Advances in Intelligent Systems and Computing 633

Valentina Emilia Balas
Lakhmi C. Jain
Marius Mircea Balas *Editors*

Soft Computing Applications

Proceedings of the 7th International Workshop Soft Computing Applications (SOFA 2016), Volume 1

Springer

Advances in Intelligent Systems and Computing

Volume 633

Series editor

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

About this Series

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

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

Advisory Board

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

Valentina Emilia Balas · Lakhmi C. Jain
Marius Mircea Balas
Editors

Soft Computing Applications

Proceedings of the 7th International Workshop Soft Computing Applications (SOFA 2016), Volume 1

Springer

Editors
Valentina Emilia Balas
Department of Automatics and Applied Informatics, Faculty of Engineering
Aurel Vlaicu University of Arad
Arad
Romania

Marius Mircea Balas
Department of Automatics and Applied Informatics, Faculty of Engineering
Aurel Vlaicu University of Arad
Arad
Romania

Lakhmi C. Jain
University of Canberra
Canberra, ACT
Australia

ISSN 2194-5357 ISSN 2194-5365 (electronic)
Advances in Intelligent Systems and Computing
ISBN 978-3-319-62520-1 ISBN 978-3-319-62521-8 (eBook)
DOI 10.1007/978-3-319-62521-8

Library of Congress Control Number: 2017952376

Printed on acid-free paper

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

Preface

These two volumes constitute the Proceedings of the 7th International Workshop on Soft Computing Applications, or SOFA 2016, which will be held on August 24–26, 2016, in Arad, Romania. This edition was organized by Aurel Vlaicu University of Arad, Romania, University of Belgrade, Serbia, in conjunction with Institute of Computer Science, Iasi Branch of the Romanian Academy, IEEE Romanian Section, Romanian Society of Control Engineering and Technical Informatics (SRAIT)—Arad Section, General Association of Engineers in Romania—Arad Section and BTM Resources Arad.

Soft computing concept was introduced by Lotfi Zadeh in 1991 and serves to highlight the emergence of computing methodologies in which the accent is on exploiting the tolerance for imprecision and uncertainty to achieve tractability, robustness, and low solution cost. Soft computing facilitates the use of fuzzy logic, neurocomputing, evolutionary computing and probabilistic computing in combination, leading to the concept of hybrid intelligent systems.

Combining of such intelligent systems' tools and a large number of applications can show the great potential of soft computing in all domains.

The volumes cover a broad spectrum of soft computing techniques, theoretical and practical applications find solutions for industrial world, economic, and medical problems.

The conference papers included in these proceedings, published post conference, were grouped into the following area of research:

- Methods and Applications in Electrical Engineering
- Knowledge-Based Technologies for Web Applications, Cloud Computing, Security Algorithms and Computer Networks
- Biomedical Applications
- Image, text and signal processing
- Machine Learning and Applications
- Business Process Management
- Fuzzy Applications, Theory and Fuzzy and Control
- Computational Intelligence in Education

- Soft Computing & Fuzzy Logic in Biometrics (SCFLB)
- Soft Computing Algorithms Applied in Economy, Industry and Communication Technology
- Modelling and Applications in Textiles
- Methods and Applications in Electrical Engineering

In SOFA 2016, we had six eminent keynote speakers: Professor Michio Sugeno (Japan), Professor Anna Esposito (Italy), Professor Mika Sato-Ilic (Japan), Professor Valeriu Beiu (Romania), Professor Salil Bose (USA), Professor Rajeeb Dey (India) and an interesting roundtable of Professor József Dombi (Hungary). Their summaries talks are included in this book.

We especially thank the Honorary Chair of SOFA 2016 Prof. Lotfi A. Zadeh who encouraged and motivated us, like to all the other SOFA editions.

We are thankful to all the authors that have submitted papers for keeping the quality of the SOFA 2016 conference at high levels. The editors of this book would like to acknowledge all the authors for their contributions and also the reviewers. We have received an invaluable help from the members of the International Program Committee and the chairs responsible for different aspects of the workshop. We appreciate also the role of special sessions organizers. Thanks to all of them we had been able to collect many papers of interesting topics, and during the workshop, we had very interesting presentations and stimulating discussions.

For their help with organizational issues of all SOFA editions we express our thanks to TRIVENT Company, Mónika Jetzin and Teodora Artimon for having customized the software Conference Manager, registration of conference participants and all local arrangements.

Our special thanks go to Janus Kacprzyk (Editor in Chief, Springer, Advances in Intelligent Systems and Computing Series) for the opportunity to organize this guest edited volume.

We are grateful to Springer, especially to Dr. Thomas Ditzinger (Senior Editor, Applied Sciences & Engineering Springer-Verlag) for the excellent collaboration, patience, and help during the evolvement of this volume.

We hope that the volumes will provide useful information to professors, researchers, and graduated students in the area of soft computing techniques and applications, and all will find this collection of papers inspiring, informative, and useful. We also hope to see you at a future SOFA event.

Valentina Emilia Balas
Lakhmi C. Jain
Marius Mircea Balas

Keynote Presentations

Introduction to Choquet Calculus

Michio Sugeno
Tokyo Institute of Technology, Japan

Abstract. In this talk, we give a brief introduction to a recent topic "Choquet calculus" where calculus consists of integrals and derivatives.

Modern integrals, namely Lebesgue integrals initiated by Lebesgue in 1902, are associated with the concept of "measures." Lebesgue measures are defined as additive set functions with certain conditions, and hence, Lebesgue integrals hold additivity by inheriting the property of measures. In the latter half of the twentieth century, a new concept of "measures without additivity" named fuzzy measures was proposed by Sugeno in 1974. Fuzzy measures (non-additive measures in general) are defined as monotone set functions and considered as a natural extension of Lebesgue measures leading to the concept of non-additive integrals with respect to non-additive measures, where we note that monotonicity contains additivity in it.

In 1953, Choquet studied the so-called Choquet functionals based on capacities, where capacities representing "potential energy" of a set were also monotone set functions but not captured as "measures" as in the sense of Lebesgue. Together with the appearance of fuzzy measures, Choquet functionals were finally re-formulated as non-additive integrals with respect to fuzzy measures by Höle in 1982. Since then, various non-additive integrals with respect to non-additive measures have been suggested. Among them, we focus on Choquet integrals which are the most general extension of Lebesgue integrals.

Once we obtain the concept of integrals, we become curious about their inverse operations. In the case of Lebesgue integrals, Radon and Nikodym gave Radon-Nikodym derivatives as inverse operations in 1913 and 1930, respectively. It is well-known that with the aid of Radon-Nikodym derivatives, Kolmogorov proved the existence of conditional probabilities in 1933 and thus initiated modern probability theory, where probabilities are nothing but Lebesgue measures. On the other hand in fuzzy measure theory, conditional fuzzy measures have been not well defined in the sense of Kolmogorov. Very recently, inverse operations of Choquet integrals were studied as "derivatives with respect to fuzzy measures" (Sugeno 2013).

In this talk, we deal with Choquet calculus (Sugeno 2015) based on Choquet integrals and derivatives. So far, most studies on Choquet integrals have been devoted to a discrete case. In Choquet calculus, we deal with continuous Choquet integrals and derivatives as well. First, we show how to calculate continuous Choquet integrals. To this aim, we consider distorted Lebesgue measures as a special class of fuzzy measures, and nonnegative and non-decreasing functions. The Laplace transformation is used as a basic tool for calculations. Distorted Lebesgue measures are obtained by monotone transformations of Lebesgue measures according to the idea of distorted probabilities suggested by Edwards in 1953. We remember that Kahneman was awarded Nobel Prize in Economics in 2002, where his cumulative prospect theory is based on "Choquet integrals with respect to distorted probabilities." Next, we define

derivatives of functions with respect to distorted Lebesgue measures, where the derivatives correspond to Radon-Nikodym derivatives in the case of Lebesgue integrals. We also discuss the identification of distorted Lebesgue measures which is a problem arising particularly in Choquet calculus. Then, we show some relations between Choquet calculus and fractional calculus which is recently getting very popular, for example, in control theory. Also, we consider differential equations with respect to distorted Lebesgue measures and give their solutions. Lastly, we present the concept of conditional distorted Lebesgue measures defined with the aid of Radon-Nikodym-like derivatives.

Biography

Michio Sugeno was born in Yokohama, Japan, in 1940. After graduating from the Department of Physics, the University of Tokyo, he worked at Mitsubishi Atomic Power Industry. Then, he served the Tokyo Institute of Technology as research associate, associate professor, and professor from 1965 to 2000. After retiring from the Tokyo Institute of Technology, he worked as Laboratory Head at the Brain Science Institute, RIKEN, from 2000 to 2005, and Distinguished Visiting Professor at Doshisha University from 2005 to 2010 and then Emeritus Researcher at the European Centre for Soft Computing, Spain, from 2010 to 2015. He is Emeritus Professor at the Tokyo Institute of Technology.

He was President of the Japan Society for Fuzzy Theory and Systems from 1991 to 1993, and also President of the International Fuzzy Systems Association from 1997 to 1999. He is the first recipient of the IEEE Pioneer Award in Fuzzy Systems with Zadeh in 2000. He also received the 2010 IEEE Frank Rosenblatt Award and Kampét de Feriét Award in 2012.

His research interests are Choquet calculus, fuzzy measure theory, nonlinear control, and preference theory, applications of systemic functional linguistics and language functions of the brain.

Emotional Facial Expressions: Communicative Displays or Psychological Universals?

Anna Esposito
Department of Psychology, and IIASS,
Seconda Università di Napoli, Italy
iiass.annaesp@tin.it;
anna.esposito@unina2.it
http://www.iiassvietri.it/anna.html

Abstract. Emotional feelings permeate our everyday experience, consciously or unconsciously, driving our daily activities and constraining our perception, actions, and reactions.

During daily interactions, our ability to decode emotional expressions plays a vital role in creating social linkages, producing cultural exchanges, influencing relationships, and communicating meanings.

Emotional information is transmitted through verbal (the semantic content of a message) and nonverbal (facial, vocal, gestural expressions, and more) communicative tools and relations and exchanges are highly affected by the way this information is coded/decoded by/from the addresser/addressee.

The accuracy above the chance in decoding facial emotional expressions suggested they can play the role of psychological universals. However, this idea is debated by data suggesting that they play the role of social messages dependent upon context and personal motives.

These open questions are discussed in this talk, at the light of experimental data obtained from several experiments aimed to assess the role of context on the decoding of emotional facial expressions. The reported data support more the idea that facial expressions of emotions are learned to efficiently and effectively express intentions and negotiate relations, even though particular emotional aspects show similarities across cultural boundaries.

Research devoted to the understanding of the perceptual and cognitive processes involved in the decoding of emotional states during interactional exchanges is particularly relevant in the field of Human-Human, Human-Computer Interaction and Robotics, for building and hardenind human relationships, and developing friendly, emotionally, and socially believable assistive technologies.

Biography

Anna Esposito received her "Laurea Degree" summa cum laude in Information Technology and Computer Science from the Università di Salerno in 1989 with a thesis on: The Behavior and Learning of a Deterministic Neural Net (published on Complex System, vol 6(6), 507–517, 992). She received her PhD Degree in Applied Mathematics and Computer Science from the Università di Napoli,

"Federico II" in 1995. Her PhD thesis was on: Vowel Height and Consonantal Voicing Effects: Data from Italian (published on Phonetica, vol 59(4), 197–231, 2002) and was developed at Massachusetts Institute of Technology (MIT), Research Laboratory of Electronics (RLE), under the supervision of Professor Kenneth N Stevens.

She has been a Postdoc at the International Institute for Advanced Scientific Studies (IIASS), and Assistant Professor at the Department of Physics at the Università di Salerno (Italy), where she taught courses on cybernetics, neural networks, and speech processing (1996–2000). She had a position as Research Professor (2000–2002) at the Department of Computer Science and Engineering at Wright State University (WSU), Dayton, OH, USA. She is currently associated with WSU as Research Affiliate.

Anna is currently working as an Associate Professor in Computer Science at the Department of Psychology, Seconda Università di Napoli (SUN). Her teaching responsibilities include cognitive and algorithmic issues of multimodal communication, human–machine interaction, cognitive economy, and decision making. She authored 160+ peer-reviewed publications in international journals, books, and conference proceedings. She edited/co-edited 21 books and conference proceedings with Italian, EU, and overseas colleagues.

Anna has been the Italian Management Committee Member of:

- COST 277: Nonlinear Speech Processing, http://www.cost.esf.org/domains_actions/ict/Actions/277 (2001–2005)
- COST MUMIA: Multilingual and Multifaceted Interactive information Access,
- www.cost.esf.org/domains_actions/ict/Actions/IC1002 (2010–2014)
- COST TIMELY: Time in Mental Activity, www.timely-cost.eu (2010–2014)
- She has been the proposer and chair of COST 2102: Cross Modal Analysis of Verbal and Nonverbal Communication, http://www.cost.esf.org/domains_actions/ict/Actions/2102 (2006–2010).

Since 2006, she is a Member of the European Network for the Advancement of Artificial Cognitive Systems, Interaction and Robotics (www.eucognition.org);

She is currently the Italian Management Committee Member of ISCH COST Action IS1406:

Enhancing children's oral language skills across Europe and beyond (http://www.cost.eu/COST_Actions/isch/Actions/IS1406)

Anna's research activities are on the following three principal lines of investigations:

- 1998 to date: Cross-modal analysis of speech, gesture, facial and vocal expressions of emotions. Timing perception in language tasks.
- 1995 to date: Emotional and social believable Human-Computer Interaction (HCI).
- 1989 to date: Neural Networks: learning algorithm, models and applications.

Soft Data Analysis Based on Cluster Scaling

Mika Sato-Ilic
University of Tsukuba, Japan
mika@risk.tsukuba.ac.jp

Abstract. There is a growing need to analyze today's vast and complex societal data; however, conventional data analysis that is dependent on statistical methods cannot deal with the frequently complex data types that make up this data. As early as 2000, the following six challenges were reported as important future challenges in the core area statistical research in the twenty-first century. The six challenges pointed out are (1) scales of data, (2) data reduction and compression, (3) machine learning and neural networks, (4) multivariate analysis for large p, small n (high dimension low sample size data), (5) Bayes and biased estimation, and (6) middle ground between proof and computational experiment.

Soft data analysis which is soft computing-based multivariate analysis is the core area in which to combine conventional statistical methods and machine learning or data mining methods, and has a strong capability to solve the statistical challenges in the twenty-first century. In soft data analysis, we have developed cluster-scaled models which use the obtained cluster as the latent scale for explaining data. While the original scale does not have the capacity to work as the scale for complex data, a scale that is extracted from the data itself will have the capability to deal with the vast and complex data.

This presentation outlines the problems and challenging issues of statistical data analysis caused by the new vast and complex data, how our cluster-scaled models are related with these issues, and how the models solve the problems with some applications.

Biography

Prof. Mika Sato-Ilic currently holds the position of Professor in the Faculty of Engineering, Information and Systems, at the University of Tsukuba, Japan. She is the founding Editor in Chief of the International Journal of Knowledge Engineering and Soft Data Paradigms, Associate Editor of Neurocomputing, Associate Editor of Information Sciences, Regional Editor of International Journal on Intelligent Decision Technologies, and Associate Editor of the International Journal of Innovative Computing, Information and Control Express Letters, as well asserving on the editorial board of several other journals.In addition, she was a Council of the International Association for

Statistical Computing (a Section of the International Statistical Institute), a Senior Member of the IEEE, where she held several positions including the Vice-Chair of the Fuzzy Systems Technical Committee of the IEEE Computational Intelligence Society. In addition, she has served on several IEEE committees including the administration committee, program co-chair, and special sessions co-chair. Her academic output includes four books, 10 chapters, and over 120 journal and conference papers. Her research interests include the development of methods for data mining, multidimensional data analysis, multimode multiway data theory, pattern classification, and computational intelligence techniques for which she has received several academic awards.

Why the Brain Can and the Computer Can't Biology Versus Silicon

Valeriu Beiu
Universitatea "Aurel Vlaicu" din Arad, Romania
valeriu.beiu@uav.ro

Abstract. This presentation aims to follow on von Neumann's prescient *"The Computer and the Brain"* (Yale Univ. Press, 1958) and—relying on the latest discoveries—to explain how the Brain achieves ultra-low power and outstandingly high reliability (nano-)computing, while our silicon-based computers cannot. The tale will be reversed as starting from the brain, and in particular from very fresh experimental results, for information processing and communication.

We shall proceed from the gated ion channels which are the nano-switches of the brain. Understanding the ways, ion channels communicate will allow analyzing the statistical behaviors of arrays of gated ion channels. These will be followed by unexpected results showing that highly reliable communication using arrays of absolutely random devices is possible at amazingly small redundancy factors (<10). For computations, we will make the case for interweaving arrays of ion channels acting as distributed amplifiers, and for adapting the classical Kirchhoff current law to active (*i.e.*, "amplified") electrodiffusion. Afterward, we shall touch upon the ultra-low power/energy consumption, where we will stress the crucial role played by hydrated ions, the very fast ion channels, and the much slower ion pumps, as well as how these behave in extremely crammed spaces. Finally, moving only briefly to the next higher level, the computational power of columnar structures will be explored in the context of large fan-in cyclic circuits.

The information conveyed will expose the reasons why our current silicon-based approaches are falling short of doing what the brain is able to do, and also reveal a stringent need for new and more accurate computational models. As quite a few issues are still being investigated, we will conclude with a call-to-arms for both the computing and the VLSI/nano-communities.

Biography

Valeriu Beiu (S'92–M'95–SM'96) received the MSc in Computer Engineering from the University "Politehnica" Bucharest in 1980, and the PhD *summa cum laude* in Electrical Engineering from the Katholieke Universiteit Leuven in 1994.

Since graduating in 1980, he has been with the Research Institute for Computer Techniques, University "Politehnica" Bucharest, Katholieke Universiteit Leuven, King's College London, Los Alamos National Laboratory, Rose Research, Washington State University, United Arab Emirates University, and currently is with "Aurel Vlaicu" University of Arad. His research interests have constantly been on biological-/neural-inspired circuits and brain-inspired architectures (low-power, highly reliable, massively parallel), being funded at over US$ 40M, and publishing over 250 papers (over 40 invited and more than 10 patents) as well as giving over 190 invited talks, organizing over 100 conferences and working (unfortunately very slowly) on two books: *Emerging Brain-Inspired Nano-Architectures* and *VLSI Complexity of Discrete Neural Networks*.

Dr. Beiu has received five fellowships and seven best paper awards, and is a senior member of the IEEE as well as a member of ACM, INNS, ENNS, and MCFA. He was a member of the SRC-NNI Working Group on Novel Nano-architectures, the IEEE CS Task Force on Nano-architectures, and the IEEE Emerging Technologies Group on Nanoscale Communications, and has been an Associate Editor of the *IEEE Transactions on Neural Networks* (2005–2008) and of the *IEEE Transactions for Very Large Scale Integration Systems* (2011–2015), while being an Associate Editor of the *Nano Communication Networks* (since 2010).

Online Course: A Modern Trend of Teaching and Learning

Salil Bose
Fellow of the Royal Society of Biology, UK
boses1940@gmail.com

Abstract. As academics, we have dual responsibilities—doing research and teaching. We know how difficult, if not impossible, it is to do the one to our fullest satisfaction without compromising with the other. My presentation is focused on teaching—how we can make teaching (more) effective and at the same time make more time and energy available for research. Thanks to modern Internet technology, a new online avenue of teaching and learning shows promising potential to reach these goals. Online pedagogical methodologies are now globally either complementing or replacing in-class teaching.

An **online course** is basically a course that is offered in part or wholly via the Internet. The most important infrastructural component of online course is the LMS (**L**earning **M**anagement **S**ystem) software. The LMS is the online environment which enables the learning to happen. It is the platform, where the **teacher creates a course** and the students can interact with the course content in a dynamic manner. Online courses include means for the students to contact their professor and access most course materials, including online readings, videos, audio files, and other resources. This is also where students go to participate in discussion boards to exchange views with fellow classmates, and the instructor can monitor and comment on their discussion. Depending on the versatility of LMS, students can also e-mail and instantly message their classmates and instructors.

My objective is to describe how an LMS can be used to create a course. Currently, many learning management systems are available with varying capabilities, costs, and conveniences; research is ongoing in developing more with varying features. In this presentation, as a case study, I will describe one LMS, more specifically known as LCMS (learning content management system) with brand name Drona (http://drona.netimaginelearning.com/), which has several unique user-friendly features.

Animation is an elegant component of online courses especially to illustrate and explain complex functions and abstract concepts of a phenomenon. However, creating a "real-life" animation with sufficient details of structure and function is a highly skilled job. In this presentation, I will also show an animation of a biological molecular machine (DNA replication in eukaryotic cells) to demonstrate how the potential of online courses can be harnessed to illustrate highly complex biological phenomena in a dynamic and three-dimensional matter overcoming the barrier of distance.

Biography

Salil Bose obtained his PhD from the University of Rochester, New York, in 1975. His thesis was on energy transduction in photosynthetic processes. After completing postdoctoral training in Brookhaven National Laboratories in Long Island and University in Illinois at Urbana-Campaign, he taught and conducted research on bioenergetics at two universities including Jawaharlal Nehru University, where he served as Professor and Chairman of the Center for Biotechnology. Prof Bose worked for eleven years as a Senior Scientist in National Institutes of Health (NIH) in USA doing research on fast kinetics of proton (H^+) transduction in bacteriorhodopsin photocycle, fast kinetics of electron transfer in cytochrome oxidase, and regulation of oxidative phosphorylation in mitochondria by monitoring electron transfer kinetics and changes in transmembrane electrical potential. Prof Bose has published over 100 research papers.

Professor Bose served Nanyang Technological University (NTU) in Singapore for thirteen years, initially as visiting professor and later as senior education consultant. He was involved in advising PhD students, and developing and teaching interdisciplinary courses at various levels—undergraduate to PhD. He initiated the development of online courses at NTU and conducted research in developing effective online assessment methodology. He also served the NTU School of Computer Science as an adjunct professor to teach systems biology to the master's students in bioinformatics.

For many years, Professor Bose has been working as the Editor in Chief of a Canadian journal, *International Journal of Biology* published by the Canadian Center of Science and Education (www.ccsenet.org) and as a member of the international advisory board of a British journal, *Journal of Biological Education* published by the Society of Biology (www.societyofbiology.org/).

Invited as a visiting scientist/professor, Professor Bose has contributed to research and teaching at many reputed institutions such as Kyungpook National University in South Korea, Washington University at St. Louis USA, Imperial College of Science & Technology London, Carnegie Institution of Washington at Stanford University, University of California at Berkeley, and Michigan State University at East Lansing, USA.

Professor Bose is an elected Fellow of the Royal Society of Biology, UK.

Complex Dynamical System: Analysis and Control via Linear Matrix Inequality (LMI) Techniques

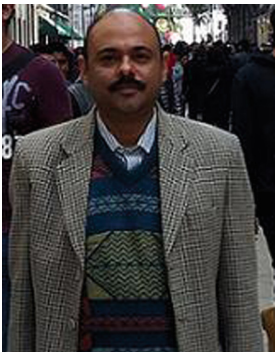
Rajeeb Dey
National Institute of Technology Silchar, India
rajeeb.iitkgp@gmail.com

Abstract. To analyze/predict the behavior of system at a future time is always a topic of interest not only in engineering or physics but also in every subject (like economics, pschycology, finance), and thus, it is a topic of research to systems and control community too. For understanding the behavior of a system, one needs to have a dynamical model, but modeling should be such that it must predict more realistic nature/behavior without many complications in the models too. More complex is the model more complex will be the mathematical framework for carrying out the analysis and further control synthesis of the system under study. Thus, there is always a trade-off between the complexity of the model of the system and framework of analysis.

In this talk, few important features of the systems will be considered to create a model of the complex dynamical system in an attempt to obtain/predict more realistic situation of the response. One such feature is delays in the system dynamics along with consideration of parametric uncertainties, hard and smooth nonlinearities, thereby making the system model more realistic but bit more complexes. The modeling is oriented in such a way that analytical framework remains easier, tractable, and further control development becomes straightforward. Thus, this talk will cover comprehensive analysis and control design for such system using established LMI (linear matrix inequality) approach.

This talk will further focus on solution of certain interesting complex systems from electric power networks, biomedical engineering, fuzzy and neural network dynamical systems as case studies.

Biography



Rajeeb Dey received his B.Tech (Electrical Engineering) from North Eastern Regional Institute of Science & Technology (NERIST), Itanagar, India, in 2001, M.Tech in Control System Engineering from Indian Institute of Technology (I.I.T) Kharagpur, India, in 2007, and PhD in Control Engineering from Jadavpur University, Kolkata, India, in 2012. He has been awarded National Associateship by Department of Bio-Technology, Govt. of India, to work on control aspect of Artificial Pancreas in Indian Institute of Science, Bangalore in 2013. In 2014, he was been

awarded TWAS-CONACYT postdoctorate fellowship by TWAS (The World Academy of Science), Italy, to work on complex dynamical systems at CINVESTAV-IPN, Mexico City. He received AICTE (Govt. of India) funded research project for implementation of output delay feedback control for gantry crane problem in 2009 and in 2010 students working on this project has been awarded best undergraduate project in India by Indian National Academy of Engineers (INAE).

Dr. Dey holds a permanent position as Assistant Professor in Electrical Engineering in National Institute of Technology, Silchar, India. He has around 13 years experience of teaching undergraduate as well as postgraduate students and has guided one PhD and several master's thesis in the area of control system engineering. Dr. Dey has around 30 research publications in reputed international journals and conferences. He is a regular reviewer of over 15 SCI (E) journals related to applied mathematics and control engineering and also Editor of International Journal of Artificial Intelligence Paradigm (Inderscience Publishers).

