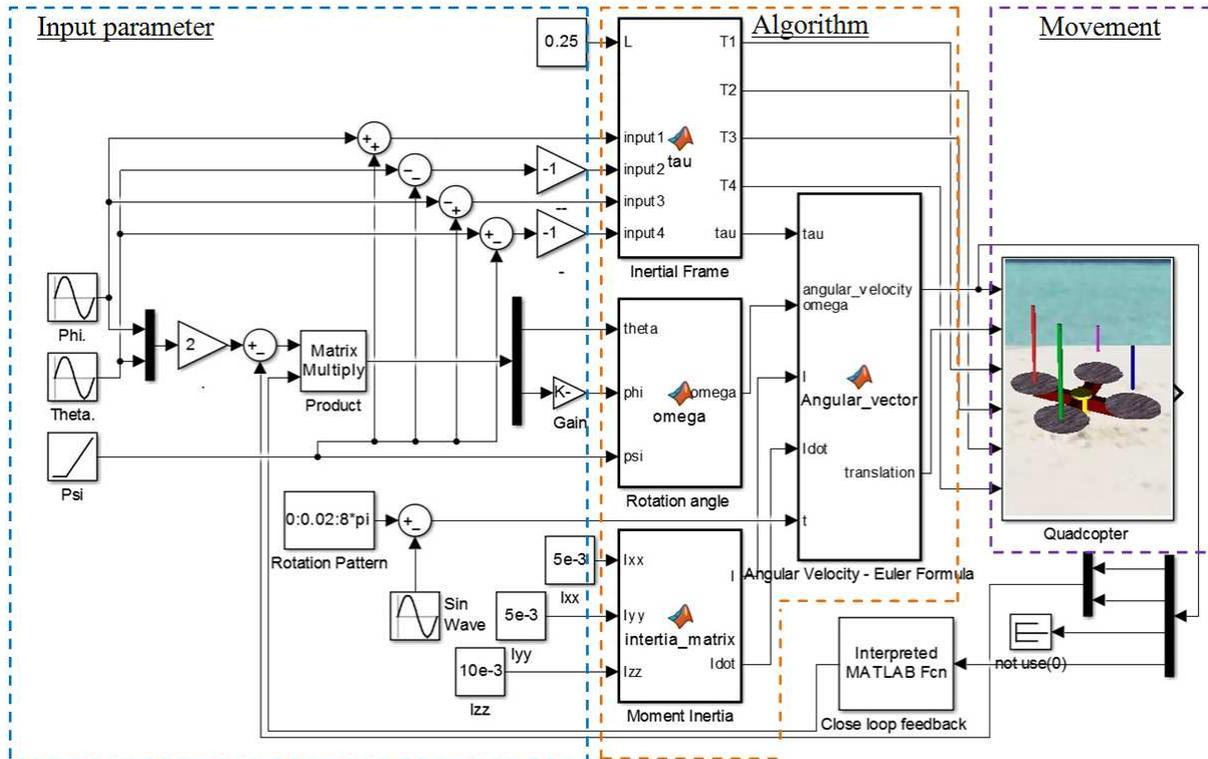


Figure 3: Simulation block diagram

3 Controller architecture

Analysis of the matrices function for every attitude reveals the fact that the quadcopter model requires a close feedback linearization control system for balancing and localization. This control architecture is necessary to enhance the dynamic control response of the quadcopter with respect to the quadcopter body frame. In this study, the movement on a circular trajectory was investigated and to achieve stability, PID and Fuzzy-PID control systems have been used [2]. Generally, the control system for the quadcopter moving on a circular trajectory, in fly/hover mode, or in other attitude movement can be done with the same control algorithm to attain stability. Fig. 4 shows the quadcopter control system with close loop feedback and linearized PID or Fuzzy-PID control. In this paper, the controller for X and Y position will compute the desired roll and pitch angles depending on the desired values for quadcopter movement on each axis. These angles are fed into the "Attitude" block and with a closed-loop feedback the controller can continuously do a linearization process against any possible error factor.

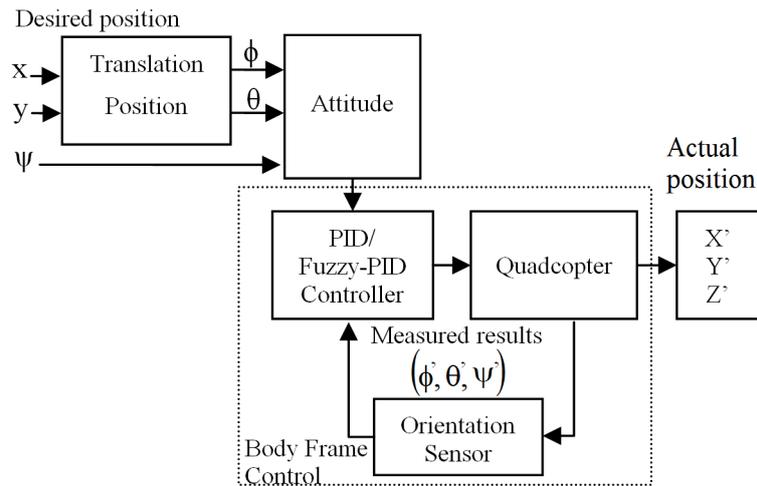


Figure 4: Quadcopter control system diagram

3.1 PID control parameter

Implementation of PID controller tuned by Ziegler-Nichols rules to the linear model of quadcopter can be done by adopting the parameter feedback linearization items which are acting on the quadcopter position. The advantages of the PID controller is its simple control structure and the ease of implementation [4]. A PID controller regards an "error" value as the difference between a measured variable and a desired set-point. The equation of a PID controller is:

$$U(t) = K_p e(t) + K_i \int_0^t e(\tau) d\tau + K_d \frac{d}{dt} e(t) \quad (10)$$

$$e(t) = X_d(t) - X(t) \quad (11)$$

In which $U(t)$ is the control output, $e(t)$ is the difference between the desired stated $X_d(t)$ and the actual state $X(t)$, and K_p, K_i, K_d are the proportional, integral, and derivative gains, respectively. Alteration in pitch and roll angles may occur based on the torque generated torque by the rotors on X-axis or Y-axis on the body frame. In this modeling, it was sufficient just to manage the K_p and K_d control parameters (P and D control), as K_i was used to reduce the final error of the system and thus it is not needed [10]. The nominal coefficients of PD used to obtain the robust stability equation are based on the angular velocity equation of the quadcopter [14]. Hence, the PD controller for the quadcopter in horizontal position can be calculated using Eq.(12). based on two relations, i.e., the torque generated by each horizontal axis (shown in Eq.(9).) and also the equation of angle alteration (shown in Eq.(4).). The total mass of the quadcopter will have impact on the generated thrust, whereas the torque will be influenced by the moments of inertia (I).

$$\tau_{x,y} = \begin{bmatrix} L(T_{rotor4} - T_{rotor2}) \\ L(T_{rotor3} - T_{rotor1}) \end{bmatrix} = \begin{bmatrix} (K_{\phi,p}(\phi_p' - \phi') + K_{\phi,D}(\phi_d' - \phi'))I_{xx} \\ (K_{\theta,p}(\theta_p' - \theta') + K_{\theta,D}(\theta_d' - \theta'))I_{yy} \end{bmatrix} \quad (12)$$

The performance of PD controller was tested using stability simulation of quadcopter during the movement on the circular trajectory. The PD controller parameters can be seen in Fig. 5. where the values of parameters K_p and K_d were obtained based on manual tuning which was done to get magnitude stability of the attitude value. Parameter setting principles are based on the Ziegler-Nichols PD. The amount of error that occurs on the roll angle is not significant, so

it only takes the proportional gain value to adjust the generated output variable. The purpose of the stabilization is to obtain the same output of angular velocity in any desired position.

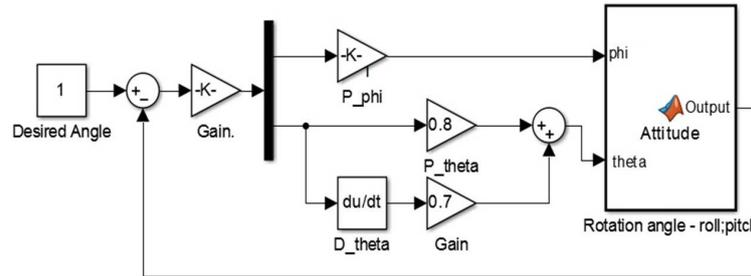


Figure 5: Quadcopter PID control system Simulink block diagram

3.2 Fuzzy-PID control parameter

As for the combination between PID control theory and Fuzzy control theory, the Fuzzy-PID control aims to establish a two variable continuous function between the PID parameters, absolute error value and absolute error value changes [11]. It has a self-tuning characteristic to automatically tune the control parameter to obtain optimal output [21]. Optimal output can be obtained by adjusting the gain values of $K_{p-Fuzzy}$, $K_{i-Fuzzy}$, $K_{d-Fuzzy}$ based on the comparison of the input in form of error and error rate values. The input will be analyzed using a Fuzzy system in which if-then rules are applied as tuning parameter set of the received input [1], [19]. In accordance with the typical PID controller, the gain values of the Fuzzy-PID controller can be adjusted with the parameters as seen on Eq.(13). [8].

$$U_{Fuzzy-PID} = K_{p-Fuzzy}e(t) + K_{i-Fuzzy} \int_0^t e(\tau)d\tau + K_{d-Fuzzy} \frac{d}{dt}e(t) \quad (13)$$

A Fuzzy-PID controller is proposed in this work to control the quadcopter stability in terms of attitude position on circular trajectory, considering that the quadcopter movements are heavily influenced by non-linear factors and dynamic coupling. The Fuzzy-PID control attitude on pitch and roll positions are designed separately in order to accurately control the PID gain values. As it is known through PID control parameter, in this design, the PD gain values are enough to control the quadcopter movement on a horizontal path.

The position of the quadcopter on a circular trajectory depends on the acceleration of the quadcopter and is based on the change in pitch and roll angles. Each angle changes will be monitored and controlled using a Fuzzy-PID controller. Each controller has 2 membership functions where the first one, represented by In1, receives the current angle which is obtained from the orientation sensor reading of the quadcopter attitude; The other one (In2) receives the error rate which is obtained from comparing current angle value with the reference angle value. Through these 2 membership functions, the Fuzzy PID system with If-Then logic or Fuzzy rule parameters can be applied to reduce the position error value on output in the form of quadcopter movement. Mamdani-type inference is used for the inference engine with centroid defuzzification method [16] and it lays out the foundation rules for the Fuzzy system [20]. Each input and output are divided into 7 groups of linguistic variables, i.e., negative large (NL), negative medium (NM), negative small (NS), zero (Z), positive small (PS), positive medium (PM), and positive large (PL). These linguistic variables express the degree of error and also the error rate in roll and pitch angle while moving on the circular trajectory. The details of these variables can be seen in Fig. 6. Two inputs and one output are normalized in the same interval [-1, 1] with a

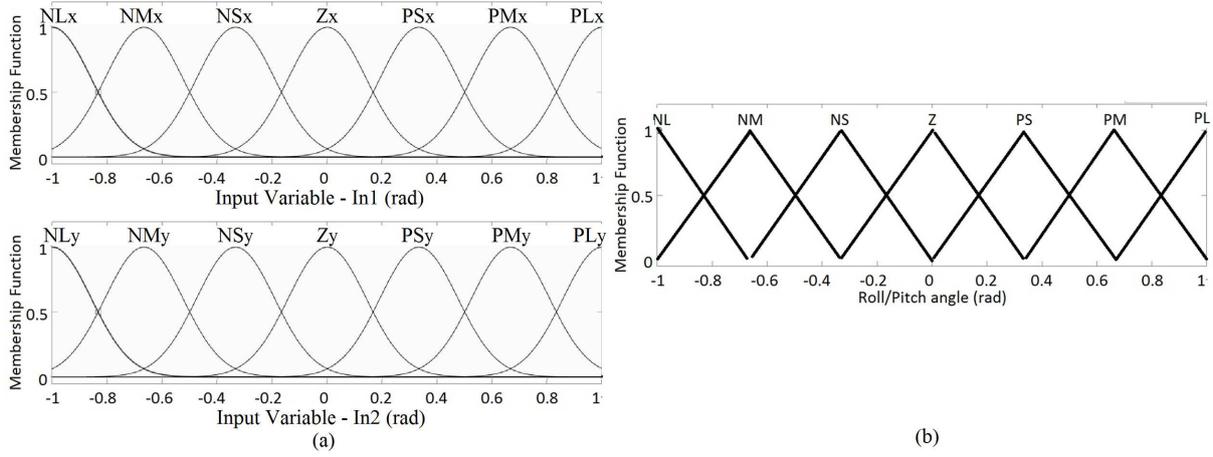


Figure 6: The membership function of the Fuzzy logic controller: (a) Input (b) Output

scaling factor variable of 0.2 for every linguistic group. The interval value must be equivalent with the circular trajectory movement values in the designed roll and pitch angle so that each scale will cover detailed changes. Table 2 shows 49 fuzzy inference rules of this control action with the possible combinations of the two input values.

Table 2: Fuzzy inference rules for quadcopter

In1/In2	NLy	NMy	NSy	Zy	PSy	PMy	PLy
NLx	NL	NL	NL	NL	NM	NS	Z
NMx	NL	NL	NL	NM	NS	Z	PS
NSx	NL	NL	NM	NS	Z	PS	PM
Zx	NL	NM	NS	Z	PS	PM	PL
PSx	NM	NS	Z	PS	PM	PL	PM
PMx	NS	Z	PS	PM	PL	PL	PL
PLx	Z	PS	PM	PL	PL	PL	PL

In1 = derivative value/error position; In2 = error rate

After establishing the fuzzy parameter values and its rules, the Fuzzy-PID system as seen in Fig. 7 can be designed. Two identical Fuzzy logic controllers are designed to control the movement in roll and pitch angles on the circular trajectory. Inputs received by the Fuzzy system are measured through the error normalization factor (GE) and the change in measurement normalization factor (GCE). This factor can be calculated based on the values of $K_{p-Fuzzy}(K_{pf})$, $K_{i-Fuzzy}(K_{if})$, $K_{d-Fuzzy}(K_{df})$, and the allowed maximum error as seen on Eq.(14). and Eq.(15).

$$GE = \frac{1}{max.error} \quad (14)$$

$$GCE = GE * (K_{pf} - \sqrt{K_{pf}^2 - 4K_{if}K_{df}} \frac{K_{if}}{2}) \quad (15)$$

Output responses of the Fuzzy are measured through response de-normalization factor (GU) and change in the response de-normalization factor (GCU) as seen in Eq.(16) and Eq.(17). This scaling factor is depending on gain information of conventional PID.

$$GU = \frac{K_{df}}{GCE} \quad (16)$$

$$GCU = \frac{K_{if}}{GE} \quad (17)$$

In this design, output normalization was done by directly triggering the changes of proportional and derivative action in which only the GU variable is used. The amount of designed error normalization factor is 0.5 with a change in measurement normalization factor of 0.3. The output of the PID controller will be used in Euler integration for the quadcopter attitude on a circular trajectory.

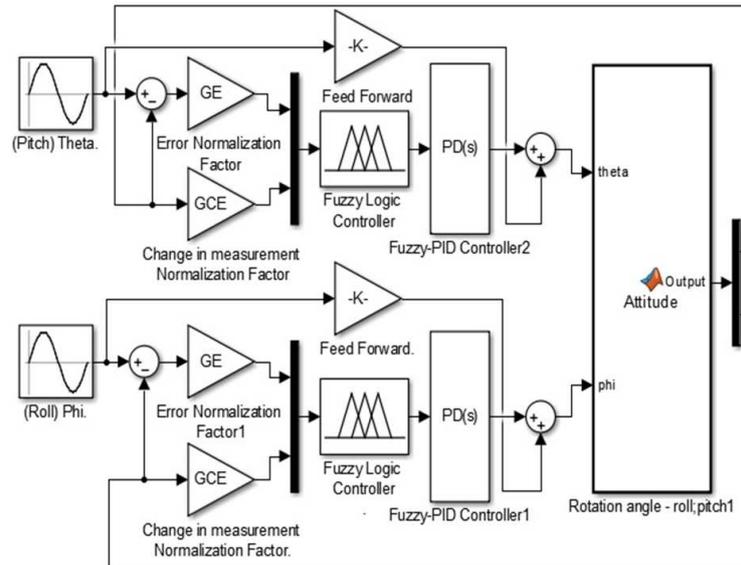


Figure 7: Quadcopter Fuzzy-PID control system Simulink block diagram

4 Simulation results

The circular movement of the quadcopter was simulated using Simulink modeling which demonstrated the applicability of designed control in terms of control system stability on roll and pitch axes with steady high performance. The results are represented as angular velocity and position control responses of the control algorithm. Circular trajectory modeling was performed by adjusting pitch and roll angle in horizontal position as shown in Fig. 8(a). which was also used as reference angle. In quadcopter movement, the angular velocity algorithm can be developed using Eq.(9). and the modeling was done with a closed loop system. Practically, output angles for roll and pitch were obtained from attitude reading or orientation sensor data and the outcome was compared with the reference angles to observe how much error occurred. Fig. 8(b). shows the output roll and pitch angles, and the differences between the desired angles and the outputs.

Due to the symmetry of the quadcopter, the same attitude error occurred on pitch and roll angles where noticeable error was seen in the settling time of 3-4 seconds and also in the settling time 7-8 seconds with an overshoot greater than 20 %. The differences between the generated attitude and the reference angles proved that the control algorithm is crucial to lessen the error. PID and Fuzzy-PID control systems are essential to cope with this problem and the differences between their usages are seen clearly from the output. The tuning of the PID control parameters for both roll angle and pitch angle loops can be accomplished by adjusting the proportional and derivative coefficient. The proportional gain coefficient is used to manage acceleration and stability, while the integral coefficient is used to reduce permanent fault, and lastly the derivative

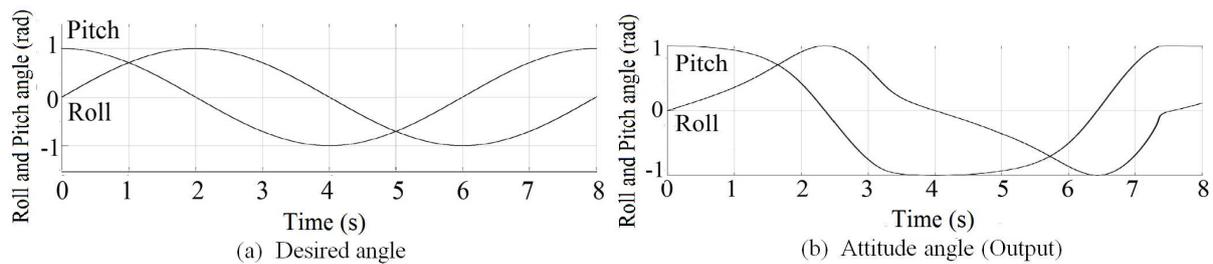


Figure 8: Quadcopter roll and pitch angle

coefficient is used to reduce oscillation range and increase stability [17]. No steady state error was observed on roll and pitch angles, thus the PID controller could be simplified into a PD controller with integral gain $K_i = 0$. The complete position and attitude controller was developed to follow a circular trajectory with values K_p of 0.8 and K_d of 0.7, for which the Simulink block diagram is presented in Fig. 5. These gains of the PD controller were used as a set of operation points for the roll and pitch control. Fig. 9. shows the comparison between the performance of the system with PD controller and the reference angle. The variable position output is explained on each axis in which x-axis position is called roll position and y-axis position is called pitch position.

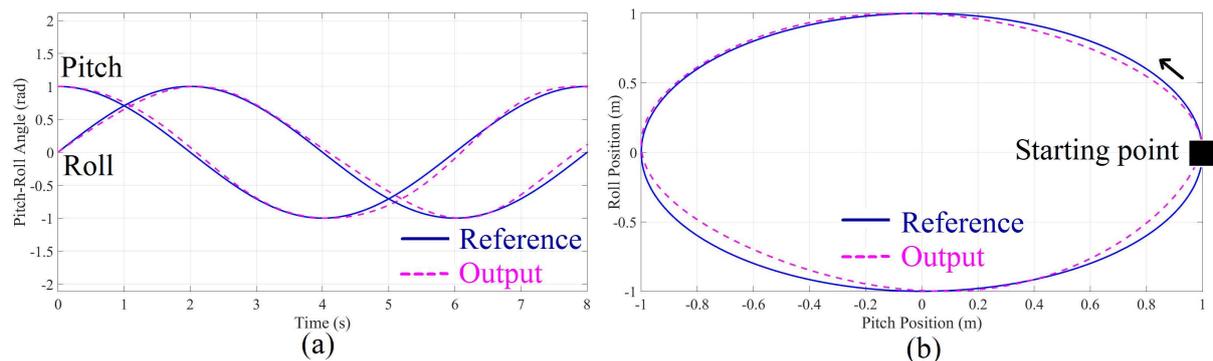


Figure 9: Quadcopter Attitude with PD control system. (a) Pitch and roll angle (b)Trajectory position

The influence of using PD control is very clear where overall output is almost identical to the reference, establishing an proper movement on circular trajectory. This data attest that the final values of proportional and derivative gain control can maintain proper operation on pitch and roll angles with maximum overshoot value no more than 7 % which occurs on the third quadrant, between 4 to 6 seconds of settling time. This circumstance is producing a shift in position on the trajectory as large as 0.12 m.

The system using the PD control on roll and pitch axis has achieved a satisfying response in circular movement, and then PD control system was replaced with the Fuzzy-PID control system to study the difference. A Fuzzy-PID controller structure with feed forward loop was designed to obtain stabilization at steady state. The Fuzzy rules for pitch and roll angles are alike and the values of Fuzzy-PID sets used for the controllers are also identical. These values can be determined by an adaptive trial and error method. The values of Fuzzy sets for K_{pf} and K_{df} are 0.288 and 0.33 respectively. Fig. 10. shows the modeling result of horizontal balancing control system on the pitch and roll axis. In the settling time between 0 to 8 seconds, the horizontal balancing control system on the roll axis can maintain the position/angle as well as on the yaw axis. This result shows that the Fuzzy-PID control system is better than the PID control system

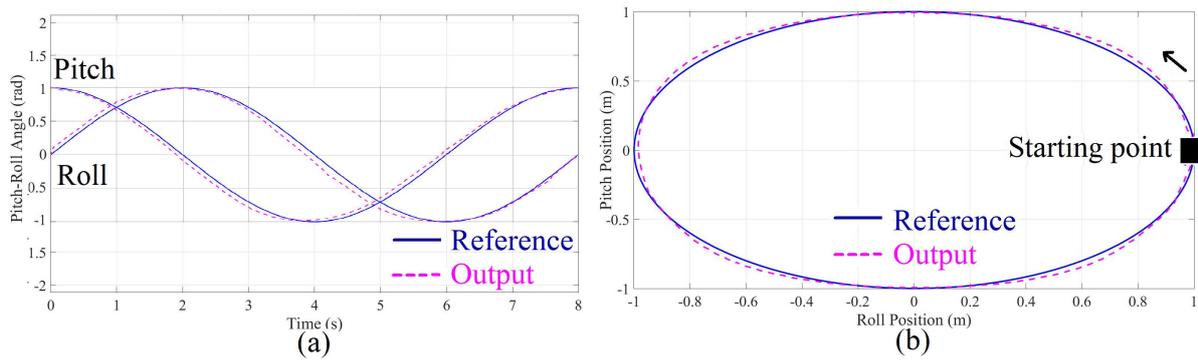


Figure 10: Quadcopter attitude with Fuzzy-PID control system. (a) Pitch and roll angle (b) Trajectory position

with a position shift less than 0.05 m and good flight stability on circular movement. This means that the fuzzy logic system can be used to change the output position to the different moments of control coefficients which can be done in the presence of fuzzy control along with the PID control [6]. It can be concluded that both systems are usable in this system and the difference between PID and Fuzzy-PID controller only causing a small error in a no-disturbance condition, allowing the quadcopter to move on the desired circular trajectory.