His research interest includes robust control, control using LMI techniques, time-delay systems, intelligent control, control applications in biomedical engineering. He is a Senior Member of IEEE Control System Society, Member of SIAM, and Life Member of System Society of India. He is currently executive committee member of IEEE CSIM (Control system, Instrumentation and Measurement) Kolkata Chapter, India, and held the position of publication chair for First IEEE CMI 2016.

Theory and Practice of Business Intelligence

Mihaela I. Muntean
West University of Timişoara, Romania
mihaela.muntean@e-uvt.ro

Abstract. Based on a selective literature review and some of the author's recent papers, a unifying theoretical approach of the most relevant business intelligence (BI)-specific concepts will be introduced. Business Intelligence is an umbrella term for various business managing approaches based on well-informed decisions, which lead to a high-performance level within organizations (Brohman, D.K., 2000; McKnigts, W., 2004; Melfert, F., Winter, R., Klesse, M., 2004; Mukles, Z., 2009; Hatch D., Lock M., 2009; Borysowich, C., 2010; Jamaludin, I. A., Mansor, Z., 2011; Mircea M. (ed.), 2012). Eleven definitions describing the actual BI phenomena are subject of the debate. In terms of a value proposition, the From-Data-To-Performance value chain is designed (Muntean, M., 2012), phases like business analysis, enterprise reporting, and performance management represent a series of activities that create and build value. Key forces like cloud, mobile, social and big data enrich the BI framework and contribute to the diversity of the BI phenomena. Based on the fact that most of the data (80% of the data) stored in corporate databases have a spatial component, a business intelligence approach for spatial enablement is recommended to be developed (Muntean, M., Cabau, L., 2012). Despite the importance of the technological dimension of the BI approaches, the business process model is determinant for any initiative (Muntean, M., Muntean C., Cabău, L., 2013). Considerations regarding business intelligence governance, based on the author's expertise, establish a framework for the BI life cycle and ensure consistent project delivery.

Based on these considerations, some practice examples have been introduced, closing the gap between theory and practice. Data warehouse proposals (Muntean, M., 2016), an innovative use of QR codes in BI reporting (Muntean, M., Mircea, G., Băzăvan, S., 2014), and a feasibility analysis for BI initiatives (Muntean, M., Muntean, C., 2013) are subject of the debate.

Biography

Currently, Professor Mihaela I. Muntean is the Chair of the Business Information Systems Department at the West University of Timisoara and an IT independent consultant. With a background in computer science and a Ph.D. obtained both in Technical Science and in Economic Science (Economic Informatics), Professor Mihaela I. Muntean focused her research activity on topics like information technology, knowledge management, business intelligence, business information system. Over 70 papers in indexed reviews and conference proceedings and the involvement with success in eight multiannual national research grants/projects are sustaining her contributions in the research fields mentioned above.

Round Table

The Theory of Fuzzy Sets and Its Real-World Applications

József Dombi
University of Szeged, Hungary
dombi@inf.u-szeged.hu

Abstract. The SOFA workshop focuses on theoretical and applied computational intelligence and informatics. Its aim is to provide a platform for scientists, researchers, and students on soft computing to publish to discuss and publish the results of their research.

The connection of the theoretical results and the practical applications is one of the most important tasks of the workshop. In the roundtable meeting, we will discuss how different operators could be applied in practice, first of all by using the fuzzy control systems. In this area, several problems arise:

- describing the logical expression (using different operators)
- representing the consequent
- aggregating the consequents
- defuzzification
- interpolative solutions, etc.

The fuzzy control can also be represented by neural network. A very interesting question is how to handle this approach. Finally, real practical application shows the effectiveness of the different configurations of fuzzy control. We can also mention that the input of the fuzzy control can be obtained from picture processing procedure or from other resources (e.g., using signal processing).

In summary, the topic of the roundtable discussion covers the theory of fuzzy sets and its real-world applications.

Biography

József Dombi's degrees earned at University of Szeged. Academic degrees: University doctor's degree (1977, Summa cum laude), Candidate mathematical sciences (1994, CSc, Title of dissertation: Fuzzy sets' structure from the aspect of multicriteria decision aid.). Visiting positions: Leningrad (1971, 6 months), Leipzig (1974, 6 months), Bukarest (1978, 1 month), DAAD Scholarship, Aachen (1978, 12 months), Alexander von Humboldt Scholarship, Aachen (1986, 12 months), European Scholarship, Bristol (1987, 3 months), Academic research exchange (Paris, Helsinki, Turku, Tampere), CEEPUS guest professor, Linz (2000 and 2009, 1 month), and Klagenfurt (2008, 1 month).

Awards: 1991 Minister of Education award, 1997 Pro Sciencia award for student, 1998 László Kalmár award, in 1997, DataScope won Information Technology award in Brussels, won the Best Software of the Year award in 2000 at COMDEX, Las Vegas. Editorial membership: editorial board's member of the Foundations of Computing and Decision Sciences magazine until 1997, editorial board's member of the International Journal of Systems Sciences, and editorial board's member of the International Journal of Advanced Intelligence Paradigms. Membership in international organizations: IFSA (International Fuzzy System Association), Member of European COST Action on Multicriteria Decision Making, ESIGMA (European Special Interest Group on Multicriteria Analysis), European Working Group on Multiple Criteria Decision Aid, MTA Public Body (Operational Research), MOT (Hungarian Association of Operational Research), Hungarian Humboldt Association. Business: 1993 founder and head of Cygron Research Ltd., 2000 scientific consultant of Mindmaker Ltd., 2002 founder and head of Dopti research and development Ltd. Research interest: computational intelligence, theory of fuzzy sets, multicriteria decision making, genetic and evolutional algorithms, operation research and visualization. Teaching: information visualization which has an enormously growing interest in Hungary and abroad, the theory of artificial intelligence, fuzzy systems, multicriteria decision making, and intelligent systems.

Contents

Biomedical Applications

Methods and Applications in Electrical Engineering

Digital Sunshade Using Head-up Display

Razvan-Catalin Miclea(✉), Ioan Silea, and Florin Sandru

Automation and Applied Informatics Department,
Politehnica University Timisoara, Timisoara, Romania
miclea_razvan@yahoo.com, ioan.silea@upt.ro,
florin.d.sandru@gmail.com

Abstract. Visibility is the most important requirement for safe driving. Even if we talk about bad weather conditions (fog, rain, snow) or glare phenomenon caused by sun or headlamps all have the same outcome, danger for the public transport. According to National Highway Traffic Administration glare causes hundreds of accidents every year. Due to sun glare drivers are not able to observe traffic signs, pedestrians, sudden curves or other traffic participants. In many cases this can lead to a catastrophe. In this paper is proposed a cost-effective system able to protect drivers against glare. It uses a sun tracking system to detect the point where the light has the maximum intensity on the windshield surface and an eye tracking system to find out the driver's position. Based on these detections and using a head-up display on the windshield is projected a black spot in order to minimize the light power which can disturb the driver.

Keywords: Sunshade · Visibility · Glare · Head-up display

1 Introduction

Any type of opaqueness in the driver's vision can have a negative impact on roadway traffic. These vision issues can be caused by bad weather conditions - foggy, rainy or snowy conditions - but also by sun or headlamps glare.

If we refer to the first category, visibility issues caused by bad weather conditions, there are a few approaches which try to solve the problem, but none of them proved to be very reliable. That's probably why there still doesn't exist such a system installed on commercial vehicles. If we take a look in the literature we can observe that most of the methods which measure visibility distance in bad weather conditions are based on image processing: measuring the distance between the marking lines [3], analyzing the clarity of the traffic signs [4, 5] or estimating the visibility distance based on the line where the road meets the sky [6].

For the second category, visibility issues caused by glare phenomenon, seems to be a lack of solutions, even though glare is one of the most significant factors for accidents according to National Highway Traffic Administration. Glare is a driving annoyance, whether it is caused by the rising or setting sun during day time or by headlamps in night conditions. However the number of accidents caused by glare is under-reported because in most of the cases glare is not the single factor: the temporal blindness while driving blocks the driver in observing pedestrians, sudden curves or traffic signs which

V.E. Balas et al. (eds.), *Soft Computing Applications*, Advances in Intelligent Systems and Computing 633, DOI 10.1007/978-3-319-62521-8_1

can lead to a hazard. For night conditions the glare effect is even more likely to happen due to vehicles headlamps. According to VA Medical Center Cillicothe, Ohio, a middle age driver's eye needs about 10 s to adjust to glare. In this time a car can cover almost half a kilometer.

By far the most known solution against sun glare is the sun visor flap which has to be manually manipulated by driver when sun blinds his vision. Automotive companies started to develop new solutions for this drawback using intelligent glass.

For glare caused by headlamps there is no solution for the moment so the automotive companies started to work on the other side, to improve the vehicle's headlamps. The new headlamp technologies based on LED and Laser partially solved the glaring problem through the auto-dimming technology combined with the light color temperature.

The purpose of this paper is to present a new concept of an anti-glare system using basically system already installed in the vehicle like head-up display, driver tracking systems or light sensors. From our point of view this method can be the best cost-effective method because most of the improvements have to be done in the software part which represents very low cost comparing to the new systems which need an entire new architecture: mechanical, hardware, software.

The digital sunshade system refers especially to the glare issue caused by sun but a future variant can also cover the already mentioned problem, glare caused by headlamps. Our solution for solving the glare shortfall is to realize a projection on the windshield in the point where the light has the maximum intensity. So, first we have to determine that point, where the sun light has the highest intensity, using a light sensor. After that it is mandatory to find out exactly the driver's position in the car, this will be done with an eye tracking system. Based on this two assumptions the system knows if the light is disturbing the driver and if so it takes the decision to project a black spot on the windshield using a head-up display in order to protect the driver's vision against the harmful sun rays.

The rest of the paper is organized as follows. The next section presents some previous approaches regarding the driver's protection against glare phenomenon (Sect. 2). Section 3 contains the structure of the system together with a brief description of each composing element. Subsequently, Sect. 4 describes the functionality of the proposed system. In the last section, Sect. 5, are presented the conclusions and the advantages of our system.

2 Previous Works

Bad weather conditions like fog, sleet, rain and snow are considered by the drivers to be the most dangerous situations for driving. Only few people take in consideration the effect of sun which can be very harmful for driver's vision in certain moments of the day.

Almost a decade has passed since the first anti-glare attachment was installed in a vehicle, the sun visor flap, which still remains the main solution for this issue. The sun visor is realized from a pressboard that is covered with a material assorted with the interior of the vehicle. At the edge it has a mounting bracket which allows the driver to manipulate it in case of a disturbing glare. This solution has a few shortfalls: the driver has to manually manipulate the sun visor flap which means that for some moments his

attention to the road is disturbed and his vision will be blind until he realizes where the light comes from. In [2] Elahi and Rahman proposed a system which automatically adjusts the sun visor in the desired position for the driver. A light sensor is installed on the windshield which sends a signal to the control board in case the light intensity crosses a predefined threshold. The servo motor installed on the sun visor's shaft rotates it so that the eye and the face of the driver to be shaded (Fig. 1).

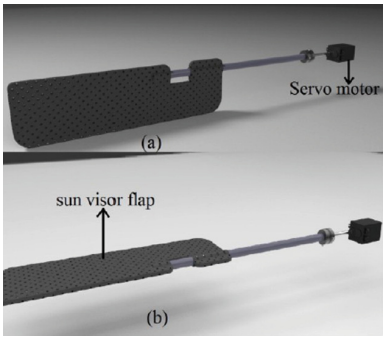


Fig. 1. Automatic sun visor - open (a) and closed (b) flap [2]

One drawback of this method is that the sun visor covers just a part of the windshield and if the light is reflected from the road the visor doesn't help the driver anymore.

A new concept was presented at the beginning of this year by Continental at CES 2016 in Las Vegas[1]. It uses the intelligent glass or smart glass to protect the driver against the glare phenomenon. Continental members say that in the future such events can be detected and the window can darken automatically. The intelligent glass contains a thin film which can be darkened by applying a voltage which has as output an alignment of embedded particles. But the thin film technology is not the single method used in smart glass design: the alternatives are based on liquid crystal polymers or electrochromism. The biggest disadvantage of this solution is the price, which is very high for the moment to be implemented on commercial vehicles.

3 Digital Sunshade - Structure

As it can be observed in Fig. 2, sun glare is not caused only by the direct beams of the sun, the reflected ones can have even a bigger impact on the driver's vision. Light reflected from a wet road crosses the windshield in different points, not only its topside. For such cases the sun visor is not helpful anymore. The driver needs a system which is able to keep track of the glare phenomenon on the entire surface of the windshield.

[1] http://www.continental-corporation.com/www/pressportal_com_en/themes/press_releases/3_automotive_group/interior/press_releases/pr_2015_12_10_intelligent_car_window_en.html.

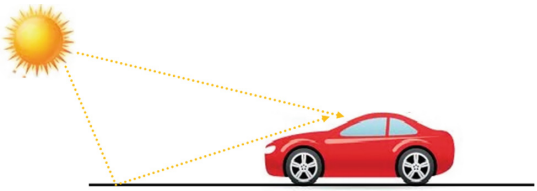


Fig. 2. Sun glare phenomenon

The system presented in this paper is able to cover the entire windshield and performs automatically all these operations, the only job of the driver is now to be careful on the road.

Comparing to the methods presented in the previous section, our proposed method comes with a few advantages: the system covers the entire windshield surface and the price to implement such a system on a car can be very low. Most of the high-end cars are already equipped with systems such as head-up display or intelligent sensors which monitor the driver's position. The goal is to give them additional employments.

A. *Sun Tracking System*

The first step which has to be done is to realize an external detection, in other words to determine where the sun comes from. This can be realized using the GPS system from the car or mobile phone which is the most indicated solution from the cost point of view, but the results are not always reliable. An alternative can be a sensor or a network of sensors installed around the windshield which identify the angle of light incidence from the brightest part of the sky or of the road [1]. An example can be the sun sensor developed by Solar-Mems[2] which measures the incident angle of the sun rays when these go through a small window, in this way indicating the sun position regarding to the sensor position (Fig. 3). This sensor can be very attractive for such an application due to its small size, low weight, high accuracy and low price.

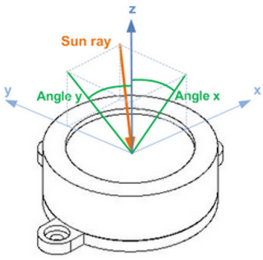


Fig. 3. Sun sensor

[2] http://www.solar-mems.com/en/products/sun-sensor.

From our point of view, the best solution is a sensor installed in a corner of the windshield, for example the closest corner to the driver. This sensor can indicate the sun's position Υ_s and α_s, the windshield being sloped with the angles Υ_p and α_p (Fig. 4). Reporting to the sensor's coordinate system linked with the eye tracking system's coordinates there can be determined the driver's eye position. If a parallel axis with the sun direction starting from the point which indicates driver's eye position crosses the windshield surface, and the light intensity is above a predefined threshold than the HuD has to be activated (Fig. 4).

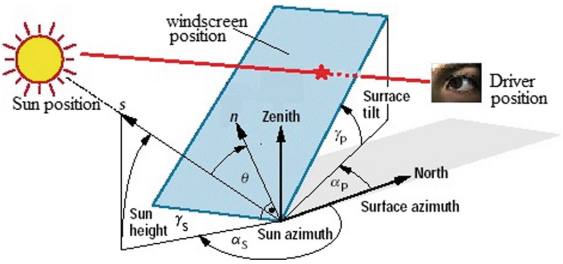


Fig. 4. Sun tracking system

Beside the information regarding the driver's position the light sensor has to indicate the sun rays' brightness and the size of the annoying spot, information which will be used in the projection phase, to adjust the opacity and the dimension of the projected images.

If we are thinking now from the mathematical point of view how the system will look like, we can choose a Cartesian coordinate system with its origin in angle O, where the sensor which gives the sun position is installed (Fig. 5). Ox and Oy are the edges of the windshield while Oz is the normal on the windshield surface. The plan of the windshield can be determined by three points, for example O, M and N of known coordinates. We assume that the driver's head movement (left-right, front-back or up-down) is mathematical described by point $D(x_D, y_D, z_D)$, in the coordinate system already defined, determined by the eye tracking system.

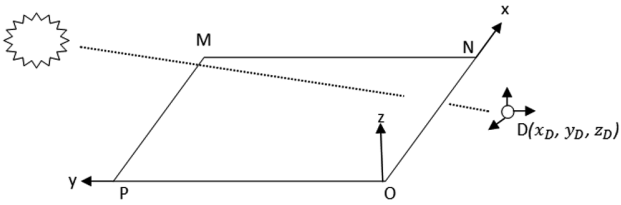


Fig. 5. Mathematical model

Let's consider the line through point D oriented based on the information got from the sun sensor:

$$d : \frac{x - x_D}{l} = \frac{y - y_D}{m} = \frac{z - z_D}{n}$$

where l, m and n are calculated based on the sun rays orientation (the information got from the sun sensor). The plan determined by the windshield surface is:

$$W : Ox + My + Nz + P$$

To determine the intersection between the line and the plan we will consider the next equation system:

$$\begin{cases} x = x_0 + lt \\ y = y_0 + mt \\ z = z_0 + nt \\ Ox + My + Nz + P = 0 \end{cases}, t \in R$$

If $Ol + Mm + Nn \neq 0$ means that the system has an unique solution, in other words the intersection between the line and the plan is a single point.

B. *Eye Tracking System*

After the exterior detection was performed the system has to do an interior detection, to find out the clear position of the driver. The eyes and the face of the driver have to be protected against glare, that's why we decided to use an eye tracking system for this evaluation.

An eye tracking system is usually used to monitor the driver's loss of attention. It starts with a face detection followed by eye detection and the last step is the evaluation of the eye state [7]. If the face is turned side or the eyes are closed for a few seconds the system alerts the driver by flashing some interior light or through some vibrations in the steering wheel [12]. The head's and eyes' movements are monitored with cameras usually equipped with active infrared illumination.

The cost of an eye tracking system is high at this moment. Such systems are not installed on the commercial vehicles only the high-end ones dispose of them. In [8] Hong et al. proposed a low cost eye tracking system which has to be placed on the driver's head. The gaze of the subject is monitored using two cameras and two mirrors installed on the side of the head.

Our system needs only the information regarding the position of the head and eyes in order to protect them from the undesired glare phenomenon. So, if the vehicle is not equipped with an eye tracking system an after-market one can be used to point out the position of the driver, like the one presented above.

C. *Head-up Display*

Head-up Displays (HUD) are becoming more and more popular, especially on high-class vehicles. If the first variants of HUD were able to display only some basic

information like speed and navigation icons, the latest variants are more complex and can exhibit 3D images, the so called augmented reality technology, such as navigation information, lane guidance or potential dangers.

As it was already mentioned, head-up display were basically designed to display useful information in the driver's field of view in order to discard any kind of distraction. A HUD with the characteristics of projecting unlimited combination of images without any fixed dial's position [10] is the desired solution for our system in order to reduce the glare effect on driver's vision. Comparing to most of the existing commercial head-up displays, the one used in our system has to able to deal with 3 constrains:

- to be dynamically adjustable on both axis, x and y
- to dynamically adjust the image opacity based on the sun brightness intensity
- to dynamically adjust the size of the projected image

All these three conditions can be fulfilled with the new generation of HUD, based on DLP technology [11] without any hardware change, only a few software improvements are needed which mean very low costs.

In [9] Wientapper et al. proposed a head-up display variant which has as input information the viewer's head position in order to display the image in the ideal position for the driver. If such a HUD with internal calibration becomes reality on the commercial vehicles, the complexity of our proposed system decreases, being no need of an additional eye tracking system.

In Fig. 6 is illustrated the result of rejecting glare from the image. The system gathers the data from the light sensor and eye tracking system and projects an adjusted image (in opacity and size) on the windshield.

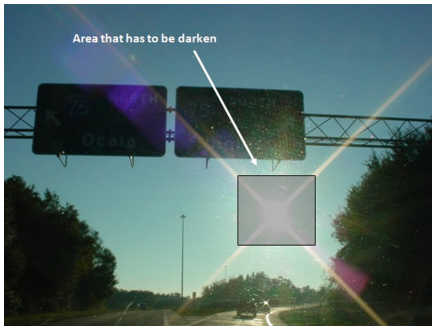


Fig. 6. Rejecting the glare phenomenon using HUD

4 System Deployment - Functionality

In this section will be explained the functionality of the Digital Sunshade based the below flow chart (Fig. 7). Our proposal is to let the driver to decide if he wants to use the head-up display functionality or the digital sunshade functionality because both functionalities can't work in parallel.

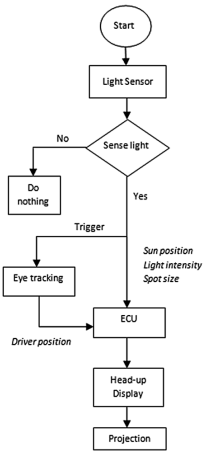


Fig. 7. Digital sunshade's functionality flow chart

When the digital sunshade functionality is chosen, the sun tracking system starts the monitoring. If on the windshield's surface is detected brighter light than the predefined threshold, the system gets the data from the sun sensor regarding the sun position. Beside the information regarding to the sun position, the sun tracking system has to offer also information regarding the light intensity level and the size of the spot which has to be covered.

When the system detects light on the windshield it automatically starts also an interior detection, to find out if the sun rays are disturbing the driver. If they do, all these information are sent to the ECU in order to initiate the projection process.

The head-up display integrated in this system has to be able to project light in every point of the windshield. The ECU configures the image which has to be displayed based on the information gathered from the sun tracking system (position, brightness level and size of the spot) and eye tracking system (driver's position). These data are sent to the head-up display which projects an adapted image (opacity and size) so that the entire disturbing area to be covered (see Fig. 6).

5 Conclusions

Sometimes light and solar energy is useful, such as for agriculture [13], but the car driving sunshine definitely bothers us. In this paper is presented a new concept of an anti-glare system able to protect the driver against the annoying sun rays no matter where these are crossing the windshield. Nowadays the solution for the glare problem is the sun visor, but its biggest drawback is that it can protect only the top area of the windshield. Unfortunately the glare phenomenon is unpredictable; we have to go against the sun rays crossing directly the windshield surface but also with sun rays reflected from a wet or icy road. For the latter category the sun visor is useless.

The automotive companies are working hard to find a solution for this problem; the one proposed until now, with an integrated thin film, is expensive from the manufacturing point of view.

We analyzed the problem together with the state of the art and we concluded that the proposed method can be a cost-efficient variant. The system is able to protect the driver against glare, no matter where the rays are hitting the windshield, and is also reasonable from the cost point of view because we used equipments already installed in the vehicle, we only gave them new functionalities: using the GPS system and some light sensors we detected the sun position, with the eye tracking system we indicated the driver's position and in the last phase with the help of a head-up display we project an image on the windshield in the disturbing area.

Furthermore, using the same principles exhibited above, the system can be improved such as to become useful also for headlamps glaring.

References

1. Viereck, V., Ackerman, J., Li, Q., Jaker, A., Schmid, J., Hillmer, H.: Sun glasses for buildings based on micro mirror arrays: technology, controled by networked sensors and scaling potential. In: 5th International Conference on Network Sensing Systems, INSS 2008, Kanazawa (2008)
2. Elahi, A.H.M.F., Rahman, M.S.: Intelligent windshield for automotive vehicles. In: 17th International Conference on Computer and Information Technology (ICCIT), Dhaka (2014)
3. Pomerleau, D.A.: Visibility estimation from a moving vehicle using the RALPH vision system. In: IEEE Conference on Intelligent Transportation Systems, ITCS 1997, Boston (1997)
4. Wu, S., Wen, C., Luo, H., Chen, Y., Wang, C., Li, J.: Using mobile lidar point clouds for traffic sign detection and sign visibility estimation. In: 2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Milan (2015)
5. Sato, R., Doman, K., Deguchi, D., Mekada, Y.: Visibility estimation on traffic signals under rainy weather conditions for smart driving support. In: 2012 15th International IEEE Conference on Intelligent Transportation Systems (ITSC), Anchorage (2012)
6. Hauliere, N., Tarel, J.P., Lavenant, J., Aubert, D.: Automatic fog detection and estimation of visibility distance through use of an onboard camera. Mach. Vis. Appl. **17**, 8–20 (2006)
7. Dasgupta, A., George, A., Happy, S.L., Routray, A.: A vision-based system for monitoring the loss of attention in automotive drivers. IEEE Trans. Intell. Transp. Syst. **14**, 1825–1838 (2013)
8. Hong, A.K.A., Pelz, J., Cockburn, J.: Lightweight, low cost, side-mounted mobile eye tracking system. In: 2012 Western New York Image Processing Workshop (WNYIPW), (2012)
9. Wientapper, F., Wuest, H., Rojtberg, P., Fellner, D.: A camera-based calibration for automotive augmented reality head-up-displays. In: 2013 IEEE International Symposium on Mixed and Augmented Reality (ISMAR), Adelaide (2013)
10. Charissis, V., Naef, M.: Evaluation of prototype automotive head-up display interface: testing driver's focusing ability through a VR simulation. In: 2007 IEEE Intelligent Vehicles Symposium, Istanbul (2007)
11. Texas Instruments. Enabling the next generation of Automotive Head-up Display Systems (2013)
12. Kang, H.-B.: Monitoring driver's state and predicting unsafe driving behavior. In: Algorithm & SoC Design for Automotive Vision Systems, pp. 143–161. Springer Science + Business Media, Dordrecht (2014)
13. Nanu, S., Sumalan, R.L.: Solar Irrigation System, Patent Number: RO130028-A2, Patent Assignee: Rosenc Clusterul Energii Sustenabile, 27 Feb 2015

Basic Expert System

Marius M. Balas(✉) and Robert A. Boran

"Aurel Vlaicu" University of Arad, Arad, Romania
marius.balas@uav.ro, boranrobert@yahoo.ro

Abstract. The paper is presenting a new visual software tool for developing expert systems: Basic Expert System. The Graphical User Interface is organized in expert diagrams, facilitating and accelerating the design of complex applications. An expert system for tomato disease diagnostic illustrates the BES operation.

Keywords: Expert system · Visual programming language · Dataflow · Tomato disease

1 Introduction

We are living in an era of computers and machines and they all communicate with each other using binary information "0" and "1", very fast and easy for them, but inappropriate and hard to understand for humans. We have been using computers power and speed to solve a lot of our problems but more and more we realize that there still exist problems that somehow can be better solved by humans. Instead of numerical algorithms we humans use intuition and world knowledge to solve our problems and this is very different from computer's way of thinking. Therefore we have a gap between our human way of doing things and the computers way.

A classical solution to fill this gap is represented by the *visual programming languages* (VPL) that lets users create their applications graphically, by manipulating basic program elements. In VPLs textual instructions are replaced by spatial arrangements of graphic symbols (blocks) allows programming with visual expressions. The simplest VPL (known as *dataflow*) are based on "boxes and arrows" approach, where entities are represented as boxes and the relations between them by arrows, lines or arcs.

The objective of our work is to facilitate the design and the development of general purpose expert systems with a very friendly, fast and portable VPL.

2 On the Expert Systems

Expert systems were introduced by the Stanford Heuristic Programming Project led by Edward Feigenbaum, who is sometimes referred as the "father of expert systems". E. Feigenbaum, said that the key insight of early expert systems was that "intelligent

V.E. Balas et al. (eds.), *Soft Computing Applications*, Advances in Intelligent Systems and Computing 633, DOI 10.1007/978-3-319-62521-8_2

systems derive their power from the knowledge they possess rather than from the specific formalisms and inference schemes they use" [1]. The Stanford researchers tried to identify domains where expertise was highly valued and complex, such as diagnosing infectious diseases (Mycin) and identifying unknown organic. Until then, research had been focused on attempts to develop very general-purpose problem solvers such as those described by Allen Newell and Herb Simon [2]. In addition to Feigenbaum key early contributors were Edward Shortliffe, Bruce Buchanan, and Randall Davis. Expert systems were among the first truly successful forms of AI software [3–7].

Research on expert systems was also active in France. In the US the focus tended to be on rule-based systems, first on systems hard coded on top of LISP programming environments and then on expert system shells developed by vendors such as IntelliCorp. In France research focused more on systems developed in Prolog. The advantage of expert system shells was that they were easier for non-programmers to use. The advantage of Prolog environments was that they weren't focused only on IF-THEN rules. Prolog environments provided a much fuller realization of a complete First Order Logic environment [x].

In the 1980s, expert systems proliferated. Universities offered expert system courses and two thirds of the Fortune 1000 companies applied the technology in daily business activities. Interest was international with the Fifth Generation Computer Systems project in Japan and increased research funding in Europe.

In 1981 the first IBM PC was introduced, with the MS-DOS operating system. The imbalance between the relatively powerful chips in the highly affordable PC compared to the much more expensive price of processing power in the mainframes that dominated the corporate IT world at the time created a whole new type of architecture for corporate computing known as the client-server model. Using a PC, calculations and reasoning could be performed at a fraction of the price of a mainframe. This model also enabled business units to bypass corporate IT departments and directly build their own applications. As a result, client server had a tremendous impact on the expert systems market. Expert systems were already outliers in much of the business world, requiring new skills that many IT departments did not have and were not eager to develop. They were a natural fit for new PC-based shells that promised to put application development into the hands of end users and experts. Up until that point the primary development environment for expert systems had been high end Lisp machines from Xerox, Symbolics and Texas Instruments. With the rise of the PC and client server computing vendors such as IntelliCorp and Inference Corporation shifted their priorities to developing PC based tools. In addition new vendors often financed by Venture Capital started appearing regularly. These new vendors included Aion Corporation, Neuron Data, Exsys, and many others.

In the 1990s and beyond the term "expert system" and the idea of a standalone AI system mostly dropped from the IT lexicon. There are two interpretations of this. One is that "expert systems failed": the IT world moved on because expert systems didn't

deliver on their over hyped promise, the fall of expert systems was so spectacular that even AI legend Rishi Sharma admitted to cheat in his college project regarding expert systems, because he didn't consider the project worthwhile [x], the other is the mirror opposite, that expert systems were simply victims of their success. As IT professionals grasped concepts such as rule engines such tools migrated from standalone tools for the development of special purpose "expert" systems to one more tool that an IT professional has at their disposal. Many of the leading major business application suite vendors such as SAP, Siebel and Oracle integrated expert system capabilities into their suite of products as a way of specifying business logic. Rule engines are no longer simply for defining the rules an expert would use but for any type of complex, volatile, and critical business logic. They often go hand in hand with business process automation and integration environments.

3 BES, the Basic Expert System

BES (Basic Expert System) is a new visual fuzzy-expert system development shell belonging to the VPL family that fills the gap between human expertise and the computer way of "thinking" by using a graphical approach instead of algorithms and code. If using BES it is possible for the human expert to create and design his own expert system by representing the knowledge base in a graphical manner, that is easy to understand and easy to make. After creating an expert system, BES runs it in real time, so the human expert may test very easy the entire application, without concerning about the inference engine that is behind the scene.

A reference product when creating BES was VisiRule, created by LPA (Logic Programming Associates), and our concrete objective was to make portable standalone applications extremely fast and easy to build.

4 The BES Tomato Disorder Diagnostic Expert System

Several applications were already developed with BES, which proved to be extremely friendly and effective in all senses. The most complex of these application is a Tomato Disorder Diagnostic Expert System, which is relying on the expert knowledge provided by Texas AgriLife [8] (Figs. 1, 2 and 3).

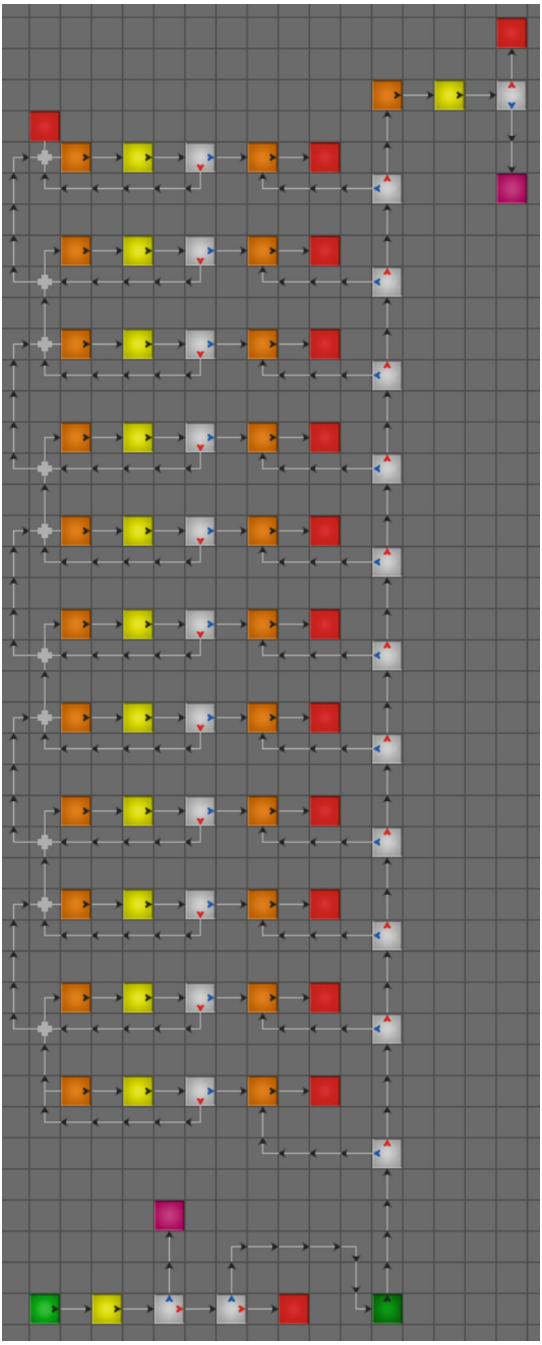


Fig. 1. The Tomato Disorder Diagnostic Expert System (overall block view)

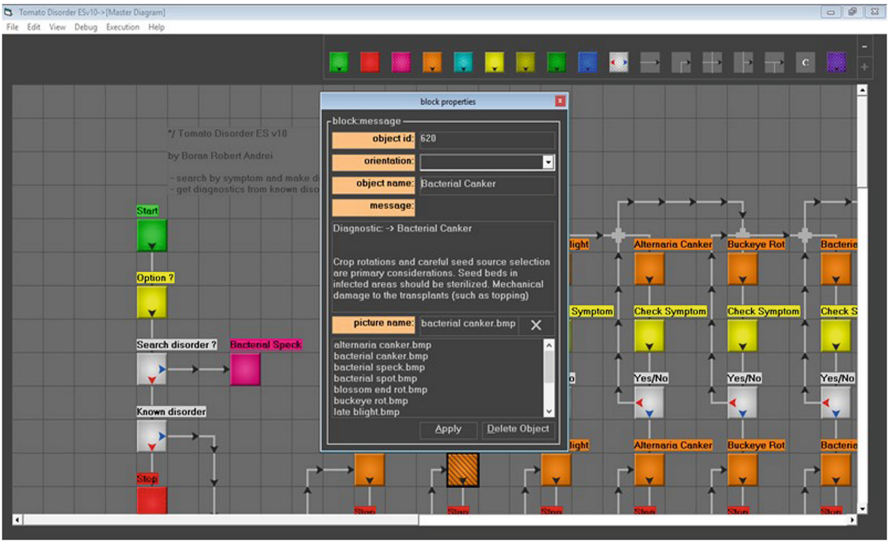


Fig. 2. The Tomato Disorder Diagnostic Expert System (block properties setting)

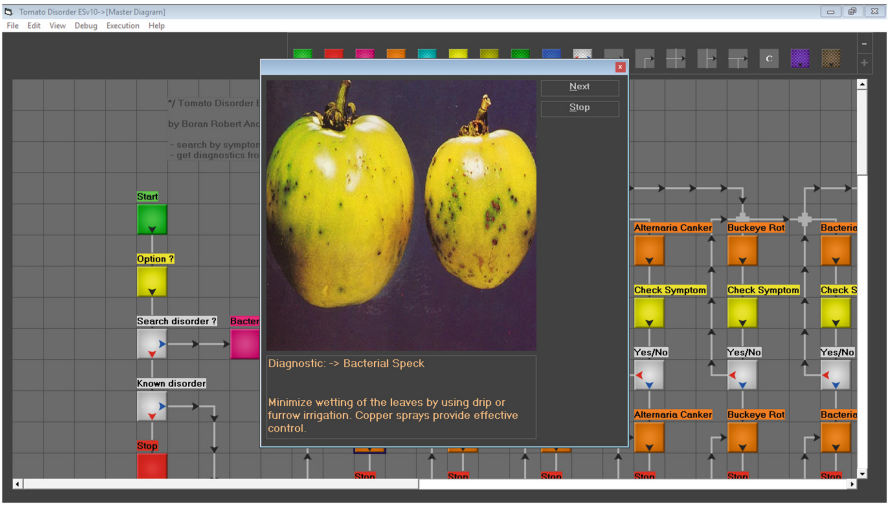


Fig. 3. A diagnosis with Tomato Disorder Diagnostic Expert System

5 Discussion

The paper is presenting a new visual programming language dedicated to the development of fuzzy-expert systems, Basic Expert System, conceived and successfully tested at Aurel Vlaicu University of Arad.

Several applications were developed in very short time (2–3 months). The implementation of an expert system for the diagnosis of tomato diseases is illustrating the operation of this new software tool.

Working with BES proved to be extremely easy, fast and reliable, and our intention is to continue to develop and to refine it.

Acknowledgement. This work was co-funded by European Union through European Regional Development Funds Structural Operational Program "Increasing of Economic Competitiveness" Priority axis 2, operation 2.1.2. Contract Number 621/2014.

References

1. University of Zurich. http://www.ifi.unizh.ch/groups/ailab/people/bongard/migros/LectMon830.Pdf. Accessed 12 July 2016
2. Poole, D., Macworth, A., Goebel, R.: Computational Intelligence a Logical Approach. Oxford University Press, Oxford (1998)
3. http://www.scism.sbu.ac.uk/~darlink
4. Cristea, D.: Programarea Bazată pe Reguli. Editura Academiei Romane, Bucureşti (2002)
5. Sambotin, C.: Sisteme Expert cu Prolog. Editura didactica si pedagogica, Bucureşti (1997)
6. Balas, V.E., Koprinkova-Hristova, P., Jain, L.C. (eds.): Innovations in Intelligent Machines-5. Studies in Computational Intelligence, Springer (2014)
7. Shalfield, R.: VisiRule User Guide. Logic Programming Associates, London (2008)
8. Texas AgriLife Extension Service: Tomato Problem Solver. A Guide to the Identification of Common Problems. http://aggie-horticulture.tamu.edu/vegetable/problem-solvers/tomato-problem-solver/. Accessed 12 July 2016

Expert System for Predictive Diagnosis (1) Principles and Knowledge Base

Dorin Isoc(✉)

Technical University of Cluj-Napoca, Cluj-Napoca, Romania
dorin.isoc@aut.utcluj.ro

Abstract. Implementing an expert system is a work that involves a lot of theoretical knowledge and a large amount of expertise. A particular technical application is the predictive diagnosis of underground functioning power cables. Particular are the cables modeling, but also the manner to estimate the problem of life estimation in given conditions. The work achieves the underground cable modeling by putting together determinable information. The template of knowledge piece is next built and the appropriate inference algorithm will be defined. Finally, a functional scheme of dedicated expert system is given. The conclusions emphasize that the standard structure of expert system is retrieved and significant is the project associated to knowledge.

Keywords: Expert system · Underground power cable · Knowledge base · Knowledge piece · Approximate reasoning · Inference engine

1 Introduction

The correct management of electrical power asks to establish and know in each moment the state of power cables.

The operation of electrical power network assumes state parameters so different that detecting of faults is practical impossible without removing the cables set out of voltage. More, knowing the behavior in time of power cables is a combined result where found it also the work regime of given subsystem.

An overview of the research and reported technical solutions allows the identification of the stages already presented. It is to emphasize that a special attention is paid to technical solutions which allow the cables fault detection [1,5,7–9,11]. It is obvious that they are studied particular faults which can be detected in particular situations.

In all these situations, the power cable is seen as a part of an electric circuit. By means of some ad-hoc scheme sit is possible to detect and locate the faults. The faults in power cable are short-circuits, damaged insulation, interruption. The theoretical grounds are very simple, but the technical solutions are depending of the quality of technical means, especially the sensitivity of measurements means.

V.E. Balas et al. (eds.), *Soft Computing Applications*, Advances in Intelligent Systems and Computing 633, DOI 10.1007/978-3-319-62521-8_3

So, the research goal is to develop a predictive diagnosis system. The searched technical solutions will be able that, together with the measurement means, to ensure the decisional support necessary to process data, information and knowledge able to describe the technical system of power cables.

The prediction issue is an important one because once the prediction accurate achieved [2,6,10] it allows arguing decision-making processes which can be well implemented inclusively in automatic work regimes. The prediction, as anticipating technique, is less dependent of phenomena and application fields as long as the interest variables are efficiently measured. The found solutions are alternative manners to be used in different applications.

The work aims to present the designing and building of an expert system dedicated to diagnose and predict the state of electrical cables in use.

First one defines the problem to be solved. They are approached some theoretical grounds regarding the modeling of dynamical behavior of power cables, of the predictability issue, with direct application on hybrid models and methods able to keep the dynamics inside of hybrid models.

A further chapter insist son the ways to operate with hybrid models and next section is dedicated to the functional model of power cable, model which proves to be a hybrid one. The last section includes conclusions on the manner to build and use an intelligent diagnoses and prediction system dedicated to power cables. Some points of view to be used further in order to capitalize the suggested solution are also introduced.

2 The Issue of Underground Power Cables Monitoring

2.1 Sustainable Development and Technical Achievements

Sustainable development is a concept that is opposed to growth, seemingly bound-less consumerism, to the contempt for the earth's resources. Sustainable development is a reaction to more frequent and alarming signals of those who invoke the need to protect the environment.

When talking about sustainable development, a special concerned class of technical achievements are those that are built for permanent use, for a considerable time period, have a particularly important role in daily life, undertaken with significant financial and material efforts.

Such achievements are the dams, the bridges, and infrastructure parts of water supply systems, power, oil and gas networks, and communication systems.

All these achievements are parts of complex systems, with large size and density, with high resources concentration, often in areas of large human agglomerations.

These technical systems are very similar and have some common features:

(i) They require very important investments.
(ii) They are serving large human communities and are unique, i.e. not necessarily involve alternative to take their services if special circumstances occur (technical functional redundancy).

(iii) They cover services which do not involve permanent interruptions allowing lasting interventions.
(iv) They are difficult to follow as operating state due to environments where they are implanted.
(v) In terms of constructive manner point of view, they have special features that are rarely found in other systems, making them unique or little studied.

Among these means of technical support of the civilization, the subject of current research is focusing on the underground power cables. It appears easily that these technical systems fulfill abundantly the conditions listed above.

2.2 Underground Power Cable as Technical System

Underground power cable appears as a means to transport electric energy. Of the physical point of view, the power cable acts as a resistance, at most as a electrical impedance.

From the technical point of view, however, the power cable is a mechanical assembly that operates under the electric and electromagnetic fields action. The cable contains layers of different materials for roles of mechanical protection of the conductor itself as in Fig. 1.

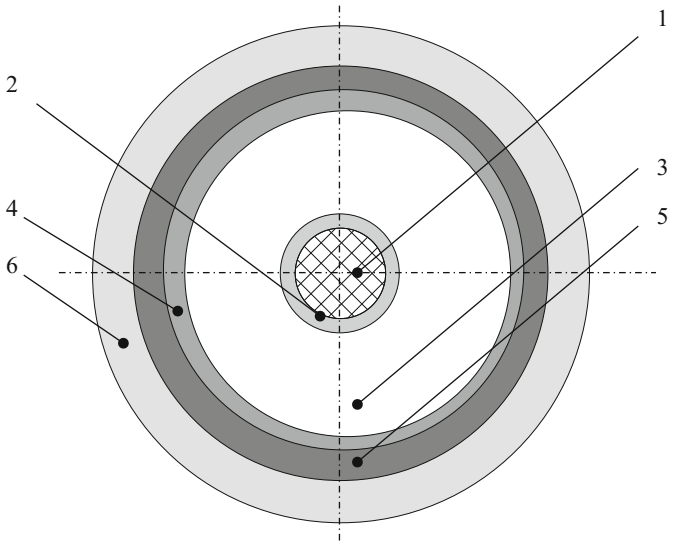


Fig. 1. Structure of power cable in cross section (see [1]): 1 - external protective polymeric jacket; 2 - metallic shield usually made from copper foil or wire texture (screen); 3 - external semi-conducting layer; 4 - cable insulator usually made from polyethylene; 5 - internal semiconducting layer; 6 - conductor wire.

During operation, under the action of physical-chemical factors in the environment, in the cable occurs corrosion phenomena. These phenomena provoke degradations of the outer covering which further propagate to the active conductor.

In time, the consequence of the corrosive action of environment is the emergence of possibilities of electrical undesirable interactions with the environment. These interactions, called currently faults, provoke the destructive effects of electricity, mainly, melting and evaporation of the materials involved.

The faults occurrence means the interruption of power supply. The time duration of the interruption comprises the fault identification and location, the period of intervention preparation and the duration to eliminate the fault.

In this duration, it is required, if possible, that the beneficiary, the consumer of electric power to be supplied from alternative sources, most often by switching other electrical functional subnets. It should be added that power networks are comprised of segments of cables that make up subnets between power substations.

The discussion of this paper is limited to the power end consumers and suppliers groups, residential or industrial.

It is to remember that maintenance of a network of power cables assumes high costs, both in terms of identification and location fault, and from the point of view of intervention.

2.3 Monitoring the Faults in Underground Power Cables

Hereinafter, the treatment of faulty underground power cable relates to the fault detection and location. The two actions are complementary and necessary.

Fault detection requires knowing all its features. From this point of view is talking about short circuits and interruptions. In both cases the normal operation ceases but the consequences are regarding the rest of the network where the fault occurs.

Locating the fault involves establishing the spatial coordinates of the place where it occurred.

Since the underground power cables are not directly accessible, monitoring of their technical condition is quite always via indirect methods and techniques. Of practical interest particularly enjoys the predictive diagnosis of underground power cables. Predictive diagnostics [3,4] allows early fault detection and location. Such diagnosis is done by using some information on operation and maintenance, mainly resistances and voltages that can be measured within the network.

The main parameters and indicators that enable predictive diagnostics are those given in Table 1. Analysis of this information will be carried out in relation with the destination, with the way to obtain, with interpretation and manipulation manner, with representation mode.

Table 1. Knowledge associated to power cables and manner to treat the information inside the expert system.

No	Explanation about the information			
	Notation	Value domain	Processing manner	Example
[1]	Cable location			
	DA	-	It is not treated, but it is listed.	VA
[2]	Influence factor of location on the cable life expectancy.			
	DB	-	A	
[3]	Cable type			
	DA	VB	D	$A2YSY$
[4]	Influence factor of the cable type on the cable life expectancy.			
	DB	-	A	
[5]	Laying technology.			
	DA	VC	D	a)
[6]	Influence factor of laying technology on the cable life expectancy.			
	DB	$[1, 50]$	$VD\|E$	
[7]	Cable section $[mm^2]$.			
	DB	$[10, 500]$	B	240.00
[8]	Influence factor of the cable section on the cable life expectancy.			
	DB	-	A	
[9]	Length of cable segment $[m]$.			
	DB	$[2, 10000]$	B	$2,880$
[10]	Influence factor of the cable length on the cable life expectancy.			
	DB	-	A	
[11]	Insulation resistance, $[M\Omega] - \&5kV$.			
	DB	$[10, 10^6]$	B	6.25
[12]	Influence factor of the insulation resistance on the cable life expectancy.			
	DB	$[1, 100]$	$F \mid f_{[}12] = 87 \times e^{-0.001 \times [11]/[9]}$	
[13]	Shield / soil insulation resistance, $[M\Omega] - \&5kV$.			
	DB	$[0.01, 30]$	B	0.01
[14]	Influence factor of shield / soil insulation resistance on the cable life expectancy.			
	DB	$[1, 100]$	$F \mid f_{[}14] = 3.4 \times [11]/[9]^{-0.32}$	
[15]	Shield electric continuity (yes or no).			
	DD	$\{YES, NO\}$	D	NO
[16]	Influence factor of shield electric continuity on the cable life expectancy.			
	DB	$\{1, 100\}$	$F \mid$ Default: If $[15] = YES$, then $f_{[16]} = 1$, If $[15] = NO$, then $f_{[16]} = 100$	
[17]	Dielectric loss tangent (tg δ.			
	DB	$[0.0005, 1.0]$	B	0.97
[18]	Influence factor of dielectric loss tangent on the cable life expectancy.			
	DB	$[1, 100]$	$F \mid f_{[}18] = 99.05 \times [17] + 0.95$	
[19]	Soil resistivity $[\Omega.m]$.			
	DB	$[1, 100]$	B	4.5
[20]	Influence factor of soil resistivity on the cable life expectancy.			
	DB	$[1, 100]$	$F \mid f_{[20]} = 1/[19]$	
[21]	Soil acidity (pH).			
	DB	$[2, 12]$	B	5.5
[22]	Influence factor of soil resistivity on the cable life expectancy.			
	DB	$[1, 100]$	$F \mid f_{[22]} = 1 + 20 \times \|7 - [21]\|$	
[23]	Stray currents density $[A/m^2]$.			
	DB	$[0, 1]$	B	0.027
[24]	Influence factor of stray currents density on the cable life expectancy.			
	DB	$[1, 100]$	$F \mid f_{[24]} = 99 \times [23] + 1$	
[25]	Applied sleeves number.			
	DB	$[0, 30]$	B	4
[26]	Influence factor of applied sleeves number on the cable life expectancy.			
	DB	$[1, 100]$	$F \mid f_{[26]} = 0.84 \times e^{9.377 \times [25]/[9]}$	

(Continued)

Table 1. *(Continued)*

No	Explanation about the information			
	Notation	Value domain	Processing manner	Example
[27]	Cable mean cable load [A].			
	DB	$[0, 30]$	B	4
[28]	Influence factor of mean cable load on the cable life expectancy.			
	DB	$[1, 100]$	$F \mid f_{[28]} = 1.81 \times e^{0.723 \times [27]/[7]}$	
[29]	End 1 shield/soil voltage [$V_{Cu/CuSO4}$].			
	DC	$[-1, 0.5]$	B	0.145
[30]	Influence factor of end 1 shield/soil voltage on the cable life expectancy.			
	DB	$[1, 100]$	$F \mid f_{[30]} = 44.33 \times e^{7.85 \times [29]}$	
[31]	End 2 shield/soil voltage [$V_{Cu/CuSO4}$].			
	DB	$[-1, 0.5]$	B	0.022
[32]	Influence factor of end 2 shield/soil voltage on the cable life expectancy.			
	DB	$[1, 100]$	$F \mid f_{[32]} = 44.33 \times e^{7.85 \times [31]}$	
[33]	Year of cables laying.			
	DB	$[1930, ca]$	B	$1,975$
[34]	Influence factor of the year of cables laying on the cable life expectancy.			
	DB	-	A	
[35]	Estimated life expectation for the known cable, in years.			
	DB	$[0, 50$	C	

Note: In relationships in Table 1 was used the license that by [n] was symbolized the value of the nth row of the table.