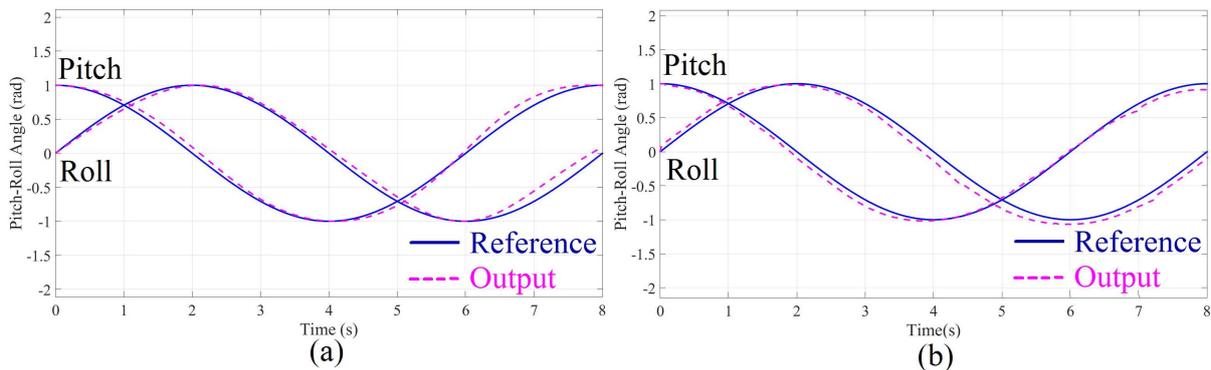


Figure 11: Attitude control system with disturbance parameter. (a) PID (b) Fuzzy-PID

To find out more about the difference between the two control systems on the quadcopter, a disturbance parameter of wind with 2 m/s velocity was added with various combination of wind direction. Discrete gust wind models were used with gain values of 3.5, 3.5, 3.5 m/s and gust wind lengths of 120, 120, 80 m on each axis. Gust wind length is the rise time needed to build up gust wind which expressed with length in each axis. The result of changes in pitch and roll angles with wind directions on each axis can be seen in Fig. 11. Disturbances affect the angle alteration in pitch and roll movement, thus resulting in greater tilt angle of the quadcopter. For a deeper analysis, Fig. 12. shows the detail of changes in quadcopter flight position with various wind directions. During the circular movement with wind effect in x- and y- directions, each control system has different behavior. In case of the PID control system, the shift in pitch and roll angles of movement happens almost at every settling time, but significant error in position occurs after 5 seconds of settling time. Overall the fourth quadrant witnesses the maximum angle alteration of 0.14 radians from the reference angle and a maximum position shift of 0.1 m. In the case of the Fuzzy-PID control, it can be clearly seen that the angle alteration happens in each position. This results in a wider flight position than the determined reference. The larger the settling time, the

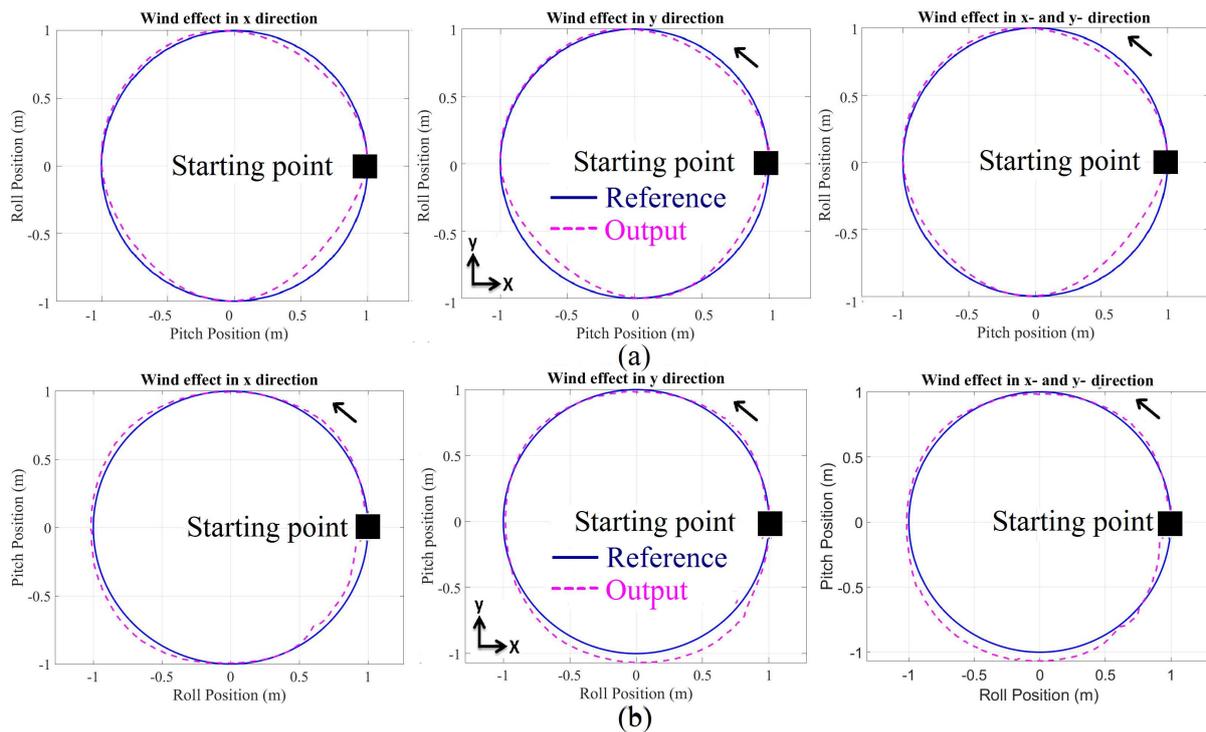


Figure 12: Results obtained with attitude control system with disturbance on each axis. (a) PID (b) Fuzzy-PID

greater the error of position shift will be. This principle applies for every combination of wind directions. Under wind effects in x- and y- direction, the position shift occurs accordingly and this result shows that the quadcopter can still maintain the position of the movement on circular trajectory. Details of position shift are as follows: maximum angle alteration of 0.17 radians and a position shift of 0.08 m, occurred in the third and fourth quadrant. Due to the symmetrical shape of quadcopter between X- and Y-axis, the angle alteration between roll and pitch must be linear so as to keep the position on a circular trajectory. When the control system is exposed with disturbance parameter, which is the common occurrence in real environments, the Fuzzy-PID is more reliable because it has the response system design through if-then logic parameters combined with adaptive PD control for error adjustment.

5 Conclusions

Quadcopter attitude movement modeling on a circular trajectory can be performed based on the Newton Euler kinematic model. A control system for continuous movement of the quadcopter is necessary to guarantee the quadcopter stability when maneuvering, where there are continuous change in pitch and roll angles. Without a control system, large amounts of overshoot were distinctly observable in the attitude movement in horizontal plane. To reduce the error, PID or Fuzzy-PID control can be added into the quadcopter movement modeling algorithm. In the case of the PID control system, only components proportional and derivative were needed to maintain quadcopter flying position stability. PD controller has maximum position shift of 0.12 m on the circular trajectory, whereas in the case of Fuzzy-PID control system the shift is only 0.05 m. This means that the quadcopter still moves on circular trajectory without significant error. Wind disturbances affect both control systems and significant error occurs in the third and

fourth quadrant. Under wind disturbance of 2 m/s, the Fuzzy-PID control system can maintain the position on the circular trajectory better than the PID control system, with just 0.03 m of maximum position shift when compared to no disturbance. The Fuzzy-PID control system works as an adaptive PID controller, the tracking response is automatically adjusted based on the Fuzzy inference rules which allow comparing two inputs received by the Fuzzy control system so it can adapt to possible errors and avoid overshoot. Conclusively, the Fuzzy-PID controller is superior to the PID controller in which manual gain adjustment has a crucial role in quadcopter flight stability.

Acknowledgment

This work has been funded under the LEADERS - Erasmus Mundus Grant (agreement number 2014-0855/001-001) by European Commission, through the Education, Audiovisual and Culture Executive Agency, in the Action Plan 2 for the years 2014-2018, and supported by PNCDI III Programme P2 - Transfer of knowledge to the economic operator (Bridge Grant PN-III-P2 2.1 BG-2016-0296) funded by UEFISCDI, Romania.

Bibliography

- [1] Ahmed S.F. et al (2015); Attitude Stabilization of Quad-rotor (UAV) System Using Fuzzy PID Controller, *Second International Conference on Computing Technology and Information Management (ICCTIM)*, 99-104, 2015.
- [2] Argentim L. M. et al (2013); PID, LQR and LQR-PID on a Quadcopter Platform, *International conference on Informatics, Electronics and Vision (ICIEV)*, 1-6, 2013.
- [3] Benic Z. et al (2016); Mathematical Modeling of Unmanned Aerial Vehicles with Four Rotors, *Interdisciplinary Description of Complex System Journal*, 14, 88-100, 2014.
- [4] Bolandi H. et al (2013); Attitude Control of a Quadrotor with Optimized PID Controller, *Intelligent Control and Automation Journal*, 4(3), 335-342, 2013.
- [5] Cai G. et al (2014); A Survey of Small-Scale Unmanned Aerial Vehicles: Recent Advances and Future Development Trends, *Unmanned System Journal*, 2, 1-25, 2014.
- [6] Du Z.B., Lin T.C., Zhao T. B. (2015); Fuzzy Robust Tracking Control for Uncertain Non-linear Time-Delay System, *International Journal of Computers Communications & Control*, 10(6):812-824, 2015.
- [7] Fu C. et al (2013); UAS See-And-Avoid Strategy using a Fuzzy Logic Controller Optimized by Cross Entropy in Scaling Factors and Membership Functions, *International Conference on Unmanned Aircraft System (ICUAS)*, 532-541, 2013.
- [8] Gautam D., Ha C. (2013); Control of a Quadrotor Using a Smart Self-Tuning Fuzzy PID Controller, *International Journal of Advanced Robotic Systems*, 10:1-9, 2013.
- [9] He Z., Zhao L. (2014); A Simple Attitude Control of Quadrotor Helicopter based on Ziegler-Nicols Rules for Tuning PD Parameters, *The Scientific World Journal*, 1-13, 2014.
- [10] Kotarski D. et al (2016); Control Design for Unmanned Aerial Vehicle with Four Rotors, *Interdisciplinary Description of Complex System Journal*, 14(2), 236-245, 2016.

-
- [11] Lai J.G. et al. (2016); A New Adaptive Fuzzy PID Control Method and Its Application in FCBTM, *International Journal of Computers Communications & Control*, ISSN 1841-9836, 11(3):394-404, 2016.
- [12] Magnussen O. et al (2013); Experimental Validation of a Quaternion-based Attitude Estimation with Direct Input to A Quadcopter Control System, *International Conference on Unmanned Aircraft System (ICUAS)*, 480-485, 2013.
- [13] Mian A., Wang D. (2008); Modeling and Backstepping-based Nonlinear Control strategy for a 6 DOF Quadrotor Helicopter, *Chinese Journal of Aeronautics*, 21, 261-268, 2008.
- [14] Moghaddam H. F., Vasegh N. (2014); Robust PID Stabilization of Linear Neutral Time-Delay System, *International Journal of Computers Communications and Control*, 9(2):201-208, 2014.
- [15] Patel K., Barve J. (2014); Modeling, Simulation and Control Study for The Quadcopter UAV, *The 9th International Conference on Industrial and Information Systems (ICIIS)*, 1-6, 2014.
- [16] Sanchez E.N. et al. (2006); Combining Fuzzy, PID and Regulation Control for an autonomous mini helicopter, *Information Sciences Journal*, 177:1999-2022, 2006.
- [17] Seidabad E.A. et al. (2014); Designing Fuzzy PID Controller for Quadrotor, *International Journal of Advanced Research in Computer Science and Technology*, 2:221-227, 2014.
- [18] Stevanovic S. et al (2012); Robust Tracking Control Of A Quadrotor Helicopter Without Velocity Measurement, *Annals of DAAAM for 2012 and Proceedings of the 23rd International DAAAM Symposium*, 595-600, 2012.
- [19] Susnea I.; Vasiliu G. (2016); A Fuzzy Logic Software Tool and a New Scale for the Assessment of Creativity, *International Journal of Computers Communications & Control*, ISSN 1841-9836, 11(3):441-449, 2016.
- [20] Xu, H.; Vilanova, R. (2015); PI and Fuzzy Control for P-removal in Wastewater Treatment Plant, *International Journal of Computers Communications & Control*, ISSN 1841-9836, 10(6):936-951, 2016.
- [21] Yang, S. et al (2009); Design and Simulation of the Longitudinal Autopilot of UAV Based on Self-Adaptive Fuzzy PID Control, *International Conference on Computational Intelligence and Security*, 634-638, 2009.
- [22] Ziegler J. G., Nichols N. B. (1942); Optimum Settings for Automatic Controllers, *Transactions of the American Society of Mechanical Engineers*, 64:759-768, 1942.

Automatic Generation Control by Hybrid Invasive Weed Optimization and Pattern Search Tuned 2-DOF PID Controller

N. Manoharan, S.S. Dash, K.S. Rajesh, S. Panda

Neelamegam Manoharan

Department of Electrical Engineering,
Sathyabama University, Chennai, India
haran_mano_2000@yahoo.com

Subhransu Sekhar Dash*, Kurup Sathy Rajesh

Department of Electrical Engineering,
SRM University, Chennai, India
munu_dash_2k@yahoo.com, rajeshks.srm@gmail.com
*Corresponding author: munu_dash_2k@yahoo.com

Sidhartha Panda

Department of Electrical Engineering,
VSSUT, Burla-768018, Odisha, India
panda_sidhartha@rediffmail.com

Abstract: A hybrid invasive weed optimization and pattern search (hIWO-PS) technique is proposed in this paper to design 2 degree of freedom proportional-integral-derivative (2-DOF-PID) controllers for automatic generation control (AGC) of interconnected power systems. Firstly, the proposed approach is tested in an interconnected two-area thermal power system and the advantage of the proposed approach has been established by comparing the results with recently published methods like conventional Ziegler Nichols (ZN), differential evolution (DE), bacteria foraging optimization algorithm (BFOA), genetic algorithm (GA), particle swarm optimization (PSO), hybrid BFOA-PSO, hybrid PSO-PS and non-dominated shorting GA-II (NSGA-II) based controllers for the identical interconnected power system. Further, sensitivity investigation is executed to demonstrate the robustness of the proposed approach by changing the parameters of the system, operating loading conditions, locations as well as size of the disturbance. Additionally, the methodology is applied to a three area hydro thermal interconnected system with appropriate generation rate constraints (GRC). The superiority of the presented methodology is demonstrated by presenting comparative results of adaptive neuro fuzzy inference system (ANFIS), hybrid hBFOA-PSO as well as hybrid hPSO-PS based controllers for the identical system.

Keywords: Automatic generation control, interconnected power system, governor, dead - band non linearity, 2 degree of freedom PID controller, invasive weed optimization, pattern search.

1 Introduction

Automatic generation control (AGC) loop in a power system calculates the required change in the generation based on the system frequency and tie-line flow deviations, and adjusts the set position of the generators in each area to maintain the time average of frequency and tie-line power changes at a low value [11], [5]. The researchers in the world over are developing a number of control strategies for AGC to keep the tie-line flow system and frequency at their desired values both in normal and disturbed conditions [29]. In recent times, soft computing based methods have been applied to tune the parameters of the controller [25] [1] ; [28]; [19]. [14] applied DE to

tune the controller parameters for a multi-source power system and relative performances of different classical controllers were compared. [2] employed teaching learning based optimization (TLBO) technique for the design of I/PID controllers for a multi-units multi-sources power system and superiority of TLBO algorithm was demonstrated over DE and optimal output feedback controller. [6] proposed an intelligent controller based on emotional learning for LFC of an interconnected power system with generation rate constraint (GRC) and demonstrated the advantage their approach over PI, fuzzy logic, hybrid Neuro Fuzzy (HNF) controller. A Firefly Algorithm (FA) with on line wavelet filter was employed by [15] for AGC of inter connected unequal three area power system. [8] employed artificial bee colony (ABC) algorithm for AGC and superiority of the approach was demonstrated over PSO algorithm. [22] used gravitational search algorithm (GSA) to optimize PI/PIDF controller parameters with conventional integral based objective functions for AGC system and compared the results with DE, BFOA and GA to show the superiority. A Teaching Learning Based Optimization algorithm has been applied by [24] for Automatic generation control of multi-area power systems with diverse energy sources.

It is seen in literatures that the performance of power system depends on the tuning technique, controller structure and choice of cost function. In this regard, it is observed that, a two degree of freedom controllers provide better performance than a single degree of freedom controller [21]. Having known all this, in the present work, an ideal 2 degree freedom of PID (2-DOF-PID) controller for AGC of multi-area power systems. Generally, all population centered heuristic optimization techniques offer acceptable results but there is no guarantee that a particular technique will give a better performance than other techniques in all optimizing problems [30]. Hence, suggesting and realizing novel heuristic techniques are always desired. Each heuristic technique has its own advantages and disadvantages. Hybrid algorithms taking the advantage of two or more algorithms have been recently proposed in literature. [16] proposed a hybrid BFOA-PSO algorithm to tune the controller parameters for AGC systems. A hybrid PSO-PS algorithm is proposed by [23] to tune the fuzzy PI parameters. A modified DE optimized fuzzy PID controller for load frequency control with thyristor controlled series compensator has been proposed by [20].

In recent times, invasive weed optimization (IWO), a novel biologically motivated optimization technique was proposed by [12]. IWO is a robust, stochastic and derivative free optimization algorithm for the solution of complex real world problems. It is based on the invasive habits of growth of weeds in nature and having excellent exploration and exploitation ability in the search area. IWO has been successfully employed to a number of engineering problems such as recommender system design [18], antenna system design [10], state estimation of nonlinear systems [13], unit commitment problem solution [27] and economic load dispatch of power systems [3]. To get excellent performance using any optimization technique, a balance of exploitation as well as exploration throughout the search procedure is to be maintained. IWO being a global search technique, searches the wide search area and may not give best solution if employed alone. Alternatively, local search methods such as Pattern Search (PS) exploits the local but cannot perform extensive search [4]. Owing to their individual strengths, there is a scope for hybridization of these algorithms [31]. In view of the above, a hybrid IWO-PS technique is suggested in this work for tuning the parameters of 2-DOF-PID controller for AGC of interconnected systems.

In the present study, a two area thermal as shown in Figure 1 is considered as the system under study. The same system is extensively used in literature for proposing new AGC approaches [1]; [19]; [17]; [23].

In Figure 1, B_1 & B_2 represent the frequency bias parameters; ACE_1 & ACE_2 stands for area control errors; u_1 & u_2 are the control outputs; R_1 & R_2 represent the regulation parameters in pu Hz; T_{G1} & T_{G2} are the time constants of governor in sec; ΔP_{G1} & ΔP_{G2} are the incremental valve positions (pu); T_{T1} & T_{T2} are the time constant of turbine in sec; ΔP_{T1} & ΔP_{T2} are the incremental

The transfer function generator and load is given by:

$$G_p(s) = \frac{K_p}{1 + sT_p} \quad (6)$$

Where $K_p = \frac{1}{D}$ and $T_p = \frac{2H}{fD}$

The output $\Delta f(s)$ of generator load system has two inputs $\Delta P_T(s)$ is given by:

$$\Delta f(s) = G_p(s)[\Delta P_T(s) - \Delta PG(s)] \quad (7)$$

2 Ideal two degree of freedom PID controller

Depending on the number of closed-loop transfer functions which can be controlled individually, the degree of freedom of a control system is classified. In a control system design problem, numerous performance criteria are to be satisfied thus a 2-degree-of-freedom (2-DOF) controller offers some advantages over the single degree of freedom control system [26]. The 2-DOF controller calculates a weighted difference signal for each of the control actions as per the set point weights and gives an output signal which is the sum of the control actions on the respective difference signals [21]. A derivative filter is used for improved system performance in presence of noise or random error in the measured process variable. It also limits the huge controller output changes which derivative action causes due to presence of measurement noise and helps to lessen the controller output variations which may result in wear in the control parts. The structure of proposed ideal 2-DOF-PID controller is given in Figure 2 where $R(s)$ represents the reference signal, $Y(s)$ is the feedback signal and $U(s)$ represents the output signal, K_p , K_i & K_d are the controller gains, PW & DW are the set point weights, and N is the filter coefficient of derivative term. A 2-DOF-PID control system is given in Figure 3 where $C(s)$ a one degree

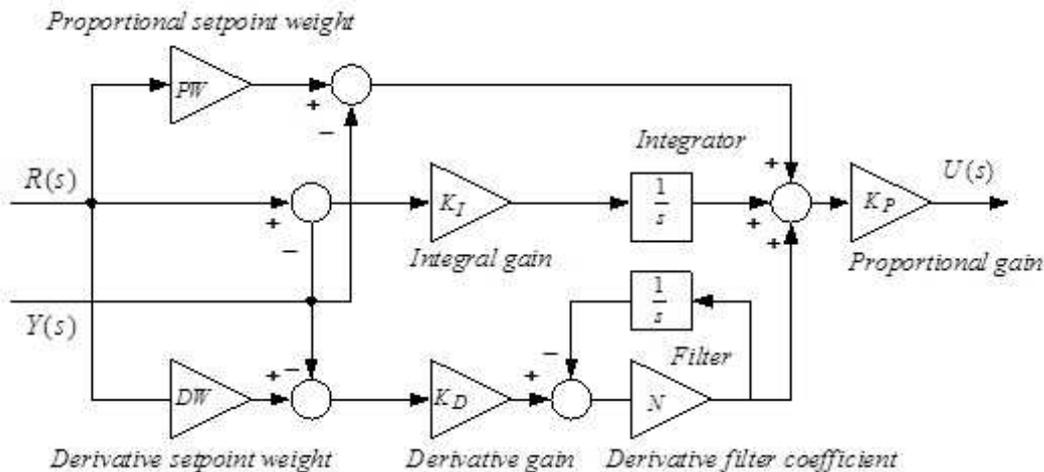


Figure 2: Two degree of freedom (TDOF) PID control structure

of freedom controller, $D(s)$ is the load disturbance and $F(s)$ is the filter acting on the reference signal. In an ideal 2-DOF-PID controller, and are specified by:

$$F(s) = \frac{(PW + DWK_D)s^2 + (PWN + K_I)s + (K_I N)}{(1 + K_D N)s^2 + (N + K_I)s + (K_I N)} \quad (8)$$

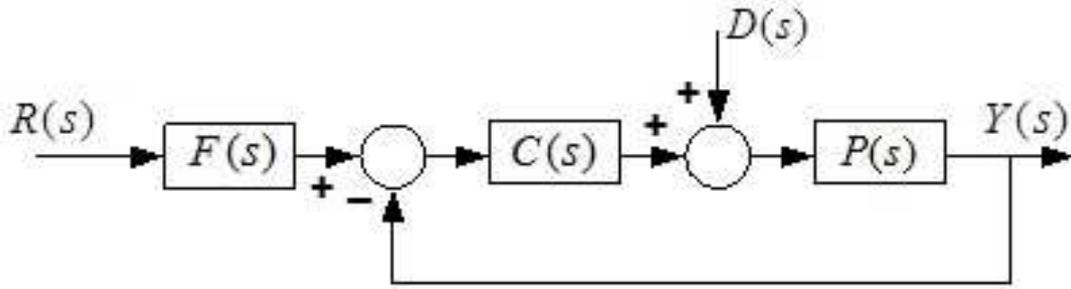


Figure 3: TDOF control system

$$C(s) = K_p \frac{(1 + K_D N)s^2 + (N + K_I)s + (K_I N)}{s(s + N)} \quad (9)$$

An integral of time multiplied absolute error (ITAE) based objective function given in Eq. (10) is chosen in this paper to design the proposed controllers. ITAE is chosen over other integral based objective functions because it gives less overshoots and settling times compared to other criterion such as integral of squared error (ISE), and integral of absolute error (IAE). Other integral squared criteria such as integral of time multiplied squared error (ITSE) and integral of squared time multiplied error (ISTE) based design produces huge controller output when there is a sudden variation in reference which is not desirable.

$$J = ITAE = \int_0^{t_{sim}} \omega_1 (|\Delta f_1| + |\Delta f_2| + |\Delta P_{tie}|) t dt \quad (10)$$

Where, ω_1 and ω_2 are the system frequency changes; ΔP_{tie} is the change in tie-line power and t_{sim} is the simulation time. The optimization problem can be expressed as:

$$\text{Minimize } J \quad (11)$$

Subject to

$$\begin{aligned} K_{pmin} \leq K_p \leq K_{pmax}, K_{Imin} \leq K_I \leq K_{Imax}, K_{Dmin} \leq K_D \leq K_{Dmax} \\ PW_{min} \leq PW \leq PW_{max}, DW_{min} \leq DW \leq DW_{max}, N_{min} \leq N \leq N_{max} \end{aligned} \quad (12)$$

Where K_{pmin}, K_{pmax} ; K_{Imin}, K_{Imax} and K_{Dmin}, K_{Dmax} are the lower and upper bounds of the control parameters. PW_{min}, DW_{min} and PW_{max}, DW_{max} are the lower and upper bounds of set point weights, N_{min}, N_{max} and are lower and upper bounds of derivative filter coefficient. The bounds of controller parameters, set point weights and filter coefficient are taken as -2 & 2, 0 & 5 and 10 & 300 respectively.

3 Overview of invasive weed optimization technique

Invasive weed optimization (IWO) is a novel population based stochastic, derivative free optimization technique inspired from the biological growth of weed plants. It was first developed and designed by [12]. The IWO algorithm is based on the colonizing actions of weed plants [18]. Some of the interesting characteristics of weed plants that are invasive, fast reproduction and distribution, robustness and self adaptation to the changes in climate conditions.

The significant characteristic of the IWO algorithm is that it lets all the plants to contribute in the reproduction procedure. Fitter plants yield more seeds than less fit plants and this results

in the algorithm to converge. Additionally, it is still probable that some of the less fit plants carry beneficial information in iteration process as compared to the some fitter plants. Thus IWO algorithm provides an opportunity to the lesser fit plants to participate in reproduction process. If the seeds formed by lesser fit plants have better fitness in the colony, they can survive [12]. Another significant characteristics of IWO algorithm is that reproduction is done without mating and every weed can yield new seeds, individually. This reproduction without mating characteristics augments a new quality to the technique as each agent may have not the same number of variables during the optimization process and the number of variables can be selected as one of the optimization parameters in IWO. Also, IWO algorithm has more chance to avoid local minima points compared to GA and PSO due to its continuous and normally distributed dispersal structure over search space which has a decreasing variance parameter centered on each parent plant [18]. The steps of the proposed approach are mentioned below:

Step 1: Based on the number of chosen variables (d) of the assumed problem, the seeds are initialized. The initial seed are distributed uniformly over the entire solution space.

Step 2: Create each seed set, after generating all the selected variables of the given problem randomly within their effective lower and upper limits. Thus, in the search space, each seed contains random values for all variable. Each seed set represents a potential solution of the given problem. Generate several seed set to create a Seed matrix (S) of size ($Pop_{max} \times d$). The total number of plants in the population is selected as (Pop_{max}) after satisfying their limits.

Step 3: The fitness value of all individuals of the current seed set (S) (each row (plant) of S) is calculated according to the cost function considered in the given problem. These individuals evolve into weed plants which are capable of creating new units.

Step 4: As per the fitness value of each plant with respect to others, each plant is ranked. Then, every weed yields new seeds depending on its rank in the set of seed. All plants are participating in reproduction process which adds a new attribute to the optimization providing chances to contribute useful information (good result) by less fit plants during iterative process.

Step 5: The number of seeds to be produced by all weed changes linearly from N_{min} to N_{max} which can be calculated by:

$$Seed\ number = \frac{F_i - F_{worst}}{F_{best} - F_{worst}}(N_{max} - N_{min}) + N_{min} \quad (13)$$

Where, F_i is the fitness associated with i^{th} weed, F_{worst} and F_{best} denotes the worst and best fitness in weed population. The created seeds are normally distributed over the field with zero average and variable standard deviation of σ_{iter} defined by

$$\sigma_{iter} = \left[\frac{iter_{max} - iter}{iter_{max}} \right]^n (\sigma_0 - \sigma_f) + \sigma_f \quad (14)$$

Where, $iter_{max}$ and $iter_{min}$ are the maximum number of iteration and current iteration, respectively. σ_0 and σ_f , are the predefined initial and final standard deviations and n represent the modulation index.

Step 6: The new seeds breed to the flowering plants when all seeds found their positions over the search area. Next, they are ranked together with their parents in the seed set matrix. Plants with lower ranking in the colony are removed and the maximum number of plants in the colony (Pop_{max}) is maintained.

Step 7: Survived plants can yield new seeds as per their ranking. The fittest individual (plant) is selected from the seed-parent combination of current seed set. If the stopping criterion is satisfied, the iterative process is terminated and the results (gain schedule) are displayed, otherwise go to Step 3 for continuation.

4 Overview of pattern search algorithm

Pattern search (PS) algorithm is an effective but simple technique applicable to the complex problems which cannot be solved by conventional optimization techniques. It has a flexible operator to fine tune the local explore capability [23]. The PS method consists of a series of polls x_k $k \in N$. A number of trial steps with $i = 1, 2, \dots, p$ are added to the polls x_k to get trial points $x_k^i = x_k + s_k^i$ at each poll. At these trial points the objective function value is calculated through a sequence of exploratory steps and compared with its previous value $J(x_k)$. The trial step s_k^* corresponding to least value of $J(x_k + s_k^i) - J(x_k) < 0$ is then selected to produce the subsequent estimation of the patterns polls $x_{k+1} = x_k + s_k^*$. The trial steps s_k^i are produced by a step length parameter $\Delta_k \in R_{in}^+$. The Δ_k value is updated in subsequent polls as per x_{k+1} value. The improvement of Δ_k , help the algorithm to converge. These elements are explained in more details in reference [4].