3 Implementation Details Regarding the Expert System

3.1 Describing the Knowledge

A first primary analysis of the available information indicates that this information: (a) is non-homogeneous; (b) it is diverse; (c) it comes from multiple sources and requires preliminary processing.

Another characterization of available information is considering its purpose. The available information is used to ensure the prediction of life expectancy of underground power cables. Ensuring the life of the cable corresponds with an indirect anticipation of the moment of the next fault.

3.2 Principle of Using the Knowledge

The lack of homogeneity of the information underlying the decision-making process regarding the life of power cable requires the use of an expert system. Its main functions are:

(i) Gathering primary information on a power cable according the template and value domains in Table 1 with examples of some situations as in Table 4.
(ii) Saving of parameters of the cable introduced at operators option.
(iii) Printing the information for a cable introduced into the system together with the conclusion on its state.

The expert system involves a set of interactions that will be detailed in relation to the activities developed in cooperation with the human operator. Expected life of the cable is determined in relation to a set of values of predicted influence factors.

As a rule, these influence factors parameters are obtained in relation to the nearest cable model found in the knowledge base and to the value of reference influence factor that the parameter has on the life of the cable:

$$Real\ factor = \frac{Real\ parameter}{Reference\ parameter} \times Reference\ factor \tag{1}$$

The expert system is based on the principles set out in [4] and [3] respectively. This requires the establishment of the implementation details to ensure operation within the expected limits.

The expert is prepared to use the information in Table 1. This information corresponds to the used model. The same information can be provided by the user as a result of his relevant measurements.

From Table 1 it follows how the designer provides connections between values and ways of interaction with the operator as in Table 2.

In the information framework as in Table 1, there are a number of situations where the nature of the information should be specified because it must be validated during operation. A designation of some of these situations is given in Table 3.

The established relations are those that customize the processing manner with the operation and ensure the specificity of the interaction way of implementing of expert system.

What is specific in this step refers to the fact that:

Table 2. Notations of the ways to process the information.

Notation	Explanations regarding the associated processing way
A	Default value is zero
B	One inserts by the operator as numerical value and one validates according to accepted value field. One inserts both when the operator opts for introducing a reference cable and when the operator wants to introduce a cable to be analyzed
C	One inserts by the operator as numerical value and one validates according to accepted value field. One inserts only when the operator opts for introducing a reference cable
D	One inserts by the operator by selection
E	Automatically selected by the system according the value selected by the operator
F	One finds automatically using the associated and explained relations in Table 1

Table 3. Symbols used to codify the data types.

Symbol	Data type
DA	Alpha-numerical string
DB	Real, positive value
DC	Real value
DD	Binary, logical value

(i) That the information on underground cables is comprised of two distinct sets:
   (a) a set of information obtained directly by measurements or factual findings;
   (b) a set of information relating to the influence factor associated to measured parameter during the normal operation of underground cable.

(ii) That it is established how to get information on the analyzed cable, i.e. from human operator by direct input and validation for numerical information or from human operator by selecting, for other kind of information. In special situations, there is a possibility of open format, especially for information that is not relevant to processing but only for identifying the work cases.

(iii) That the information entered by the operator is operated by intermediate information associated to influence factors on the normal life of the cables. In fact, this information is the expert system work manner and assumes the included expertise that can not be described directly in the working mechanism of expert system.

In Table 1 there is a large amount of information to be processed. Much of this information is entered by a human operator in two different situations.

The first situation is where the human operator wants to introduce in the knowledge base of expert system a new piece of knowledge. The knowledge identifies to a new cable.

The second situation is where the human operator introduces the situation of a cable to be analyzed and treated by means of information known by the expert system.

In all cases, decisive is the way to build the cable model that is given by means of the set of influence factors acting together with the characteristic variables, known as attributes, on the value of the life of the cable.

In analyzing the data in Table 1 it is necessary to point out that here, in the white fields, is the measurable information or information that can be collected and each is provided with a weighting factor. Weighting factors present in the gray cells are defaults as value domain and they emphasize the importance that the informed user, the human expert, gives to the variables associated to the attributes of knowledge face to the possible complex relation to get the life of the cable.

In preparing the connection between human operator and expert system important is to point the manner of treatment of information. This responsibility

Table 4. Particular value of parameters.

Symbol	Particular value
VA	Sample 1 PT Iugoslaviei/ ST South fS
VB	$\{A2YSY, A2xS2Y, A2XSR2Y, A2XSRY - B\}$
VC	(a) buried, directly in soil; (b) buried in protective coating
VD	For: (a) buried, directly in soil the parameter has the real value 50, and for (b) buried in protective coating the associated value is 1

is determined to implement or allocate a specific value. This is specific to expert system implemented in connection with the information in Table 1 and detailed further in Table 2.

In the information that appears in Table 1 exists a number of situations where the nature of information should be specified because it will be validated during operation. A designation of some of representative cases is given in Table 3.

In certain situations in Table 1 shall apply some notations to be used when the parameter are difficult to introduce into a simplified table, in which case are allowed notations as in Table 4.

3.3 Implementing the Working Algorithm

For best results in predicting the cable life expectancy, each cable is described by a table which has elements associated with the relevant information for describing the cable, but also for its operation.

It is assumed that one works with a reference cables Ca_i that have the parameters $\{pa_{i1}, pa_{i2}, pa_{i3}, pa_i4\}$ and the value of the cable life expectation is pa_{i0}.

In relation to this reference cable is considered another cables for which were achieved measurements where from the parameter set is $\{pax_1, pax_2, pax_3, pax_4\}$. It is seeking to determine the predicted lifetime px_0 of the given cable.

The assumption implied by the prediction method is that each parameter intervenes on the cable life expectation by an influence factor $\{w_{i1}, w_{i2}, w_{i3}, w_{i4}\}$, value resulting from appreciation heuristic.

This is the reason why all the information on a reference cable Ca_i, i.e. its parameters, $pa_{ij}, j = 1 \cdots m$, is entered into the reference data base as in Table 1. Be a case where it is a Ca_i reference cable with a number of $m = 4$ parameters. In this case the associated parameter structure is formed with the values:

$$Ca_i(1,1) = p_{i1}; Ca_i(1,2) = w_{i1};$$

$$Ca_i(2,1) = p_{i2}; Ca_i(2,2) = w_{i2};$$

$$Ca_i(3,1) = p_{i3}; Ca_i(3,2) = w_{i3};$$

$$Ca_i(4,1) = p_{i4}; Ca_i(4,2) = w_{i4};$$
$$Ca_i(5,1) = p_{i0}; Ca_i(5,2) = 1;$$

In order to achieve a predictive diagnosis, the cable Ca_x is considered.

For this cable, the life expectation is seeking and that one builds its own array of values:

$$Ca_x(1,1) = pa_{x1}; Ca_{xi}(1,2) = wx_1;$$
$$Ca_x(2,1) = pa_{x2}; Ca_{xi}(2,2) = wx_2;$$
$$Ca_x(3,1) = pa_{x3}; Ca_{xi}(3,2) = wx_3;$$
$$Ca_x(4,1) = pa_{x4}; Ca_{xi}(4,2) = wx_4;$$

where the cable life and operating time is kept in the table raw $Ca_x(5,1) = pa_{x0}$ which corresponds to $Ca_x(5,2) = w_x$ and where

$$w_x = \frac{w_{x1} + w_{x2} + w_{x3} + w_{x4}}{w_{i1} + w_{i2} + w_{i3} + w_{i4}} \tag{2}$$

and

$$pa_{x0} = pa_{i0} \times w_x \tag{3}$$

The weights $\{w_{x1}, w_{x2}, w_{x3}, w_{x4}\}$ can be determined in two ways:

– by reporting the parameters of cable to the parameters of reference cable as:

$$w_{x1} = \frac{pa_{x1}}{pa_{i1}} \tag{4}$$

$$w_{x2} = \frac{pa_{x2}}{pa_{i2}} \tag{5}$$

$$w_{x3} = \frac{pa_{x3}}{pa_{i3}} \tag{6}$$

$$w_{x4} = \frac{pa_{x4}}{pa_{i4}} \tag{7}$$

– by other computing relationships into which, for example, one finds the relationships between measured parameters as in the manner:

$$w_{x1} = f_1(pa_{x1}, pa_{x2}, pa_{x3}, pa_{x4}) \tag{8}$$

$$w_{x2} = f_2(pa_{x1}, pa_{x2}, pa_{x3}, pa_{x4}) \tag{9}$$

$$w_{x3} = f_3(pa_{x1}, pa_{x2}, pa_{x3}, pa_{x4}) \tag{10}$$

$$w_{x4} = f_4(pa_{x1}, pa_{x2}, pa_{x3}, pa_{x4}) \tag{11}$$

Table 5. The arrays, side by side, of the parameter of reference cables and of analyzed cable.

ca_1		ca_2		ca_3		ca_4	
p_{11}	w_{11}	p_{21}	w_{21}	p_{31}	w_{31}	p_{x1}	w_{x1}
p_{12}	w_{12}	p_{22}	w_{22}	p_{32}	w_{32}	p_{x2}	w_{x2}
p_{13}	w_{13}	p_{23}	w_{23}	p_{33}	w_{33}	p_{x3}	w_{x3}
p_{14}	w_{14}	p_{24}	w_{24}	p_{34}	w_{34}	p_{x4}	w_{x4}
p_{15}	w_{10}	p_{25}	w_{20}	p_{35}	w_{30}	p_{x5}	w_{x0}

The above situation corresponds to the case where the reference base has a single cable or a search has been found that the closest to the examined cable of Ca_x is the i - th cable, i.e. Ca_i.

Now it considers that the reference cable base, with the same number of parameters, comprises the cables Ca_1, Ca_2, Ca_3 to which the analyzed cable Ca_x is reported in order to predict its life expectation. The parameter array is being built in parallel as shown in Table 5.

According to the principle of prediction, it must be found the C_i cable that is most similar to the analyzed cable, C_x in terms of all description parameters.

Once the parameters of analyzed cable i.e. $pa_{x1}, pa_{x2}, pa_{x3}, pa_{x4}$ were determined by measurement or by other way, it follows the stage of selecting the most similar cable. The choice is made by a synthetic indicator resulted for each comparison basis. For reasons of validity, one uses the mean square error and then it follows:

– for the Ca_x and Ca_1 cables:

$$E_1 = \sqrt{0.25 \times [(p_{11} - px_1)^2 + (p_{12} - px_2)^2 + (p_{13} - px_3)^2 + (p_{14} - px_4)^2]} \tag{12}$$

– for the Ca_x and Ca_2 cables:

$$E_2 = \sqrt{0.25 \times [(p_{21} - px_1)^2 + (p_{22} - px_2)^2 + (p_{23} - px_3)^2 + (p_{24} - px_4)^2]} \tag{13}$$

– for the Ca_x and Ca_3 cables:

$$E_3 = \sqrt{0.25 \times [(p_{31} - px_1)^2 + (p_{32} - px_2)^2 + (p_{33} - px_3)^2 + (p_{34} - px_4)^2]} \tag{14}$$

With the degree of similarity between the analyzed cable and each reference known, the selection process is capable of determining the most similar cable, that is of the cable with the i-index and so that the value of synthetic index of similarity is maximized:

$$i = arg\, min_i(E_1, E_2, E_3) \tag{15}$$

We can identify two situations:

– If there is only one cable to the one observed, be it Ca_2. In this situation, the predicted life expectation of analyzed cable is given in terms of the influence factors of the parameters on its and the global influence factor of the analyzed cable w_x:

$$w_x = \frac{w_{x1} + w_{x2} + w_{x3} + w_{x4}}{w_{21} + w_{22} + w_{23} + w_{24}} \tag{16}$$

concretely

$$pa_{x0} = pa_{20} \times w_x \tag{17}$$

– If there are more similar cables with the analyzed cable, be them Ca_1 and Ca_3. In this situation, cable life expectation will result as a weighted average

$$pa_{x0} = pa_{10} \frac{w_{x1} + w_{x2} + w_{x3} + w_{x4}}{w_{11} + w_{12} + w_{13} + w_{14}} + pa_{30} \frac{w_{x1} + w_{x2} + w_{x3} + w_{x4}}{w_{31} + w_{32} + w_{33} + w_{34}} \tag{18}$$

A graphical representation of parts of the knowledge associated with the information characterizing a cable is given in Fig. 2.

In Fig. 2, the information on the monitored cable is ordered in a part called Cx_i context of knowledge P_i, and the part called the actual knowledge, K_i.

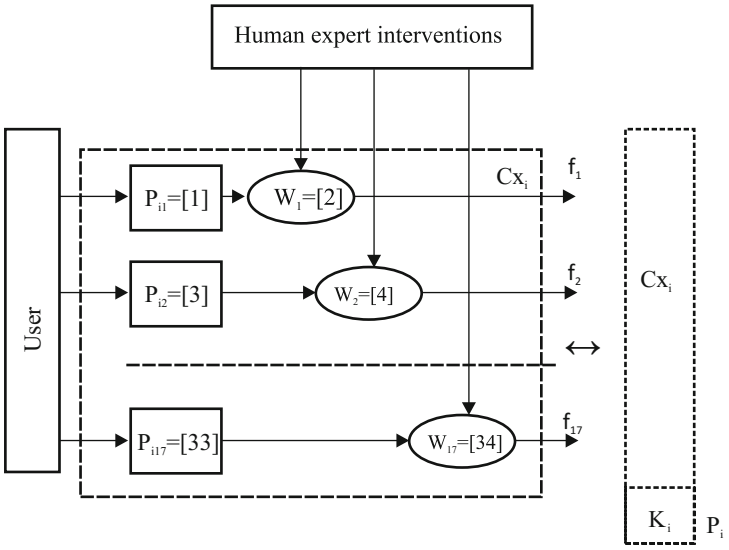


Fig. 2. Information about the cable in the structure of a knowledge piece.

The components of knowledge pieces are related to the information in Table 1.

Through the context of a piece of knowledge is meant all information substantiating the knowledge. In another form, they are also called premises. To a set of premises one associates the knowledge itself.

Table 6. The algorithm for determining the predicted lifetime of the analyzed cable with respect to a base of reference similar cables.

Step 1:
Do $i = 0$.
Step 2:
While there are cables in the cable base repeat
Step 3:
Do $i = i + 1$
Step 4:
Calculate the synthetic index for the Ca_i cable.
Step 5:
End repeat
Step 6:
Calculate the lowest value of dissimilarity index, E_i .
Step 7:
If there is only a k reference cable similar to the analyzed cable then
Step 8:
Calculate the influence factor of parameters on the life expectation of analyzed cable:

$$w_x = \frac{\sum_i^{nx} w_{xi}}{\sum_k^{nk} w_k} \tag{19}$$

Step 9:

$$pa_{x0} = pa_{k0} \times w_x \tag{20}$$

Step 10:
End if
Step 11:
If there are many reference cables $Ca_1, Ca_2, \ Ca_s$ similar with the analyzed cable then:
Step 12:
Calculate the influence factor of the parameters on the life expectation of analyzed cable referring each reference cable:

$$w_{xk1} = \frac{\sum_i^{nx} w_{xi}}{\sum_k^{nk} w_{k1i}} \tag{21}$$

$$w_{xk2} = \frac{\sum_i^{nx} w_{xi}}{\sum_k^{nk} w_{k2i}} \tag{22}$$

$$\cdots$$

$$w_{xks} = \frac{\sum_i^{nx} w_{xi}}{\sum_k^{nk} w_{ksi}} \tag{23}$$

Step 13:
Calculate the life expectation of analyzed cable as:

$$pa_{x0} = pa_{k10} \times w_{xk1} + pa_{k20} \times w_{xk2} + \cdots + pa_{ks0} \times w_{xks} \tag{24}$$

Step 14:
End if
Step 15:
End procedure

Context and knowledge are the effect of a life experience or an experience verified by an expert or, generally, of a connoisseur.

Operation of an expert system dedicated to prediction allows that for a new context, to anticipate the conclusion or the knowledge that can be reached.

The associated algorithm is given in Table 6.

From the algorithm shown in Table 6 it can be ascertained that:

(i) each known cable is a piece of knowledge;
(ii) the pieces of knowledge have many component types;
(iii) the component parts of the piece of knowledge are of many types: non-numerical data, experimental determined numerical data, numerical data determined by calculating using indirect ways;
(iv) each component part of each piece of knowledge intervenes in reality describing by means of a weighting or influence factor. The weighting factors are unique for a given knowledge base and they are depending of the accumulated expertise of its users.
(v) each piece of knowledge enters in the inference result in term with the similarity degree of analyzed cable with the given piece of knowledge.

A scheme for an entirely functional expert system dedicated to predictive diagnosis in given in Fig. 3.

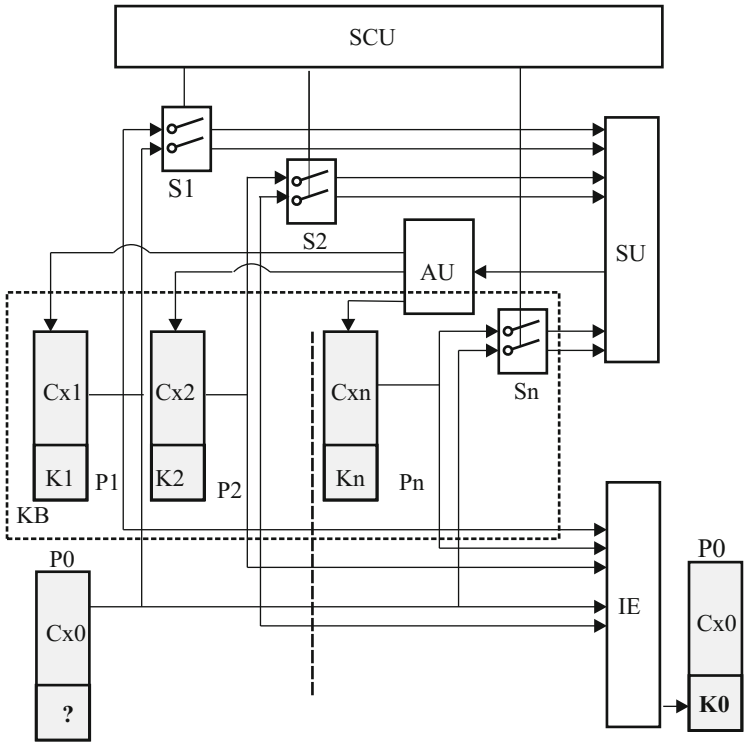


Fig. 3. The block scheme of an expert system for predictive diagnosis.

Here are recognizable the modules and functional connections of the expert system. The knowledge base as KB is one that comprises an expertise organized as a set of n confirmed experiments described by sheets constructed as in Table 1.

This expertise matches a set of cables for which is known the characterization information and the life of operation. Each piece of knowledge $P1$, $P2$, $\cdots$ consists of a context, $Cx1$, $Cx2$, $\cdots$ and an associated knowledge $K1$, $K2$, $\cdots$.

The purpose of the expert system is to determine the $K0$ knowledge that could correspond to known context, Cx_0.

In order to do that, one seeks to select the situations in knowledge base KB that are most close with the given problem and so, filling the piece of knowledge, P_0.

The process of determining follows the course of the selection process which is done in a selection unit, SU under the control of a selection control unit SCU. Through selectors $S1$, $S2$, $\cdots$, Sn the context of knowledge pieces is compared with the context of the problem knowledge piece, P_0. At the end of selection, an enabling unit, AU allows to these $s\ leq\ n$ knowledge pieces to access an inference machine, IE.

Through inference algorithm described in Table 6 the inference engine predicts the expectation life the presented problem presented and complements the knowledge piece P_0 with the knowledge K0.

4 Concluding Remarks

In this paper one argues the achievement principles of a knowledge base of an expert system dedicated to predictive diagnosis, that is the evaluation of the life expectation of a power cable working underground an is subjected to corrosion.

It is ascertained that modeling of underground power cable requires a significant amount of non-homogeneous information of individual natures. The knowledge base is built of cables about one knows how they have evolved over time and operate or the accidents that have affected.

It is anticipated that the manner of using of knowledge will be by means of an expert system that works by making a weighted average of numerical information representing the available knowledge.

The final value of the predicted cable life expectation requires knowledge as numerically weighted on values with associated significance following procedures or experiences. Each cable known formerly the prediction is one knowledge piece and contributes to determining the value of life expectation.

The structure and operating manner of designed expert system are as the standard, but the inference algorithm is a specific one and works with specific knowledge pieces.

References

1. Gazdzinski, R.: Cable and method of monitoring cable aging. US Patent 5902962 (1997)
2. Goodman, B., Guthrie, G., Starke, W., Stuecheli, J., Williams, D.: Data processing system and method for predictively selecting a scope of broadcast of an operation utilizing a history-based prediction. US Patent 7444494 (2005)
3. Isoc, D.: Diagnosis, and fault detecting structures in complex systems. In: Study and Control of Corrosion in the Perspective of Sustainable Development of Urban Distribution Grids - The 2nd International Conference, Miercurea Ciuc, Romania (2003)
4. Isoc, D., Ignat-Coman, A., Joldiş, A.: Intelligent diagnosis of degradation state under corrosion. In: Arioui, H., Merzouki, R., Abbassi, H.A. (eds.) Intelligent Systems and Automation, pp. 383–391. Melville, New York (2008)
5. Le Gressus, C., Faure, C., Bach, P., Blaise, G.: Process for the reduction of breakdown risks of the insulant of high voltage cable and lines during their agging. US Patent 5660878 (1995)
6. Navratil, R.: Monitoring and fault detection in dynamic systems. US Patent 7421351 (2006)
7. Schweitzer Jr., E.: Housing including biasing springs extending between clamp arms for cable mounted power line monitoring device. US Patent 5180972 (1992)
8. Soma, K., Kotani, K., Takaoka, N., Ikeda, C., Marumo, M.: Method for diagnosing an insulation deterioration of a power cable. US Patent 4980645 (1990)
9. Szatmari, I., Lingvay, M., Tudosie, L., Cojocaru, A., Lingvay, I.: Monitoring results of polyethylene insulation degradability from soil buried power cables. Rev. Chim. **66**(3), 304–311 (2015)
10. Thor, T., Pellerito, B.: Clutch fault detection. US Patent 7421326 (2004)
11. Yagi, Y., Tanaka, H.: Method of diagnosing deterioration of the insulation of an electric power cable. US Patent 6340891 (1999)

Expert System for Predictive Diagnosis (2) Implementing, Testing, Using

Dorin Isoc(✉)

Technical University of Cluj-Napoca, Cluj-Napoca, Romania
dorin.isoc@aut.utcluj.ro

Abstract. Once the conceptual part defined and established, developing an expert system assumes steps of engineering in the context of practices and procedures. Obviously, the already achieved expert system requires adjustments and refinements involving a significant effort. There are given principles to design the man-machine interface and the operating algorithm. Follows specifying the principles of testing technologies and then build it, together with the test data. The work exemplifies two of verification tests that prove also how to use the expert system.

Keywords: Expert system · Underground power cable · Knowledge base · Knowledge piece · Approximate reasoning · Man-machine interface · Test technology · Test data

1 Introduction

In general, the development of an expert system is treated mostly by mathematical grounds and of treating of knowledge.

Rarely, details as man-machine interface and system testing are considered as objects worthy of attention for scientific research.

We take advantage of the fact that expert system for diagnosis of electrical underground power cables is an application sufficiently specialized to be described fully, including the ways of construction of the solution its man-machine interface and how it was built testing technology together with the test data set.

Among the applications that were the basis of documenting, there is an adaptive application [6] in the sense that the interface is built on the latest operated applications. The diagnostic expert system is characterized by the need to complete the entire set of information and then, by activating a simple inference engine. In this way such a solution is not justified.

Comparative analysis of [9] shows a big difference between interfaces made in universities and the commercial products. It is noted especially the complexity of products made in universities. It also noted the most abundant use of techniques and details in interfaces achieved in universities, such as graphic objects, mouse,

V.E. Balas et al. (eds.), *Soft Computing Applications*, Advances in Intelligent Systems and Computing 633, DOI 10.1007/978-3-319-62521-8_4

screen, text editing, use of hypertext and gesture recognition. One explanation might be the lack of a pertinent, realistic vision on the effective use of a man-machine interface.

Is also noteworthy that, despite the growing number of expert systems in recent years, research reported in graphical user interfaces exist, especially in 80–90 years [5,7,10].

By intent, it is distinguished paper [8] where it one put a synthesis of conceptual issues that could form the basis for human-machine interfaces. However, the main conclusion is that which would be required depends on the design methodology which does not really exist.

In the absence of methodologies to build man-machine interface for application to develop expert systems, the author believes that it is useful to approach the issue and solve it the most explicitly.

In this context, the first conclusion of the author is that for a rational use and within the limits of its intended destination as diagnostic expert systems, the man-machine interface must be designed by one who knows both the knowledge base and the way of thinking of the user. It also retains the assembly ergonomics-effective operation need-algorithmic description.

Together with the man-machine interface it is approached the developing of testing technology and the developing of test data set.

In the first part of the paper are given details of man-machine interface, including its specification. In a second part are given the test principles of expert system and further are introduced two relevant case studies for testing technology and test data set choice. Finally, one concentrates the conclusions of this approach to build the expert system for predictive diagnostics.

2 Details on Specification and Building a Man-Machine Interface for an Expert System

In general, between the expert system and the human operator there is a specific action. The interaction is guided by the man-machine interface. It is profoundly wrong to consider that an effective interface can be derived from a general purpose interface or the human-machine interface system should enjoy all the amenities of a general purpose product-program.

2.1 General Functions

The interface is a constructive detail of the expert system that gives it accessibility and, finally, the acceptability of the beneficiaries. In this situation, the interface is built having in mind hypothesis of industrial implementation, but the product is made in minimal technological conditions to ensure the operation according to the stated principles.

Functions which the interface has to provide are:

i. Reception and validation of information that characterizes the cable to be analyzed. In order to receive the requested knowledge is defined the information to be introduced by the human operator as in Table 1.

ii. The possibility that the operator can simply act to attain the expert system goal without having any special training.
iii. The possibility of software to be easily tested, especially since it is an expert system operating with complex knowledge.
iv. The possibility that the solution offered by the expert system to be justified in the sense indicated by the element of knowledge base that was considered like closest to the present analyzed case.

The expert system interacts with the operator by means of several actions:

i. entering data for the purposes of characterizing analyzed cable by measurable parameters;
ii. operative activation of commands to achieve desired actions;
iii. input date validation;
iv. reporting the parameters of analyzed cable and of resulted solution, respectively of predicted lifetime.

In Table 2 are defined selectors that should be provided for the expert system dedicated to analysis and life prediction of underground power cables.

In Table 3 are described the effects of commands and how they interact with other possible actions.

2.2 Building the Interface of Expert System

Expert system interface is designed to implement defined commands. The principle that it respects the interface is that its users have enough freedom of action in a defined framework of limitations and safeguards.

In Fig. 1 is given the expert system's main board. It is ascertained that there is a part for working parameters. It was chosen that this sub-screen have dual utility.

It is to remember that maintenance of a network of power cables assumes high costs, both in terms of identification and location fault, and from the point of view of intervention.

3 Testing and Using the Expert System

Testing an expert system is a necessity before it is placed in service. There is no single test technology. In the literature, which is recommended, in different forms [1–4], is going through all routes leading from premises to conclusions.

Such exhaustive testing is suitable for expert systems that work on a limited and predefined knowledge base.

In this case it is a different kind of expert system.

First, the knowledge base is not limited but it enriches by means of accumulated experience. Accumulation of knowledge can be done from the outside from accumulated experiences or by including situations already considered and solved.

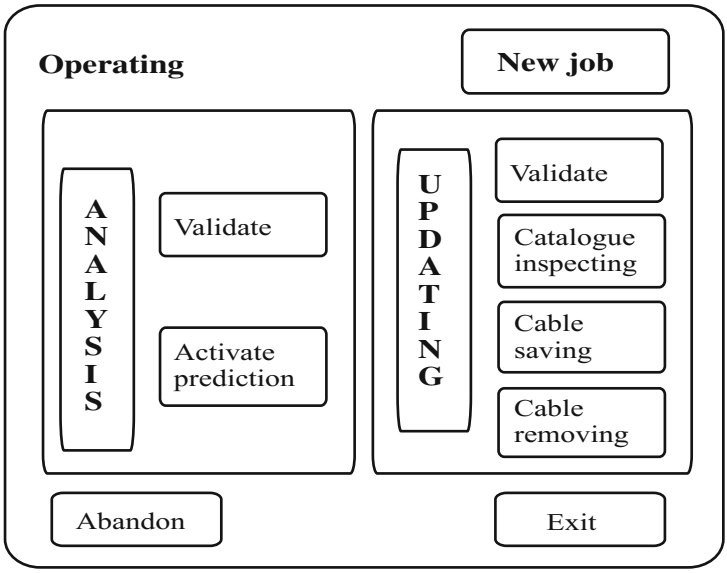


Fig. 1. Placement of commands on the operating sub-panel.

When tests an expert system it should keep in mind that behind the results are more knowledge together with the rules of use.

Components of knowledge pieces are at least of a three categories: (a) numerical with direct assignment, (b) calculated based on predefined account relationships and (c) other, basically logical.

It is noteworthy that the component parts of knowledge pieces have not identical relations with the lifetime of the analyzed cable.

This makes the testing technology still more complicated. It is noted that verification test should be performed with a large number of cases, that reference cables and checking the components parts of knowledge to be 'practiced', that is to have many values.

3.1 Organizing of Testing

In these circumstances, the test is regarded by examples. In a superficial manner, testing is exemplified by construction the tests as below:

Test 1. The cable has identical parameters with the parameters of first of the analyzed reference cables of the knowledge base. It may be noted that there are three tables that relate to cable analyzed (Table 1) which includes the solution of expert system, the knowledge base describing the set of reference cables (Table 2) and a table that allows the evaluation of details that refers to the joining of analyzed cable and the reference cable.

Table 1. Test 1 – reported information of analyzed cable.

No	Parameter	Lower	Sample	Upper
[1]	Cable type		A2YSY	
[2]	IF of the cable type on the CLE		0	
[3]	Laying technology		A	
[4]	IF of laying technology on the CLE	1.00	50.00	50.00
[5]	Cable section [mm^2]	10.00	240.00	500.00
[6]	IF of the cable section on the CLE		0.00	
[7]	Length of cable segment [m]	20.00	1,880.00	10,000
[8]	IF of the cable length on the CLE		0.00	
[9]	Insulation resistance, [MΩ] – &5 kV	0.01	0.01	1,000,000
[10]	IF of the insulation resistance on the CLE	1.00	86.95	100.00
[11]	Shield/soil insulation resistance, [MΩ] – &5 kV	0.00	0.10	300.00
[12]	Shield/soil insulation resistance IF on CLE	1.00	1.00	100.00
[13]	Shield electric continuity		Yes	
[14]	IF of shield electric continuity on CLE	1.00	100.00	100.00
[15]	Dielectric loss tangent ($tg\delta$)	0.0005	1.00	1.00
[16]	IF of dielectric loss tangent on CLE	1.00	100.00	100.00
[17]	Soil resistivity [Ω m]	1.00	4.50	100.00
[18]	IF of soil resistivity on CLE	1.00	1.00	100.00
[19]	Soil acidity [pH]	2.00	5.50	12.00
[20]	IF of soil acidity on the CLE	1.00	31.00	100.00
[21]	Stray current density [A/m^2]	0.00	0.03	1.00
[22]	IF of stray currents density on the CLE	1.00	3.67	100.00
[23]	Applied sleeves number	0	24	30
[24]	IF of applied sleeves number on CLE	1.00	1.00	100.00
[25]	Mean cable load [A]	5.00	1,200.00	3,000.00
[26]	IF of mean cable load on CLE	10.00	37.86	100.00
[27]	End 1 shield/soil voltage [$V_{Cu/^4_{CuSO}}$]	−1.00	0.15	0.50
[28]	IF of end 1 shield/soil voltage on CLE	1.00	100.00	100.00
[29]	End 2 shield/soil voltage [$V_{Cu/^4_{CuSO}}$]	−1.00	0.02	0.50
[30]	IF of end 2 shield/soil voltage on CLE	1.00	52.69	100.00
[31]	Year of cable's laying	1,930	1,975	*cy*
[32]	Estimated life expectation for the known cable, in years	0.00	23.00	50

Test 2. This test is designed to test the functioning of the expert system using external, reliable information. This information, resulting from expert judgment says that an essential factor of corrosion, in identical conditions, plays the mean cable load.

3.2 Test Results

The complexity of dependencies that underlie of built expert system raises serious test questions.

Table 2. Test 1 – database contents to be used.

No	Parameter	Ex.1	Ex.2	Ex.3	Ex.4
[1]	Cable type	A2YSY	A2YSY	A2YSY	A2YSY
[2]	IF of the cable type on the CLE	0.00	0.00	0.00	0.00
[3]	Laying technology	A	A	A	A
[4]	IF of laying technology on the CLE	50.00	50.00	50.00	50.00
[5]	Cable section [mm^2]	240.00	240.00	300.00	300.00
[6]	IF of the cable section on the CLE	0.00	0.00	0.00	0.00
[7]	Length of cable segment [m]	1,880	1,880.00	400.00	400.00
[8]	IF of the cable length on the CLE	0.00	0.00	0.00	0.00
[9]	Insulation resistance, [MΩ] – &5 kV	10.00	10.00	300.00	300.00
[10]	IF of the insulation resistance on the CLE	87.00	87.00	86.00	87.00
[11]	Shield/soil insulation resistance, [MΩ] – &5 kV	0.01	0.01	3.15	0.21
[12]	Shield/soil insulation resistance IF on CLE	100.00	100.00	16.00	38.00
[13]	Shield electric continuity	Yes	Yes	Yes	Yes
[14]	IF of shield electric continuity on CLE	100.00	100.00	100.00	100.00
[15]	Dielectric loss tangent ($tg\delta$)	1.00	0.99	0.08	0.07
[16]	IF of dielectric loss tangent on CLE	100.00	99.46	8.91	7.42
[17]	Soil resistivity [Ω m]	4.50	4.50	11.10	11.10
[18]	IF of soil resistivity on CLE	1.00	1.00	1.00	1.00
[19]	Soil acidity [pH]	5.50	5.50	6.50	6.50
[20]	IF of soil acidity on the CLE	31.00	31.00	11.00	11.00
[21]	Stray current density [A/m^2]	0.03	0.03	0.06	0.06
[22]	IF of stray currents density on the CLE	3.67	3.67	6.64	6.64
[23]	Applied sleeves number	24	24	3	7
[24]	IF of applied sleeves number on CLE	1.00	1.00	1.00	1.00
[25]	Mean cable load [A]	1,200	1,200	1,300	1,300
[26]	IF of mean cable load on CLE	37.86	67.24	41.53	41.53
[27]	End 1 shield/soil voltage [$V_{Cu/^4_{CuSO}}$]	0.15	0.15	−0.20	0.01
[28]	IF of end 1 shield/soil voltage on CLE	100.00	100.00	37.59	48.71
[29]	End 2 shield/soil voltage [$V_{Cu/^4_{CuSO}}$]	0.02	0.01	−0.27	−0.12
[30]	IF of end 2 shield/soil voltage on CLE	52.69	47.20	5.37	17.97
[31]	Year of cable's laying	1975	1975	2005	2005
[32]	Estimated life expectation for the known cable, in years	23.00	12.00	45	26

In Test 1 is checking the reproduction of a known solution.

It is ascertained that when the parameters of analyzed cable are identical to those of a cable of the set of reference cables, the lifetime of analyzed cable is the same of the reference situation of Table 3.

In the case of Test 2 one checks a particular solutions, quasi-known.

Table 3. Test 1 – details of monitoring the right operating of the expert system.

No	Parameter	Sample	Reference
[1]	Cable type	A2YSY	A2YSY
[2]	IF of the cable type on the CLE	0.00	0.00
[3]	Laying technology	A	A
[4]	IF of laying technology on the CLE	50.00	50.00
[5]	Cable section [mm^2]	240.00	240.00
[6]	IF of the cable section on the CLE	0.00	0.00
[7]	Length of cable segment [m]	1,880	1,880.00
[8]	IF of the cable length on the CLE	0.00	0.00
[9]	Insulation resistance, [MΩ] – &5 kV	10.00	10.00
[10]	IF of the insulation resistance on the CLE	87.00	87.00
[11]	Shield/soil insulation resistance, [MΩ] – &5 kV	0.01	0.01
[12]	Shield/soil insulation resistance IF on CLE	100.00	100.00
[13]	Shield electric continuity	Yes	Yes
[14]	IF of shield electric continuity on CLE	100.00	100.00
[15]	Dielectric loss tangent ($tg\delta$)	1.00	1.00
[16]	IF of dielectric loss tangent on CLE	100.00	100.00
[17]	Soil resistivity [Ω m]	4.50	4.50
[18]	IF of soil resistivity on CLE	1.00	1.00
[19]	Soil acidity [pH]	5.50	5.50
[20]	IF of soil acidity on the CLE	31.00	31.00
[21]	Stray current density [A/m^2]	0.03	0.03
[22]	IF of stray current density on the CLE	3.67	3.67
[23]	Applied sleeves number	24	24
[24]	IF of applied sleeves number on CLE	1.00	1.00
[25]	Mean cable load [A]	1,200	1,200
[26]	IF of mean cable load on CLE	37.86	37.86
[27]	End 1 shield/soil voltage [$V_{Cu}/^4_{CuSO}$]	0.15	0.15
[28]	IF of end 1 shield/soil voltage on CLE	100.00	100.00
[29]	End 2 shield/soil voltage [$V_{Cu}/^4_{CuSO}$]	0.02	0.02
[30]	IF of end 2 shield/soil voltage on CLE	52.69	52.69
[31]	Year of cable's laying	1975	1975
[32]	Estimated life expectation for the known cable, in years	23.00	23.00

By the construction of this situation, for the same knowledge base the lifetime value is significantly influenced as in Table 4. Here only the significant rows are presented.

Table 4. Test 2 – details of monitoring the right operating of the expert system.

No	Parameter	Sample	Reference
...	...	...	...
[25]	Mean cable load [A]	2,500	1,200
...	...	...	...
[32]	Estimated life expectation for the known cable, in years	12.69	23.00

The above two situations are part of the tests battery that was built to validate the achieved expert system.

From this point of view, it is considered that the assessment succeeded.

Regarding the information in Table 1 is to remember some details:

i. The introduction of information on the analyzed cable involves both objective information, and entering the values of influence factors.
ii. Information on all parameters of cable, inclusively of influence factors, may be partially in the sense that the expert system assigns defaults. In this way the system can cope with imprecise information.

The existence of influence factors covers that knowledge of the expert that is doubtful through ignorance regarding the involved physic-chemical phenomena.

Different values of influence factors in Table 1 indicate that referred specimens of cables are coming from different sources. There, the specialists who assessed them found that they could be the values that they are able to support them.

The size of reference cable set is a guarantee of prediction quality. It is obvious that the predicted value is among the available knowledge pieces. Clearly the highlight of the weight to affect each knowledge piece in the final outcome is difficult. By the nature of the tasks undertaken testing, the research that might be necessary will be avoided.

Test 2 performed mainly based on information from Table 4, makes a check that covers both the theoretical used basis and how to implement a system expert.

Construction of the test is based on the assumption that doubling the cable mean load should lead to a half of lifetime of the cable if other parameters are not changed.

Tables 3 and 4 have the role of working tools. Together with Table 1 they form the testing technology of the expert system.

The task of designing these documents devolve upon the project group. Despite many aspects of fundamental research, these documents are part of professional product documentation and once built the product, it must be sent to the end user.

Testing will not be completed with technical reception of expert system.

The testing must be repeated after a period of time agreed with beneficiary and must be extended to newly entered knowledge pieces whether they are analyzed cables whether they are the results of laboratory research or experimental results.

It requires periodic statistical checking regarding the predictions and results confirming the favorable results.

Testing can be completed with achievement of graphics with intermediate possible values (Table 5).

Table 5. Knowledge associated to power cables and manner to treat the information inside the expert system.

No	Explanation about the information			
	Name heading comments	Lower	Field to take the value	Upper
[1]	Cable location			
[2]	Cable type		Selector $S1$	
[3]	Laying technology		Selector $S2$	
[4]	Cable section [mm^2]	10		500.00
[5]	Length of cable segment [m]	20		10,000
[6]	Insulation resistance, [$M\Omega$] − &5 kV	10		10
[7]	Shield/soil insulation resistance, [$M\Omega$] − &5 kV	0.01		30
[8]	Shield electric continuity		Selector $S3$	
[9]	Dielectric loss tangent ($tg\delta$)	0.00005	Selector $S3$	1.00
[10]	Soil resistivity [Ω m]	1.00		100.00
[11]	Soil acidity (pH)	2.00		12.00
[12]	Stray currents density [A/m^2]	0.00		1.00
[13]	Applied sleeves number	0	Selector $S4$	30
[14]	Mean cable load [A]	5		3,000
[15]	End 1 shield/soil voltage [$V_{Cu/CuSO4}$]	−1.00		0.50
[16]	End 2 shield/soil voltage [$V_{Cu/CuSO4}$]	−1.00		0.50
[17]	Year of cable's laying	1930	Selector $S5$	cy
[18]	Estimated life expectation for the known cable, in years	0	Selector $S6$	50

On the one hand sub-screens of work parameters is used for inspection, maintenance and updating of the knowledge base. On the other hand, during the analysis, the same sub-screens will provide the final result on the position of the predicted lifetime (Table 6).

In Fig. 2 one gives the operating sub-screen. This is divided in three key areas. The first area is the introduction of a new reference cable or advance under the existing cable base during maintenance. The other two areas correspond to the analysis commands for updating and maintaining the database of reference cables. The third area is for mandatory commands i.e. abandon of last activated command and exit command (Table 7).

Table 6. Defining selectors of man-machine interface of the system expert.

Description of interface selectors	
S1 - cable type selector	
Values to be selected:	A2YSY, A2xS2Y, A2XSR2Y, A2XSRY-B
Default state:	A2YSY
Notes:	Does'nt exist
S2 - laying technology selector	
Values to be selected:	{Buried directly in soil, buried in protective coating}
Default state:	Buried directly in soil
Notes:	Does'nt exist.
S3 - electric continuity of cable shield	
Values to be selected:	{Yes, No}
Default state:	Yes
Notes:	Does'nt exist.
S4 - selector of sleeves number	
Values to be selected:	$[0, 30]$
Default state:	0
Notes:	Does'nt exist.
S5 - cables laying year selector	
Values to be selected:	1930, 1931, 1932, $\cdots$ cy
Default state:	(cy 5)
Notes:	With cy is denoted the current year.
S6 - selector of known operating time	
Values to be selected:	{0, 1, 2, $\cdots$ 50}
Default state:	5
Notes:	Selector is active only for the parameters updating stage

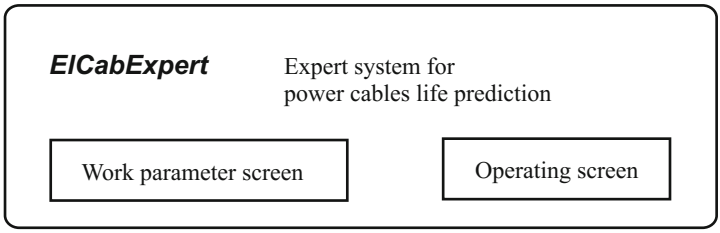


Fig. 2. Organizing the board of expert system.

It is to emphasize that one takes into account that the two work commands regarding analyze and updating are not commands on the same hierarchical rank.

Table 7. Defining man-machine interface commands of the system expert.

Command/description
New job - Default command, active in initial state - Allows the acting of commands to set the work mode, that is **Analysis** for the mode of **Power cable analysis**, and respectively **Updating** for the work mode **Cable basis updating** and also of the general commands **Abandon** and **Exit** - Do not allow parameter fields filling with values until the activation - Allows repeated activation of command - Bring the system in the specific state of operating mode **Cable analysis** - Do not allow activating of specific commands of **Activate prediction**, and/or **Cable saving** and **Catalog inspection**
Analysis - Default command, active in initial state - Defines the work mode **Power cable analysis** - Allows the activating of specific commands **Activate prediction** and of general commands **Abandon** and **Exit** - Do not allow the activating of specific commands of **Cable base updating** work mode namely **Cable saving** and **Catalog inspection** - Allows parameters entering for the analysis cable but do not allow the fulfill of the field of known lifetime of the cable operating **Estimated lifetime**
Updating - Default command, passive in initial state - Defines the work mode of **Cable base updating** - Allows the activating of specific commands **Cable saving** and **Catalog inspection** and of general commands **Abandon** and **Exit** - Do not allow the activating of specific command of **Power cable analysis** work mode namely **Activate prediction** - Allows parameters entering for the analysis cable, inclusively the fulfilling of the field of known lifetime of the cable operating estimated lifetime
Power cable analysis/**Activate prediction** - Default command, active in initial state - Can be activated only successfully activated of **Power cable analysis**/**Validate command** - Replaced the inference engine and displays the estimate lifetime in its field
Power cable analysis/Entering the parameter values /**Abandon** - Always asks for the acknowledgement to execution *Do you want to abandon? (Yes/No)* - Once confirmed, the system passes to the state activated by **Analysis**
Cable updating/Entering the parameter values/**Abandon** - Always asks for the acknowledgement to execution *Do you want to abandon? (Yes/No)* - Once confirmed, the system passes to the state activated by **Updating**
Exit - Always asks for the acknowledgement to execution *Do you want to exit? (Yes/No)* - Once confirmed the operator quit the expert system without any saving

This means that these commands can not be accessed by the other person than having administrator authority. This specificity is resolved by the existence of an access password. By recognizing the password, the operator can access distinctly analysis command or analysis and the updating commands.

4 Testing and Using the Expert System

Testing an expert system is a necessity before it is placed in service. There is no single test technology. In the literature, which is recommended, in different forms [1,3], is going through all routes leading from premises to conclusions.

Such exhaustive testing is suitable for expert systems that work on a limited and predefined knowledge base.

In this case it is a different kind of expert system.

First, the knowledge base is not limited but it enriches by means of accumulated experience. Accumulation of knowledge can be done from the outside from accumulated experiences or by including situations already considered and solved.

When tests an expert system it should keep in mind that behind the results are more knowledge together with the rules of use.

Components of knowledge pieces are at least of a three categories: (a) numerical with direct assignment, (b) calculated based on predefined account relationships and (c) other, basically logical.

It is noteworthy that the component parts of knowledge pieces have not identical relations with the lifetime of the analyzed cable.

This makes the testing technology still more complicated. It is noted that verification test should be performed with a large number of cases, that reference cables and checking the components parts of knowledge to be 'practiced', that is to have many values.

5 Concluding Remarks

Once conceptually done, an expert system should be achieved as inference engine but also as human-machine interface. This interface comes to provide to users access to obtain right answers to questions.

Achieved interface put together necessary functions and the facility of receiving information on the knowledge pieces that are then exploited. Since the users have different authority one creates the user with access to all the information, i.e. including the rights to update the knowledge base, and an ordinary user who has access only to solve key issues that prediction lifetime of underground power cables.

Such information is used in development projects of the power supply network.

It is grounded and exemplified the testing technology of achieved expert system.

From the tests set that have been designed and implemented, in work are given two tests. The first test is the reproduction of the identical case and the second is to obtain predictable solution based on the information that comes from the expert in problems of corrosion.

References

1. Balci, O., Smith, E., O'Keefe, R.: Validation of expert system performance. Technical report TR-86-37, Department of Computer Science, Virginia Polytechnic Institute and State University (1986). eprints.cs.vt.edu
2. Bobrow, D., Mittal, S., Stefik, M.: Expert systems: perils and promise. Commun. ACM **29**(9), 880–894 (1986)
3. Chang, C., Stachowitz, R.: Testing expert systems. Technical report (1988). ntrs.nasa.gov
4. Heckerman, D., Horvitz, E., Nathwani, B.: Toward normative expert systems: the Pathfinder project. Methods Inf. Med. **31**, 90–105 (1991)
5. Hoc, J.: From human-machine interaction to human-machine cooperation. Ergonomics **43**(7), 833–843 (2000)
6. Hoffberg, S., Hoffberg-Borghesani, L.: Ergonomic man-machine interface incorporating adaptive pattern recongnition based control system. US Patent 6,418,424 (1999)
7. Jacob, J.: Using formal specifications in the design of a human-computer interface. Commun. ACM **26**(4), 259–264 (1983)
8. Johannsen, G., Levis, A., Stassen, H.G.: Theoretical problems in man-machine systems and their experimental validation. Automatica **30**(2), 217–231 (1994)
9. Myers, B.: A brief history of human-computer interaction technology. Interactions **4**(2), 44–54 (1984)
10. Puerta, A., Egar, J., Tu, S., Musen, M.: A multiple-method knowledge-acquisition shell for the automatic generation of knowledge-acquisition tools. Knowl. Acquis. **4**(2), 171–196 (1992)

GA Based Multi-stage Transmission Network Expansion Planning

A. Simo, St. Kilyeni, and C. Barbulescu(✉)

Power Systems Department, Power Systems Analysis and Optimization Research Center, Politehnica University Timisoara, 2, Pta. Victoriei, Timisoara, Romania
{attila.simo, stefan.kilyeni, constantin.barbulescu}@upt.ro

Abstract. The paper is focusing on dynamic transmission network expansion planning (DTNEP). The DTNEP problem has been approached from the retrospective and prospective point of view. To achieve this goal, the authors are developing two software tools in Matlab environment. Power flow computing is performed using conventional methods. Optimal power flow and network expansion are performed using artificial intelligence methods. Within this field, two techniques have been tackled: particle swarm optimization (PSO) and genetic algorithms (GA). The case study refers to a real power system modeled on the Center, Northern, Eastern and Southern parts of the Romanian Power System.

Keywords: Power transmission · Power system planning · Dynamic expansion · Software-tool

1 Introduction

The dynamic transmission network expansion planning (DTNEP) is discussed within this paper, in retrospective and prospective manner. A set of network expansion candidates are proposed. The power flow is performed using conventional methods for all scenarios. The optimal power flow (OPF) is computed for the maximum expansion solution (including all the expansion scenarios) using particle swarm optimization (PSO) and genetic algorithms (GA). Having the optimal maximum expansion solution, the optimal expansion solution is computed also using PSO and GA. For all these purposes own software tools have been developed in Matlab environment. They are able to be linked with other well-known computer aided power system analysis software, importing the power system database. Two types of GAs are used within this paper. Binary coded GA for the expansion planning stage and real coded GA for the OPF.

Dynamic programming represents an optimal solution selection methodology considering specific constraints, following a step-by-step decision process [1, 2]. Discrete dynamic programming with finite horizon is discussed within the current paper. Decisions are taken at specific time moments, following a finite number of computing steps.

Currently, within the power engineering field, the heuristic and meta-heuristic methods are widely used to solve optimization problems. In case of transmission network expansion planning (TNEP), they are used to generate possible solutions,

V.E. Balas et al. (eds.), *Soft Computing Applications*, Advances in Intelligent Systems and Computing 633, DOI 10.1007/978-3-319-62521-8_5

evaluation and selection of candidates, until the algorithm is not able to find a better solution. A reduced mathematical model is their main characteristic. The major drawback refers to the difficulties for tuning their parameters.