5 Results

5.1 Application of hIWO-PS algorithm

As the two areas are assumed identical, similar controllers are assumed in each area. The objective function is calculated by applying a 10percent step load disturbance in area-1. A series of runs are executed to properly select the algorithm parameters. Number of search agents and iterations are taken as 20 and 50 respectively. The optimization process was repeated 10 times and the best solution obtained in 10 runs is selected as final controller parameters. In the next step, the proposed hIWO-PS algorithm is applied to optimize the controller parameters. In hIWO-PS algorithm, initially optimal IWO is executed for 40 iterations and then PS is employed for 10 iterations. The final solution corresponding to the minimum objective function value provided by optimal IWO is used as the beginning points of PS algorithm. For the implementation of PS algorithm, the following parameters are used: mesh size=1, mesh expansion factor=2, mesh contraction factor= 0.5, max. no. of function estimations=10, max. no. of iterations = 10. The optimized 2-DOF-PID parameters are provided in Table 1. For comparison, the optimized PI controller parameters are also specified in Table 1.

Table 1: Tuned controller parameters

Controller Technique	Controller parameters
IWO: PI	KP=-0.3005, KI=0.4551
hIWO-PS	KP =-0.3106, KI =0.4524
IWO: 2-DOF-PID	KP =1.764, KI =1.764, KD =0.4785, PW =7.201, DW =4.3767, N =298.2108
hIWO-PS: 2-DOF-PID	KP =1.889, KI =1.9398, KD =0.4941, PW =7.2088, DW =2.7742, N =318.1317

5.2 Result analysis

A 10 percent step load disturbance in area-1 is considered at t=0.0 sec. The ITAE values with IWO and hIWO-PS optimized PI/2-DOF-PID controllers are shown in Table 2.

To demonstrate the efficiency of the proposed hIWO-PS technique, results are compared with genetic algorithm: GA, bacteria foraging optimization algorithm: BFOA [1], differential evolution: DE [19], particle swarm optimization: PSO, hybrid BFOA-PSO cite16, non dominated shorting GA-II: NSGA-II optimized PI controllers, NSGA-II optimized PID Controller with derivative filter cite17, pattern search: PS, PSO, and hybrid PSO-PS cite23 optimized fuzzy PI controllers for the same interconnected power system. It is obvious from Table 2 that

Table 2: ITAE values with different controllers and optimization techniques

Controller	Tuning method/ Optimization technique	ITAE Value
PI	Hybrid Invasive Weed Optimization (IWO)- Pattern Search	1.1761
PI	Invasive Weed Optimization (IWO)	1.1763
PI	Ziegler Nicholas (Ali, & Abd-Elazim, 2011)	3.7568
PI	GA (Ali, & Abd-Elazim, 2011)	2.7475
PI	BFOA (Ali, & Abd-Elazim, 2011)	1.7975
PI	DE (Rout, Sahu, & Panda, 2013)	1.2551
PI	PSO (Panda, Mohanty, & Hota, 2013)	1.2142
PI	Hybrid BFOA-PSO (Panda, Mohanty, & Hota, 2013)	1.1865
PI	NSGA-II (Panda, & Yegireddy, 2013)	1.1785
2-DOF-PID	Hybrid IWO-PS	0.1037
2-DOF-PID	IWO	0.1311
PIDF	NSGA-II (Panda, & Yegireddy, 2013)	0.387
Fuzzy PI	PS (Sahu, Panda, & Sekher, 2015)	0.6334
Fuzzy PI	PSO (Sahu, Panda, & Sekher, 2015)	0.4470
Fuzzy PI	Hybrid PSO-PS (Sahu, Panda, & Sekher, 2015)	0.1438

with the same PI controller, tuned using the same ITAE objective function, lowest ITAE value is obtained with proposed hIWO-PS technique (ITAE=1.1761) compared to IWO (ITAE=1.1763), Z-N tuning (ITAE=3.7568), GA (ITAE=2.7475), BFOA (ITAE=1.7975), DE (ITAE=1.2551), PSO (ITAE=1.2142), Hybrid PSO-PS technique (ITAE=1.1865) and NSGA-II (ITAE=1.1785). In the above evaluation, identical interconnected power system with two similar PI controllers is assumed and the controller parameters are tuned using an ITAE objective function. Therefore, it can be concluded that proposed hIWO-PS technique provides better performance than IWO, GA, BFOA, DE, PSO, Hybrid PSO-PS NSGA-II techniques as lowest ITAE value is achieved using hIWO-PS technique. From Table 2, it is furthermore apparent that, value of ITAE is considerably reduced (ITAE=0.1311) with IWO tuned 2-DOF-PID controller. The ITAE value is reduced (ITAE=0.1037) with hIWO-PS tuned 2-DOF-PID controller. It is also evident from Table 2 that hIWO-PS tuned 2-DOF-PID controller gives minimum ITAE value compared to IWO optimized 2-DOF-PID controller (ITAE=0.1349), Pattern Search (PS) tuned fuzzy PI controller (ITAE=0.6334), PSO tuned fuzzy PI controller (ITAE=0.447), Hybrid PSO-PS tuned fuzzy PI controller (ITAE=0.1438) and NSGA-II tuned PIDF controller (ITAE=0.387).

In the next step, a Step Load Perturbation (SLP) of 10 percent is applied at $t = 0$ sec in area-1 and time domain simulation results are plotted. The system dynamic responses are shown in Figures 4-6. The results of some recently published approaches like DE [19], BFOA [1], hBFOA-PSO [16] tuned PI controller and PSO fuzzy PI, PS fuzzy PI & hPSO-PS fuzzy PI [23] controllers for the identical system are also provided in Figures 4.

It can be seen from Figure 4 that, considerable improvement is achieved with hIWO-PS tuned 2-DOF-PID controller compared other methods. For a better illustration of advantage of proposed approach over various approaches proposed in recent times, ITAE values as well as settling times in tie-line power and frequency deviations for the above disturbance is summarized in Table 3. It is evident from Table 3 that best system performance in terms of minimum ITAE values and settling times are obtained with proposed hIWO-PS tuned 2-DOF-PID controller as related to other recent methods.

The dynamic response of the system for a concurrent 10percent SLP in area 1 as well as 20 percent SLP in area 2 at $t = 0$ s is assumed and the system dynamic responses are shown in

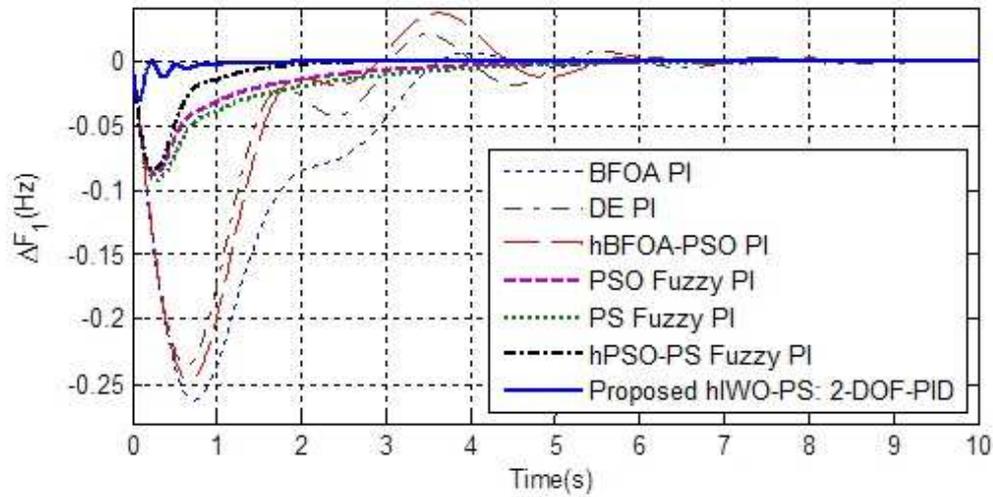


Figure 4: Change in frequency of area-1 for 10% step load increase in area-1

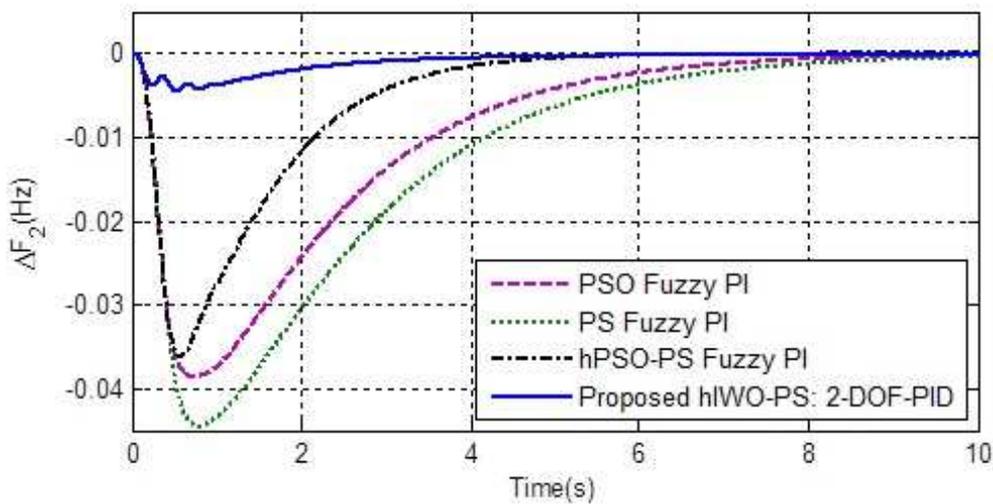


Figure 5: Change in frequency of area-2 for 10% step load increase in area-1

Figures 7-9. It is obvious from Figures 7-9 that the proposed controllers perform satisfactorily with change in the location and step size of the disturbance. Better dynamic responses are obtained with proposed hIWO-PS tuned 2-DOF-PID controller as compared to other recently reported methods in all the cases.

Robustness analysis is done to investigate the usefulness of the system when there are wide deviations in the loading conditions and parameters of the system. These parameters (loading condition and time constants) of speed governor, turbine, tie-line power are varied one after another from their initial values by +50 percent to -50 percent in steps of 25percent. The performance index under changed conditions are provided in Table 4. The above sensitivity analysis is performed by assuming a SLP of 10 percent in area-1 at $t=0$ sec. To demonstrate the advantage of the proposed approach, results are compared with hPSO-PS tuned fuzzy PI controller [23] under the same varied conditions. In this comparison, hPSO-PS tuned fuzzy PI controller values are selected for comparison as least ITAE value is attained with, hPSO-PS

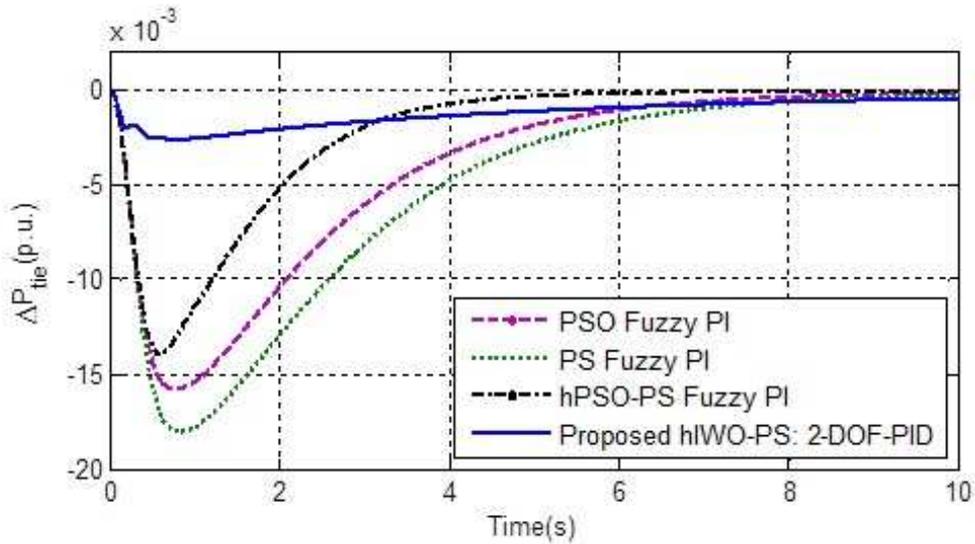


Figure 6: Change in tie line power for 10% step load increase in area-1

Table 3: Performance comparison with recent AGC approaches

Performance	ITAE	Settling Time		
	Value	ΔF_1	ΔF_2	ΔP_{tie}
Conventional ZN: PI (Ali, & Abd-Elazim, 2011)	3.7568	45	45	28
GA: PI (Ali, & Abd-Elazim, 2011)	2.7475	10.59	11.39	9.37
BFOA: PI (Ali, & Abd-Elazim, 2011)	1.7975	5.52	7.09	6.35
DE: PI (Rout, Sahu, & Panda, 2013)	0.9911	8.96	8.16	5.75
PSO: PI (Panda, Mohanty, & Hota, 2013)	1.2142	7.37	7.82	5.0
hBFOA-PSO: PI (Panda, Mohanty, & Hota, 2013)	1.1865	7.39	7.65	5.73
NSGA-II: PI (Panda, & Yegireddy, 2013)	1.1785	6.49	7.54	5.79
NSGA-II: PIDF (Panda, & Yegireddy, 2013)	0.387	3.03	4.86	4.34
PS: Fuzzy PI (Sahu, Panda, & Sekher, 2015)	0.6334	6.05	7.10	5.56
PSO: Fuzzy PI (Sahu, Panda, & Sekher, 2015)	0.4470	5.13	6.22	4.83

tuned fuzzy PI controller related to other methods. It is obvious from the simulation results that the system performances remain more or less the same with varied loading condition and system parameters. Thus it can be concluded that, the hIWO-PS tuned 2-DOF-PID controller offers a robust and efficient control strategy. Also, the controller parameters which are tuned at the nominal system conditions, need not be retuned when there are wide variations in the system parameters.

5.3 Extension to three unequal area non-linear hydro thermal power system

To establish the capability of the proposed method to deal nonlinearity and several tie-lines, the method is applied to a unequal three area non-linear thermal hydro power system ([16]; [10]; [23]) as shown in Figure 10. In this case different controllers are assumed in each area as the areas are unequal. A GRC (Generation Rate Constraints) of 3% min is assumed for thermal units. A GRC of 270% min for rising and 360% min for lowering generation are considered for hydro unit. The related system parameters are specified.

Three-area hydro thermal power system with generation rate constraints:

$$B_1 = B_2 = B_3 = 0.425 pu MW/Hz; R_1 = R_2 = R_3 = 2.4 Hz/pu MW; T_{G1} = T_{G2} =$$

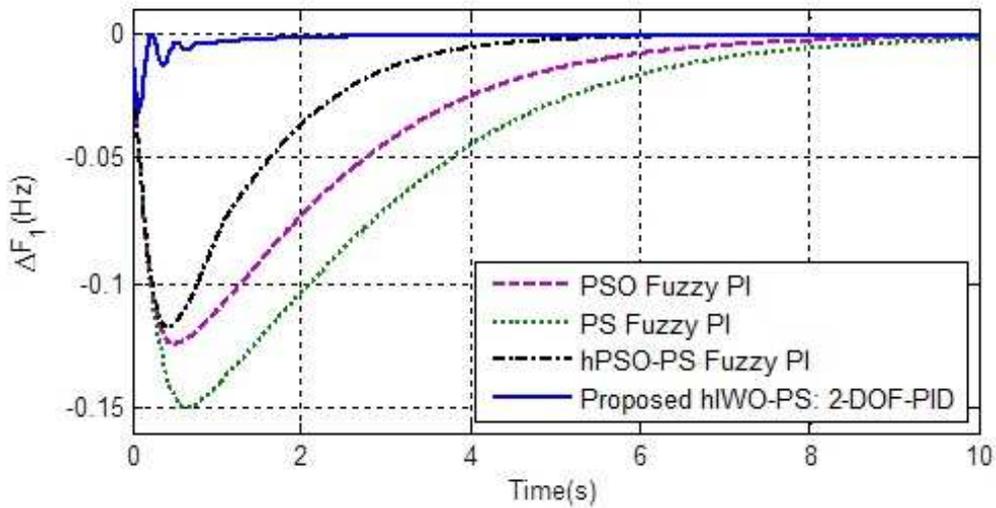


Figure 7: Change in frequency of area-1 for 10% step load increase in area-1 and 20% step load increase in area-2

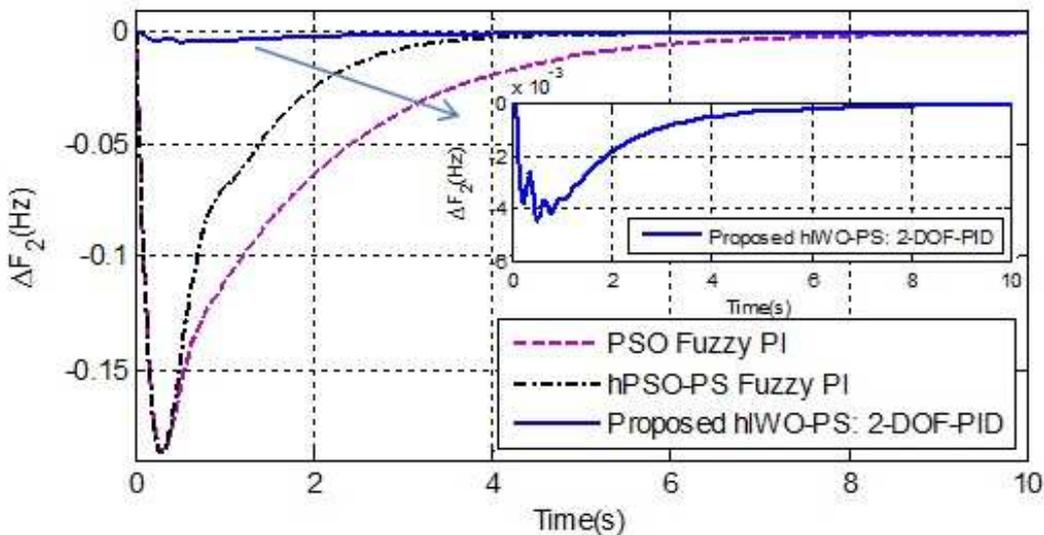


Figure 8: Change in frequency of area-2 for 10% step load increase in area-1 and 20% step load increase in area-2

$0.08s; T_{r1} = T_{r2} = 10.0s, T_{T1} = T_{T2} = 0.3s; T_W = 1.0s; T_R = 5s; K_{PS1} = K_{PS2} = K_{PS3} = 120Hz/p.u.MW; T_{PS1} = T_{PS2} = T_{PS3} = 20s; T_{12} = T_{23} = T_{31} = 0.086pu; a_{12} = a_{23} = a_{31} = -1$

The objective function in this case is defined by:

$$J = ITAE = \int_0^{t_{sim}} (|\Delta f_1| + |\Delta f_2| + |\Delta f_3 + |\Delta Pt_{12}| + |\Delta Pt_{13}| + |\Delta Pt_{23}|)t.dt \quad (15)$$

Where $\Delta f_1, \Delta f_2$ and Δf_3 are the frequency deviations and $\Delta Pt_{12}, \Delta Pt_{13}$ and ΔPt_{23} are the tie-line power deviations between individual areas. The final parameters obtained using proposed hIWO-PS algorithm are:

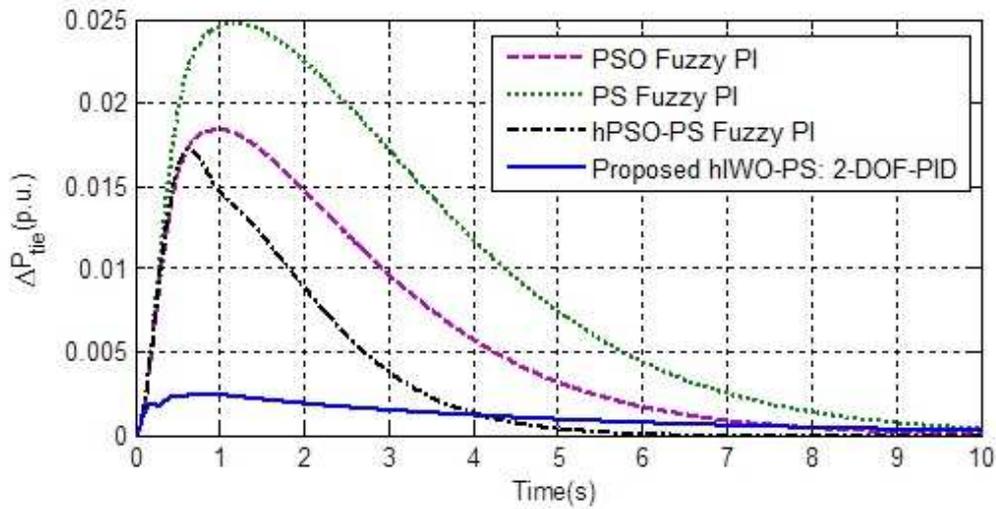


Figure 9: Change in tie line power for 10% step load increase in area-1 and 20% step load increase in area-2

Table 4: Robustness analysis for two area two unit system

Parameter Variation	Percent Change	Performance index with hPSO-PS optimized fuzzy PI (Sahu, Panda, & Sekher, 2015)				Performance index with Proposed hIWO-PS optimized 2-DOF-PID			
		Settling time Ts(Sec)			ITAE Value	Settling time Ts(Sec)			ITAE Value
		ΔF_1	ΔF_2	ΔP_{tie}		ΔF_1	ΔF_2	ΔP_{tie}	
Nominal	0	2.26	3.74	2.94	0.1438	1.39	1.97	2.01	0.1037
Loading condition	+50	2.26	3.75	2.94	0.1438	1.39	1.97	2.01	0.1037
	+25	2.26	3.75	2.94	0.1438	1.39	1.97	2.01	0.1037
	-25	2.26	3.74	2.94	0.1437	1.39	1.97	2.01	0.1037
	-50	2.26	3.74	2.94	0.1437	1.39	1.97	2.01	0.1037
T_g	+50	2.21	3.64	2.81	0.1321	1.29	1.9	1.99	0.1029
	+25	2.22	3.70	2.88	0.1386	1.44	1.96	2.0	0.1034
	-25	2.28	3.76	2.96	0.1460	1.39	1.97	2.01	0.1046
	-50	2.31	3.77	2.97	0.1469	1.41	1.99	2.03	0.1053
T_t	+50	1.98	3.61	2.80	0.1348	1.02	1.81	1.89	0.991
	+25	2.16	3.69	2.88	0.1409	1.21	1.86	1.92	0.1015
	-25	2.33	3.76	2.95	0.1422	1.48	2.03	2.06	0.1064
	-50	2.39	3.74	2.91	0.1354	1.58	2.12	2.13	0.1089
T_{12}	+50	2.73	3.51	2.70	0.1361	1.42	1.89	2.08	0.0919
	+25	2.56	3.60	2.80	0.1399	1.39	1.92	2.05	0.0967
	-25	1.92	3.98	3.14	0.1513	1.4	2.0	1.87	0.1163
	-50	3.02	4.48	3.53	0.1917	1.51	1.97	1.87	0.1403

$K_{P1} = 1.8539, K_{I1} = 1.7880, K_{D1} = 0.1682, PW_1 = 14.5646, DW_1 = 11.2175, N_1 = 495.2406$

$K_{P2} = 1.6238, K_{I2} = 1.8195, K_{D2} = 0.2642, PW_2 = 10.1788, DW_2 = 14.7692, N_2 = 125.3624$

$K_{P3} = 0.0711, K_{I3} = 1.9101, K_{D3} = 0.0240, PW_3 = 0.0122, DW_3 = 18.9549, N_3 = 139.4454$

A 1% SLP is applied at the same time in all the three areas at $t=0$ sec. The system dynamic responses are given in Figures 11-13.