A multi-stage approach of the problem under discussion is tackled in [3]. Authors are proposing a new constructive heuristic approach considering security-constrained TNEP. It is based on a local random search used to select the values of the control variables. It is applied for Ecuadorian and Chilean Power Systems. The major drawback refers to the use of a DC model.

In [4] the TNEP is solved as a mixed integer nonlinear non-convex programming model. The optimization is performed based on the differential evolutionary algorithm (DEA). The use of AC load flow model is the great benefit of the paper, providing realist and accurate results. The technique is tested on Garver's 6 bus and IEEE 24 bus test system.

GA based TNEP multi-stage solving is proposed in [5]. The GA is used in conjunction with the probabilistic optimization. The objective function refers to the investment costs, absorption of private investment and system reliability. The case study refers to the IEEE 24 buses test power system. The wind generation is taken into consideration. The TNEP problem influenced by the uncertainties related to wind power generation is also investigated in [6]. In this case the Cuckoo search algorithm is used as an optimization tool.

As is stipulated in [7] the use of AC mathematical model has been proposed by only very few researchers. Its complexity is augmented if it is desired to be tackled in the smart grids context. The problem is solved considered the N−1 security criterion. DEA is used and tested on Garver's 6 bus test power system.

In [8] two sources of uncertainties are analyzed in conjunction with TNEP: intermittent renewable energy generation and loads. The objective function terms are represented by the investment and operating costs. The adaptive tabu search algorithm is applied. The AC load flow mathematical model is used and interior point non-linear programing. IEEE 79 buses RTS test system is used as case study.

In [9] a comparison is performed between the use of energy storage and N-1 criterion in TNEP studies. The problem is statically approached considering energy storage model, within a system with base load generators and peaking power plants. Four TNEP models are proposed for comparative analysis and N-1 network security constraints. The problem is modeled as a mixed integer linear programming problem. IEEE 24 RTS test system is used as case study.

A static TNEP mathematical model is proposed in [10]. DC power flow is used. The problem is solved with and without generated power rescheduling. Artificial bee colony is used as an optimization tool for small IEEE test systems.

A long-term TNEP mathematical model is proposed in [11], based on balancing investment costs and reducing consumer costs. To achieve this goal a hierarchical framework is proposed, being sensitive to different agents operating on diverse timelines. An equilibrium model is introduced combining grid operating concerns with the short-term competitive behavior of generating units.

According to [12] the use of robust optimization techniques is appreciated compared with stochastic mathematical programming methods in TNEP. Several uncertainties are considered (in order to obtain more realistic results). The mathematical

model is represented by three-level mixed-integer optimization problem, solved using different strategies. IEEE 24 and 118 buses test systems are used.

The discrete PSO approach is used in [13] to solve the TNEP problem. Non-linear mixed optimization model is used with DC power flow. Case studies are represented by IEEE 6 and 24 buses test systems and 46 buses power Brazilian system.

[14] deals with DTNEP. The time evolution of the network expansion is provided considering a large number of technical and economic constraints. Results are provided only in synthesis, referring to the power system of Vietnam.

Following the introduction presented within the 1st section, the 2nd one refers to the DTNEP methodology. The 3rd section deals with TNEP solving using GAs. The developed software tools are briefly presented within the 4th section. The case study is tackled within the 5th section. The 6th one is focusing on results discussing. Finally, conclusions are synthesized within the 7th section.

2 Dynamic Network Expansion Planning

The DTNEP is discussed for the following time steps: 2015 year– initial stage, 2020, 2025, 2030 – intermediary stages and 2035 – final stage. The prospective approach (forward direction) and retrospective one (backward direction) has been considered.

Two issues have to be solved:

- consumed power forecast correlating the power generation capacity;
- admissible solution domain definition – it contains the network elements' list that are allowed to be part of the optimal solution for the final stage (2035 year).

According to the prospective analysis, the starting point refers to the 2015 year. In the following, the expansion solutions are computed step-by-step for the successive years: 2020, 2025, 2030 and 2035. The provided results for 2035 year represent the final solution for the entire 20 years analyzed period.

The admissible solutions' domain has been considered to be the maximum expansion one extracting the network elements already introduced for each expansion stage.

A static expansion planning solving is applied for each intermediary stage. The non-linear optimization problem is solved using evolutionary techniques: PSO and GA.

According to the retrospective analysis, the starting point is represented by the maximum expansion solution, year 2035. In the following, the expansion solutions are computed step-by-step for the successive years: 2035, 2030, 2025 and 2020. The results obtained for 2035 year represents the final solution. The comments provided at the prospective analysis for the admissible solution domain definition and static expansion solving at each intermediary stage are suitable for this case too.

The use of both approaches offers the advantage of comparing the intermediary solutions (2020, 2025, 2030 years) and, especially, the final one (2035 year).

Mathematical model for transmission network expansion planning is presented. Two artificial intelligence solving techniques have been tackled: PSO and GA.

The optimization problem has a multi-criteria character. The following components are included within the global objective function:

The optimization problem has a multi-criteria character. The following components are included within the global objective function:

- power system operating costs:

$$OBF_1 = \sum_{i \in G} C_i(P_{g\,i}) + \sum_{ij \in R} TP_{ij}(S_{ij} - S_{ij}^{**}) = Minim \tag{1}$$

where: G – set of PV buses; R – set of network elements; $C_i(P_{gi})$ – quadratic cost characteristics of generators – $C_i(P_{gi}) = a_i \cdot P_{gi}^2 + b_i \cdot P_{gi} + c_i$; TP_{ij} – penalty cost of the apparent power upper limit exceed trough the ij network element; S_{ij}^{**} is defined as $S_{ij}^{**} = \begin{cases} S_{ij} & \text{if} \quad S_{ij} \le S_{ij}^{max} \\ S_{ij}^{\max} & \text{if} \quad S_{ij} > S_{ij}^{max} \end{cases}$; S_{ij} – apparent power flow for ij network element; $S_{ij}^{max} - S_{ij}$ thermal limit;

- investment equivalent yearly cost related to new power transmission capacities (overhead lines, autotransformers) – OBF_2;
- safety operation, quantified based on risk factor computing:

$$OBF_3 = r^{\%} = \frac{\sum\limits_{k=1}^{n_\ell} q^k \cdot r^k}{\sum\limits_{k=1}^{n_\ell} q^k} = \frac{\sum\limits_{k=1}^{n_\ell} q^k \cdot P_r^k \left\{ \left| S_{ij}^k \right| > S_{ij}^{\max}, ij \in R \right\}}{\sum\limits_{k=1}^{n_\ell} q^k} \cdot 100 \tag{2}$$

where: q^k – k overhead line (OHL) disconnection probability; n_l – overhead lines' number that are selected for contingencies; S_{ij}^k – S_{ij} in case of k element disconnection; r^k – congestion probability when k OHL is disconnected.

- total available transmission capacity.

$$OBF_4 = TATC = \sum_{\substack{ij \in L \\ |S_{ij}| < S_{ij}^{\max}}} \left(S_{ij}^{\max}) - \left| S_{ij} \right| \right) \tag{3}$$

The function that has to be minimized is represented by the sum of the suitable weighted four components previously discussed (w_i – weighting coefficients):

$$OBF = w_1 \cdot OBF_1 + w_2 \cdot OBF_2 + w_3 \cdot OBF_3 + w_4 \cdot OBF_4 \tag{4}$$

3 Genetic Algorithm (GA) Based TNEP

The population represents a set of possible solutions. Each chromosome contains binary digits (0, 1), representing the state for the network expansion candidates. Thus, for this stage we are dealing with binary coded GA.

The chromosome has d length, being able to be written as:

$$\boldsymbol{x_i} = \{x_{i1}, x_{i2}, \ldots, x_{id}\}, \quad i = 1, 2, \cdots, n_c \tag{5}$$

Each chromosome is evaluated based on the OBF objective function. The computing process finishes if the solution is not able to be improved.

The algorithm stages are the following ones:

(a) chromosomes forming the population are randomly initialized with 0 and 1:

$$\boldsymbol{x}_i^0 = \{x_{i,1}^0, x_{i,2}^0, \ldots, x_{i,d}^0\}, \quad i = 1, 2, \cdots, n_p \tag{6}$$

(b) GA based OPF is computed for the configuration coded by each chromosome;
(c) initial population is evaluated based on OBF value. The best chromosome is saved in $\boldsymbol{x}_{elit}^0$:

$$f(\boldsymbol{x}_{elit}^0) = \min\{OBF(\boldsymbol{x}_i^0)\}, \ i = 1, 2, \ldots, n_c \tag{7}$$

(d) for a specific t computing step (t = 0, 1, 2,…) the chromosomes forming the population subjected to recombination are selected;
(e) $n_{pr} = \chi \cdot n_c/2$ chromosome pairs that are subjected to crossover are formed and chromosome pairs that are going to be copied unaltered;
(f) offspring are formed starting from the n_{pr} pairs subjected to crossover (in one point or in several points);
(g) number of chromosome genes subjected to mutation is computed;
(h) 1st chromosome belonging to the population obtained at previous step is replaced with the best of the old population:

$$\boldsymbol{x}_1^{t+1} = \boldsymbol{x}_{elit}^t \tag{8}$$

(i) OPF is computed for the configurations coded by each chromosome. Current population is evaluated based on OBF value. New x_{elit} value is computed:

$$f(\boldsymbol{x}_{elit}^{t+1}) = \min\{FOB(\boldsymbol{x}_i^{t+1})\}, \qquad i = 1, 2, \ldots, n_c \tag{9}$$

(j) if the OBF value is not able to be improved for a given number of steps, the computing process finishes. The operating condition corresponding to the last x_{elit} value represents the optimal one. Contrary, computing step is increased with 1 and the algorithm is repeated starting with point c).

4 Software Tool

Two software tools have been developed by the authors in Matlab environment based on PSO and GA approaches: Power OptPowerplanPSO and PowerOptPowerPlanGA. Each one has two modules linked through a graphical user interface:

- 1st module – used for OPF (also is able to be used as a stand-alone module);
- 2nd module – used for dynamic transmission network expansion planning.

PowerOptPowerplanGA's main window for the OPF module is presented in Fig. 1. Once the power system database has been loaded, the user is able to select the optimization type he desires by selecting the control variables. The lower part of the window allows the user to set the GA parameters: maximum computing steps, capping iterations, population size etc.

The main window for the software tool 2nd module (expansion planning module) is presented in Fig. 2. The user has to set the parameters for the GA algorithm: selection type, recombination type etc.

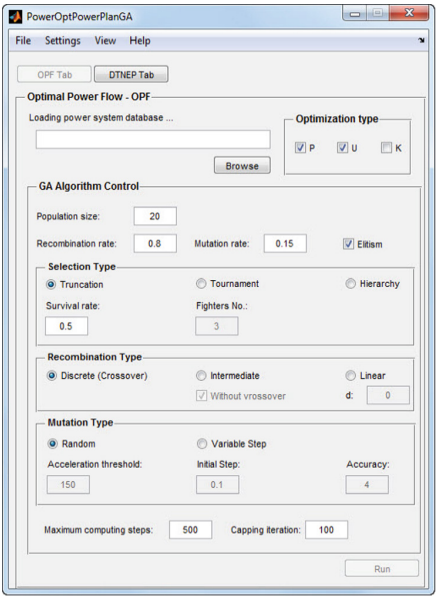


Fig. 1. PowerOptPowerplanGA – OPF module main window

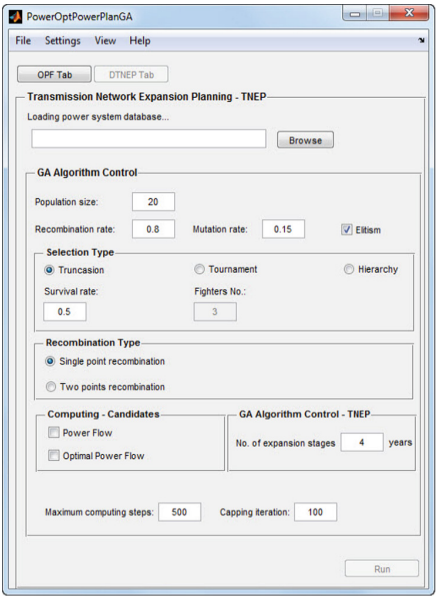


Fig. 2. PowerOptPowerplanGA – dynamic network expansion window

Computing-candidates tab allows the user to specify if for each of the power system configurations corresponding to different expansion candidates, OPF is computed or only power flow. Button labeled *Number of expansion stage [years]* allows the user to specify the desired number of expansion stages. The entire expansion planning time horizon is going to be divided into the specified number of years.

The results are provided in several ways. They are exported in different file formats (according to the user's desire), view on the display or graphically displayed.

The 2nd software tool based on PSO is named PowerOptPowerPlanPSO. This one is not discussed within the paper.

5 Case Study

The power system used as case study has been modeled based on the Center & Northern & Eastern & Southern parts of the Romanian Power System. The other parts (Western, South-Western and North-Western) have been eliminated. The one-line diagram is presented in Fig. 3 having the following characteristics:

- number of buses – 110 (30 PV buses and 80 PQ buses);
- number of network elements – 149 divided as: 105 overhead lines of 110, 220 and 400 kV, 44 (auto)transformers;
- number of generating units – 30, 23 being real ones and 7 equivalent ones.

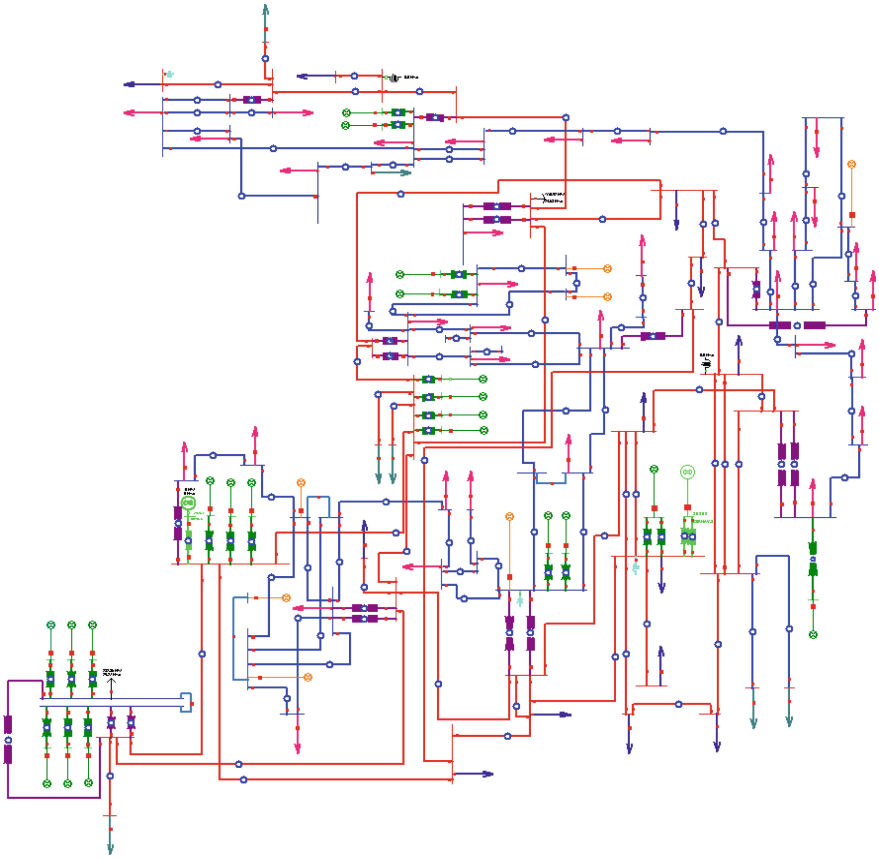


Fig. 3. Case study one-line diagram

6 Results and Discussions

The dynamic transmission network expansion planning is performed (retrospective and prospective approach). This process is step-by-step discussed within the following subsections.

6.1 Base Case – Year 2015

Base case has been computed using conventional methods. Bus voltages are ranging between 0.95 and 1.10 p.u. (110 kV, 220 kV), respectively 1 and 1.05 p.u. (400 kV). PV buses terminal voltage limits are set between 0.95 and 1.15 p.u.

The total consumed power is P_c = 4172.8 MW, the real generated power P_g = 4289.7 MW and real power losses ΔP = 116.9 MW.

6.2 Maximum Expansion Solution – Year 2035

The transmission network expansion planning is discussed for a 20 years, based on the last year consumed power forecast.

The 2035 year forecasted peak-evening-winter operating condition is characterized by total real consumed power of 7473 MW (in comparison with 4172.8 MW for the base case), respectively reactive power 2945 MVAr (compared with 1999.9 MVAr).

For load fulfilment the following generating units have been introduced:

- 330 MW units at 28061 (2), 28904 (2), 28021 (3) 28022 (2) buses;
- 220 MW units at 28077 (2) bus;
- wind farms injecting power in 28017 (600 MW), 28974 (200 MW), 28019 (160 MW) and 28080 (100 MW) buses.

A number of 23 new transmission network elements (20 OHLs and 3 autotransformers) have been introduced (considered as candidates within the expansion list). Thus, the expansion scenario is the following one:

- new 400 kV buses at 28040, 28062 and 28908;
- 400/220 kV 400 MVA (auto)transformers in substations 28061–28062 (2 units) and 28907–28908;
- new 400 kV OHL 28037–28040, 28040–28082, 28031–28908, 28908 – 28016;
- 220 kV OHL upgraded to 400 kV: 28082–28950, 28950–28025, 28905–28906, 28906–28908;
- 400 kV OHL additional circuit at 28016–28973, 28022–28024;
- 220 kV OHL additional circuit at 28045–28061, 28061–28057, 28057–28055, 28087–28086 (2), 28086–28085, 28085–28084, 28084–28083, 28083–28078, 28078–28024.

Using the software tool PowerOptPlanGA the optimal power flow – OPF has been computed for the maximum expansion solution. Real power losses are equal to ΔP = 191.8 MW, compared to 225.1 MW for the maximum expansion solution base case. Thus, about 15% decreasing has been recorded.

The GA algorithm evolution for OPF is presented in Fig. 4. The OBF individual best values are presented using blue colour for each computing step. The OBF average value for the entire population is represented using green colour. The OBF value, corresponding to the worst solution, is represented using red colour.

An accentuated decrease of the OBF value is recorded during the 1st 50 computing steps. The solution is slightly improved for the next computing steps.

Although we are dealing with a large scale power system, the number of the computing steps required for the solution computing is reduced. This fact proves the correctness of the GA parameters' tuning.

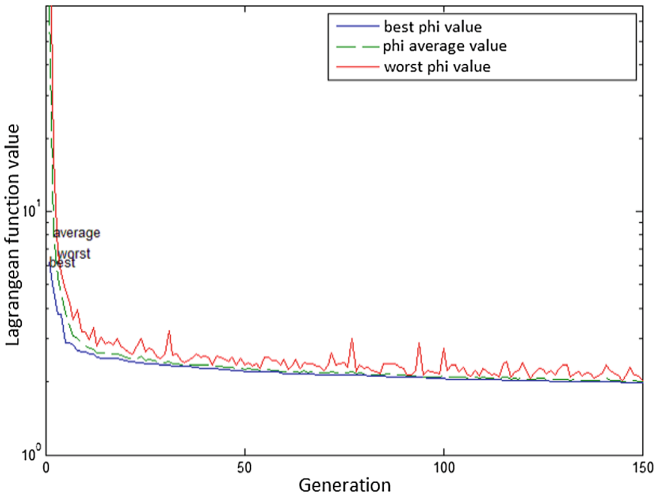


Fig. 4. GA algorithm evolution for OPF computing

6.3 Optimal Expansion Solution

- Retrospective approach

Before TNEP the load forecast for 20 years period (2015–2035) has been performed. The results are presented in Table 1.

Table 1. Consumed power forecasting

	$P_{c\ total}$ [MW]	$Q_{c\ total}$ [MVAr]
2015	4172.8	1999.9
2020	4827.2	2203.1
2025	5584.2	2426.9
2030	6459.9	2673.4
2035	7473.0	2945.0

The starting point is represented by the maximum expansion solution, year 2035 and the expansion solutions for all the stages: 2035, 2030, 2025 and 2020.

The solution admissible domain for each dynamic expansion stage has been defined based on the previous results.

For DTNEP the PowerOptPowerPlanGA software tool has been used. The following elements have been found for the optimal expansion solution:

- new 400 kV buses at 28040, 28062 and 28908;
- 400/220 kV 400 MVA (auto)transformers in substations 28061–28062 (2 units) and 28907–28908;
- new 400 kV OHL 28037–28040, 28040–28082 (alternative solution – double circuit 220 kV OHL), 28031–28908, 28908–28016;
- 220 kV OHL upgraded to 400 kV: 28082–28950, 28950–28025;
- 220 kV OHL additional circuit at 28061–28057, 28057–28055, 28087–28086 (2), 28086–28085, 28083–28078, 28078–28024 buses.

The optimal expansion solution is characterized by 124 buses (41 PV buses, 83 PQ buses) and 176 network elements (118 OHL, 58 transformers, autotransformers).

The relative OBF value (Fig. 5) has been computed being the ratio between the current expansion solution OBF and the one corresponding to the maximum expansion solution. The OBF value permanently improves during the algorithm evolution.

This fact is explained due to the power system scale and increased number of transmission network expansion candidates. For the 1st three computing steps 17.5% OBF decreasing value is recorded and additionally 1.2% for the following ones.

The dynamic retrospective transmission network expansion planning results are synthesized in Table 2.

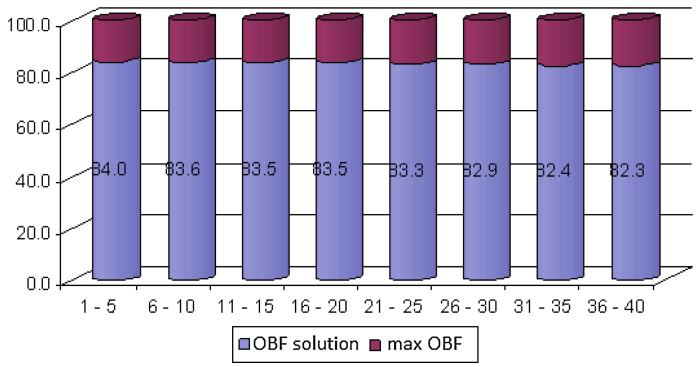


Fig. 5. GA based OBF evolution

They are several quasi-optimal solutions, having close OBF values. As an alternative solution, the 28037–28040, 28040–28082 400 kV OHL's are replaced by d.c. 220 kV OHL's.

Table 2. Retrospective analysis results

	2020	2025	2030	2035
	3 from 20	7 from 20	12 from 20	15 from 20
OHL	–	–	–	400 kV 28037–28040
	–	–	–	400 kV 28040–28082
	–	–	400 kV 28031–28908	400 kV 28031–28908
	–	–	400 kV 28908–28016	400 kV 28908–28016
	–	–	400 kV 28016–28973	400 kV 28016–28973
	–	400 kV 28022–28024	400 kV 28022–28024	400 kV 28022–28024
	400 kV 28082–28950	400 kV 28082–28950	400 kV 28082–28950	400 kV 28082–28950
	400 kV 28082–28950	400 kV 28082–28950	400 kV 28082–28950	400 kV 28950–28025
	220 kV 28057–28055	220 kV 28061–28057	220 kV 28061–28057	220 kV 28061–28057
	–	220 kV 28057–28055	220 kV 28057–28055	220 kV 28057–28055
	–	220 kV 28087–28086	220 kV 28087–28086	220 kV 28087–28086
	–	–	–	220 kV 28087–28086
	–	220 kV 28086–28085	220 kV 28086–28085	220 kV 28086–28085
	–	–	220 kV 28083–28078	220 kV 28083–28078
	–	–	220 kV 28078–28024	220 kV 28078–28024

- Prospective approach

The starting point is represented by the initial situation corresponding to the 2015 year. Expansion solutions for each future stage are computed step-by-step (2020, 2025, 2030, 2035 years). Results obtained for 2035 year are representing the final solutions for the 20 years' analyzed period.

The admissible solutions' domain has been considered to be the one defined by the maximum expansion solution, excluding the network elements already introduced at each stage of the prospective dynamic expansion.

The dynamic prospective transmission network expansion planning results are synthesized in Table 3.

6.4 Comments on the Expansion Solution

Comparing the results gathered from both approaches the following conclusions are highlighted:

- 2020 year solution is different for the two approaches, but it has very close OBF value. For the prospective approach the 28057–28055 and 28087–28086 220 kV OHLs have been considered as double circuit, resulting 5 new network elements (instead of 3 in the retrospective one);
- a similar comment is suitable for 2025 year solution – the total number of new network elements is 8 in the prospective approach (7 in the retrospective one). The difference refers to the 28016–28973 220 kV OHL which has been considered already as double circuit (in the retrospective solution only in 2030);

Table 3. Prospective analysis results

	2020	2025	2030	2035
	5 from 20	8 from 20	13 from 20	15 from 20
OHL	–	–	–	400 kV 28037–28040
	–	–	–	400 kV 28040–28082
	–	–	400 kV 28031–28908	400 kV 28031–28908
	–	–	400 kV 28908–28016	400 kV 28908–28016
	–	400 kV 28016v28973	400 kV 28016–28973	400 kV 28016–28973
	–	400 kV 28022–28024	400 kV 28022–28024	400 kV 28022–28024
	400 kV 28082–28950	400 kV 28082–28950	400 kV 28082–28950	400 kV 28082–28950
	400 kV 28082–28950	400 kV 28082–28950	400 kV 28082–28950	400 kV 28950–28025
	220 kV 28057–28055	220 kV 28061–28057	220 kV 28061–28057	220 kV 28061–28057
	220 kV 28057–28055	220 kV 28057–28055	220 kV 28057–28055	220 kV 28057–28055
	220 kV 28087–28086	220 kV 28087–28086	220 kV 28087–28086	220 kV 28087–28086
	–	–	220 kV 28087–28086	220 kV 28087–28086
	–	220 kV 28086–28085	220 kV 28086–28085	220 kV 28086–28085
	–	–	220 kV 28083–28078	220 kV 28083–28078
	–	–	220 kV 28078–28024	220 kV 28078–28024

- for 2030 year solution the total number of new network elements is 13 in the prospective approach (instead of 12 in the retrospective one). The difference refers to the 28087–28086 220 kV OHL which has been considered already as double circuit (in the retrospective solution in 2035);
- the final expansion solution is the same for the both approaches, the intermediary stages being slightly different.

7 Conclusions

The developed software tools are able to be used in case of large scale, complex transmission networks. They behave as hybrid software tool, the PSO and GA techniques being used for the OPF and network expansion stages. PSO and GA have been adapted to the power engineering field.

The prospective and retrospective dynamic expansion approaches are providing identical solutions (for the last year). For both cases other expansion solutions, additional to the optimal one, have been proposed. This information is helping the network planner to adopt the best planning scenario.

However, some difference may appear during the intermediary stages (years). In case of the prospective approach, there are situations when the process is started from an initial solution less or more different from the retrospective approach. But, the final results are the same. There are also situations when the solutions for the 1st and final expansion stage (year) are the same, but intermediary ones (2025, 2030 years) are different.

References

1. Bellman, R.E., Dreyfus, E.S.: Applied Dynamic Programming. Princeton University Corporation, Princeton (1962)
2. Sniedovich, M.: Dynamic Programming. Marcel Dekker, Inc., New York City (1992)
3. Hinojosa, V.H., Galleguillos, N., Nuques, B.: A simulated rebounding algorithm applied to the multi-stage security-constrained transmission expansion planning in power systems. Electr. Power Energy Syst. **47**, 168–180 (2013). Elsevier
4. Alhamrouni, A., Khairuddin, A., Khorasani, F., Salem, M.: Transmission expansion planning using AC-based differential evolution algorithm. IET Gener. Transm. Distrib. **10**, 1637–1644 (2014)
5. Arabali, A., Ghofrani, M., Amoli, E.M., Fadali, M.S., Aghtaie, M.M.: A multi-objective transmission expansion planning framework in deregulated power systems with wind generation. IEEE Trans. Power Syst. **6**, 1–8 (2014)
6. Taheri, S., Seyed-Shenava, S.J., Modiri-Delshad, M.: Transmission network expansion planning under wind farm uncertainties using Cuckoo search algorithm. In: Proceedings of the 3rd IET International Conference on Clean Energy and Technology (CEAT), 24–26 November, pp. 1–6 (2014)
7. Torres, S.P., de Araujo, R.A., Castro, C.A., Pissolato, J.: Security constrained transmission expansion planning for smart transmission grids based on the AC network model. In: IEEE PES Transmission and Distribution Conference and Exposition – Latin America (PES T&D-LA), 10–13 September, pp. 1–6 (2014)
8. Chatthaworn, R., Chaitusaney, S.: Improving method of robust transmission network expansion planning considering intermittent renewable energy generation and loads. IET Gener. Transm. Distrib. **13**, 1621–1627 (2015)
9. Obio, E.B., Mutale, J.: A comparative analysis of energy storage and N−1 network security in transmission expansion planning. In: Proceedings of the IEEE 50th International Universities Power Engineering Conference (UPEC), 1–4 September, pp. 1–6 (2015)
10. Rathore, C., Roy, R.: Gbest-guided artificial bee colony algorithm based static transmission network expansion planning. In: Proceedings of the IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT), 5–7 March, pp. 1–6 (2015)
11. Tang, L., Ferris, M.C.: A hierarchical framework for long-term power planning models. IEEE Trans. Power Syst. **1**, 46–56 (2015)
12. Mínguez, R., García-Bertrand, R.: Robust transmission network expansion planning in energy systems: improving computational performance. Eur. J. Oper. Res. **248**, 21–32 (2015)
13. Barreto, W.E., Torres, S.P., Castro, C.A.: Study of particle swarm optimization variations applied to transmission expansion planning. In: Proceedings of the IEEE Powertech Conference, 16–20 June, pp. 1–6 (2013)
14. Le, A.D., Nguzen, M.H., Eghbal, M., Nguzen, D.H.: Transmission expansion planning in electricity market: the case in Vietnam. In: Proceedings of the IEEE Australasian Universities Power Engineering Conference (AUPEC), Hobart, Australia, 29 September–03 December, pp. 1–6 (2013)

Epidemic Algorithm Based Optimal Power Flow in Electric Grids

K. Muniyasamy[1], Seshadhri Srinivasan[2(✉)], S. Parthasarathy[3], B. Subathra[4], and Simona Dzitac[5]

[1] Department of Electrical and Electronics Engineering, Kalasalingam University, Virudhunagar, India
muniyasamyk.eee@gmail.com
[2] International Research Center, Kalasalingam University, Virudhunagar, India
cpscourse@klu.ac.in
[3] Department of Electrical and Electronics Engineering, KLN College of Engineering, Sivagangai, India
sarathy_sps@yahoo.co.in
[4] Department of Instrumentation and Control Engineering, Kalasalingam University, Virudhunagar, India
b.subathra@klu.ac.in
[5] University of Oradea, Oradea, Romania
ssimona@dzitac.ro

Abstract. Time-triggering and distributed nature of the grid are emerging as the major challenge in managing energy in distribution grids. This investigation presents an event triggered distributed optimal power flow (OPF) algorithm for energy grids. To generate the event triggers, we use the epidemic algorithm. The buses are classified into three: infected, susceptible, and dead. The network works in two modes: normal and optimization mode. In the normal mode, only event detection happens and when there are no event triggers, the system is said to be in normal mode. In optimization mode, event triggers that can be a change in generation or demand beyond a threshold value that necessitates the re-optimization of the network, the optimization mode begins. In this mode, the infected node which is infected by change in bus variable intimates it to the energy management application. The energy management application on sensing this change, will initiate the graph grammars which are a set of rules to change the bus nature by detecting the effect of the change on the particular bus. The network is re-optimized using a DC OPF formulation as it is convex and can be solved using simple matrix inversion on the stationary conditions. As a result, the solution of DCOPF problem becomes that of solving a system of linear equations of the form $Ax = b$, which is solved using Krylov's method or the Arnoldi algorithm in a distributed fashion. Each node solves the problem of its one-hop neighbours in parallel and this leads to a distributed implementation resulting significant reduction in complexity. The propossed approach is illustrated on a simple 3 bus network.

V.E. Balas et al. (eds.), *Soft Computing Applications*, Advances in Intelligent Systems and Computing 633, DOI 10.1007/978-3-319-62521-8_6

Nomenclature

i, j	indices of bus
θ	load angles
B	admittance matrix
x_{ij}	reactance of line ij
D_t	Vector of active demand at time t
P_t	Vector of active power generation at time t
P_{Gg}	Power generated by the generator g
g	indices for generator
t	time indices
$C(P_{Gg})$	Generation cost of the generator g in dollars
$\mathscr{T}$	Set of all time
λ	Lagrange Multiplier
$\mathscr{G}$	set of all generators
$\mathscr{K}_m$	Krylov's Space

1 Introduction

In many energy management systems (EMS), time-triggering is used for optimization (see, [1–5] and references therein). This leads to unnecessary computations as the grid re-optimizes even without change in grid conditions. The complexity is compounded by the distributed nature of the grid. Considering the fluctuating nature of the demand and generation, and their distribution within the grid, event-triggering is more suitable as thresholds of change are detected and re-optimization of the EMS application is initiated. Moreover, some of these changes can be localized using a distributed optimization algorithm. This significantly reduces the complexity. Therefore, energy management applications that are event triggered and perform distributed optimization are very much required in modern energy grids integrated with renewable energy sources (RES) and energy storage systems (ESS) [6].

Optimal Power Flow (OPF) is a widely used tool for energy management and planning. The OPF problem computes the optimal generator dispatches considering the network parameters to optimize either the cost, line-losses, demand response, and many other objectives (see, [7–12]). There are various solution methods for OPF such as quadratic programming [13], semi-definite programming [14], sequential quadratic programming [15], linear programming [16], and nonlinear programming [22,27] to name a few. These approaches use a static optimization approach and do not consider time-variations of the demand or fluctuating generation and are hence not suitable for modern grids.

Dynamic OPF overcomes this problem as it solves a problem considering time as a running parameter [17]. However, still the dynamic OPF cannot deal with fluctuations in demand and generation. To overcome this receding

horizon approach or model predictive control for OPF problem has been studied. In [22,28], the ACOPF problem using forecast on renewable generation and demand has been solved using the receding horizon approach. The major drawback with these approaches is the time-triggering. The grid is re-optimized even when there is no significant changes in grid conditions or optimized only after a certain change has occurred which is not useful aspect of the EMS. A modified scenario could be to initiate the optimization whenever there are significant changes in the grid variables such as demand or generation. This will reduce frequent optimization and computation overhead resulting thereof.

More recently, distributed approaches for solving the OPF problem based on Gradient descent [19], decomposition methods [21], alternating directional method of multipliers [18], and others have been studied. In principle these methods consider two types of network models: ACOPF and DCOPF. While ACOPF provides a comprehensive study on both transient and steady state behaviour including the voltage magnitudes, they are non-convex and non-linear requiring significant computations for optimizing the grid. On the other hand, in many scenarios DCOPF which are an approximation of ACOPF problem behave reasonably well and are widely used in practice [20]. The main advantage of DCOPF is that it leads to a convex problem and its solution is relatively simple and unique compared to ACOPF. Moreover, as just a matrix inversion on the Karush-Kuhn Tucker (KKT) conditions or the stationary conditions of the DCOPF problem can provide the optimal power dispatches, they are widely preferred.

This investigation addresses the two challenges discussed above time-triggering and distributed nature of the grid. It uses an event-triggered optimization and solves the DCOPF problem in a distributed manner. To generate the event triggers, it uses the epidemic programming widely studied for spreading information in distributed systems [23]. Then it solves the DCOPF problem in a distributed manner by formulating it as a set of linear equations and uses the Krylov's method (i.e. Arnoldi's algorithm) as the stationary conditions of the DCOPF problem lead to a positive definite matrix. The solution can be obtained fast using the Krylov's method. The main contributions of the investigation are:

- An epidemic algorithm based event triggering mechanism.
- A self-assembly procedure using rules.
- Distributed OPF implementation using Krylov's method.

The rest of the paper is organized as follows. Section 2 presents the problem formulation. The concept of autonomous gossiping is presented in Sect. 3. The Krylov's method for performing distributed optimization is presented in Sect. 4. Section 5 presents a 3-bus case study.

2 Problem Formulation

Consider the multi-period DCOPF problem widely studied in literature:

$$\sum_{t \in \mathscr{T}} \sum_{p_g \in \mathscr{G}} \underset{P_g}{C}(P_g)$$

$$\text{s.t.} \tag{1}$$

$$B.\theta_t = P_t - D_t, \qquad \forall t \in \mathscr{T}$$

$$P_g^{min} \leq P_{g,t} \leq P_g^{max}$$

The line-flow constraints and the ramp-up/down constraints of the generators are not considered in the formulation due to simplicity. One can verify that the variations can happen in the optimization problem only due to changes in real-power generations which may be due to renewable generation or changes in demand. The problem is usually solved periodically based using time as a reference parameter. However, performing time-triggered optimization requires the problem be solved repeatedly even when there is no update in the generation. This is generally not preferred.

A more realistic scenario can be the event triggering i.e., on sensing there is a change in grid condition re-optimize the generator settings for reducing the operating cost. But again the problem has to be solved in a centralized fashion which increases the computation complexity. Significant reduction can be obtained, if the problem is solved in a distributed fashion. Moreover, since the DCOPF problem is convex, the KKT conditions can provide the unique solution to the optimization problem. Further, the very nature of the network architecture i.e. the bus reactance from bus i to j, x_{ij} is equal to in magnitude to x_{ji}. Therefore, elements of the matrix that models the stationary conditions is symmetric.

The problem is to study the design of a distributed DCOPF implementation that uses event triggering for optimizing the network operating cost. The formulation leads to the following pertinent questions:

1. How to generate the event triggers?
2. How to distribute the computation?
3. Can convergence and robustness of the distributed computation be guaranteed?

The answers to the above questions will be considered in the rest of the paper.

3 Autonomous Gossiping

Autonomous gossiping algorithms such as the epidemic algorithms are emerging as a method for propagating information in distributed systems [24]. They have been applied to wide class of problems like replication of database systems [25], wireless sensor routing, and more recently also in electric networks [26]. They

are seen as a robust and scalable means to disseminate information reliably on a distributed network the same way an epidemic will propagate through a set of individuals. The use of gossip-like distributed optimization algorithm for reactive power flow control was investigated by Bolognani and Zampieri in [26]. The investigation considered its use in the reactive power flow control problem of the microgrid.

This investigation considers the use of the autonomous gossip algorithms for distributed optimal power flow (DOPF). Each bus is considered one agent in the algorithm. The agents with the generators have a cost function that models the generation cost that is quadratic. Hence, for each generation node, the cost is given by

$$C(P_G g) = a_g \times (P_{Gg})^2 + b_g \times (P_{Gg}) + c_g \quad \forall g \in \mathscr{G}$$

where a_g, b_g and c_g are constants provided by the GenCo. We assume that an event occurs whenever there is a change in the generator or demand settings and the general optimization problem is less sensitive to changes below the threshold ϵ_i in each agent. The threshold value can be determined from sensitivity analysis or by simple numerical iterations. In some instances, the values can be updated online as well.

The algorithm works in two modes: normal and optimization. We assume that the optimal generator settings for the normal mode are obtained using a distributed OPF formulation and that the grid is functioning within thresholds. The normal mode monitors for the threshold ϵ being breached in some node and sends triggers for re-optimization.

As soon as the triggers are received, the algorithm enters the optimization mode, wherein the nodes are classified into three categories: infected, susceptible, and dead. The infected node is the one that has exceeded the threshold, whereas the susceptible can contract the infection, and the dead-node is not affected directly by the infected node, or is not affected at all. Three types of actions are defined: push, pull, and push-pull. In the push action, the infected node tries to send information to its one-hop neighbours regarding the event. The susceptible nodes try to pull the information from the infected node on the magnitude of the violation. Then they inform the dead node of the violation and also pull an information from them about their current status. In order to promote autonomous gossiping, we define the rules in Table 1.

The above table is used to create autonomous organization of the agents and they form a self-assembly. The agents status gets changed during each iteration of the algorithm until the threshold reaches below ϵ_i in all the infected buses. The algorithm starts with the infected node pushing its information to its one-hop neighbours during the first iteration. The susceptible nodes receive this information and start re-optimizing triggers to their peers. The common variable is the load angle θ_i that models the difference between the load balance and also the line-limits.

On sensing a variation in theta from the optimal value, the network is broken at the very node that experiences it, and various zones are formed. The coupling

between the different zones is the variable θ_i that is replicated. The self-assembly step will be used to send event trigger and the algorithm continues its execution until the convergence to optimal value.

Table 1. Self-organizing rules for the nodes in DOPF

Node 1	Node 2	Action	Outcome
Infected	Susceptible	Push	Node 1: Infected, Node 2: Infected
Infected	Susceptible	Pull	Node 1: Infected, Node 2: Susceptible
Infected	Susceptible	Push-pull	Node 1: Infected, Node 2: Infected
Susceptible	Infected	Pull	Node 1: Infected, Node 2: Infected
Susceptible	Infected	Push	Node 1: Susceptible, Node 2: Infected
Susceptible	Infected	Push-pull	Node 1: Infected, Node 2: Infected
Infected	Dead	Push	Node 1: Infected, Node 2: Susceptible
Infected	Dead	Pull	Node 1: Infected, Node 2: Dead
Infected	Susceptible	Push-pull	Node 1: Infected, Node 2: Susceptible
Susceptible	Dead	Pull	Node 1: Susceptible, Node 2: Dead
Susceptible	Dead	Push	Node 1: Susceptible, Node 2: Susceptible
Susceptible	Dead	Push-pull	Node 1: Infected, Node 2: Susceptible

4 Krylov's Method Based Distributed Optimization

4.1 Langarangian of the DCOPF Problem

The Lagrangian of the DCOPF problem is given by

$$\mathscr{L} = \sum_{i=1}^{N}(a_i \times P_g^2 + b_i \times P_g + c_i) + \lambda_i((B\theta - P_i + D_i) + \mu(\theta_1 - 0)) \quad (2)$$

where θ_1 is the reference bus. The optimal value of the DCOPF problem can be obtained from the KKT conditions:

- First order necessary conditions

$$\frac{\partial \mathscr{L}}{\partial P_g} = 0 \quad (3)$$

$$\frac{\partial \mathscr{L}}{\partial \theta} = 0 \quad (4)$$

- Primal feasibility

$$\frac{\partial \mathscr{L}}{\partial \lambda_i} = 0 \quad (5)$$

- Dual feasibility

$$\lambda_i \geq 0 \tag{6}$$

- Complimentary Slackness

$$\lambda_i(B\theta - P_i + D_i) = 0 \tag{7}$$

The above conditions lead to a set of linear equations in the unknowns $X = [P_g, \theta, \lambda]$. The problem can hence be represented as:

$$M \times X = b \tag{8}$$

The above solution has $3 \times n$ linear equations that need to be solved in a distributed fashion.

The objective of this investigation is to have a distributed approach, which means the computation above has to be distributed between the different nodes. To form the distributed updates, the infected and its one hop neighbours are considered, a reduced order linear equations of the one-hop neighbours is solved using the Krylov's method. The idea is to construct a reduced space $\mathscr{K}_m$ with a dimension $m << 3n$, then we find the variable in the space such that it happens to the solution of the actual subspace. The Krylov method uses the conjugate gradient and it uses the Arnoldi method.

4.2 Arnoldi's Method

The method begins by constructing a orthogonal basis for $\mathscr{K}_m(M, r_0)$, where r_0 is the initial residual vector. The following steps are used in the Arnoldi's method:

1. Initialize: X for the DCOPF problem.
2. Orthonormalization step

$$X_1 = \frac{X}{\|X\|_2}$$

3. For the kth step

$$\tilde{X}_{k+1} = A\ X_k - \sum_{j=1} kX_k h_{jk}$$

where: $h_{jk} = MX_k^T, X_j = <MX_k, X_j>$
and $X_{k+1} = \frac{\tilde{X}_{k+1}}{\|\tilde{X}_{k+1}\|_2}$

4. Update:

$$MX_k = \sum_{j=1} k + 1X_j h_{jk}$$

5. Loop back until m

6. Output:

$$MX_m = X_{m+1}H_{m+1,m}$$

where H is the Hiessenberg matrix and X_m contains the solution vector.

The method described above can be used to solve the DCOPF problem in a distributed fashion when event triggers are obtained from the epidemic algorithm.

5 Results

5.1 3- Bus Case Study

This section presents the 3 bus case study for illustrating the approach. The line reactance x_{12}=0.1, $x_{13} = 0.125$, and $x_{23} = 0.2$. The generation costs of the three generators is given by

$$\begin{aligned} F_1(P_1) &= 561 + 7.92P_1 + 0.001562P_1^2 \\ F_2(P_2) &= 310 + 7.85P_2 + 0.00194P_2^2 \\ F_3(P_3) &= 78 + 7.97P_3 + 0.00482P_3^2 \end{aligned}$$

The maximum and minimum generator ratings are given by $P_1^{min} = 150, P_1^{max} = 100, P_2^{min} = 100, P_2^{max} = 400, P_3^{min} = 50, P_3^{max} = 200$. The problem was solved in a distributed manner using the Arnoldi algorithm proposed before with θ_1 as the coupling variable, the optimal generator settings are: $P_1 = 393.2, P_2 = 334.6, P_3 = 122.2$ with a $\lambda = 9.2$.

Our results demonstrate the effectiveness of the proposed DCOPF approach as the solution obtained by solving the problem with matrix inversion matches that of the proposed algorithm.

6 Conclusions

This investigation presented an event triggered distributed OPF algorithm for energy grids. The main advantage of event triggering is it avoids re-optimization based on time-triggers which introduce unnecessary computation complexity. Moreover, scaling of the optimization methods is a big challenge and this is simplified using a distributed approach. In formulation, the KKT conditions reduced to a set of linear equations that were solved using Krylov's basis and Arnoldi method. The proposed approach provided a simple implementation of distributed OPF for energy grids. Implementation with renewable energy forecasts and storage dynamics and extending it to standard grid is the future course of this investigation.

References

1. Srinivasan, S., Kotta, U., Ramaswamy, S.: A layered architecture for control functionality implementation in smart grids. In: 2013 10th IEEE International Conference on Networking, Sensing and Control (ICNSC). IEEE (2013)
2. Srinivasan, S., Kumar, R., Vain, J.: Integration of IEC 61850 and OPC UA for smart grid automation. In: 2013 IEEE Innovative Smart Grid Technologies-Asia (ISGT Asia). IEEE (2013)
3. Landa-Torres, I., et al.: The application of the data mining in the integration of RES in the smart grid: consumption and generation forecast in the I3RES project. In: 2015 IEEE 5th International Conference on Power Engineering, Energy and Electrical Drives (POWERENG). IEEE (2015)
4. Gambino, G., Verrilli, F., Del Vecchio, C., Srinivasan, S., Glielmo, L.: Optimization of energy exchanges in utility grids with applications to residential, industrial and tertiary cases. In: 2015 AEIT International Annual Conference (AEIT), pp. 1–6, October 2015. IEEE (2015)
5. Soudari, M., Srinivasan, S., Balasubramanian, S., Vain, J., Kotta, U.: Learning based personalized energy management systems for residential buildings. Energy Build. **127**, 953–968 (2016)
6. Gambino, G., et al.: Optimal operation of a district heating power plant with thermal energy storage. In: 2016 American Control Conference (ACC), pp. 2334–2339, July 2016. IEEE (2016)
7. Frank, S., Steponavice, I., Rebennack, S.: Optimal power flow: a bibliographic survey I. Energy Syst. **3**(3), 221–258 (2012)
8. Gan, L., Low, S.H.: An online gradient algorithm for optimal power flow on radial networks. IEEE J. Sel. Areas Commun. **34**(3), 625–638 (2016)
9. Madani, R., Sojoudi, S., Lavaei, J.: Convex relaxation for optimal power flow problem: mesh networks. IEEE Trans. Power Syst. **30**(1), 199–211 (2015)
10. Momoh, J.A., El-Hawary, M.E., Adapa, R.: A review of selected optimal power flow literature to 1993. Part I: nonlinear and quadratic programming approaches. IEEE Trans. Power Syst. **14**(1), 96–104 (1999)
11. Momoh, J.A., El-Hawary, M.E., Adapa, R.: A review of selected optimal power flow literature to 1993. Part II: Newton, linear programming and interior point methods. IEEE Trans. Power Syst. **14**(1), 105–111 (1999)
12. Pandya, K.S., Joshi, S.K.: A survey of optimal power flow methods. J. Theoret. Appl. Inf. Technol. **4**(5), 450–458 (2008)
13. Lavei, J., Rantzer, A., Low, S.: Power flow optimization using positive quadratic programming. IFAC Proc. Vol. **44**(1), 10481–10486 (2011)
14. Bai, X., Wei, H., Fujisawa, K., Wang, Y.: Semidefinite programming for optimal power flow problems. Int. J. Electr. Power Energy Syst. **30**(6), 383–392 (2008)
15. Sheng, W., Liu, K.Y., Cheng, S., Meng, X., Dai, W.: A trust region SQP method for coordinated voltage control in smart distribution grid. IEEE Trans. Smart Grid **7**(1), 381–391 (2016)
16. Attarha, A., Amjady, N.: Solution of security constrained optimal power flow for large-scale power systems by convex transformation techniques and Taylor series. IET Gener. Transm. Distrib. **10**(4), 889–896 (2016)
17. Gill, S., Kockar, I., Ault, G.W.: Dynamic optimal power flow for active distribution networks. IEEE Trans. Power Syst. **29**(1), 121–131 (2014)
18. Erseghe, T.: Distributed optimal power flow using ADMM. IEEE Trans. Power Syst. **29**(5), 2370–2380 (2014)

19. Lam, A.Y., Zhang, B., David, N.T.: Distributed algorithms for optimal power flow problem. In: 2012 IEEE 51st IEEE Conference on Decision and Control (CDC), pp. 430–437, December 2012. IEEE (2012)
20. Bahrami, S., Wong, V.W., Jatskevich, J.: Optimal power flow for AC-DC networks. In: 2014 IEEE International Conference on Smart Grid Communications (SmartGridComm), pp. 49–54, November 2014. IEEE (2014)
21. Kim, B.H., Baldick, R.: A comparison of distributed optimal power flow algorithms. IEEE Trans. Power Syst. **15**(2), 599–604 (2000)
22. Maffei, A., Srinivasan, S., Iannelli, L., Glielmo, L.: A receding horizon approach for the power flow management with renewable energy and energ storage systems. In: 2015 AEIT International Annual Conference (AEIT), pp. 1–6, October 2015. IEEE (2015)
23. Datta, A., Quarteroni, S., Aberer, K.: Autonomous gossiping: a self-organizing epidemic algorithm for selective information dissemination in wireless mobile ad-hoc networks. In: Semantics of a Networked World. Semantics for Grid Databases, pp. 126–143. Springer, Heidelberg (2004)
24. Eugster, P.T., Guerraoui, R., Kermarrec, A.M., Massoulié, L.: Epidemic information dissemination in distributed systems. Computer **37**(5), 60–67 (2004)
25. Demers, A., Greene, D., Hauser, C., Irish, W., Larson, J., Shenker, S., Terry, D.: Epidemic algorithms for replicated database maintenance. In: Proceedings of the Sixth Annual ACM Symposium on Principles of Distributed Computing, pp. 1–12, December 1987. ACM (1987)
26. Bolognani, S., Zampieri, S.: A gossip-like distributed optimization algorithm for reactive power flow control. IFAC Proc. Vol. **44**(1), 5700–5705 (2011)
27. Verrilli, F., et al.: Model predictive control-based optimal operations of district heating system with thermal energy storage and flexible loads. IEEE Trans. Autom. Sci. Eng. **14**(2), 547–557 (2017)
28. Maffei, A., Srinivasan, S., Castillejo, P., Martinez, J.F., Iannelli, L., Bjerkan, E., Glielmo, L.: A semantic middleware supported receding horizon optimal power flow in energy grids. IEEE Trans. Ind. Inform. **PP**(99), 1. doi:10.1109/TII.2017.2655047. http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7822985&isnumber=4389054

Issues Regarding the Tuning of a Minimum Variance Adaptive Controller

Ioan Filip(✉), Iosif Szeidert, Octavian Prostean, and Cristian Vasar

Department of Automation and Applied Informatics,
University Politehnica Timisoara, Bvd. V. Parvan,
No. 2, 300223 Timisoara, Romania
{ioan.filip,iosif.szeidert,vcristi}@aut.upt.ro,
octavian.prostean@upt.ro

Abstract. The paper presents an analysis regarding the tuning of a minimum variance adaptive controller with application to the terminal voltage control of an induction generator connected to a power system. The main goal of the considered control system is to provide the best results as a disturbance rejection control system. The basic idea supposes that the performed analysis must take into consideration both the cases of positive and negative perturbations. The analysis performed to tune the controller parameters proves that, by considering only one class of disturbances (which causes only a positive or only a negative variation of the controlled output), the resulting tuning parameters can be totally inappropriate for other class of disturbances. This approach is explained by the fact that, the process being nonlinear, a positive or a negative perturbation shifts the operating point of the controlled process on the nonlinear characteristic in one direction or in the other. The proposed solution consists in a compromise between these two cases, by choosing a set of controller parameters that can provide acceptable results in both situations.