The responses with ANFIS based controller [10], hBFOA-PSO tuned PI controller [16] and hPSO-PS based fuzzy PI controller [23] are also shown in Figures 11-13 for comparison. Figures 11-13 clearly establishes that system performance is appreciably enhanced with hIWO-PS tuned

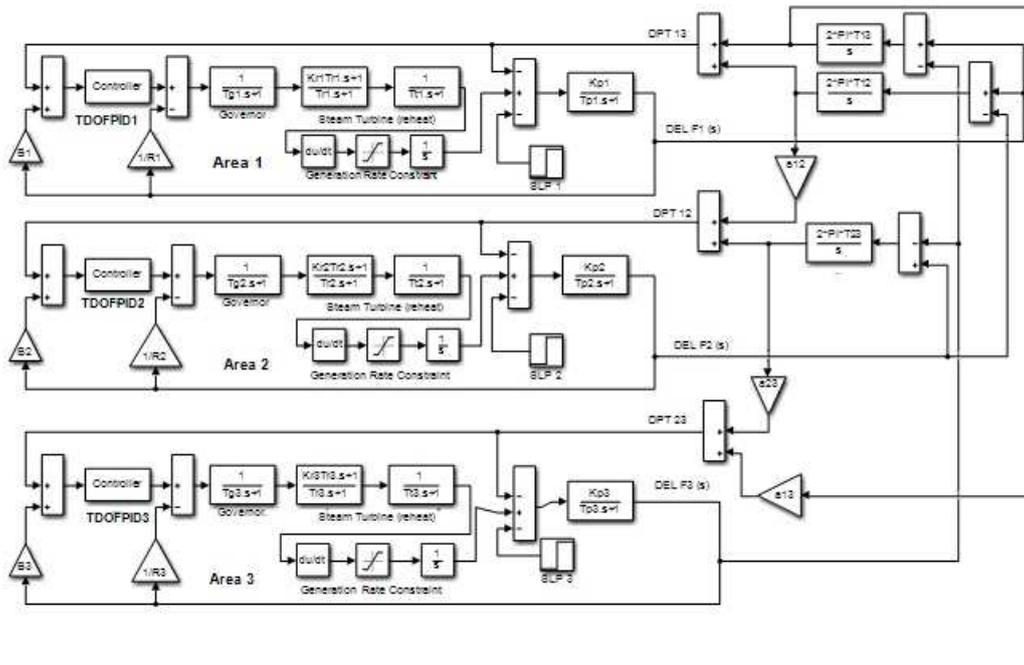


Figure 10: Transfer function model of model of three-area hydro-thermal system with generation rate constraint

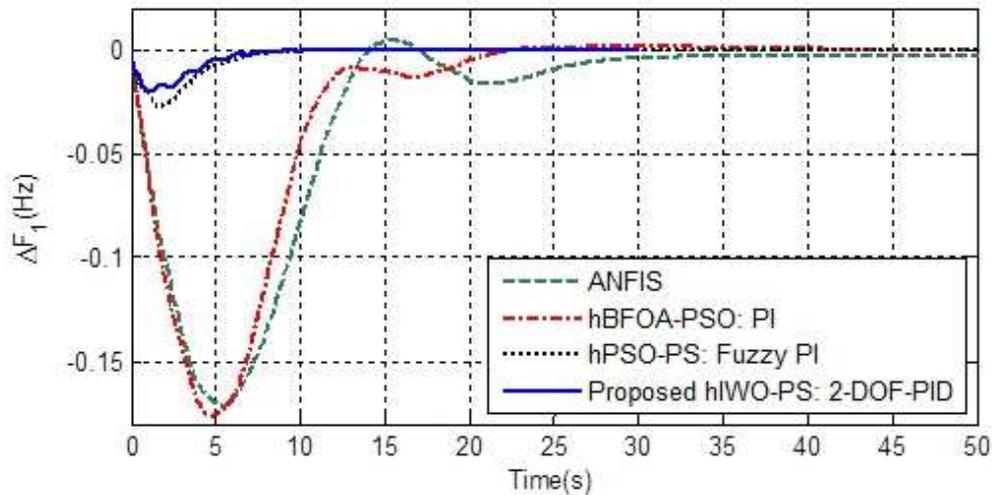


Figure 11: Change in frequency of area-1 for 1% step load increase in all areas

2-DOF-PID controller compared to ANFIS based controller, hBFOA-PSO tuned PI controller and hPSO-PS based fuzzy PI controller. Finally, to demonstrate the advantage of the hIWO-PS tuned 2-DOF-PID controller, ITAE values are compared with some recently reported optimization approaches [23]. In all the cases 1% SLP is applied in all the areas at the same time. The results are briefed in Table 5. It is obvious from Table 5 that least ITAE value is achieved with proposed hIWO-PS tuned 2-DOF-PID controller (ITAE=0.8378) compared to hPSO-PS technique (ITAE=1.3999), RCGA (ITAE=2.4873), GSA (ITAE=1.7805), DE (ITAE=1.6857) and FA (ITAE=1.5344) algorithms. It is evident from Table 9 that, hIWO-PS algorithm out

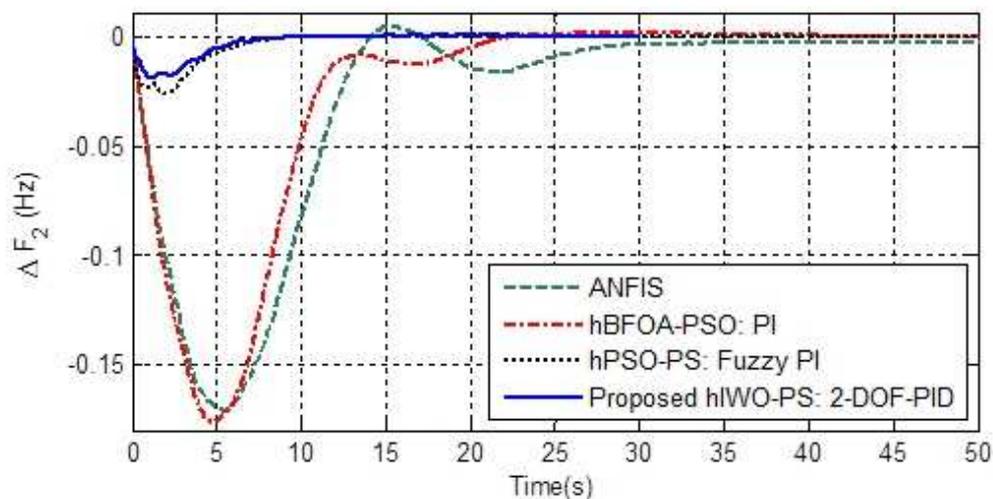


Figure 12: Change in frequency of area-2 for 1% step load increase in all areas

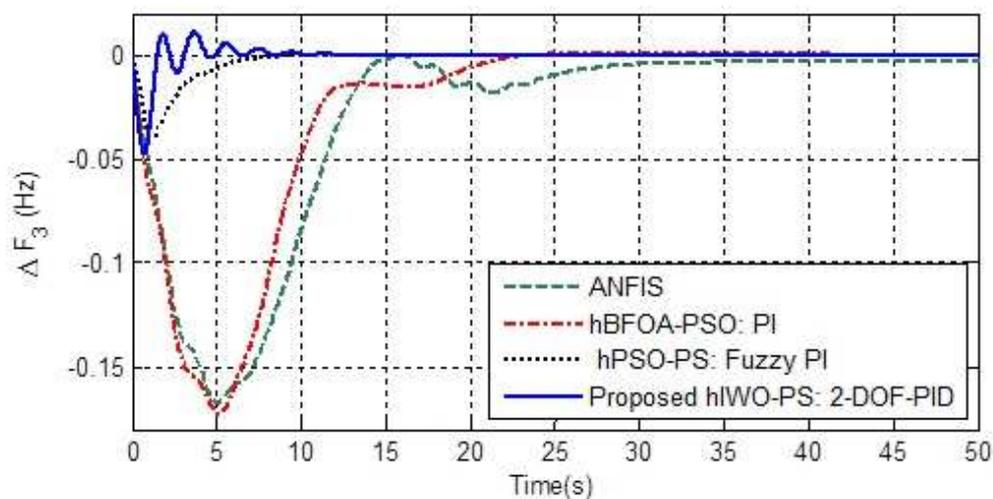


Figure 13: Change in frequency of area-3 for 1% step load increase in all areas

Table 5: Comparison of ITAE values with different approaches for three area system

Techniques	Controller ITAE	Value
GA (Sahu, Panda, & Sekher, 2015)	Fuzzy PI	2.4873
GSA (Sahu, Panda, & Sekher, 2015)	Fuzzy PI	1.7805
DE (Sahu, Panda, & Sekher, 2015)	Fuzzy PI	1.6857
FA (Sahu, Panda, & Sekher, 2015)	Fuzzy PI	1.5344
hPSO-PS (Sahu, Panda, & Sekher, 2015)	Fuzzy PI	1.3999
hIWO-PS	2-DOF-PID	0.8378

performs RCGA, GSA, DE, FA and hPSO-PS techniques.

6 Conclusion

A 2 degree freedom of PID (2 - DOF - PID) controller for automatic generation control (AGC) of multi-area power systems is presented in this paper. The controller parameters are tuned by hybrid invasive weed optimization and pattern search (hIWO - PS) technique. An extensively used standard two area thermal system test system which is considered at the first instance for the AGC design. At the outset, the superiority of hIWO-PS over IWO, Ziegler Nichols (ZN), genetic algorithm (GA), bacteria foraging optimization algorithm (BFOA), differential evolution (DE), particle swarm optimization (PSO), hybrid BFOA - PSO, hybrid PSO - PS and non-dominated sorting GA - II (NSGA - II) is established. It is observed that proposed hIWO - PS tuned 2-DOF-PID controller achieves better system dynamic performances compared to several AGC approaches reported in recent times. Furthermore, robustness analysis is carried out and it is shown that the hIWO - PS tuned 2-DOF-PID controller perform satisfactorily when there are extensive variations in system parameters and operating load conditions. Lastly, the proposed method is applied to three unequal area non-linear hydro thermal system. It is observed that proposed hIWO - PS tuned 2- DOF - PID controller gives better dynamic response than ANFIS, hybrid hBFOA- PSO and hybrid hPSO - PS based approaches for the same power system.

Bibliography

- [1] Ali E.S., Abd-Elazim S.M.(2011), Bacteria foraging optimization algorithm based load frequency controller for interconnected power system, *International Journal of Electric Power & Energy Systems*, 33, 633 – 638, 2011.
- [2] Barisal A.K. (2015), Comparative performance analysis of teaching learning based optimization for automatic load frequency control of multi-sources power systems, *International Journal of Electric Power & Energy Systems*, 66, 67 – 77, 2015.
- [3] Barisal A.K., Prusty R.C. (2015), Large scale economic dispatch of power systems using oppositional invasive weed optimization, *Applied Soft Computing*, 29, 122–137, 2015.
- [4] Dolan E.D., Lewis R.M., Torczon V. (2003), On the local convergence of pattern search, *SIAM Journal of Optimization*, 14, 567–583, 2003.
- [5] Elgerd O.I. (2000), *Electric Energy Systems Theory - An Introduction*, Tata McGraw Hill, New Delhi, India, 2000.
- [6] Farhangi R., Boroushaki M., Hosseini S.H. (2012), Load frequency control of inter-connected power system using emotional learning based intelligent controller, *International Journal of Electric Power & Energy Systems*, 36, 76–83, 2012.
- [7] Gozde H., Taplamacioglu M.C. (2011), Automatic generation control application with craziness based particle swarm optimization in a thermal power system, *International Journal of Electric Power & Energy Systems*, 33, 8–16, 2011.
- [8] Gozde H., Taplamacioglu M.C., Kocaarslan I. (2012), Comparative performance analysis of Artificial Bee Colony algorithm in automatic generation control for interconnected reheat thermal power system, *International Journal of Electric Power & Energy Systems*, 42, 167–178, 2012.
- [9] Karimkashi S., Ahmed A.K. (2010), Invasive weed optimization and its features in electromagnetics, *IEEE Trans Antenna Propagation*, 58, 1269-1278, 2010.

-
- [10] Khuntia S.R., Panda S. (2012); Simulation study for automatic generation control of a multi-area power system by ANFIS approach, *Applied Soft Computing*, 12, 333–341, 2012.
- [11] Kundur P. (2009), *Power System Stability and Control*, Tata McGraw Hill, New Delhi, India, 2009.
- [12] Mehrabian A.R., Lucas C. (2006), A novel numerical optimization algorithm inspired from weed colonization, *Ecological Informatics*, 1, 355–366, 2006.
- [13] Mohamadreza A., Hamed M., Roozbeh I.Z. (2012), State estimation of nonlinear stochastic systems using a novel meta-heuristic particle filter, *Swarm Evolutionary Computation*, 4, 44–53, 2012.
- [14] Mohanty B., Panda S., Hota P.K. (2014), Controller parameters tuning of differential evolution algorithm and its application to load frequency control of multi-source power system, *International Journal of Electric Power & Energy Systems*, 54, 77–85, 2014.
- [15] Naidu K., Mokhlis H., Bakar A.H.A., Terzija V., Ilias H.A. (2014), Application of firefly algorithm with online wavelet filter in automatic generation control of an interconnected reheat thermal power system, *International Journal of Electric Power & Energy Systems*, 63, 401–413, 2014.
- [16] Panda S., Mohanty B., Hota P.K. (2013), Hybrid BFOA-PSO algorithm for automatic generation control of linear and nonlinear interconnected power systems, *Applied Soft Computing*, 13, 4718–4730, 2013.
- [17] Panda S., Yegireddy N.K. (2013), Automatic generation control of multi-area power system using multi-objective non-dominated sorting genetic algorithm-II, *International Journal of Electric Power & Energy Systems*, 53, 54–63, 2013.
- [18] Rad S.H., Lucas, C. (2007), A recommender system based on invasive weed optimization algorithm, *Proceedings 2007 IEEE congress on evolutionary computation*, CEC2007, 4297–4304, 2007. doi: 10.1109/CEC.2007.4425032
- [19] Rout U.K., Sahu R.K., Panda S. (2013), Design and analysis of differential evolution algorithm based automatic generation control for interconnected power system, *Ain Shams Engineering Journal*, 4, 409–421, 2013.
- [20] Sahoo D.K., Sahu R.K., Gorripotu T.S., Panda S. (2016), A novel modified differential evolution algorithm optimized fuzzy proportional integral derivative controller for load frequency control with thyristor controlled series compensator, *J. Electr. Syst. Inform. Technol.*, <http://dx.doi.org/10.1016/j.jesit.2016.12.003>
- [21] Sahu R.K., Panda S., Rout U.K. (2013), DE optimized parallel 2-DOF PID controller for load frequency control of power system with governor dead-band nonlinearity, *International Journal of Electric Power & Energy Systems*, 49, 19–33, 2013.
- [22] Sahu R.K., Panda S., Padhan S. (2014), Optimal gravitational search algorithm for interconnected power systems, *Ain Shams Engineering Journal*. 5, 721–733, 2014.
- [23] Sahu R.K., Panda S., Sekher G.T.C. (2015), A novel hybrid PSO - PS optimized fuzzy PI controller for AGC in multi-area interconnected power system, *International Journal of Electric Power & Energy Systems*, 64, 880–893, 2015.

-
- [24] Sahu, R.K., Panda, S., & Gorripotu, T.S. (2016), Automatic generation control of multi-area power systems with diverse energy sources using Teaching Learning Based Optimization algorithm. *Eng. Sci. Technol. Int. J.*, 19 (1), 113–134, 2016.
- [25] Saikia L.C., Nanda J., Mishra S. (2011), Performance comparison of several classical controllers in AGC for multi-area interconnected thermal system. *International Journal of Electric Power & Energy Systems*, 33, 394–401, 2011.
- [26] Sanchez J., Visioli A., Dormido S. (2011), A two-degree-of-freedom PI controller based on events, *Journal of Process Control*, 21, 639–651, 2011.
- [27] Saravanan B., Vasudevan E.R., Kothari D.P. (2014), Unit commitment problem solution using invasive weed optimization algorithm, *International Journal of Electric Power & Energy Systems*, 55, 21–28, 2014.
- [28] Shabani H., Vahidi B., Ebrahimpour M. (2012), A robust PID controller based on imperialist competitive algorithm for load-frequency control of power systems, *ISA Transactions*, 52, 88–95, 2012.
- [29] Shayeghi H., Shayanfar H.A., Jalili A. (2009), Load frequency control strategies: A state-of-the-art survey for the researcher, *International Journal of Energy Conversion and Management*, 50, 344–353, 2009.
- [30] Wolpert D.H., Macready W.G. (1997), No free lunch theorems for optimization, *IEEE Transactions on Evolutionary Computation*, 1, 67–82, 1997.
- [31] Xiao H. H. (2014), A novel hybrid optimization algorithm of pattern search and IWO, *Applied Mechanics and Materials*, 687-691, 1557-1559, 2014.

A Solution for Interoperability in Crisis Management

F.J. Pérez, M. Zambrano, M. Esteve, C. Palau

Francisco J. Pérez*, Marcelo Zambrano, Manuel Esteve, Carlos Palau

Communications department

Distributed Real Time Systems and Applications Lab

Universitat Politècnica de València

46022 Valencia, Camino de Vera s/n, Spain,

*Corresponding author: frapecar@upvnet.upv.es

Abstract: To effectively manage a crisis, is necessary the participation of multiple agencies related to protection and public safety, which allow actions in accordance with the demands of environment and requirements of all those affected. Interoperability is the key to comprehensive and comprehensive Crisis Management (CM), that allow to face any type of disaster, at any time or place. For this, it is necessary the permanent exchange of information that enables all the agencies involved, coordinate its operations and collaborate to manage the situation in the best way possible. This article describes the approach to interoperability in CM carried out by Secure European Common Information Space for the Interoperability of First Responders and Police (SECTOR) European Project, which has one of its main objectives, the development of an interoperability platform that allows the agencies involved in the management of a crisis, to achieve a joint, coordinated and collaborative response. The platform has a core, a Common Information Space (CIS), which manage as a single storage entity, all information of the Information Systems (ISs) connected to the platform, regardless of the model of data and computer applications used by each one of them.

Keywords: interoperability, distributed architectures, command and control, crisis management, data models.

1 Introduction

A crisis can be defined as an unforeseen situation that endangers the environment, property and / or life of people. The CM refers to those resources and processes required to deal with a crisis in the best possible way. One of its most important characteristics is its multi-agency nature, which allows meet the demands of a critical environment and the requirements of all those affected, through the resources, skills and knowledge coming from different protection and public safety agencies involved [3] [14] [15]. It presents four domains of operation: physical, information, cognitive and social.

- The physical domain is related to the environment where the crisis develops (extension and geographical location, environmental conditions, available resources, affected, etc.).
- The domain of information refers to all those aspects related to the information needed to describe the operations environment (sensor, measurement, storage, publication, etc.).
- The cognitive domain makes use of this information, to create situation awareness from which can be made decisions in line with reality.
- The social domain deals with inter-agency interaction, promoting collaboration between them and covering the three domains described above. It allows the effective exchange of information (information domain); the creation of a common situational awareness and

consensual decision making (cognitive domain); and the coordinated execution of these decisions, which bring about the desired effects on the environment (physical domain) [16] [2].

The capability of a system to exchange and understand information from other systems is defined as Interoperability [9]. In heterogeneous and complex environments such as CM, interoperability is the key to achieve the orchestration of the resources involved, and to allow a coordinated and collaborative response [20] [22] [24].

This article details the solution proposed by the European project SECTOR for interoperability in CM. SECTOR is a part of the Seventh Framework Program of the European Union for Research and Technological Development (FP7), and the Universitat Politècnica de València (UPV) has actively collaborated in its development [21]. One of its objectives is the design and implementation of an interoperability platform that allows the agencies involved integrating within a communications infrastructure to exchange and share information related to the management of a crisis. The platform has as core, a middleware layer that adapts the information coming from different ISs, to the model and format data defined in a CIS, which manages as a single storage entity, all the information coming from those ISs.

The feasibility and operability of the platform has been validated by means of functional tests on a prototype implemented based on the architecture described in this article, and within a simulated scenario for a crisis. Each of the ISs integrated to the platform, shared in the CIS, information related to the state of operations environment and the resources deployed both inside and outside himself. This information made it possible to create accurate and real situational awareness, which in turn enabled an effective planning and coordination of response and recovery operations.

The paper is divided into five sections: firstly, an introduction to the proposal and methodology used; Second, it describes the motivation and the oeuvres that have been taken as main references for the development of SECTOR project; Third, it details the propose architecture and the functionalities of each of its components: fourth, it describes the tests of functionality and the obtained results; And finally, presents the conclusions and final notes for this work.

2 Motivation and related work

One of the most important factors to be considered regarding interoperability in the CM is the heterogeneity of data and the way users access them. Many Crisis Management Systems enable inter-agency interoperability through the use of a proprietary set of IT tools and a standardized data model, that enabling the transparent exchange of information between applications and agencies (e.g. Coordcom [18], Atos [1] or DESTRIERO [19]). However, this scheme is limited by the usability and scope of those tools designed by the manufacturer, which do not necessarily achieve the particular scope and requirements of all agencies. In general, users are reluctant to use external computer tools, either by affinity with the applications they use regularly or by mistrust and expertise lack with computer applications that they do not know.

This work focuses on the CM social domain, and its main contribution lies in the capabilities of the platform, to allow the exchange of information between the agencies involved, regardless of data model, systems and applications computer used in each of them. Agencies use their own tools and ISs to exchange and share information between them.

The architecture detailed in this article has taken as main references for its development to the National Information Exchange Model (NIEM) and the non-relational Data Base (DB) (No Structured Query Language (NoSQL) DB). NIEM was created by the United States Department of Defense (DoD) and the Department of Homeland Security (DHS) in order to interconnect

communities with the need to exchange information to fulfill a common goal. It proposes the development of a middleware layer and the adoption of a normalized data model to allow the ISs involved to exchange information independently of their data model and proprietary applications (Figure 1) [7].

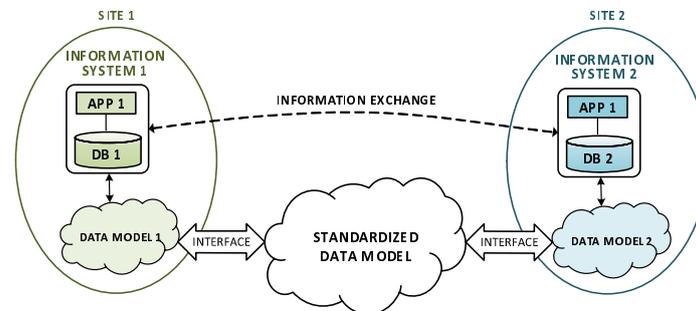


Figure 1: NIEM Schema

On the other hand, the NoSQL DBs allows to store each object with an independent data schema, solving the deficiencies of relational DBs in how many to the management of heterogeneous data [6] [13]. Figure 2 shows a diagram that summarizes the proposal made by SECTOR in terms of data management with NoSQL DBs.

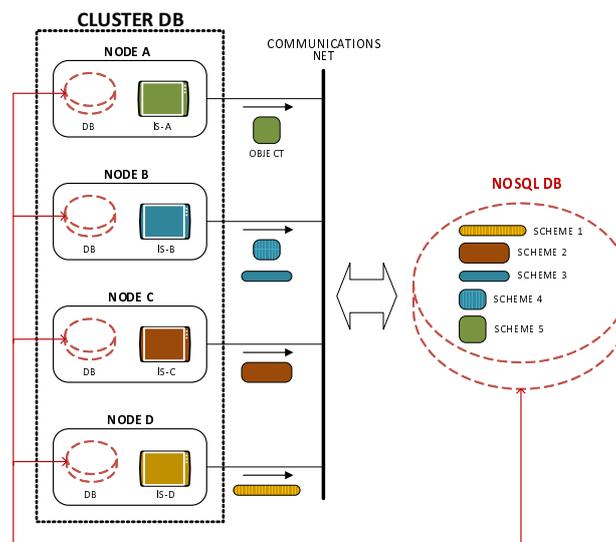


Figure 2: No Relational Distributed Data Base Schema

3 Architecture

The architecture has been designed to support the requirements of availability, scalability and information heterogeneity present in the CM. It is based on a Communications Distributed Infrastructure (CDI) and middleware layer, which enable ISs to exchange and share information regardless of their data model, systems and computer applications. The middleware layer adapts the information coming from the ISs to the data model defined in the CDI. The information that needs to be shared is stored locally on a NoSQL DB, in order to later distribute it said information on the other nodes DBs that make up the CDI. The sharing and exchange of information are

done through a messaging and notification mechanism (detailed below), under the eXtensible Markup Language (XML) data schema defined in NIEM and a Data Distribution Service (DDS) protocol.

The architecture has been developed entirely under free software, eliminating dependence with manufacturers (maintenance, upgrades, etc.) and leaving an open-door for customization and development of new functionalities. Linux has been chosen as the Operating System for the communication nodes, MongoDB as NoSQL DB Manager [23], Java as a programming language for the development of the various modules of the platform, and AngularJS [13] as framework JavaScript for the development of the Human Machine Interface (HMI). To facilitate its design, implementation and scalability, the architecture has been divided into three main elements (Figure 3):

- Information Systems: devices, tools or computer systems, registered on the platform to contribute and / or obtain information from it.
- CDI: responsible for providing connectivity between the nodes that make up the platform. Its topology, as well as the technology used in the communication links, are transparent to the middleware layer, which facilitates the implementation and scalability of the platform.
- Middleware layer fulfills the functions of transceiver between the CDI and the ISs. It is responsible for allowing the exchange of information between ISs integrated to the platform. It is subdivided into four components: Interoperability Boxes (IBXs), communications nodes, CIS and HMI.

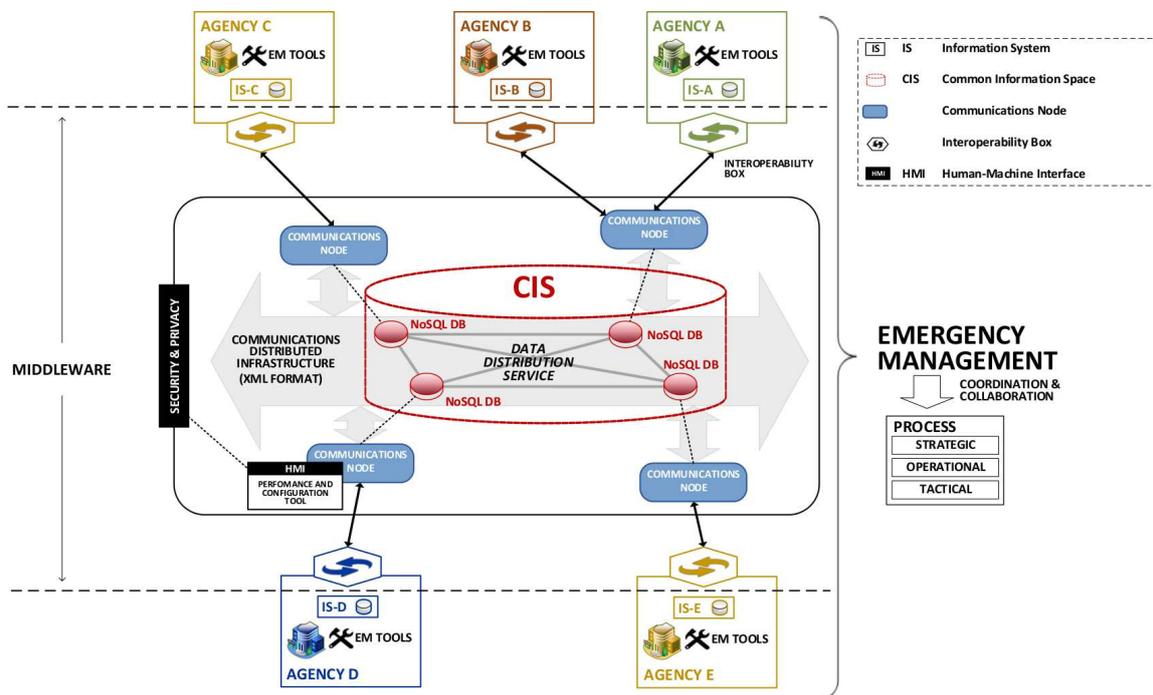


Figure 3: Architecture

3.1 Interoperability boxes

They are responsible for adapting the information coming from the ISs to the XML format defined in the CDI, and vice versa. Each IBX has an interface to the IS and another to the

communications node, allowing the flow of information between them through Web Services (WS) [12]. The requests / responses from and to the ISs are made through the WSs that have been defined in the IS in question (Simple Object Access Protocol (SOAP), Representational State Transfer (REST), Hypertext Transfer Protocol (HTTP), etc.), while the requests / responses from and to the communications nodes, according to platform design and following the principles of standardization proposed, are made only by SOAP.

Among the communication interfaces, there are two sublayers (transformation and communications), which allow the conversion of the data format in a bidirectional way. Due to the heterogeneity in the formats and protocols used by ISs, each of them needs their own IBX, and there will be as many IBXs as ISs will be integrated into the platform. Figure 4 (left panel) shows a general block diagram for an IBX.

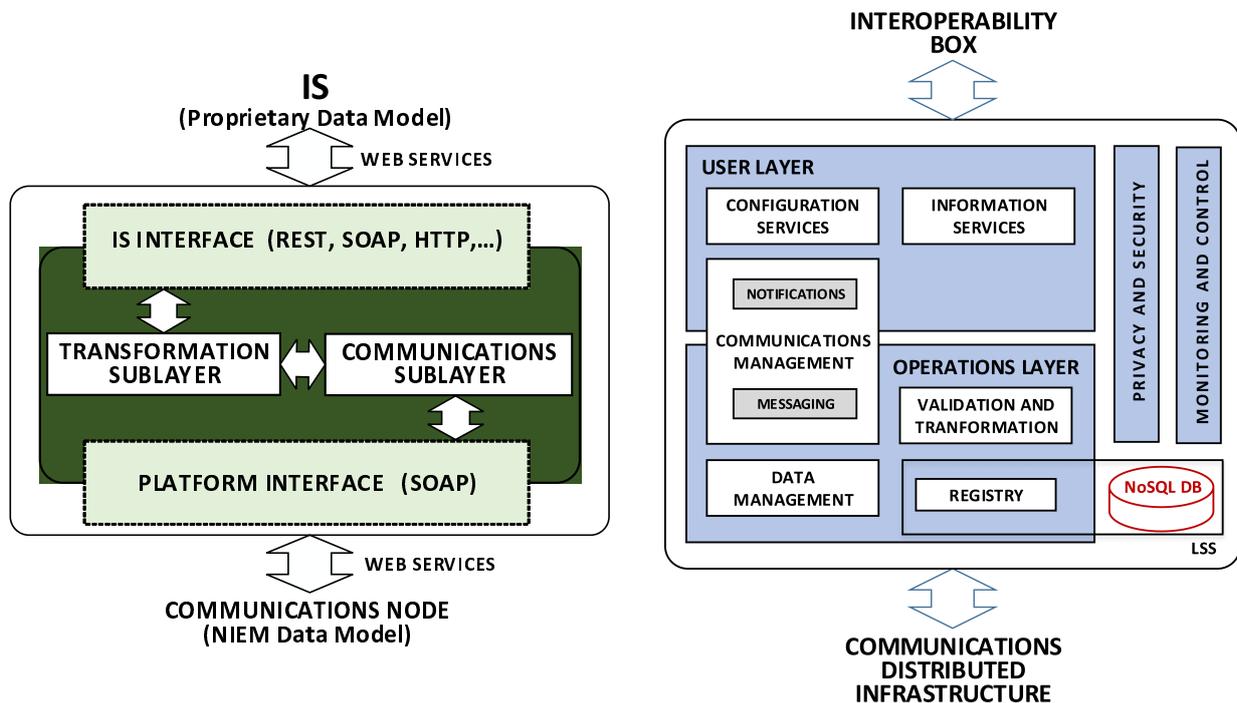


Figure 4: Interoperability Box (left panel) and Communications Node (right panel)

3.2 Communications nodes

They are the gateway to the platform and among its main responsibilities are the registration of users and systems, as well as the notification of any information changes or relevant event on the platform, to the local ISs (integrated to the platform through it) and other communications nodes.

Figure 4 (right panel) shows a block diagram summarizing the internal architecture of a communications node. Four main elements can be observed: a user layer responsible for making available to ISs, the platform configuration services (e.g. ISs registration, notifications subscription, addressing, etc.) and interaction with the CIS (Reading, writing, deleting and updating data); A layer of operations responsible for the management and functional processes inside node (e.g. data validation, node-to-node communication, data management, etc.); Two groups of transversal services responsible for the security, processes supervision and node access; And a Local Storage Space (LSS), in which is it stored a replica of the CIS and the users and systems register.

When a new node needs to be integrated into the platform, the first step is to register on the same. Once authorized its access, the node becomes part of the Cluster that shapes the CIS, and initiate communications with the other nodes through the CDI. The information received (data and metadata) is stored in the LSS, which is accessible by all ISs registered in that node, and will be forwarded to the other nodes that so require it.

The platform can be configured with the number of nodes required, and each node can serve more than one ISs, depending on the scope of the platform, the processing capacity of the node, and the specific requirements of each IS in terms of resources, security, availability and / or reliability.

3.3 Common information space

The CIS is the core of the architecture. It allows the management of all the information coming from the ISs registered in the platform, as a unique storage entity, and it is based on a NoSQL DB cluster, to enable the storage of each data object under an independent schema. The CIS presents a data model based on NIEM, with a binary documental structure type JSON (BSON), this promotes the exchange of information and the transparency of the physical objects location. Moreover it takes advantage of the architecture distributed character to replicate the information in each of the cluster nodes, balancing the data processing, and providing availability, reliability and scalability to the platform. Among its most important features, it is important highlight the decentralization, scalability and adaptability about the data heterogeneity. Both the workload as the diffusion level are parameterizable, and they should be configured following the crisis environment needs.

SECTOR uses MongoDB as DB engine, due to its multiplatform character open source and the document-orientation storage [23]. Regarding the information management, it has been selected OpenSplice (open source distribution of DDS protocol) for the implementation of Publish/Subscribe mechanism on data distribution [4]. Figure 5 summarizes a block diagram with CIS internal architecture.

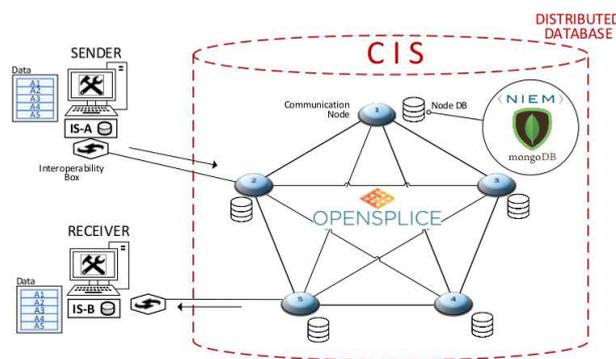


Figure 5: Common Information Space.

3.4 Communications and data management

The information exchange among nodes, and among a node and the local ISs, it is performed using a Publish/Subscribe mechanism, with a XML data schema based on NIEM. The XML datagrams, contain the data and metadata required to notify, to the ISs and the communication nodes, about any change performed in the availability of information and/or the platform status. Before the data storage, the datagrams should be validated and transformed to the binary format BSON defined in the CIS, through Validation and Transform modules available in each node.

The Publish/Subscribe mechanism has been developed taking in to account the performance optimization and the use of network resources operating in low capability environments. This is the case for CM situations with low process capacity and storage, limited bandwidth, packet loss etc. [23] [5].

The information publication and the information requests, use the DDS protocol to distribute the data and metadata through the CDI. The ISs should specify the Topics of which they will be providers, and the Topics of which they want subscribe as consumers. Every time that new information is shared through the CIS, the local node notifies to the local ISs subscribed to the Topic, about the information update via "Notification" service. If the information is relevant for a IS, the registered IBX can retrieve the information from the CIS, transform it into its proprietary data model and make the adapted information accessible to their informatics tools and/or systems. Similarly, the local node shares the update with the other nodes in the platform using the "Messaging" service, and these notify to their ISs if applicable about the new information available. This process is shown in Figure 6.

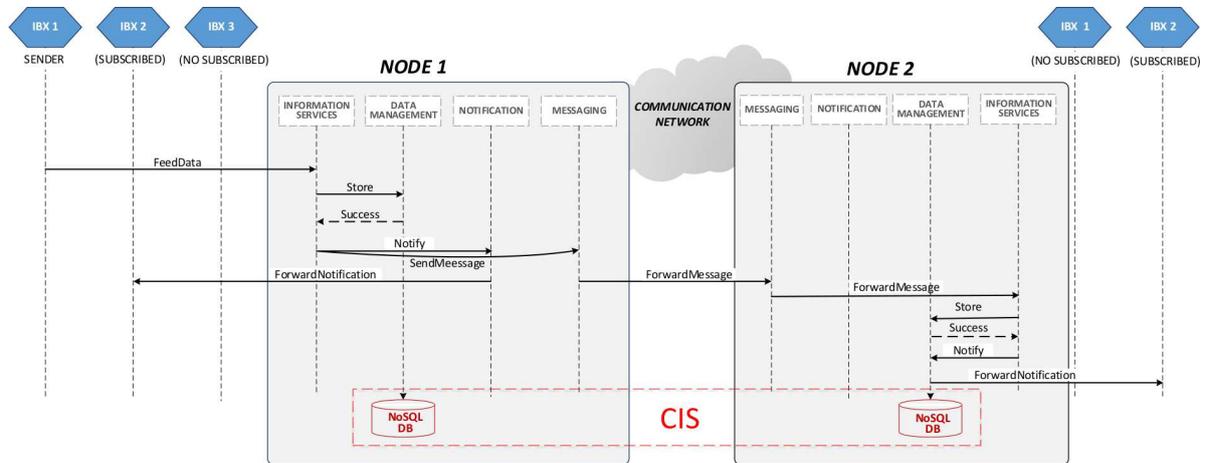


Figure 6: Notification Process.

3.5 Security and human machine interface

The platform access and privacy is managed through the HMI. It is able to create users, manage profiles and assign permissions using the "Privacy and Security" module. It also allows the maintenance and the audits of the platform via "Monitoring and Control" module.

Each node, stores a copy of information related with the users and systems registry, guaranteeing always the accessibility and the security in the platform.

Regarding to the communications security, it is managed using security protocols as Hypertext Transfer Protocol Secure (HTTPS) and Transport Layer Security (TLS), and using security certificates to enable the CIS access.

4 Validation and results

The platform feasibility and functionality, was validated using a prototype developed following the architecture described in this paper and executing a simulation in a flood scenario in Roermond city (Netherlands). Two communication nodes were deployed, and used by the involved agencies in the simulation, for sharing and interchanging information related with the

flood and the environment status. Figure 7, shows the final deployment configuration for the scenario validation.

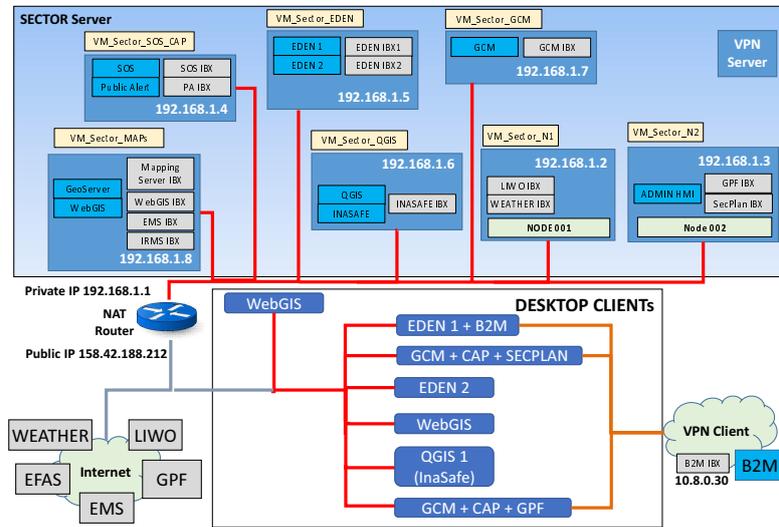


Figure 7: Platform deployment schema

The distributed architecture of SECTOR platform, divides the data processing load among all the CDI nodes and decreases the hardware and software requirements in the communication nodes. A unique HOST with standard features regarding to the hardware and the software (17 octa core processor, 32GB of RAM, 2TB of disk and VMware vSphere 6.5) was used during the tests for virtualizing the 2 communication nodes and the ISs with their respective IBXs.

Each participating agencies, added and/or got information from the CIS, using their own tools and informatics systems. Table 1, shows a summary with the participation of the agencies, and the ISs used by them during the tests.

Table 1: Scenario actors and information systems

Coordination type	Organization	Actor	System
Local	Fire Brigades	SGSP	EDEN
	Police	PSNI	InaSafe
	Red Cross	IAEMO	EDEN
	Roermond Mayor	SSV	GCM
	Directorate for National Safety and Security	External End-User1	WebGIS
International	Bavarian Police	PAS	InaSafe
	UK Red Cross	IAEMO	GPA
	AMI	External End-User2	GPF+GPA

The simulacrum starts with the information sharing related with the environment conditions and the status of the rivers Maas and Mesue, by the local CM agencies. After the information analysis, and following the obtained forecasts, the protocol of an imminent flooding is started. After the corresponding preventive actions, including reports publication and generation of alerts through the CIS (e.g. affectation perimeter, climatic and ambient conditions of the operational environment, hot spots, location and planning of first responders units, etc.). The flooding

overflows the local response capacity, so the emergency state was declared and the help of national and international neighbouring agencies was requested by the local authorities. During the response phase and the situation stabilization, the information related with the actual situation of the environment, the permanent status reports, and location of first responder units, victims and affected people were shared through the CIS. Following the information shared in the CIS, the response and recovery actions were planned, deploying five multi-agency response groups, covering all the affected zone by the flooding. Figure 8, summarizes the participation of the different actors in the exercise, and some of the tools used.

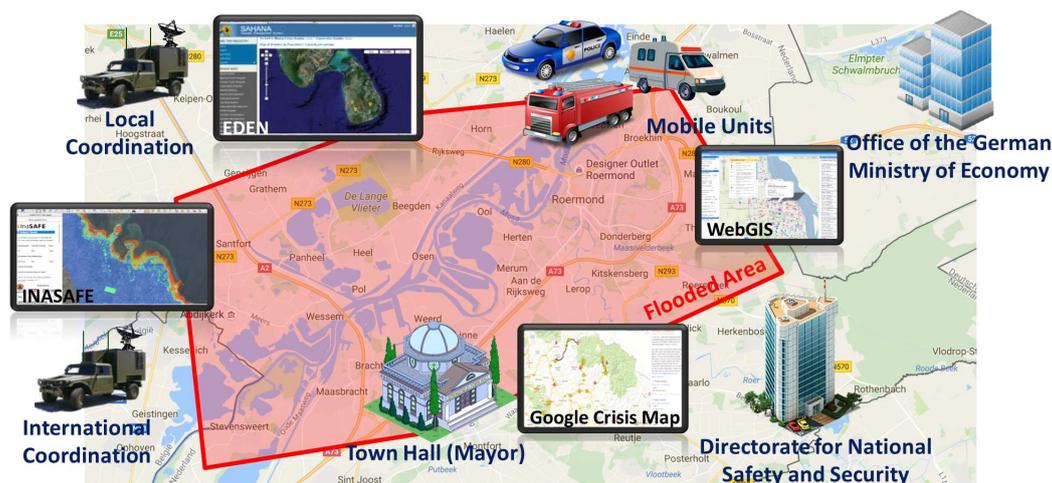


Figure 8: Test scenario

During the simulacrum, and with the final purpose of validate the platform, it was attended by collaborative personnel, both tactical and strategic, from the participant agencies, as well as representatives of several agencies related with the CM in European Union. After the simulacrum, a questionnaire was carried out to the attendant personnel, to have a general impression about the platform functionalities. The obtained results are summarized in Appendix A, Table 2.

The 100% of the respondents were agree, to a greater or less extent, with the platform usability. An 81% would use the platform, either for training or a real environment, and believes that its agency would benefit with the platform functionalities. A 3% would not use the platform, and a 16% would be undecided about its use. In addition, an 87% of the respondents would recommend the use of the platform and agrees with the proposed approach for the inter-agency interoperability.

The proposed solution has been validated for this use case, but it is possible to generalize it in any scope for CM (fires, earthquakes, etc.). The only requirement for the integration of new tools into the platform is the IBXs development, which allow the data exchange and the communication with the nodes following the NIEM format defined in the CDI.

5 Final notes and conclusions

In complex and diverse environments such as CM, the interoperability is the key for an integrated and comprehensive management, allowing a flexible and effective response to any type of disaster that may arise. For this, it is necessary a continuous exchange of information, allowing all involved agencies, to coordinate their operations and collaborate to manage the situation in the best possible way. This paper describes the approach of the European project SECTOR, for the interoperability between the agencies involved in the CM. It describes the architecture of an

interoperability platform, able to integrate into a communication infrastructure the information from the different systems, used by the involved agencies, sharing and interchanging information.

The main contribution of the platform, is found in its capability to manage heterogeneous data, and allow users from different agencies, to obtain and share information, using their own systems and informatics tools.

The core of the architecture, is in its CIS, which allows manage as unique storage entity, all the information coming from different information systems registered in the platform. It is based on a No Relational Distributed DB, and NIEM data model, allowing the heterogeneous data management and divide the work process among all the nodes in the platform.

The platform was validated using a developed prototype following the architecture described in this paper, and tested in a simulated scenario for a crisis. The platform capabilities were verified to facilitate the exchange of information among agencies, and the shared information was used by the personnel involved to plan and coordinate the response and recovery operations.

Moreover, the stored information in the CIS, may be used as support for the management of future crisis, assisting decision support and command & control systems, with relevant information regarding problems and solutions raised in previous events.

Regarding to the future work, alternatives for the implementation of the CIS are being explored, based on distributed database schemas, that allow the requirement optimization in hardware and software for the storage and the data processing [11] [10]. On the other hand, the human machine interface, currently allows administrators access to the platform configuration and monitoring tools. However, its implementation leaves an open-door for the development of other type of tools for customization or requirements compliance.

Currently a proprietary tool is being developed, and it is able to exchange audio and video information in real time, regardless of the communication technology used in the data link between the nodes.

Acknowledgment

This work is funded by European Commission as part of FP7 (No. 607821). Project Supported by LEONARDO company, Universitat Politècnica de València, ASELSAN Elektronik Sanayi ve Ticaret A.S, Totalforsvarets Forskningsinstitut, ITTI SP ZOO, Consorzio Interuniversitario Nazionale Per L'Informatica, SAADIAN Technologies Limited, E-GEOS SPA, Universitaet Stuttgart, Thales SA, Police Service of Northern Ireland, Wyższa Szkoła Policji W Szczytnie, Health Service Executive HSE, Szkoła Główna Służby Pożarniczej, Stichting Studio Veiligheid, and Eberhard Karls Universitaet Tuebingen.

Bibliography

- [1] Atos (2014); *Safeguarding your citizens and assets with emergency management*, ATOS, 2014.
- [2] Defense Department, O.F.T. (2005); *The Implementation of Network-Centric Warfare*, Government Printing Office, 2005.
- [3] Federal Emergency Management Agency; *Course IS-0230.d: Fundamentals of Emergency Management*, FEMA, 2015.
- [4] Foster A. (2016); *DDS at the Tactical Edge - Next Generation Military Fog Computing and Cloudlets Live Webcast*, PRISMTECH, 2016.

-
- [5] Garg P., Sharma M. (2016); "BIG DATA" Analysis Using Hadoop & MongoDB, *International Journal of Modern Computer Science*, 4(3), 153–156, 2016.
- [6] Han J., Haihong E., Guan L., Du J. (2011); Survey on NoSQL Database, *2011 6th International Conference on Pervasive Computing and Applications*, 363–366, 2011.
- [7] Hay D. C. (2014); *National Information Exchange Model. Core Evaluation*, Essential Strategies International, Houston, 2014.
- [8] Huang C.Y., Yang T.T., Chen W.L, Nof S. Y. (2010); Reference Architecture for Collaborative Design, *International Journal of Computers Communications & Control*, 5(1), 71–90, 2010.
- [9] IEEE; *IEEE Standard Computer Dictionary: A Compilation of IEEE Standard Computer Glossaries*, IEEE, 1990.
- [10] Li Y., Manoharan S. (2011); A performance comparison of SQL and NoSQL databases, *2013 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM)*, Victoria, BC, 15-19, 2013. doi: 10.1109/PACRIM.2013.6625441
- [11] Padhy R. P., Patra M. R., Satapathy S. C. (2011); RDBMS to NoSQL: Reviewing some next Non-relational Database's, *International Journal of Advanced Engineering Sciences and Technologies*, 11(1), 15–30, 2011.
- [12] Pirnau M. (2010); Implementing Web Services Using Java Technology, *International Journal of Computers Communications & Control*, 5(2), 251–260, 2010.
- [13] Tudorica B. G., Bucur C. (2011); A comparison between several NoSQL databases with comments and notes, *2011 RoEduNet International Conference 10th Edition: Networking in Education and Research*, IEEE, 1–5, 2011.
- [14] UNCHR (2012); *Manual for Emergency Situations*, United Nations High Commissioner for Refugees, 2012.
- [15] Wayne B. (2008); *Guide to Emergency Management and related terms, definitions, concepts, acronyms, organizations, programs, guidance, executive orders & legislation*, FEMA, 2008.
- [16] Williams A. P. (2010); *Agility and Interoperability for 21st Century Command and Control*, vol. 4, CCRP, 2010.
- [17] [Online]. Available: <https://angularjs.org/>, Accessed on 31 March 2017.
- [18] [Online]. Available: <http://www.carmenta.com/en/products/carmenta-coordcom/>, Accessed on 2 May 2017.
- [19] [Online]. Available: <http://www.destriero-fp7.eu/>, Accessed on 2 May 2017.
- [20] [Online]. Available: <http://www.fema.gov/>, Accessed on 14 April 2017.
- [21] [Online]. Available: <http://www.fp7-sector.eu/>, Accessed on 14 April 2017.
- [22] [Online]. Available: <http://www.isotc223.org/>, Accessed on 12 April 2017.
- [23] [Online]. Available: <https://www.mongodb.com/json-and-bson>, Accessed on 31 March 2017.
- [24] [Online]. Available: <http://www.unisdr.org/>, Accessed on 11 April 2017.

Appendix A: Questionair

Table 2: Questionair for tactical and strategic personnel

Question	(5)	(4)	(3)	(2)	(1)
The key features of the platform are suitable for dealing with crisis response.	8	18	5		
The SECTOR platform offers added value to support faster decision making in response activities.	10	20	1		
The SECTOR platform offers added value to support improved decision making regarding resource and activity planning in crisis response.	11	16	4		
SECTOR enhances the COP for the current Crisis Management Tools.	8	23			
Using the SECTOR platform can be effective to speed up response time.	9	20	2		
The SECTOR platform combines useful Neutral 3rd party applications and information sources for crisis response.	8	18	5		
SECTOR boosts interoperability and information exchange among different organizations.	16	14	1		
Access to the data from the integrated IT Systems and tools is available quickly and efficiently.	6	20	5		
SECTOR provides new capabilities, which enhance the use of the Crises Management Tools.	6	21	4		
I would recommend a fully developed SECTOR platform to a colleague.	8	14	9		
I would use the SECTOR platform in "real life".	5	16	8	2	
I would use the SECTOR platform for field exercises.	9	20	2		
I would also use the SECTOR platform during the recovery phase of a crisis.	7	13	10	1	
I consider my organization's efforts in the field could benefit from the coordination facilitated by the SECTOR platform.	6	15	9	1	
A tool based on SECTOR would be a valuable asset in crises and emergency management.	6	23	2		
Notwithstanding budget constraints, my organization would see value in funding its own SECTOR service or becoming a funding partner in a multi-agency SECTOR service.	2	11	17	1	
A pay-per-use model is a good way for organizations to fund a SECTOR service.	1	10	15	2	3
A fixed, flat annual fee is a good way for organizations to fund a SECTOR service.	3	15	13		

Coverage Hole Recovery Algorithm Based on Molecule Model in Heterogeneous WSNs

X.L. Song, Y.Z. Gong, D.H. Jin, Q.Y. Li, H.C. Jing

Xiaoli Song*

1. State Key Laboratory of Networking and Switching Technology,
Beijing University of Posts and Telecommunications, Beijing 100876, China
2. Henan University of Science and Technology, Luoyang 471023, China

*Corresponding author: song_xiaoli@163.com

Yunzhan Gong, Dahai Jin

State Key Laboratory of Networking and Switching Technology,
Beijing University of Posts and Telecommunications, Beijing 100876, China

Qiangyi Li, Hengchang Jing

Henan University of Science and Technology, Luoyang 471023, China

Abstract: In diverse application fields, the increasing requisitions of Wireless Sensor Networks (WSNs) have more and more research dedicated to the question of sensor nodes' deployment in recent years. For deployment of sensor nodes, some key points that should be taken into consideration are the coverage area to be monitored, energy consumed of nodes, connectivity, amount of deployed sensors and lifetime of the WSNs. This paper analyzes the wireless sensor network nodes deployment optimization problem. Wireless sensor nodes deployment determines the nodes' capability and lifetime. For node deployment in heterogeneous sensor networks based on different probability sensing models of heterogeneous nodes, the author refers to the organic small molecule model and proposes a molecule sensing model of heterogeneous nodes in this paper. DSMT is an extension of the classical theory of evidence, which can combine with any type of trust function of an independent source, mainly concentrating on combined uncertainty, high conflict, and inaccurate source of evidence. Referring to the data fusion model, the changes in the network coverage ratio after using the new sensing model and data fusion algorithm are studied. According to the research results, the nodes deployment scheme of heterogeneous sensor networks based on the organic small molecule model is proposed in this paper. The simulation model is established by MATLAB software. The simulation results show that the effectiveness of the algorithm, the network coverage, and detection efficiency of nodes are improved, the lifetime of the network is prolonged, energy consumption and the number of deployment nodes are reduced, and the scope of perceiving is expanded. As a result, the coverage hole recovery algorithm can improve the detection performance of the network in the initial deployment phase and coverage hole recovery phase.

Keywords: Coverage hole recovery algorithm, molecule model, data fusion, heterogeneous wireless sensor network

1 Introduction

The past few years have witnessed the development of wireless sensor technology and manufacturing microelectronic technology. Wireless sensor networks consist of a large number of micro sensor nodes which have perception, calculation, and communication abilities that are applied to the military or civilian areas, such as environmental monitoring, industrial control, battle-field surveillance, detection of high-risk environments, biological medicine, intelligent household, health monitoring and forecast systems [24, 25].

WSNs are composed of a large number of autonomous nodes which are densely deployed in the target region using a multi-hop, wireless communication mode, and self-organizing, large-scale high density wireless sensor network system organization. The effective deployment of nodes in a wireless sensor network is an important prerequisite for the normal operation of the WSNs [1, 16].

Underwater wireless sensor network (UWSN) is widely applied to information and data collection in many fields such as marine pollution monitoring, marine disaster prevention, underwater aided navigation, marine resources exploration, battlefield surveillance, mine detection, and underwater target detection, tracking and positioning. Thus UWSN causes intensive attention and has been one of the current research hotspots [1, 24, 25]. A large amount of research has been conducted in many applications, for example, goals locating and tracking et al, referring to UWSNs and communication protocols [15–17]. However, in view of deployment optimization research on the underwater wireless sensor nodes, the progress is still slow [9, 21, 22]. Compared with onshore wireless sensor nodes deployment, deployment researches on underwater wireless sensor nodes have more particularities, as the underwater environment is relatively complicated and the mobile wireless sensor nodes can move with water flows frequently. Therefore, it's essential to enhance the studies on how to adjust the position of the underwater wireless sensor nodes to improve the monitoring effect, according to the changeable underwater environment and considering the mobility of monitored targets [8, 12, 30].

The coverage problem is an essential problem in nodes deployment of wireless sensor network; under the condition that the residual energy and ability of perception, communication, and computing of sensor node are limited, using a certain strategy combination of software and hardware, which can ensure the coverage area, network lifetime and good connectivity, thus realize effective awareness and monitoring, is an important indicator of network perception service quality, and is also a hot topic in sensor network research [15–17, 21, 22].

In many applications, there are usually a variety of monitoring objects in the surveillant area. For example, the sensor nodes need to monitor water temperature and salinity in water environmental monitoring, or the sensor nodes need to surveil a variety of chemical diffusions to provide early warnings about factory pollution [3, 9, 10, 12, 30].

Since the hardware cost of abundant sensor nodes is higher currently, in multi-object monitoring applications, each node assembles a variety of different types of sensors; these nodes are heterogeneous. For the limitation of node energy, the more sensors are assembled on a node, the shorter lifetime of the node. Two important problems should be considered in multiple object monitoring network coverage, namely how to use less costs to obtain the ideal network coverage performance, and how to weigh the importance of network monitoring of different objects according to different objectives [8, 14, 23, 29].

Heterogeneous characteristics of heterogeneous wireless sensor networks give expression to three aspects: node heterogeneity, link heterogeneity, and network protocol heterogeneity [13, 19, 20]. Node heterogeneity includes perception ability, calculation ability, communication ability, energy, etc. The aspects of communication ability, perception ability, and residual energy have the greatest influence on the coverage. There has been little research to date on the coverage problem of the random deployment of heterogeneous wireless sensor networks.

A coverage optimization algorithm of wireless sensor network is presented to perceive the radius of the heterogeneous networks, which can effectively improve the coverage ratio in [26]. A multi-objective evolutionary algorithm based on integer vector planning is proposed in [2], which can effectively solve the problem of multiple target monitoring. Both of the two perception models which are used in [5] and [31] are relatively simple.

A routing protocol-based evolutionary algorithm is proposed for the heterogeneous cluster node of WSNs in literature [11], which can effectively reduce the errors of cluster nodes while

handling data combination and separation, and prolong the survival time of a network, but does not give the algorithm of non-cluster nodes. An efficient dynamic clustering strategy (EDCS) is put forward in literature [4], which can effectively solve the selection problem of the multi-level heterogeneous network cluster node, effectively improve the performance of a network and prolong the survival time of a network, without giving a specific node deployment algorithm, however.

A new idea is proposed by increasing heterogeneous nodes to prolong the survival of wireless sensor network in literature [18], but the heterogeneous node perception problem is not considered. To solve the sensing radius of the heterogeneous network node deployment problem, the expanding-virtual force algorithm (EX-VFA) algorithm is put forward in [28], but the connectivity and data fusion problem is barely considered in EX-VFA.

Because of the wireless sensor network nodes' energy depletion or other reasons, the area which is not covered by any node forms "empty holes" in the target field [6]. How to repair the holes is also a hot research topic [7]. To solve the problem of covering blind areas, a node deployment coverage algorithm of heterogeneous sensor networks based on organic small molecules model is proposed in this paper. Based on the organic small molecule structure model, a heterogeneous node detection model is established. In order to improve the network coverage, extend the life of the network, and achieve the optimal deployment of sensor nodes to the monitored area, a data fusion model is proposed that combines data fusion and decision rules. Randomly deployed mobile nodes will move according to the principle of the "virtual potential field." The effectiveness of the coverage algorithm is verified by simulation experiments, which can effectively reduce the number of deployment nodes and the node energy consumption, improve the efficiency of the network coverage, prolong the network lifetime, expand the detection range, and improve the detection performance of the network.

The energy consumption of the wireless sensor network determines the lifetime of the wireless sensor network directly. Reasonably deploying the wireless sensor nodes can improve the coverage effect of the wireless sensor network and reduce the movement distance of the wireless sensor nodes. A nodes deployment algorithm is designed in this article based on perceived probability model of underwater wireless sensor network within the monitored area.

The simulation results show that the algorithm in this article is superior to the virtual force algorithm on improving the coverage effect of underwater wireless sensor network and reducing the movement distance of the wireless sensor nodes, which can achieve the goal of wireless sensor network nodes' reasonable distribution to reduce energy consumption in the initial deployment phase and coverage holes recovery phase.

The rest of this paper is organized as follows:

The background literatures are introduced in Section 1. The assumption and mathematical models for node coverage and perceived are given in Section 2. The virtual force algorithm is explained in Section 3. Our algorithm in this article is presented in Section 4. The results are analyzed in Section 5. The conclusion and some future perspectives are brought in Section 6.

2 Related work

The sensing range of each wireless sensor node is limited in the monitored area. This requires the reasonable deployment of wireless sensor nodes according to certain algorithms in order to ensure that the entire area can be monitored within the sensing range in the monitored field.

The position of wireless sensor nodes after being randomly deployed is uncertain in the coverage problem of wireless sensor network, regarding mobile wireless sensor nodes which can't satisfy the requirements in many cases.

The mobile wireless sensor nodes move according to targets in order to improve the coverage effect of wireless sensor networks.

The movement processes consume a lot of energy, so some reasonable adjustment of the location of wireless sensor nodes and reduction of the movement distance of wireless sensor nodes thereby reduce the consumption energy of wireless sensor networks.

2.1 Assumption

To simplify the calculation, a quantity N of the same mobile nodes deploy randomly in the monitored region. The mobile wireless sensor node S_j owns wireless sensor network ID number j .

The same wireless sensor nodes in the network own the same sensing radius R_s , the same communication radius R_c , and $R_c = 3R_s$.

The wireless sensor node can obtain the location information of itself and its neighbor nodes.

The mobile node owns E_{ini} initial energy which is sufficient to support the completion of the mobile node position migration process.

The mobile node sending 1 byte data consumes E_s energy and receiving 1 byte data consumes E_r energy.

The mobile node migration of 1 meter consumes E_m energy.

The distance between the node S_a and the node S_b is $d(S_a, S_b)$.

The time length of the number k movement process is t_k .

The balance distance between the two wireless sensor nodes of the number k movement process is d_k .

Among them, $d_1 < d_2 < d_3 < \dots < d_k < \dots < d_m$.

2.2 Coverage model

The monitored area is $A \times B \times C$ pixels, which means that the size of each pixel is $\Delta x \times \Delta y \times \Delta z$.

The perceived probability of the i -th pixel perceived by the wireless sensor network is $P(p_i)$, when $P(p_i) \geq P_{th}$ (P_{th} is the minimum allowable perceived probability for the wireless sensor network), the pixels can be regarded as perceived by the wireless sensor network.

The value of $P_{cov}(P_i)$ is used to measure whether the i -th pixel is perceived by the wireless sensor node, i.e.

$$P_{cov}(p_i) = \begin{cases} 0, & \text{if } P(p_i) < P_{th} \\ 1, & \text{if } P(p_i) \geq P_{th} \end{cases} \quad (1)$$

The coverage rate R_{area} is the ratio of the perceived area to the sum of the monitoring area, which is defined in this article, i.e.

$$R_{area} = \frac{P_{area}}{S_{area}} = \frac{\Delta x \times \Delta y \times \Delta z \times \sum_{x=1}^A \sum_{y=1}^B \sum_{z=1}^C P_{cov}(p_i)}{\Delta x \times \Delta y \times \Delta z \times A \times B \times C} \quad (2)$$

Among them, P_{area} is the perceived area while S_{area} is the sum of the monitoring area.

2.3 Perceived model

This article defines the event r_{ij} as the point whereby the i -th pixel p_i is perceived by the wireless sensor node with ID number j . The probability of occurrence of the event is $P(r_{ij})$,

which is equal to the perceived probability $P(p_i, s_j)$, namely, the pixel p_i is perceived by the wireless sensor node s_j , i.e.

$$P(p_i, s_j) = \begin{cases} 1, & \text{if } d(p_i, s_j) \leq R_s - R_e \\ \ln\{1 - \frac{e-1}{2R_e}[d(p_i, s_j) - R_s - R_e]\}, & \text{if } R_s - R_e < d(p_i, s_j) < R_s + R_e \\ 0, & \text{if } d(p_i, s_j) \geq R_s + R_e \end{cases} \quad (3)$$

Where: $d(p_i, s_j)$ is the distance between the i -th pixel p_i and the j -th wireless sensor node s_j ; R_s is the sensing radius of the k -th type wireless sensor node; and R_e is the perceived error range of the k -th type wireless sensor node.

The cooperative sensing monitoring method between lots of wireless sensor nodes is used in this research and the collaborate perceived probability of pixel p_i being perceived by any of underwater wireless sensor nodes is shown as:

$$P(p_i) = 1 - \prod_{j=1}^N [1 - P(p_i, s_j)] \quad (4)$$

3 Virtual force algorithm

3.1 Virtual force algorithm

Assume that the wireless sensor nodes are the particulates in the electric field and that an electric force between the wireless sensor nodes exists and moves the wireless sensor nodes under the action of the electric force as evenly as possible in order to achieve a reasonable distribution of underwater wireless sensor networks, thus improving the coverage effect of the targets in the monitoring area.

If the distance between the two wireless sensor nodes is too far, the attractive force will play a major role and bring the two nodes close to each other. If the distance between the two wireless sensor nodes is too close, the repulsion force will play a major role and make the two nodes separate from each other.

Calculating the distance $d(s_i, s_j)$ between the two nodes s_i and s_j , which determines how the mobile wireless sensor nodes should move.

The repulsion $F(p_i, s_j)$ of wireless sensor node s_i to wireless sensor node s_j can be represented as:

$$F(s_i, s_j) = \begin{cases} \frac{k_1}{d(s_i, s_j)^{\alpha_1}}, & 0 < d(s_i, s_j) < R \\ 0, & d(s_i, s_j) \geq R \end{cases} \quad (5)$$

Where: k_1 , α_1 is the gain coefficient, $d(s_i, s_j)$ is the distance between the two nodes s_i and s_j , and R is the effective distance.

The direction of the force is composed of wireless sensor node s_i to the wireless sensor node s_j . After decomposition, the component along the X-axis direction can resolve to $F_x(s_i, s_j)$, the component along the Y-axis direction can obtain $F_y(s_i, s_j)$, and the component along the Z-axis direction is found to be $F_z(s_i, s_j)$.

So the total value of the force along the X-axis direction is $F_x(s_j) = \sum_{i=1}^N F_x(s_i, s_j)$, the sum of the force along the Y-axis direction is $F_y(s_j) = \sum_{i=1}^N F_y(p_i, s_j)$, and the sum of the force along the

Z-axis direction is $F_z(s_j) = \sum_{i=1}^N F_z(p_i, s_j)$. Thus the resulting force from the circular monitoring area whose center is located in the k -th type wireless sensor nodes s_j with the radius R_k is:

$$F_{xyz}(s_j) = \sqrt{F_x^2(s_j) + F_y^2(s_j) + F_z^2(s_j)} \quad (6)$$

After force calculations according to the sum of the force, the wireless sensor node s_j moves to the new location $(x_{new}, y_{new}, z_{new})$ from the original location $(x_{old}, y_{old}, z_{old})$:

$$x_{new} = \begin{cases} x_{old}, & |F_{xyz}(s_j)| \leq F_{th} \\ x_{old} + \frac{F_x(s_j)}{F_{xyz}(s_j)} \times MaxStep \times e^{-F_{xyz}^{-1}(s_j)}, & |F_{xyz}(s_j)| > F_{th} \end{cases} \quad (7)$$

$$y_{new} = \begin{cases} y_{old}, & |F_{xyz}(s_j)| \leq F_{th} \\ y_{old} + \frac{F_y(s_j)}{F_{xyz}(s_j)} \times MaxStep \times e^{-F_{xyz}^{-1}(s_j)}, & |F_{xyz}(s_j)| > F_{th} \end{cases} \quad (8)$$

$$z_{new} = \begin{cases} z_{old}, & |F_{xyz}(s_j)| \leq F_{th} \\ z_{old} + \frac{F_z(s_j)}{F_{xyz}(s_j)} \times MaxStep \times e^{-F_{xyz}^{-1}(s_j)}, & |F_{xyz}(s_j)| > F_{th} \end{cases} \quad (9)$$

Where, F_{th} is the virtual force threshold. The wireless sensor node need not move when the virtual force which the node received is less than the value. MaxStep is the maximum allowable movement distance.

The virtual force algorithm is as follows:

Step 1: Each wireless sensor node in the monitoring area broadcasts the information which includes the node ID and location information.

Step 2: The wireless sensor node updates the neighbor table information if the wireless sensor node receives the broadcast information of the neighbor nodes.

Step 3: Formulas (5) and (6) are used to calculate the resultant force $F(s_j)$ of wireless sensor node s_j .

Step 4: Formulas (7) to (9) are used to calculate the new location where the wireless sensor nodes need to move to.

Step 5: The movement process won't proceed if the new location where the wireless sensor node needs to move to is located outside of the monitoring region in which case the wireless sensor node needn't move, so proceed to Step 6; otherwise it is moved to a new location, then go to Step 6.

Step 6: The algorithm stops when it reaches a pre-set cycle number T ; otherwise begin the next movement process, then go to Step 1.

3.2 Problems in the virtual force algorithm

The problem in the virtual force algorithm is the following:

Randomly deploy 16 wireless sensor nodes in the two-dimensional monitoring area and the result is shown in Figure 1 (the nodes are located in center of the circles).

According to the virtual force algorithm, all of the circles have to move to the new location; but in fact the red circles don't need to move because moving the red circles won't improve the percentage of coverage, but consumes the energy of the wireless sensor nodes.

The same problem also exists when the wireless sensor nodes are randomly deployed in three-dimensional monitoring areas.

So how to solve the problem in the nodes' density area is discussed in the next section.

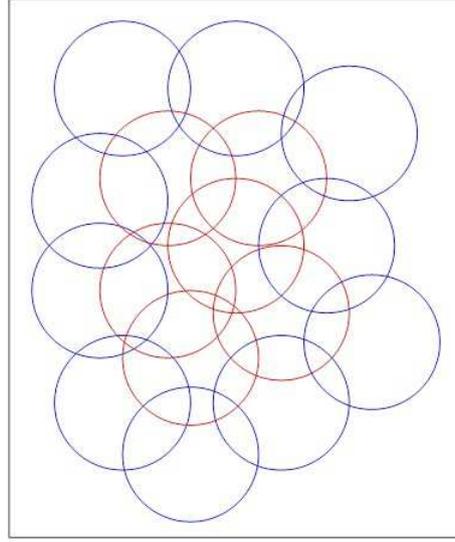


Figure 1: Sixteen randomly deployed wireless sensor nodes in the monitoring area

4 Perception model using molecular structure

We assume that the monitoring area H is rectangular, and that N kinds of heterogeneous perception abilities of wireless sensor nodes are randomly distributed within the rectangle H , and also assume that the wireless sensor network has the following properties:

(1) Heterogeneous perception ability sensor. Sensors have different perception abilities; the perception of the radius is different, and the node model perception is the probability perception model. The position of the sensor node i is (x_i, y_i) , the perception distance is r_i , and the perceived probability is p_i according to the function $p_i(r_i)$. The typical perception probability model is as follows:

$$P(S_i, Q) = \begin{cases} 0, & r + r_e \leq d_{ip}, \\ \frac{E_{ir}}{E_i} e^{(-\lambda\alpha^\beta + \alpha\delta)}, & r - r_e \leq d_{ip} \leq r + r_e, \\ 1, & r - r_e \geq d_{ip} \end{cases} \quad (10)$$

Where $P(S_i, Q)$ is the perception probability of the sensor node S_i to target node Q , d_{ip} is Euclidean distance between sensor node S_i and the target node Q , r_e ($r_e < r$) is an uncertainty perception measure of the sensors, E_i is the initial energy of the sensor nodes, E_{ir} is the remaining energy, $\alpha = d_{ip} - (r - r_e)$, λ and β are the attenuation coefficient of perception quality for the things within the scope of sensor nodes monitoring.

δ is random number meeting normal distribution, which shows the reality of various interference in the influence of the perceived probability;

(2) All sensor nodes have the ability of data fusion, and underwater communications have enough energy to finish self-positioning and can move to the specified location freely;

(3) Before the deployment algorithm is executed, all nodes have finished self-positioning accurately, and the nodes' coordinates are known;

(4) There are three kinds of working state in node life: dormancy, detection, and communication. The node energy consumption in the communication state is the largest, and is the smallest in a dormant state. The relationship of the energy consumption and the time of the nodes in the communication state is positive. Meanwhile, the larger the communication distance, the more energy is consumed.

DSmT is put forward by the French scholar Dezert in literature [27], and is later developed jointly by Dezert and Smarandache as well as other scholars. DSmT is an extension of the classical theory of evidence, but different from basic D-S theory. DSmT can combine with any type of trust function of an independent source, but is mainly concentrated in a combined uncertainty, high conflict, and inaccurate source of evidence. Especially when the conflict between sources becomes bigger or the elements are vague or relatively imprecise, DSmT can go beyond the limitations of the D-S theoretical framework to solve complex fusion problems of static or dynamic status.

On the basis of the above assumptions and DSmT related theory, we give the following definitions:

Definition 1: P-reliability coverage: Assume that point Q exists in the monitoring area H , and that the credibility of the target node at point Q in sensor networks, namely, test results $P(Q)$ meets

$$P(Q) \geq p \quad (0 < p \leq 1) \quad (11)$$

Definition 2: Effective coverage proportion: If the area H that needs to be monitored is two-dimensional, the acreage of H is $S(H)$, the acreage meeting P-reliability coverage is $S_p(H)$, and the ratio of $S_p(H)$ and $S(H)$ is

$$\eta = \frac{S_p(H)}{S(H)} \times 100\% \quad (12)$$

Where η is the effective coverage proportion of the two-dimensional monitoring area H .

If the monitoring area H is three-dimensional, its volume is $V(H)$, the volume meeting p -reliability coverage is $V_p(H)$, and the ratio of $V_p(H)$ and $V(H)$ is

$$\eta = \frac{V_p(H)}{V(H)} \times 100\% \quad (13)$$

Where η is the effective coverage proportion of the three-dimensional monitoring area H .

Definition 3: Effective detection rate: If there are $N(H)$ target nodes in the monitoring area H , in which the $N_A(H)$ target nodes are effectively monitored by the sensor network, then

$$\varphi = \frac{N_A(H)}{N(H)} \times 100\% \quad (14)$$

Where φ is the effective detection rate of the monitoring area H in the sensor network.

Organic small molecules are the organic compound whose molecular weight is under 1000, such as methane (CH₄), ethane (C₂H₆), ethanol (C₂H₅OH), benzene (C₆H₆), etc.

The main atoms of small organic molecules are carbon atoms, hydrogen atoms, oxygen atoms, or nitrogen atoms, among others et al. They are grouped together according to certain structures as shown in Figure 2. Inspired by this, the author combined different structures' sensor nodes into one sensing unit according to certain rules. The compound nodes having different complementary sensing nodes can effectively expand the scope of perception, improve the efficiency of a single node's perception, and increase the effective coverage of sensor networks by an appropriate data fusion model.

Definition 4: Perception atoms: Assuming there are N sensor nodes in the monitoring area H , including M nodes which have the ability to perceive target nodes alone, the nodes that can monitor the target and transmit data to the cluster head nodes individually are called perception atoms.

Definition 5: Perception molecules: Assume that n kinds of sensor nodes are deployed in the monitoring area, and the different kinds of sensor nodes are combined into sensing units in

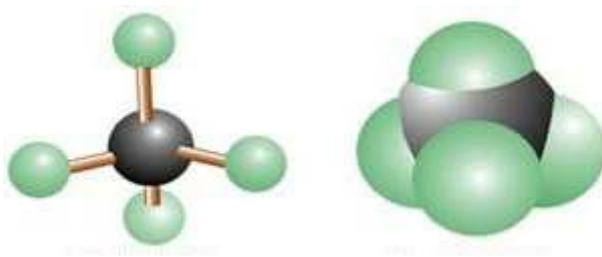


Figure 2: Structure model of methane

accordance with the appropriate data fusion model. These are called sensing unit perception molecules. Perception molecules are the smallest perception and communication unit.

Assume that there are n perception atoms in the molecule, each perception atomic data is $\theta_1, \theta_2, \dots, \theta_n$ respectively, and then the perception molecular fusion data source is

$$U = \{\theta_1, \theta_2, \dots, \theta_n\} \quad (15)$$

Considering the perceiving unit model and data fusion model, we put forward a sensor network node deployment algorithm based on a small molecule model as follows:

Step 1: Build an appropriate perception model, and determine the unit (perception molecule) perception range and the position of the atom in every perception unit according to the monitoring and coverage requirements for monitoring area of H . Compute the optimal distance between the molecules, and the number of all kinds of sensing nodes.

Step 2: Randomly deploy sensing nodes in the monitoring area H according to the number of all kinds of sensing nodes in Step 1.

Step 3: Select a starting node, which usually is a node in vertex position of the monitoring area H . Then select the corresponding node closest to the starting node in order to build a sensor molecule, and set its ID to 1 consequently.

Step 4: Choose a perception node closest to perception molecule 1, and build the second perception molecule according to Step 3, and set its ID to 2. According to the "molecular force" principle, the force between molecules is associated with the distance between the molecules. Assume that the distance between the two perception molecules is D' , the best distance for molecules is D_i , t is the threshold value;

When the distance D' between perception molecules meets

$$D' < D_i - t \quad (16)$$

The intermolecular force is repulsion, and node 2 moves far away from a unit.

When the distance D' between perception molecules meets

$$D' > D_i + t \quad (17)$$

The intermolecular force is attraction, and node 2 moves closely to a unit.

Until the perception molecular distance meets

$$D_i - t \leq D' \leq D_i + t \quad (18)$$

Node 2 moves completely.

Step 5: Using the greedy algorithm, repeat Steps 3 and 4, and adjust the molecular position until the entire molecules are adjusted.

Step 6: Each perception molecule completes the initialization for data fusion task allocation and data fusion node rotation sequence. Consequently, the perception molecular sensor nodes act as data fusion nodes and communication nodes in turn.

Step 7: The data fusion nodes collect the data of other sensors for data fusion at regular intervals.

Step 8: The nodes for data fusion send the data to sink nodes, and then specify a data fusion node and broadcast the messages to other nodes according to the data fusion node rotation order.

Step 9: Sensor nodes repeat Steps 7 and 8 after receiving the data fusion command.

5 Results analysis

We simulate our algorithm by MATLAB and assume that the monitoring area H is square whose side length is 3000m. The aim is to obtain p -reliability coverage in the monitoring area, and $p = 0.9$. The effects of target distribution and other environmental factors are not considered.

We assume that two types of sensors are used in underwater sensor networks: passive sonar sensors and giant magnetic resistance sensors. For passive sonar sensors the effective radius of perception is $R = 50\text{m}$ under the premise of probability coverage.

According to the character of giant magneto resistance sensors and the magnetic properties of the target node, the perception range of giant magneto resistance sensors is irregular, and can be assumed to be elliptic, with an elliptical long radius $R_a = 60\text{m}$ and short radius $R_b = 30\text{m}$ under the condition of meeting the probability coverage. The numbers of two kinds of sensor are deployed by a ratio of 1:4 and the total number is N .

The simulation experiments are carried out using the random deployment algorithm and based on the “virtual force” deployment algorithm respectively. These two kinds of algorithm simulation results are compared with the performance of the deployment algorithm in this paper. The higher the network coverage rate and the more effective the detection rate of the target node, the better the performance of network monitoring consequently.

In the deployment phase simulate 100 times when $N = 100, N = 120, N = 140, N = 160, N = 180, N = 200$.

The simulation results are averaged and shown in Figures 3 to 5.

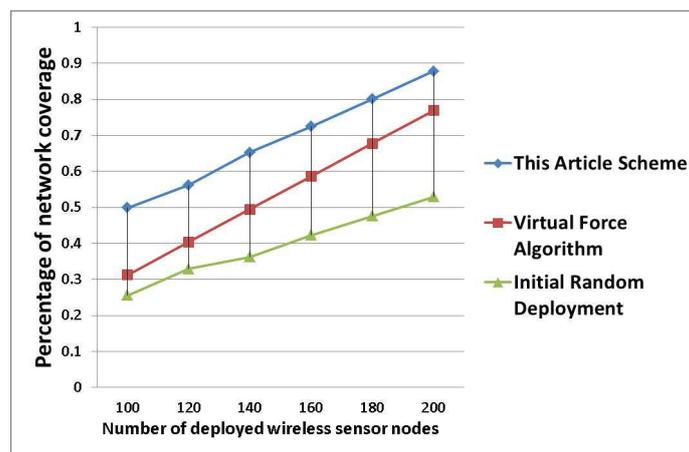


Figure 3: Percentage of coverage after deployment in different algorithms

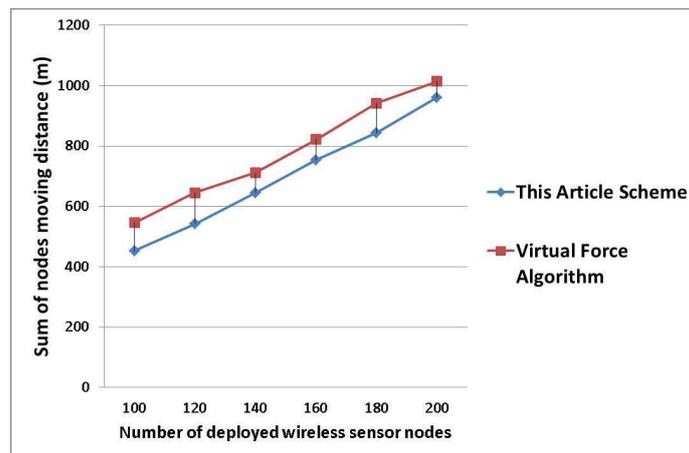


Figure 4: Sum of nodes movement distance after deployment in different algorithms

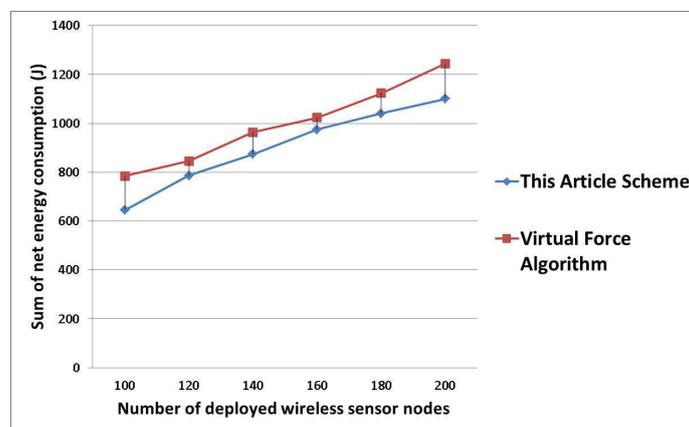


Figure 5: Sum of net energy consumption after deployment in different algorithms

Randomly lose 10 nodes after initial deployment and use their algorithms in wireless sensor networks to recover coverage holes. The results of 100 simulations are averaged and shown in Figures 6 to 8.

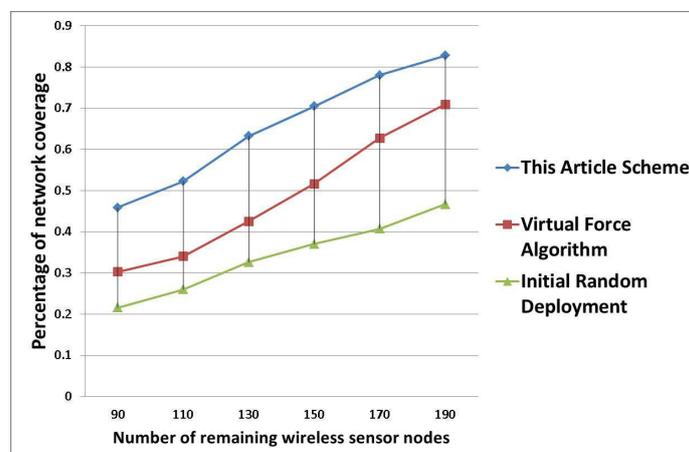


Figure 6: Percentage of network coverage after recovery coverage holes in different algorithms

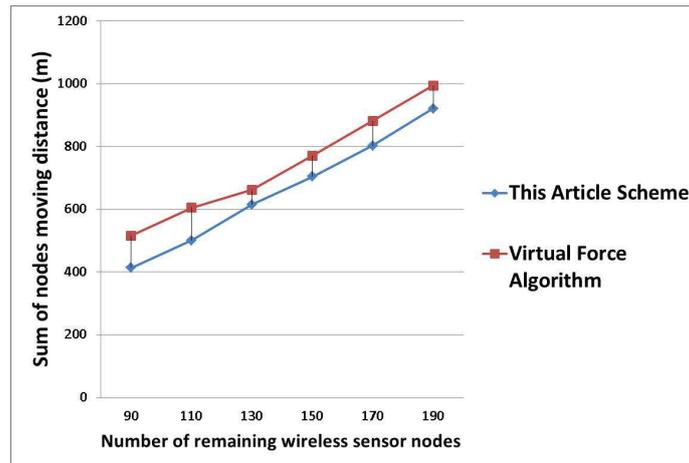


Figure 7: Sum of nodes' movement distance after recovery coverage holes in different algorithms

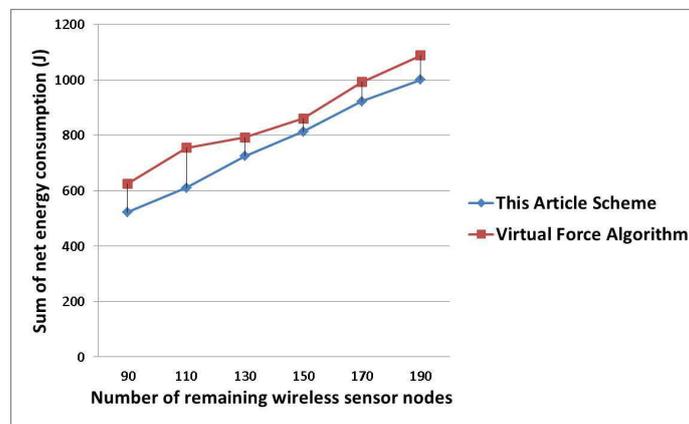


Figure 8: Sum of net energy consumption after recovery of coverage holes in different algorithms

Compared with initial nodes randomly deployed and virtual force algorithm, Figure 3 and Figure 6 show our method achieves more effective coverage rate, a single node perception is more efficient as well, which is a key target of node deployment.

The algorithm of this article reduces the blind area during perception and the scope of overlapping sense of node perception by adopting data fusion algorithm, compared with virtual force algorithm which is better than the initial randomly nodes deployed, thus obtains the best effect within the three methods mentioned above.

In our algorithm, both the number of moving perceptual nodes and the energy consumption are greater than those of the virtual force algorithm at the initial stage of deployment, nevertheless, the less energy consumption as a whole and the longer lifetime of WSN are superior to those of the rest compared methods after the complete deployment due to dormancy mechanism of redundancy node as showed in Figures 3 to 8.

The monitoring area coverage percentage increases. The sum of nodes' movement distance increases and the sum of nodes' energy consumption increases with the growing number of wireless sensor nodes as shown in Figures 3 to 8. The nodes' movement distance is larger and the coverage effect is worse when using the virtual force algorithm in areas that are dense with wireless sensor nodes, because the virtual force algorithm cannot reasonably move the wireless sensor nodes in the initial deployment phase and coverage hole recovery phase.

6 Conclusions

In this paper, a heterogeneous sensor network node deployment algorithm is presented based on the organic small molecule model and used to reduce the deployment node number as well as improve the efficiency of network coverage. We describe the details of our algorithm and compare it with other proposed schemes. The comparison and theory analysis show that the proposed scheme outperforms most of the existing schemes. Further work needs to be done in order to improve and optimize the model presented in this paper and try other fusion strategies to optimize the heterogeneous sensor network node deployment algorithm and make it more efficient.

Acknowledgements

This work was partially supported by the National High-Tech Research and Development Program of China (863 Program) under Grant No. 2012AA011201, in part by the Major Research plan of the National Natural Science Foundation of China under Grant No. 91318301, and in part by the National Natural Science Foundation of China (NSFC) under Grant No. 61202080. The authors also acknowledge the helpful comments and suggestions of the editors and reviewers gratefully, which have improved the presentation.

Bibliography

- [1] Aggarwal A., Kirchner F. (2014), Object Recognition and Localization: The Role of Tactile Sensors, *Sensors*, 14(2), 3227-3266, 2014.
- [2] Attea B.A., Khalil E.A. (2012); A New Evolutionary Based Routing Protocol for Clustered Heterogeneous Wireless Sensor Networks, *Applied Soft Computing*, (12), 1950-1957, 2012.
- [3] Cajal C., Santolaria J., Samper D., Garrido A. (2015), Simulation of Laser Triangulation Sensors Scanning for Design and Evaluation Purposes, *Int. Journal of Simulation Modelling*, 14(2), 250-264, 2015.
- [4] Cardei M. et al. (2005); Energy-efficient Target Coverage in Wireless Sensor Networks, *24th Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM 2005)*, Miami, 1976-1984, 2005.
- [5] Chen J., Du Q., Li X., Ding F. (2012), Research on the Deployment Algorithm of Heterogeneous Sensor Networks Based on Probability Model, *Journal of Chinese Computer Systems*, 2012, 33(1), 50-53, 2012.
- [6] Dezert J. (2002); Foundations of a New Theory of Plausible and Paradoxical Reasoning, *Information and Security Journal*, 13(9): 90-95, 2002.
- [7] Dezert J. et al. (2006); Target Type Tracking with PCR5 and Dempster's Rules: A Comparative Analysis, *Proceedings of Fusion 2006 International conference on Information Fusion*, Firenze, Italy, 2006.
- [8] Du X., Sun L., Guo J., Han C. (2014), Coverage Optimization Algorithm for Heterogeneous WSNs, *Journal of Electronics and Information Technology*, 36(3), 696-702, 2014.

- [9] Duan H.Y. (2016); Research on Collaboration in Innovative Methods of Manufacturing Innovation Chain, *Iberian Journal of Information Systems and Technologies*, E11: 292-303, 2016.
- [10] Fichera A., Frasca M., Volpe R. (2016), On energy distribution in cities: a model based on complex networks, *International Journal of Heat and Technology*, 34(4), 611-615, 2016.
- [11] Halder S., Bit S.D. (2014), Enhancement of Wireless Sensor Network Lifetime by Deploying Heterogeneous Nodes, *Journal of Network and Computer Application*, (38), 106-124, 2014.
- [12] Hassan A.R., Gbadeyan J.A. (2015); A reactive hydromagnetic internal heat generating fluid flow through a channel, *International Journal of Heat and Technology*, 33(3), 43-50, 2015.
- [13] Hong Z., Yu L., and Zhang G. (2013); Efficient and Dynamic Clustering Scheme for Heterogeneous Multi-level Wireless Sensor Networks, *Acta Automatica Sinica*, 39(4): 454-464, 2013.
- [14] Huang S., Cheng L. (2011); A Low Redundancy Coverage-enhancing Algorithm for Directional Sensor Network Based on Fictitious Force, *Chinese Journal of Sensors and Actuators*, 24(3): 418-422, 2011.
- [15] Jing H.C. (2015), Routing Optimization Algorithm Based on Nodes Density and Energy Consumption of Wireless Sensor Network, *Journal of Computational Information Systems*, 11(14), 5047-5054, 2015.
- [16] Jing H.C. (2015), Node Deployment Algorithm Based on Perception Model of Wireless Sensor Network, *International Journal of Automation Technology*, 9(3), 210-215, 2015.
- [17] Jing H.C. (2014), Coverage holes recovery algorithm based on nodes balance distance of underwater wireless sensor network, *International Journal on Smart Sensing and Intelligent Systems*, 7(4), 1890-1907, 2014.
- [18] Kashi S.S., Sharifi M. (2012); Coverage Rate Calculation in Wireless Sensor Networks, *Computing*, 94(11): 833-856.
- [19] Kumar D., Aseri T C, Patel R B. (2009); EEHC: Energy Efficient Heterogeneous Clustered Scheme for Wireless Sensor Networks, *Computer Communications*, 32, 4, 662-667.
- [20] Li M. (2011); *Study on Coverage Algorithms for Heterogeneous Wireless Sensor Networks*, Ph.D. dissertation, Chongqing University, 2011.
- [21] Li Q., Ma D., Zhang J. (2014); Nodes Deployment Algorithm Based on Perceived Probability of Wireless Sensor Network, *Computer Measurement and Control*, 22(2), 643-645, 2014.
- [22] Li Q., Ma D., Zhang J., Fu Z. (2013); Nodes Deployment Algorithm of Wireless Sensor Network Based on Evidence Theory, *Computer Measurement and Control*, 21(6), 1715-1717, 2013.
- [23] Li M., Tang M. (2013), Information security engineering: A framework for research and practices, *International Journal of Computers Communications & Control*, 8(4), 578-587, 2013.

- [24] Moreno-Salinas D., Pascoal A., Aranda J. (2013), Sensor Networks for Optimal Target Localization with Bearings-Only Measurements in Constrained Three-Dimensional Scenarios, *Sensors*, 13(8), 10386-10417, 2013.
- [25] Moradi M., Rezazadeh J., Ismail A.S. (2012), A Reverse Localization Scheme for Underwater Acoustic Sensor Networks, *Sensors*, 12(4), 4352-4380, 2012.
- [26] Sengupta S. et al. (2013); Multi-objective Node Deployment in WSNs: in Search of an Optimal Trade-off among Coverage, Lifetime, Energy Consumption, and Connectivity, *Engineering Applications of Artificial Intelligence*, 26(1), 405-416, 2013.
- [27] Smarandache F., Dezert J. (2006); *Advances and Applications of DSMT for Information Fusion*, Rehoboth: American Research Press, 2006.
- [28] Smarandache F., Dezert J. (2005); Information Fusion Based on New Proportional Conflict Redistribution Rules, *Proceedings of Fusion 2005 Conference*, Philadelphia, 1-8, 2005.
- [29] Tang M., Li M., Zhang T.(2016), The impacts of organizational culture on information security culture: a case study, *Information Technology and Management*, 7(2), 179-186, 2016.
- [30] Xu L., Li C., Jun Y. (2014), Multi-objective Strategy of Multiple Coverage in Heterogeneous Sensor Networks, *Journal of Electronics and Information Technology*, 36(3), 692-695, 2014.
- [31] Zhen H., Yu L., Zhang G. (2013); Efficient and Dynamic Clustering Scheme for Heterogeneous Multi-level Wireless Sensor Networks, *Acta Automatica Sinica*, 39(4), 454-460.

High-speed Train Control System Big Data Analysis Based on Fuzzy RDF Model and Uncertain Reasoning

D. Zhang

Dalin Zhang

National Research Center of Railway Safety Assessment
Beijing Jiaotong University
Beijing, 100044, China
dalin@bjtu.edu.cn

Abstract: China high-speed train control system is a combination of computer, communication and control. Its events are diverse, including sensor data stream, GPS signal, GSM-R transmission data, real-time video monitoring data, train control software data, etc. These data have the typical characteristics of big data. If these data are well applied, this will be of great help to operations, maintenance, safety, passenger services, etc. This paper presents an efficient analysis method based on the fuzzy RDF model and uncertain reasoning for high-speed train control system big data. We have used the method proposed in this paper to analyze the data of the high-speed train control system. The experiment results show that the method proposed in this paper has good efficiency and scalability for the analysis of big data with different structures, types and context sensitive from high-speed train control system.

Keywords: high-speed train control system, fuzzy RDF, D-S theory, uncertain reasoning.

1 Introduction

Railway transportation system is a complex system with many factors. It has been unable to meet the high transport capacity, high efficiency, high safety, high quality of service requirements relying on traditional theory and technical methods. The intelligent rail transport system has become a trend, such as the European rail traffic management system, the Japanese Cyber Rail, the US IRS (Intelligent Railway System), etc. In recent years, the IBM Smarter Railroad [3], CISCO Smart + Connected Railway [11], SIMENS Intelligent Train [20], they have promoted the development of the intelligent railway.

High-speed train control system makes use of the information provided by the ground, the distance between the target and the route, and generated the control curve automatically on the train control equipment. China has developed the standard of Chinese train control system(CTCS) in 2008, which have 4 level totally [17]. The CTCS-3 uses the GSM-R wireless communication for bidirectional transmission of information between the train and the ground. It uses the wireless blocking center(RBC) to generate traffic permits, uses the track circuit to achieve the train occupancy check, and uses a transponder to achieve train positioning. At the same time, it has the function of CTCS-2.

High-speed train control system is a combination of computer, communication and control. Its events are diverse, including sensor generated data stream, GPS signal, GSM-R transmission data, real-time video monitoring data, train control software data, etc. These data have the typical characteristics of big data, including massive data, distributed, complex, context-sensitive, etc. If these data are well applied, this will be of great help to high-speed rail operations, maintenance, safety, passenger services, etc.

The dynamic monitoring system(DMS) is a comprehensive system for dynamic monitoring of the use status of the relevant equipment of the train control system. It can analyze the data, guide

the maintenance, and provide the basis for the decision. DMS consists of vehicle information collection, ground data center and query terminal. A variety of complex data analysis beyond our current ability range to use DMS. How to extract the useful knowledge from DMS's massive data effectively is an urgent need for the current situation. Because these excavated knowledges can help the maintenance department to achieve rapid and accurate fault diagnosis, and provide reasonable preventive maintenance decision support.

For big data, the current distributed database technology NoSQL and Hadoop can handle it well. If the semantic technology is used to analyze, the main challenge lies in the distributed query and reasoning. Such as streaming media, sensors, the current distributed computing technology can also meet this demand. The challenge of semantic technology is the flow reasoning, that is in the process of continuous arrival of the data and the incomplete information should be used for reasoning. Such as distributed reasoning can use Hadoop / Storm and the flow reasoning can use continuous(dynamic) queries [2]. There are many types of data in the DMS system, and the structure is very complex and the speed of change is very fast. It not only needs high performance but also needs management and real-time analysis. If it should be needed to meet the above three, there is no good solution.

In view of the above problems, this paper proposes a distributed real-time context-aware data processing method based on the fuzzy linked data and uncertain reasoning. This method extends the RDF model to a fuzzy RDF model, and propose real-time distributed context reasoning based D-S theory.

2 CTCS-3 data context modeling

2.1 CTCS-3 data

The system structure of CTCS-3 is shown in the following Figure 1. The real-time data of CTCS-3 can be summarized as follows:

- (1) RBC. Generating the traffic permit according to the track circuit, interlocking and other information; Accepting vehicle equipment data through GSM-R.
- (2) GSM-R network. Two-way communication between vehicle equipment and ground equipment.
- (3) Transponder. Sending positioning, level conversion information, line parameters and temporary speed-limited to the vehicle equipment.
- (4) Secure computer. Generating the dynamic speed curve and monitoring the safe operation of the train.
- (5) Track circuit. Checking the train occupancy and sending traffic permit information.
- (6) CTC system. Realization of centralized control for the station signal equipment.

2.2 Linked data

At present, more and more big data applications began to introduce semantic technology, which makes the description of the data more standardized and rich in machine comprehensible semantics. Rich semantic links make the system more open and inter-operable, and make big data analysis deep into the knowledge level, which requires big data technology can provide a wealth of related functions and simple reasoning ability. Big data technology can effectively solve the distributed environment of the web scale of unstructured information management and utilization issues. And the associated data brings the rich formalized semantics, which is a tool for cross-domain integration and intelligent analysis.

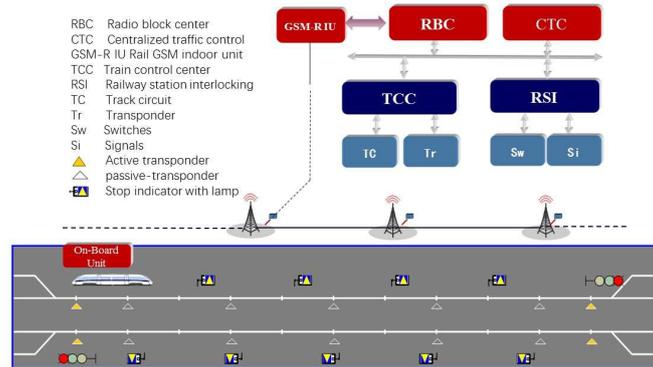


Figure 1: CTCS-3 system structure

Resource description framework(RDF). RDF [13] is a triplet $\langle subject, predicate, object \rangle$, which describes the generic model of the resource description, where the *subject* is a resource with a uniform identifier(URI), or a blank node with no name space, such as DOI, ISBN etc.; *object* could be a resource with a URI, a blank node, or a string value; and a *predicate* that represents the relationship between the *subject* and the *object*.

More and more big data applications are described and encoded using the RDF model. The meta-data and ontology make the data semantically recognizable by the machine, enrich the semantics of big data, and make the data have better interoperability. Since the linked data [7] is semantically described, it is no longer a piece of information. When we combine linked data with big data mining, which provides a powerful tool for big data analysis, which implements semantic-based integration in the data. When using the linked data technology in big data applications, we can call it as a linked big data application.

Linked data. Linked data is a data application form, which uses the URI as a data identifier, and RDF triplet structure as a data model, and based on HTTP, which is a simplified realization of the semantic web, and the intention is to build data web. The linked data is based on the existing web technology (HTTP, URL and HTML), and the web specification is further standardized and defined with basic principles.

Depending on the principle of the linked data, the elements of the triad of the linked data should be encoded as RDF as much as possible. Especially the *subject*, must be able to access in an open HTTP way, and the RDF description should contain other more data link. The data in the linked data is not an independent, context-free abstract data, but a clear knowledge unit with a URI identifier and an RDF description (including cross-domain links), which is the basic semantic unit managed in the web. Any one thing, people, institutions, places, events, concepts can be described as a linked data, so the linked data to big data development is inevitable.

2.3 Fuzzy RDF model

The purpose of context modeling is to describe the data and its environment, which plays an extremely important role in building a context-aware system. At present, the context modeling methods are key-value model, tag configuration model, graph model, object-oriented model, based on the logical model, ontology model. Ontology model has the advantages of knowledge sharing, logical reasoning, easy knowledge reuse, and so on. In this paper, the use of extended RDF modeling method is also a kind of ontology.

This study used Drupal [12] to convert CTCS-3 data into linked data. The RDF data model

is the basis of linked data, but in specific applications, will inevitably use some domain ontology, such as FOAF, Dublin Core, SKOS, OWL, SIOC and so on. Drupal also supports the import of external ontologies, which can define the mapping between these external and local content models. Drupal provides a way for mapping content types, fields, and nodes to classes, attributes, and objects in the RDF triplet model, namely *subject*, *predicate*, *object*. In the field of CTCS-3 data analysis, some concepts are not deterministic, such as speed, security, etc., so the RDF model needs to be extended. The context representation of this paper is based on the fuzzy RDF model, and it is optimized for CTCS-3 data processing.

Fuzzy RDF model(FRDF). The FRDF model FR is expressed as $FR = \langle fs, fp, fo \rangle$, where the fs is the fuzzy concept, and the fp is the set of fuzzy attributes for the subject, the fo is a set of fuzzy concept instances. The attributes fp is defined as $fp = \{p_1, p_2, \dots, p_n\}$, the p_i is an attribute that represents the relationship between the subject and the object; $fo = \{o_1, o_2, \dots, o_n\}$, the o_i represents a fuzzy instance. The function of the fuzzy degree in the FRDF model is $\mu : fo \rightarrow [0, 1]$. If $\mu_{o_i} = 1$, we call o_i concrete instance.

For example, the "the No G115 train" is a fuzzy concept, meanwhile, it is an instance of another concept "train". "the speed of No G115 train is very fast" is an example of CTCS-3 data, which can be described by the FRDF model. The "No G115", "train", "speed", "very", "fast" can total be described by the FRDF model as an instance. The "train" is a fuzzy concept, and it has the Fuzzy attribute "driving speed". The value of speed can be expressed as a group of instances $\{S(\text{slow}), M(\text{middle}), F(\text{fast})\}$.

Fuzzy contexts. According to the FRDF model, Fuzzy context FC is defined as $FC = \langle FS, FP, FO \rangle$. The FS is a set of fuzzy concepts, FP is a fuzzy attribute set that represents the relationship between the subject set FS and the object set FO , and FO is a set of fuzzy objects.

2.4 FRDF model based on D-S theory

The D-S theory [10] is a generalization of the Bayesian reasoning method, which mainly takes advantage of the Bayesian conditional probability in probability theory, and requires a priori probability. The D-S theory does not require a priori probability, can describe uncertainty, is widely used to deal with uncertain data. The D-S theory classifies the information that people are interested in some identification frameworks.

Basic probability assignment(BPA). For any of the recognition frames Θ and its subset A , if the following conditions are met: (1) $m(\phi) = 0$; (2) $\sum_{A \subseteq \Theta} = 1$; Then, $m : 2^\Theta \xrightarrow{\text{yields}} [0, 1]$ is called the BPA function. $m(A) = 0$ is the basic probability assignment of A , which indicates the degree of belief in A . The BPA is also called the mass function.

Belief function. To express the degree of trust in the proposition, D-S theory introduces the concept of belief function Bel , and its relationship with BPA satisfies $Bel : 2^\Theta \xrightarrow{\text{yields}} [0, 1]$, $Bel(A) = \sum_{B \subseteq A} m(B) = 1(\forall A \subseteq \Theta)$.

D-S rule. For n limited mass functions m_1, m_2, \dots, m_n on the one recognition frame Θ , $\forall A \subseteq \Theta$, $A = A_1 \cap A_2 \cap \dots \cap A_n$, then the synthesis of n belief functions is as follows:

$$m(A) = (m_1 \oplus m_2 \oplus \dots \oplus m_n)(A) = \frac{1}{K} \sum_A m_1(A_1) \cdot m_2(A_2) \cdot \dots \cdot m_n(A_n) \quad (1)$$

$$K = \sum_{A \neq \emptyset} m_1(A_1) \cdot m_2(A_2) \cdot \dots \cdot m_n(A_n) \quad (2)$$

The K is called the normalization factor, which reflects the degree of conflict among the evidence. When $K = 0$, it is called incomplete conflict; When $0 < K < 1$, it is called incomplete conflict; When $K = 1$, it is called full conflict.

D-S theory does not well reflect the structure of contextual information and the relationship between them, so its application is limited. To play its superiority, it must use a valid context information representation. The D-S theory uses the evidence recognition framework E and the conclusion recognition framework Θ to divide the information. The FRDF model divides the context into two layers: the context of the low-level acquisition and the context of the high-level reasoning, which can naturally be combined to establish the FRDF model based on the D-S theory. The FRDF model has a very good scalability. The FRDF model can be expanded based on the D-S theory easily.

In the FRDF model based on the D-S theory(FRDFDS), the fs is still the fuzzy concept. The fp is the set of attributes for the subject which been expanded and contains some special evidence attributes and conclusion attributes, such as *hasConclusions*, *hasEvidences*, *hasBPAAvalue* and so on. The fo is a set of instances of concept which also been expanded and contains some special evidence and conclusion instances.

At present, there are five attributes which are added to the basic FRDF model: the attribute *hasConclusions*, the attribute *hasEvidences*, the BPA numeric attribute *hasBPAAvalue*, the *Bal* value attribute *hasBal*, and the time attribute *hasTimetag*. In the context of the FRDFDS model, all data is represented as a unified FRDFDS model, which is the advantage of the linked data, and each data can be considered as a node of the FRDFDS model instance. The *hasTimetag* attribute is used to get the context information for the time. The *hasConclusions* attribute may map a single instance or multiple instance instances. Likewise, the attribute *hasEvidences* corresponds to all evidence instance collections. In the pervasive computing environment, a conclusion context information may be another conclusion of the evidence context information. In order to adapt to the context of the dynamic environment, diversity, uncertainty, we add data attribute *hasBel* and *hasBPAAvalue* for the conclusion and evidence, which are used to describe the probability of the values. The above FRDFDS model is universal, just as the basic RDF model, the relevant attributes and instances can be added to express the uncertainty information of the context.

Belief structure. Given a recognition frame Θ , the evidence space E , the mapping $F : E \rightarrow \Theta$ and the basic belief distribution value m , where Θ and E are finite sets, the quaternion $D = \langle \Theta, E, F, m \rangle$ is called a belief structure. Based on the basic FRDFDS model given in this paper, we can construct the belief structure D , and then create the instance of the uncertainty context information, and finally, use the evidence combination rule step by step reasoning process. The m is the distribution of the basic belief, which is a quantitative evaluation based on evidence. We can make a synthesis based on different sources of evidence through visiting their value of m correspondingly.

3 Multi-layer real-time uncertain context reasoning method

Context reasoning is the key to solve the problem of uncertainty reasoning, and which often depends on the context model. According to the FRDFDS model proposed in this paper, three kinds of reasoning mechanisms can be used in context reasoning [1], namely, ontology reasoning, rule reasoning and evidence theory reasoning. The rule-based reasoning is to generate new knowledge by matching existing facts with predefined rules. For example, when the sensor information is collected: "John is currently in the bedroom, the indoor curtains closed, the light intensity is dark", you can infer that "John is sleeping". The reasoning process of the ontology-based is like the rule-based reasoning process, except that the rules used are defined by the OWL language itself to obtain knowledge that is implicit in explicit definitions and declarations, such as the symmetrical property *SymmetricProperty*, the transitive property *TransitiveProperty*, etc. For the reasoning based on D-S theory, the evidence set and the conclusion set are expressed

by multiple discriminant frameworks respectively. The synthetic method provides the hierarchical reasoning model of evidence. Uncertainty is transferred layer by layer based on the synthesis rules of D-S Theory. The hierarchical model can add the evidence based on the original synthesis, greatly improving the efficiency of evidence synthesis.

CTCS-3 is the complex application of the Internet of things. The logic-based reasoning method will become more complex. To solve the problem of large data, this paper proposes a distributed multi-layer real-time uncertain context reasoning framework based on FRDFDS model. In this framework, the first layer generates the context state according to the FRDFDS model instance, in which the FRDFDS instance is generated by the complex event processing engine. Since the event processing agent may be multiple, the FRDFDS instance is distributed. CTCS-3 data is collected in real-time, so the context state is continually updated with updated data in real-time. The context state of the $(i + 1)^{th}$ layer can be inferred from the context of the i^{th} layer. The reasoning process combines the traditional rule-based reasoning and the uncertainty reasoning based on the D-S theory.

The overall algorithm framework is proposed by this paper in algorithm 1. The algorithm 1 combines traditional rule-based reasoning and the uncertainty reasoning based on the D-S theory. From the algorithm 1, we can see that all context can be expressed by a FRDFDS model instance, and this framework is multi-layer and real-time and has a good adaptability. In algorithm 1, the fuzzy degree is described by the function μ . When all reasoning is finished, the framework will trigger an `update_context()` procedure. The procedure `ds_reasoning()` and `update_context()` will be present in algorithm 2 and algorithm 3.

Algorithm 1 Multi-layer real-time uncertain context reasoning framework

Require: Input Context[layers][instances] = Context[leveal0][instances0]; //layers represent the level of context; instance is a FRDFDS model instance set;

Ensure: Output Context[layers][instances]; // the context at time t;

```

1: initialize Context[layers][instances];
2: initialize context update = false;
3: for (all layers  $\in$  N and layers > 0) do
4:   for (all FRDFDS model instance in Context[layers][instances]) do
5:     if instance is an uncertain then
6:       Call D-S theory reasoning procedure ds_reasoning();
7:     else
8:       Call rule-based reasoning Interface;
9:     end if
10:  end for
11: end for
12: if update= false then
13:   Call update_context(Context[layers][instances]);
14:   update= true;
15: end if

```

The algorithm 2 gives a complete uncertainty reasoning process based on D-S theory. The whole process is based on the belief structure which can map the FRDFDS model perfectly. The classical evidence combination formula requires that all evidence of participation in the synthesis have the same degree of importance. However, in the pervasive computing environment, the evidence obtained from different sources of evidence may differ in importance and belief, so that the basic credibility of the evidence needs to be corrected before the evidence is synthesized to reflect the different importance of the evidence and belief. The formula for correcting the belief

is as follows:

$$m'_i(A_i) = \begin{cases} Bel(E)m_i(A_i) & \text{when } A_i \neq \Theta \\ 1 - \sum m'_i(A_i) & \text{when } A_i = \Theta \end{cases} \quad (3)$$

For different context applications, the *Bel* function is different, but the choice of *Bel* generally follows the following rules: (1) the more important evidence, the higher belief, the corresponding correction factor is great, that is, we assigned a big weight for the important evidence distribution. (2) the higher evidence conflict, the lower belief, the smaller corresponding correction factor, that is, there will be a larger amendment in the future.

Algorithm 2 Uncertainty reasoning based on the D-S theory

Require: Input a context instance in instance set; //an instance in one layer context

Ensure: Output a context instance in instance set; //the context instance after running a reasoning process;

```

1: Procedure ds_reasoning(instance) {
2:   initialize conflict factor threshold value = k;
3:   for (all conclusions in instance) do
4:     // conclusion is also a FRDFDS model instance.
5:     get its evidence set and
6:     calculate the BPA numeric and the Bal value;
7:     calculate the conflict factor;
8:     if conflict_factor ≤ k then
9:       calculate the Bal value for this conclusion;
10:    else
11:      calculate the Modified Bal value for evidence set;
12:      calculate the Bal value for this conclusion;
13:    end if
14:  end for
15: }
```

The weighted combination rule is proposed to fuse the evidence, which solves the shortcomings of the evidence theory in the case of high degree of conflict between the evidence. However, the reasoning model of the method is not adaptive. Therefore, this paper improves the combination rule with the following modified belief formula:

$$m'(A) = (m'_1 \oplus m'_2 \oplus \dots \oplus m'_n)(A) = \frac{\sum_A \omega_1 m'_1(A_1) \cdot \omega_2 m'_2(A_2) \cdots \omega_n m'_n(A_n)}{\sum_{A \neq \emptyset} \omega_1 m'_1(A_1) \cdot \omega_2 m'_2(A_2) \cdots \omega_n m'_n(A_n)} \quad (4)$$

$$\omega_i = \frac{\sum_{j=1, j \neq i}^n (1 - d_{BPA}(m_i, m_j))}{\sum_{i=1}^q \sum_{j=1, j \neq i}^n (1 - d_{BPA}(m_i, m_j))} \quad (5)$$

The $d_{BPA}(m_i, m_j)$ is the distance of m_i , and m_j . The formula (5) reflects the different importance and belief of the evidence which resolves the one-vote veto problem and the robustness problem in the conflict evidence combination. It makes full use of the conflict evidence information and avoids the valid information loss.

In order to improve the speed of analysis, we do not look back at historical data, but only analyze and update the current node data based on rules, D-S theory. This paper uses the transitive property *TransitiveProperty* to update the state attribute information in Drupal module instead of using based a finite state machine method. As shown in Algorithm 3, this paper uses fuzzy similarity to update the fuzzy data on the same data node. In the distributed, multi-layer,

real-time uncertain context updating based on similarity, the former context in the different node is fused by the D-S evidence theory. Based on the similarity of fuzzy sets, the new context is fused to the existing context in the context update processing. The fuzzy similarity calculation rules is as follows.

Fuzzy set similarity. The recognition framework is $\Theta = \{A_1, A_2, \dots, A_n\}$. The fuzzy set M and N are two random fuzzy subsets, and the similarity between them is

$$\psi(M, N) = 1 - \frac{1}{|\Theta|} \sum_i^n |\mu_M(A_i) - \mu_N(A_i)| \quad (6)$$

Algorithm 3 Real-time context update based on similarity

Require: Input Context[layers][instances] at time t and t-1;// the context at time t; layers represent the level of context; instance is a FRDFDS model instance set;

Ensure: Output the updated Context[layers][instances];

```

1: Procedure update_context(Context[layers][instances]) {
2:   for (all all layers  $\in$  N) do
3:     for (all FRDFDS model instance in Context[layers][instances] ) do
4:       initialize similarity threshold value = s;
5:       get the previous t-1 instance p_instance;
6:       calculate the index of similarity between p_instance and instance;
7:       if index  $\leq$  s then
8:         make an additional with previous
9:       end if
10:    end for
11:  end for
12: }
```

4 Experiment and evaluation

4.1 DMS

As shown in Figure 2, to monitor real-timely the operation of CTCS-3 equipment, China high-speed train equipped with DMS which consists of on-board information detecting equipment, ground data center and inquiry terminal. The on-board information detecting equipment collect data from ATP, transponder, track circuit and RBC, and then transmits them to the ground center through GPRS/GSM-R/WLAN network. Through this remote monitoring method, the ground center can monitor and deal with the working states and faults of on-board signal equipment. The business logic of DMS includes data acquisition, storage and analysis, and command. The main function of data acquisition part is data sampling and aggregation. The storage and analysis part mines and analyses the running status data of all related systems comprehensively. It can realize the early warning analysis and fault diagnosis. Based on the storage and analysis, the command part provides data sharing service for daily production and emergency dispatching of the electrical department, and it can improve the efficiency of the electrical department daily work.

4.2 Data analysis system framework

At present, the capability of analysis needs to be further improved in DMS. The work of this paper can be seen as an extension of DMS which is integrated with DMS in the way as shown

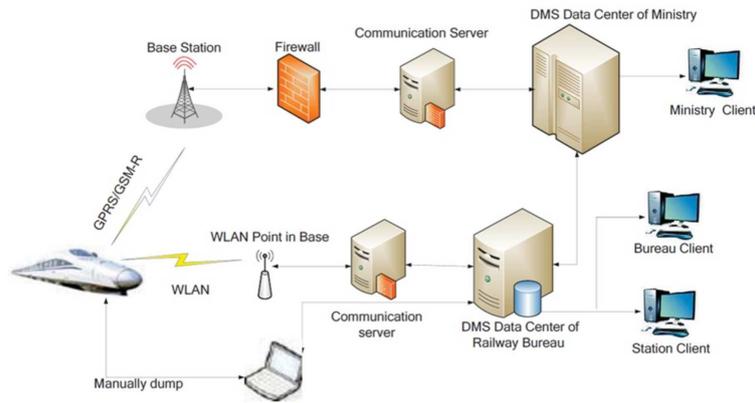


Figure 2: DMS network structure

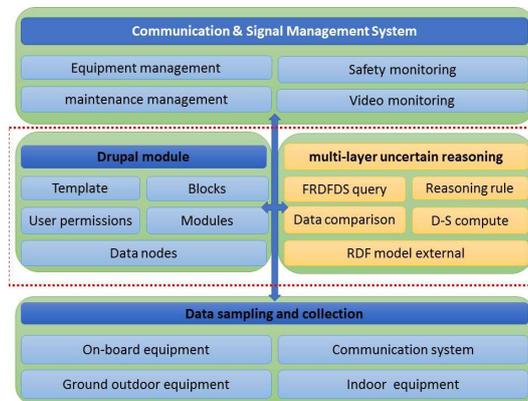


Figure 3: Data analysis integration framework

in Figure 3. In the Figure 3, the middle part of the figure is the work of this paper and which is marked with a dotted line; The top part is the DMS UI, which is a set of functions that can be used in all aspects of railway operations; The base part is the DMS data pool.

It is easy to see from the left of the middle of the Figure 3 that our analysis method is based on Drupal module [12]. The RDF is the standard framework for semantic web and also is a recommended framework for linking data. The linked data technology enables data to be interoperability, reusable, and easier to use. Drupal is the first mainstream content management system to support semantic web technology in its core, which can publish linked data by exposing content with RDF. Our main work in this paper is layout on the right of the middle of the figure. We extend the RDF to a FRDFDS model based on the D-S theory firstly. All data in the DMS data pool will be expressed as a FRDFDS data. We also customize the data reasoning rule on the nodes by our reasoning rule module. Accordingly, the query plug-in also was expanded and programmatically filters and displays content. We implemented the algorithms presented in this paper as two functional plug-ins deployed to Drupal modules.

4.3 Experiment analysis

As shown in Figure 2, there are two types of DMS, one is deployed in the railway bureau and the other in the Ministry of Railways. Our experiment was deployed DMS data center in Beijing communication and signaling section of Beijing railway bureau. We selected Beijing-Tianjin inter-city train control system data as the research object. Specifically, we selected two

work area's real-time data from 6 am to 6 pm in one day as experimental data. We collect the performance of the algorithms every half hour and analyze it. Our experiment is detailed, we try to find the various problems of our algorithm. The experiments were divided into 3 groups simultaneously on 3 IBM x3650 M4 servers respectively. The first group analyzes the data of a work area. Its purpose is to analyze the influence of the number of layers on the performance of the reasoning algorithms. The second group analyzes the data of two work areas. Its purpose is to analyze the effect of the increase of the number of nodes on the performance of the algorithm. The third group also analyzes one work area data, but its analysis algorithms were replaced by proposed in paper [10]. We have implemented the method in paper [10] and compared it with our reasoning method.

These three IBM x3650 M4 servers are connected with DMS data center through a high-speed network, each with 2 Intel Xeon E5-2600 processors and 4GB memory, and the operating system of them are ubuntu 14.04. The Drupal 8.0 with our algorithm was deployed in every server respectively which can automatically generate a variety of FRDFDS nodes.

The first group experiment tested the distributed multi-level real-time uncertain context reasoning method. The results are shown in Figure 4(Bar graph). According to Figure 4, we recorded the time of the four-tier query respectively, where the first layer is Drupal based data query time, it reflects the performance of Drupal; The second layer, the third layer, the fourth layer were a higher level of knowledge reasoning inquire. We summed up the 24 knowledge which needs to reason through the algorithm in the preliminaries. This knowledge includes certainty and uncertainty, which are distributed in these three layers. As can be seen from the figure, the results showed that the algorithm performance is not significantly decreased when the amount of data is increased to a certain value from the level0 to level2; The level3 layer calculation has increased significantly. This is because we deployed several uncertain reasoning, for example, safety, flexibility and so on.

The second set of experiments analyzes the data of two work areas, which doubled the amount of experimental data and focus on the efficiency of the algorithm in a distributed data environment. As shown in the Figure 4(Line graph), we compare the results of the first set of experiments with the experimental results of the second group (dense data and sparse data), when the number of data increases, the performance of the methods is decreased, because the increase of the number leads to more complexity of the state rule and uncertain calculation. The column chart in the figure represents the one work area sparse data, and the line chart represents the two areas dense data. But our algorithm does not increase much with the increase of experimental data, this is because the method is based on the current node, and we will update the current node based on the fuzzy similarity at the end of each calculation, as shown in Algorithm 3. This shows that our algorithm has good scalability.

To more intuitive statistics of our algorithm efficiency, we counted the average process time of each layer by our algorithms with the increase in the number of layers in different types of data. As shown in the following Figure 5, when the algorithm handles the data(data1) of one work area, the average time of level 0 is 34.0 ms; the average time of level1 is 46.9 ms; the average time of level2 is 57.1 ms; the average time of level3 is 98.3 ms. When the algorithm handles the data of two work areas(data2), the average time of level 0 is 47.5 ms; the average time of level1 is 65.0 ms; the average time of level2 is 86.2 ms; the average time of level3 is 110.3 ms. With the increase in data, level3 processing time increased by 12 ms. This also shows that it has a better performance, because our method uses a similarity-based method, avoiding the calculation of many context nodes of the top level.

We have implemented Jusselme's method [10] and compared it with the reasoning method in this paper. The purpose of experiment 3 is to evaluate the correctness of the proposed method and Jusselme's in 24 point-in-time, and the results are shown in Figure 6. In the initial period,

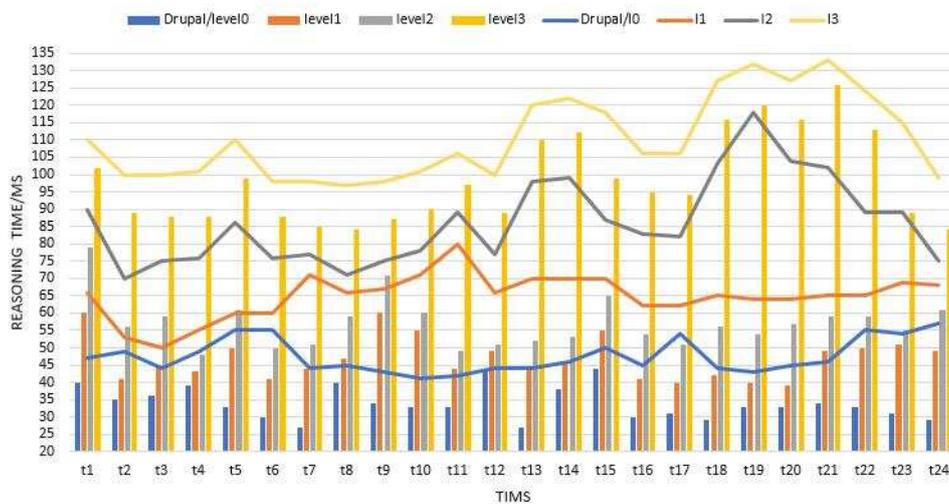


Figure 4: Performance comparison on different data sets

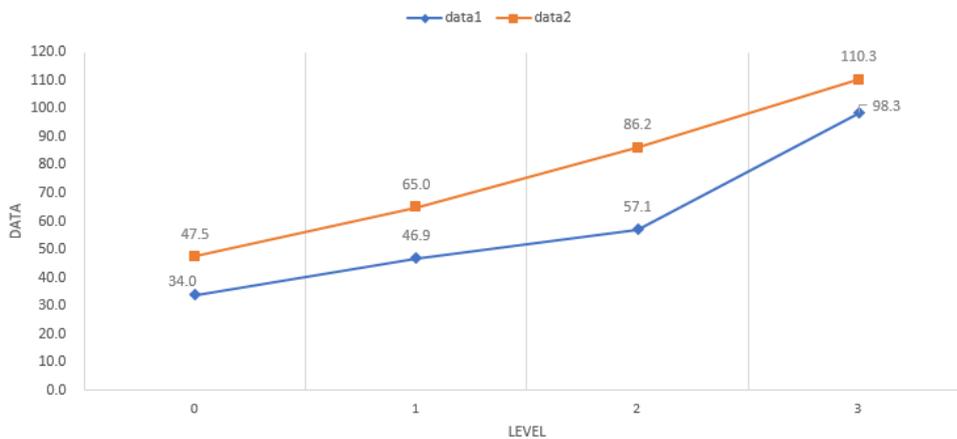


Figure 5: Performance comparison on different data density

the correctness of this algorithm is slightly lower than the paper 12 method; at time point 9, the correctness of this method begins to be higher than the paper 12 method; then, the correctness of the method is improved, and at time point 15, the correctness of this method reaches a steady state.

Due to the existence of conflicting evidence, the use of classical evidence theory for reasoning may result in incorrect results, and improved combinatorial rules can improve this situation. The paper [10] method proposed a weighted combination of rules for evidence fusion, to solve the shortcomings of high conflict between the evidence, but the lack of self-adaptability of the method. On the basis of this idea, this paper first gives the belief distribution table by the expert and then adjusts it by human intervention in the concrete calculation. If there is no human intervention, the original credit distribution table is calculated. In the first eight time points, we have a corresponding fine-tuning of the credit rating table and achieved good results.

In general, as can be seen from the above experiments, the research method in this paper is effective in dealing with the distributed context sensitive complex event of the internet of things. What's more, when in large-scale networking applications, it has better performance than the

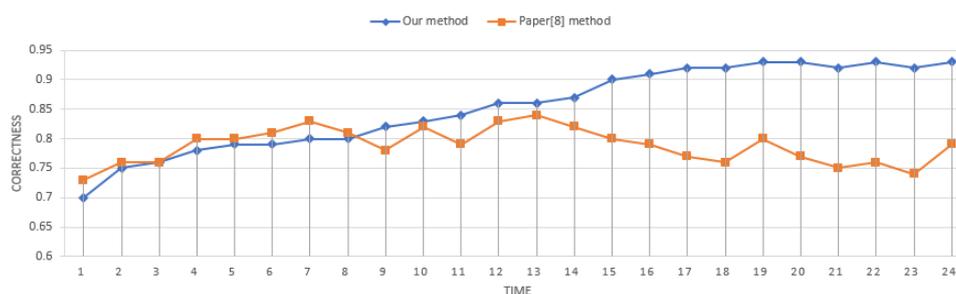


Figure 6: Evaluate the correctness of the proposed method and method in paper [10]

general method and scalability. In addition, this method has high correctness in combination with expert dynamic intervention, and the experimental results show that it has attractive usability in the field of dynamic control.

5 Related research

5.1 Railway intelligent transportation system

The railway management department has been using the "failure - repair" work mode. However, efficient railway operations require maintenance and repair equipment timely, so as to avoid the occurrence of a failure, by analyzing the evolution of equipment status [25]. The distributed inspection and monitoring system continues to collect infrastructure status data and train operating status data, which generate relevant information through conversion. The evolutionary trend of this information can be analyzed to gain knowledge of the evolution of the infrastructure state, which provides decision support for preventive maintenance of the infrastructure.

In recent years, with the development of railway information construction and intelligent transportation management, China's railway transport system [18] has gradually built a number of application-oriented management information systems. At present, these information management systems deal with different aspects of vehicle, infrastructure operation and maintenance management, which have independent organizational structure and produce different forms of data. If we want to use the data of each information system to provide support for the upper application, you must first solve the problem of data sharing. Based on the above, we can realize the intelligent processing of data, using data fusion analysis, expert system, data mining, knowledge reasoning and other technologies, in the field of warning, traffic control, integrated scheduling, resource management, operation management, fault analysis. This is one railway intelligent transportation system(RITS), which is proposed according to the actual situation of China's railway and makes each information system work together. In order to realize the RITS, many scholars have carried out many researches, such as meta-data sharing based on meta-data [5], general data mode based on XML [15], IoT [24], cloud computing technology [8], knowledge reasoning [4] and so on.

In summary, the scholars have begun to study the use of cloud computing and distributed systems to collect and analysis related information for rail transport, infrastructure management, maintenance departments [21] . However, the development of these work is mainly focused on transport planning and operation and management. In the communication, signal system, there is no relevant research and applications are reported.

The train control system and the GSM-R network used in the high-speed train are large and complex systems. The application of new technologies makes the workload and the difficulty of

work greatly increased in the infrastructure maintenance department. A variety of complex data analysis process greatly exceeded our current maintenance capacity. In addition, our maintenance personnel do not have enough time and energy to analyze the data, which leads to a lot of data idle, and has not been fully utilized. How to effectively analyze the massive data produced by various detection and monitoring devices, and get useful rules and knowledge that is an urgent task for high-speed rail. It helps the maintenance department to achieve rapid and accurate fault diagnosis and provide reasonable support for preventive maintenance and maintenance decisions.

5.2 Context sensitive event processing

The context model plays an important role in the development and application of the data analysis system in heterogeneous environments. There are various context representation models were presented [19], including key-value model, object oriented model and ontology-based model. The ontology is the best model of event context representation, but the traditional ontology cannot handle uncertain knowledge which limits its application in uncertain event processing, so in recent years there have been some fuzzy ontology model and reasoning research: The logic based fuzzy model [6] attempts to integrate fuzzy logic into ontology design structure [16]; The distributed fuzzy reasoning Petri net and fuzzy ontology [22] were used for distributed fuzzy reasoning extensively [14].

The challenge of the context sensitive system is mainly to make the right decision for the user's context in real-time. The processing context data in an intelligent way is called context reasoning. In recent years, there has been some research work on the processing of context sensitive events. Zhou et al. proposed a similarity based on context reasoning, which defines the similarity between the context models [26]. This method does not require the initial context information, thus reduces the complexity of the reasoning process. Helmer et al. described the framework of context event processing, and summarized the current context support the commonly used event processing systems [9]. Ashish et al. proposed a context sensitive and complex event processing method based on ontology [23]. However, the most of current articles discuss the idea and framework of context sensitive complex event processing, which lack the details of processing algorithms.

6 Conclusion

China High-speed train control system is a combination of computer, communication and control. Its data is diverse, including sensor generated data stream, GPS signal, GSM-R transmission data, real-time video monitoring data, train control software data, etc. This paper presents an efficient analysis method based on the fuzzy linked data and uncertain reasoning for high-speed train control system big data. We have used the method proposed in this paper to analyze the data of the high-speed train control system. The experiment results show that the method proposed in this paper has good efficiency and scalability for the analysis of large data with different structures, types and context sensitive from train control system. The work of this paper is based on the real practical application. In the future, there is still a lot of work to be done, such as the adaptive distribution of belief value, RDF representation of expert knowledge, the architecture of reasoning system based on Drupal.

Acknowledgment

This work is sponsored by the National Natural Science Foundation of China under Grant No. 61502029.

Bibliography

- [1] Anicic D. et al. (2011), Etalis: Reasoning in Event-Based Distributed Systems, Volume 347 of the series *Studies in Computational Intelligence*, Springer, 99-124, 2011.
- [2] Aniello L., Baldoni R., Querzoni L. (2013), Adaptive online scheduling in storm, *Proceedings of the 7th ACM international conference on Distributed event-based systems*, ACM, 207-218, 2013.
- [3] Dierkx K. (2009), *The Smarter Railroad: An Opportunity for the Railroad Industry*, IBM Institute for Business Value, 2009.
- [4] Feljan A.V. et al. (2017), Framework for Knowledge Management and Automated Reasoning Applied on Intelligent Transport Systems, arXiv preprint arXiv:1701.03000, 2017.
- [5] Gregor D. et al. (2016), A methodology for structured ontology construction applied to intelligent transportation systems, *Computer Standards & Interfaces*, 47, 108-119, 2016.
- [6] Gu L. et al. (2014), Trust Model in Cloud Computing Environment Based on Fuzzy Theory, *International Journal of Computers Communications & Control*, 9(5), 570-583, 2014.
- [7] Hartig O., Bizer C., Freytag J.C. (2009), Executing SPARQL queries over the web of linked data, *The Semantic Web-ISWC*, 293-309, 2009.
- [8] He S., Song R., Chaudhry S.S. (2014), Service-oriented intelligent group decision support system: application in transportation management, *Information systems frontiers*, 16(5), 939-951, 2014.
- [9] Helmer S., Poulouvassilis A., Xhafa F. (2011), Introduction to Reasoning in Event-Based Distributed Systems, *Reasoning in Event-Based Distributed Systems*, Springer Berlin Heidelberg, Vol 347, 1-10, 2011.
- [10] Jusselme A.L., Grenier D., Bosse E. (2001)O, A new distance between two bodies of evidence, *Information fusion*, 2(2): 91-101, 2001.
- [11] Kondepudi S., Baekelmans J. (2012), *Service Delivery Platform: The Foundation of Smart+ Connected Communities*, Cisco Smart+ Connected Communities Institute, 2012.
- [12] Kumar N. et al. (2016), *Drupal 8 Development: Beginner's Guide*, Packt Publishing Ltd, 2016.
- [13] Lassila O., Swick R R. (1999), Resource description framework (RDF) model and syntax specification, *W3C Recommendation*, 22 February 1999.
- [14] Liu H.C. et al. (2017), Fuzzy Petri nets for knowledge representation and reasoning: A literature review, *Engineering Applications of Artificial Intelligence*, 60, 45-56, 2017.
- [15] Medjoudj M., Yim P. (2007), Extraction of critical scenarios in a railway level crossing control system, *International Journal of Computers Communications & Control*, 2(3): 252-268, 2007.
- [16] Nadaban S. (2015), Fuzzy continuous mappings in fuzzy normed linear spaces, *International Journal of Computers Communications & Control*, 10(6), 74-82, 2015.

- [17] Ning B., Tang T., Qiu K., Gao C., Wang Q. (2004), CTCS-Chinese Train Control System, *Computers in Railways*, WIT Press, 393-399, 2004.
- [18] Ning B. et al. (2006), Intelligent railway systems in China, *IEEE Intelligent Systems*, 21(5), 80-83, 2016.
- [19] Perera C. et al. (2014), Context aware computing for the internet of things: A survey, *IEEE Communications Surveys & Tutorials*, 16(1): 414-454, 2014.
- [20] Roop S.S., Ruback L.G. (2001), Intelligent rail crossing control system and train tracking system, *U.S. Patent*, 6, 179-252, 2001.
- [21] Tan X., Ai B. (2011), The issues of cloud computing security in high-speed railway, *Electronic and Mechanical Engineering and Information Technology (EMEIT), 2011 International Conference on. IEEE*, 8, 4358-4363, 2011.
- [22] Rehman Z., Kifor C.V. (2016), An Ontology to Support Semantic Management of FMEA Knowledge, *International Journal of Computers Communications & Control*, 11(4), 507-521, 2016.
- [23] Taylor K., Leidinger L.(2011), Ontology-driven complex event processing in heterogeneous sensor networks, *ESWC'11 Proceedings of the 8th extended semantic web conference on The semantic web: research and applications* , Part II, 285-299, 2011.
- [24] Zhang N. et al. Optimization scheme of forming linear WSN for safety monitoring in railway transportation, *International Journal of Computers Communications & Control*, 9(6), 800-810, 2014.
- [25] Zheng W., Hu N. (2015), Automated test sequence optimization based on the maze algorithm and ant colony algorithm, *International Journal of Computers Communications & Control*, 10(4): 593-606, 2015.
- [26] Zhou H., Wang Y., Cao K. (2013), Fuzzy DS theory based fuzzy ontology context modeling and similarity based reasoning, *Computational Intelligence and Security (CIS), 2013 9th International Conference on, IEEE*, 707-711, 2013.

Author index

Andonie R., 475
Araújo A., 507

Barría C., 461
Braz Jr. G., 507

Caçaron A., 475
Cabrera D., 461
Cheng Q.F., 492
Chueh Y., 475
Cordero D., 461
Cubillos C., 461

Dash S.S., 533
Dzitac I., 449
Dzitac S., 519

Esteve M., 550

Feng L., 492

Gattass M., 507
Gomes Jr. D., 507
Gong Y.Z., 562
Guo S.C., 492

Jin D.H., 562
Jing H.C., 562
Jing K., 492

Kuantama E., 519

Li Q.Y., 562

Manoharan N., 533
Manolescu M.J., 449

Pérez F.J., 550
Paiva A., 507
Palau C., 550
Palma M., 461
Panda S., 533

Rajesh K.S., 533
Reis P., 507

Silva A., 507
Song X.L., 562

Tarca R., 519

Vesselenyi T., 519

Wang P.Z., 492
Wang T.T., 492

Zambrano M., 550
Zhang D., 577
Zhang D.Y., 492
Zhao Z.F., 492