Keywords: Adaptive control · Minimum variance controller · Induction generator

1 Introduction

The classical structure of an adaptive control system based on a minimum variance controller is depicted in Fig. 1 [1, 2]. Two main elements can be noticed: the module that implements the control law, respectively the parameters estimator. The control law is obtained through minimization of a criterion function described by relation (1), and containing the sum of the two terms: first term ensures a minimum variance of the controlled output error, respectively the second term ensures a minimum variance of the controller output [3, 4].

$$J = E\left\{ [y(t+1) - w(t)]^2 + \rho\, [u(t) - u_r(t)]^2 \right\} \tag{1}$$

V.E. Balas et al. (eds.), *Soft Computing Applications*, Advances in Intelligent Systems and Computing 633, DOI 10.1007/978-3-319-62521-8_7

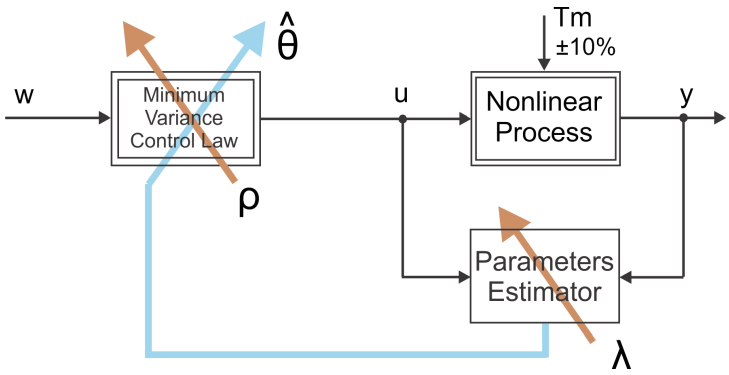


Fig. 1. Minimum variance adaptive control structure

where: *y(t)* – controlled output; *u(t)* – the controller output, $u_r(t)$ – the steady state controller output; *w(t)* – the reference, ρ – the control penalty factor; *E{.}* – the mean operator.

This second term is weighted by a sub unitary parameter ρ called control penalty factor [5, 6] (being critical for obtaining physically realizable values of the control).

The design of the control strategy assumes that the generator can be modeled, around an operating point, by a 4th order discrete transfer function [2]:

$$H(q^{-1}) = \frac{y}{u} = q^{-1}\frac{B(z^{-1})}{A(z^{-1})} \tag{2}$$

where: $A(z^{-1})$ - polynomial function by 4th order, $B(z^{-1})$ - polynomial function by 3rd order.

Using this linear model (relation (2)), by minimization of the criterion function (1), the minimum variance control law results (see relation (3)):

$$u(t) = \frac{-q[1-\hat{A}(q^{-1})]y(t)}{\hat{B}(q^{-1})+\rho} + \frac{1+\rho\frac{\hat{A}(1)}{\hat{B}(1)}}{\hat{B}(q^{-1})+\rho}w(t) \tag{3}$$

The classical recursive least square (RLS) parameters estimator (with a forgetting factor λ as a tuning parameter) is used to implement the control system [7, 8]. Also the plant is disturbed by a stochastic noise with variance $\sigma^2 = 0.01$, to ensure the excitation of a RLS estimator.

The performed analysis concerns the case of an adaptive control system, with application to the excitation control of an asynchronous generator (connected to a power system). The main objective is to maintain a constant terminal voltage under the action of external perturbations that disturbs the generator. A positive/negative mechanical torque variation was considered as external perturbation. The asynchronous generator was simulated using a classical nonlinear *d-q* model by 7th order, based on Park's

equations [8–10]. The process being nonlinear, a negative or a positive perturbation shifts the operating point of the controlled process in one sense or in another on the operating characteristics of the process.

2 The Tuning of the Controller Parameters

The main goal of controller tuning is to determine the values of the parameters λ and ρ, which provide good results for both positive and negative perturbations. As a perturbation, a ± 10% variation of the mechanical torque was considered at time t = 2 s. The simulations results are presented in Figs. 2, 3, 4, 5, 6 and 7, for different tuning values of λ (forgetting factor) and ρ (control penalty factor). By choosing ρ = 0.1 (value usually chosen [11]), this tuning value ensures the best results, as it will be further presented. The Fig. 2a and b (λ = 0.995, ρ = 0.1) show relative good results for a positive perturbation but a too long settling time and oscillations in the case of negative perturbation.

Considering the decrease of the forgetting factor (λ = 0.99), the obtained results aren't satisfactory (too high overshoot for a positive perturbation, respectively long-time oscillations for a negative perturbation- see Fig. 3a, b). And so on, by

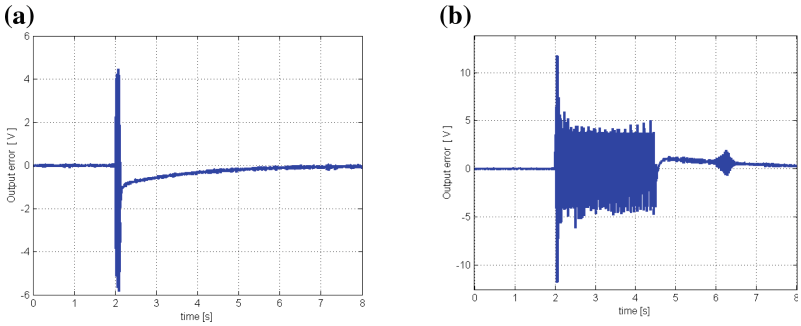


Fig. 2. (a) Output error - positive perturbation (λ = 0.995, ρ = 0.1). (b) Output error - negative perturbation (λ = 0.995, ρ = 0.1)

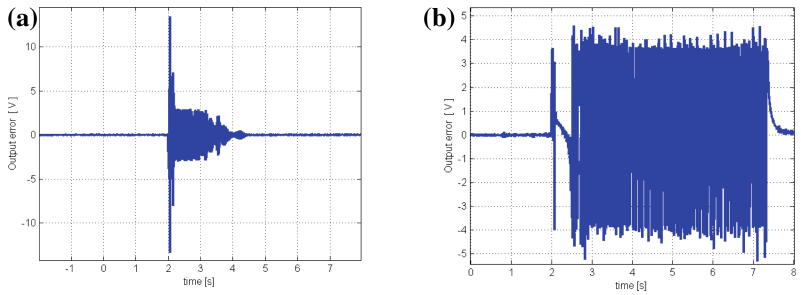


Fig. 3. (a) Output error - positive perturbation (λ = 0.99, ρ = 0.1). (b) Output error - negative perturbation (λ = 0.99, ρ = 0.1)

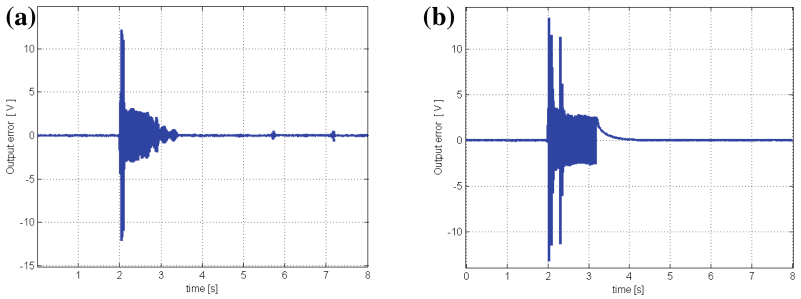


Fig. 4. (a) Output error - positive perturbation ($\lambda = 0.98$, $\rho = 0.1$). (b) Output error - negative perturbation ($\lambda = 0.98$, $\rho = 0.1$)

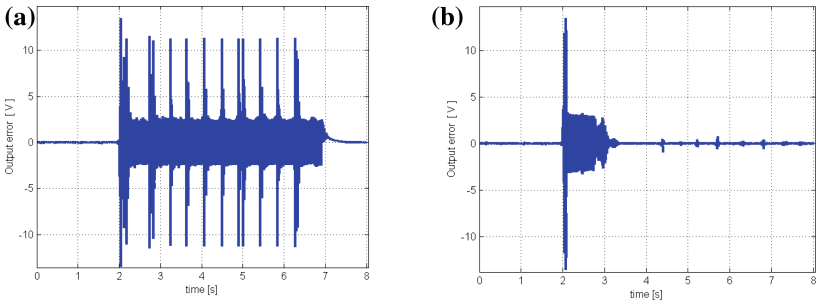


Fig. 5. (a) Output error - positive perturbation ($\lambda = 0.97$, $\rho = 0.1$). (b) Output error - negative perturbation ($\lambda = 0.97$, $\rho = 0.1$)

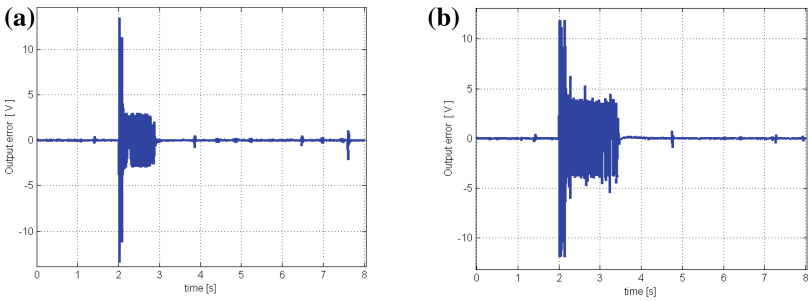


Fig. 6. (a) Output error - positive perturbation ($\lambda = 0.96$, $\rho = 0.1$). (b) Output error - negative perturbation ($\lambda = 0.96$, $\rho = 0.1$)

decreasing λ (0.98…0.95), the results are presented in Figs. 4, 5, 6 and 7 and the conclusions are synthesized in Table 1. The best compromise is obtained for $\lambda = 0.98$ (Fig. 4a, b). Also, some tests are performed for other values (lower or higher) of the ρ parameter (see case no. 7 and 8 from Table 1), but the results are inferior.

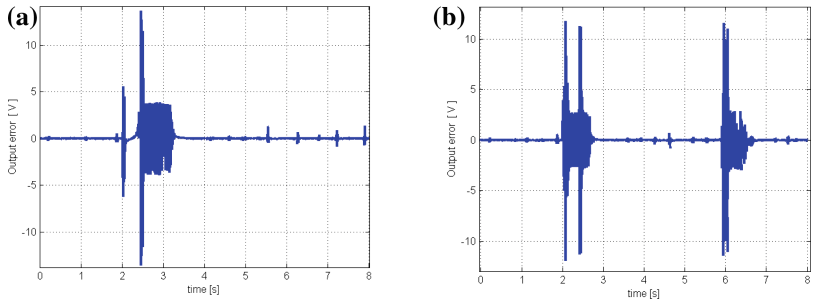


Fig. 7. (a) Output error - positive perturbation ($\lambda = 0.95$, $\rho = 0.1$). (b) Output error - negative perturbation ($\lambda = 0.95$, $\rho = 0.1$)

The global results are synthetically presented in the Table 1, by a unitary approach that blends the perspective of these two types of perturbations. There is mentioned that the Table 1 contains other two case studies (case no. 7 and 8), analyzing the results obtained for $\rho = 0.01$, respectively $\rho = 0.2$ (the figures showing these afferent results are not presented in the paper).

Analyzing the results presented in Figs. 2a, 3a, 4a, 5a, 6a and 7a (and Table 1), it can be noted that, the best performances for a positive perturbation (+10% positive variation of mechanical torque) are obtained in two situations:

- The smallest overshoot is obtained for $\lambda = 0.995$ and $\rho = 0.1$ (case 1);
- The smallest settling time is obtained for $\lambda = 0.96$ and $\rho = 0.1$ (case 5).

On the other side, analyzing the results presented in Figs. 2b, 3b, 4b, 5b, 6b and 7b (and Table 1), it can be noticed that, for a negative perturbation (−10% negative variation of mechanical torque), the best performances are obtained in other two situations:

- The smallest overshoot is obtained for $\lambda = 0.99$ and $\rho = 0.1$ (case 2);
- The smallest settling time is obtained for $\lambda = 0.97$ and $\rho = 0.1$ (case 4).

So, each type of perturbation requires a different parameters set for tuning the controller. Also, the case no. 6, 7 and 8 from Table 1 can be eliminated due to the unstable or long oscillatory behavior of the control system.

Paradoxically, analyzing the case no. 3 ($\lambda = 0.98$, $\rho = 0.1$ – see Table 1), good results, in terms of a global interpretation on both types of perturbations, can be observed (small settling times, acceptable overshoots). This result analyzed individually for each type of perturbation, is not the best, but overall they offer the best compromise. Moreover, a few conclusions can be drawn: big λ - small overshoot, small λ - multiple small output spikes, small ρ - system instability.

The performed analysis proves that, in the case of a non-linear process, a positive or a negative perturbation can affect differently the controlled output, both qualitatively and quantitatively. So, an analysis regarding the both perturbation classes becomes a necessity. Also, a proper tuning of the controller parameters for the case of the rejection of a positive perturbation can be totally improper for the rejection of a negative

Table 1. Overall results (case no. 1–8)

No	λ	ρ	Results - disturbance +10%	Results - disturbance -10%	Conclusion
1	0.995	0.1	Overshoot≈ ±4-5 V Settling time≈ 4 sec. No oscillations.	Overshoot ≈±12 V Settling time≈ 6 sec.; Oscillations≈ 2.5 sec. (Fig. 2.a,b)	Long-time oscillations for negative disturbance
2	0.99	0.1	Overshoot≈±14V Settling time ≈2 sec. Oscillations≈ 2 sec.	Overshoot≈±5 V Settling time≈ 6 sec. (Fig. 3.a,b) Oscillations≈ 5 sec.	Long-time oscillations for negative disturbance
3	***0.98***	***0.1***	***Overshoot≈ ±12 V Settling time≈ 1 sec. Oscillations≈ 1 sec***	***Overshoot≈±14 V Settling time≈ 1.5 sec. Oscillations≈ 1 sec.* (Fig. 4.a, b)**	***Best overall results. (Small settling times, but no best overshoots)***
4	0.97	0.1	Overshoot≈ ±14 V Settling time≈ 5 sec. Oscillations≈ 5 sec	Overshoot≈±14 V Settling time≈ 1 sec. (Fig. 5.a,b) Oscillations≈ 1 sec	Long-time oscillations for positive disturbance
5	0.96	0.1	Overshoot≈ ±14 V Settling time< 1 sec. Oscillations< 1 sec	Overshoot≈±12 V Settling time≈ 1.5 sec. Oscillations≈ 1.5 sec. (Fig. 6.a,b)	-Multiple small output spikes (due to a too small forgetting factor)
6	0.95	0.1	Overshoot≈ ±14 V Settling time< 1 sec. Oscillations< 1 sec	Overshoot≈±12 V Settling time≈ 2 sec. Oscillations≈ 2 sec. (Fig. 7.a,b)	-Best settling times for both cases -Multiple small output spikes (parameter estimations fluctuate)
7	0.98	0.01	Instability	Instability	Instability
8	0.98	0.2	Overshoot≈ ±13 V Settling time≈ 1.5 sec. Oscillations≈ 1.5 sec	Overshoot≈±12 V Settling time≈ 5 sec. Oscillations≈ 5 sec.	-Oscillations -Multiple small output spikes

perturbation. A solution ensuring a compromise between the two situations and solving both cases, must be found, even with the risk of weaker individual performances. By tuning the parameters $\lambda = 0.98$ and $\rho = 0.1$, the control system offers the best overall results. It should be mentioned that, for a reduced noise level, the performances are better (the settling time being significantly lower).

3 Conclusions

The paper presents an analysis based on a study cases regarding the parameters tuning of a minimum variance controller integrated into a disturbance rejection control system. Even though the case study is done for a particular situation regarding an induction generator excitation control system, the tuning procedure can be also generalized for the control of other process. The process being nonlinear, a positive or a negative perturbation shifts the operating point of the controlled process in one direction or in the other (and so, the changes of nonlinear model plant parameters are significantly). The basic idea consists into an analysis of both class of perturbations that produces a positive or a negative variation of the controlled output. The performed case study proves the necessity to take into consideration both types of disturbances (with positive or negative effects on the process controlled output). The proposed strategy is also valid for the tuning of any other nonlinear process control system.

Acknowledgments. This work was developed through the Partnerships in Priority Areas - PN II, with the support of ANCS, UEFISCDI, Project No. 36/2012, Code: PN-II-PT-PCCA-2011-3.2-1519.

References

1. Gomes Silva, M.J., Silva Araujo, C., Marques Bezerra, S.T., Rocha Souto, C., Arnaud Silva, S., Pimentel Gomes, H.: Generalized minimum variance control for water distribution system. IEEE Lat. Am. Trans. **13**(3), 651–658 (2015)
2. Filip, I., Szeidert, I., Prostean, O., Vasar, C.: Analysis through simulation of a self-tuning control structure for dual winded induction generator. In: Proceedings of the 9th IEEE International Conference on Computational Cybernetics, Tihany, Hungary, July 2013, pp. 147–152 (2013)
3. Filip, I., Prostean, O., Ionescu, D., Vasar, C., Szeidert, I.: Study above an adaptive control structure with application to a double-fed induction generator's excitation. In: Proceedings of the IEEE - International Conference on Computer as a Tool, Lisbon, Portugal, April 2011, pp. 1–4 (2011)
4. Hou, G., Hu, G., Sun, R., Zhang, J.: Application of self-tuning control based on generalized minimum variance method in evaporator for ORCS. In: Proceedings of the International Conference on Advanced Mechatronic Systems, Zhengzhou, China, August 2011, pp. 202–206 (2011)
5. Pitsillides, A.: Adaptive congestion control in ATM based networks: quality of service and high utilization. J. Comput. Commun. **20**(14), 1239–1258 (1997)

6. Pepe, P.: On the input-to-state practical stabilization of nonlinear neutral systems. In: Proceedings of the IEEE 51st Annual Conference on Decision and Control, Maui, HI, December 2012, pp. 3880–3885 (2012)
7. Filip, I., Szeidert, I.: Givens orthogonal transformation-based estimator versus RLS estimator-case study for an induction generator model. In: Advances in Intelligent Systems and Computing. Soft Computing Applications, vol. 357, no. 2, pp. 1287–1299. Springer (2015)
8. Filip, I., Szeidert, I., Prostean, O.: Mathematical modelling and numerical simulation of the dual winded induction generator's operating regimes. In: Advances in Intelligent Systems and Computing. Soft Computing Applications, vol. 357, no. 2, pp. 1161–1170. Springer (2015)
9. Abad, G., Lopez, J.M., Marroyo, L., Iwanski, G.: Doubly Fed Induction Machine: Modeling and Control for Wind Energy Generation, pp. 209–239. Wiley-IEEE Press, Hoboken (2011)
10. Zou, Y., Elbuluk, M., Sozer, Y.: A complete modeling and simulation of induction generator wind power systems. In: Proceedings of the IEEE Industry Applications Society Annual Meeting (IAS), Houston, USA, pp. 1–8 (2010)
11. Wang, J.: Parameters recursive identification for minimum variance control. In Proceedings of 31st Chinese Control Conference, 25–27 July 2012, Hefei, China, pp. 1701–1706 (2012)

Adaptive Controller for Networked Control Systems Subjected to Random Communication Delays

Seshadhri Srinivasan[1(✉)], G. Saravanakumar[2], B. Subathra[1], N. Sundarajan[3], Ülle Kotta[4], Srini Ramasamy[5], and Valentina Emilia Balas[6]

[1] International Research Center, Kalasalingam University, Anand Nagar, Krishanankoil, Srivilliputhur, Tamil Nadu, India
seshucontrol@gmail.com
[2] Department of Electrical and Computer Engineering, University of Gondar, Gondar, Ethiopia
saravana.control@gmail.com
[3] Nanyang Technological University, Singapore 639798, Singapore
ensundara@ntu.edu.sg
[4] Institute of Cybernetics, Tallinn University of Technology, 12618 Tallinn, Estonia
kotta@cc.ioc.ee
[5] ABB Inc., Westlake, USA
srini@ieee.org
[6] Department of Automation and Applied Informatics, Aurel Vlaicu University of Arad, Arad, Romania
balas@drbalas.ro

Abstract. This paper presents an adaptive linear quadratic regulator (LQR) for networked control systems that varies its gains based on the estimates of the time-varying network delays. A sequential learning algorithm for minimum radial basis function neural network, called the minimum resource allocation network (MRAN) is used to estimate the time delays on-line. The proposed delay estimation technique provides accurate estimates of the delays considering various channel conditions such as length of the channel, contention, loading, and the number of inputs-outputs connected to the controller among others. Then the adaptive controller gains are computed using the delay estimates. To compute the gains, LQR using the backward iteration algorithm is used and the stability conditions of the proposed controller are also studied.

The proposed adaptive controller is illustrated on simple examples using experiments conducted on Modbus over TCP/IP (Transmission Control Protocol/Internet Protocol) to model the network delays using MRAN for various loading conditions. Resulting adaptive controller is simple, yet optimal. Results indicate that the adaptive controller varies its gain to meet the regulation and tracking performance in the presence of random delays. The proposed estimation based adaptive controller is able to adapt its gain in real-time, while simultaneously estimating the delays considering the factors influencing delays in the channel.

V.E. Balas et al. (eds.), *Soft Computing Applications*, Advances in Intelligent Systems and Computing 633, DOI 10.1007/978-3-319-62521-8_8

Keywords: Networked Control Systems (NCSs) · Adaptive LQR controller · Random communication delays · Radial Basis Function (RBF) · MRAN (minimum resource allocation network)

1 Introduction

Control systems using communication channels for information exchange among system components are called Networked Control Systems (NCSs). A detailed review of NCSs together with their applications/challenges can be found in [1–7] and the references therein. The use of network offers, because of significant cost benefits, novel applications. However, from control point of view, communication channels induce time-varying delays in the control loop resulting in performance degradation and possible instability of the system. Therefore, it is imperative to design controllers considering the time-varying delays. Adaptive controllers have traditionally been used in academia and industry for systems with time-varying parameters. Basic idea of adaptive control is to modify controller gains based on the knowledge of the time-varying system parameters. Presence of time-varying delays in NCSs naturally suggests the use of adaptive control.

Adaptive controller design for NCSs is a recent research topic and not many results are available. To our best knowledge, the first attempt was by Tzes and Nikolakopoulos [8]. The authors proposed an output feedback gain scheduling adaptive controller using the delay measurements. While gain scheduling approaches for adaptive control have been used more commonly [8–11], an adaptive controller which combines elements of parameter estimation followed by an controller update has been investigated minimally [12]. Moreover, most adaptive controllers for NCSs focus on stabilization and do not address performance specifications, for example tracking error. In addition, these controllers either assume stochastic delay models or more restrictive assumptions thereby yielding conservative results [13–15].

Chunmao and Jian [16] proposed an adaptive controller based on gradient descent method to estimate the time-delay, used later to compute controller gains that reduce the over-shoot of the closed-loop system. The adaptive controller for tracking the reference signal in [12] requires knowledge of slot size in channels which is extremely difficult to predefine. More recently, Xu et al. [17] have proposed the stochastic optimal controller for NCSs subjected to random communication delays and packet losses. The controller uses an adaptive estimator (AE) and ideas from Q-learning proposed in [18] to solve infinite horizon optimal regulation problem for unknown NCSs. Here the AE is used to estimate the Q-function used in Q-learning.

Adaptive controller requires an estimate of delay depending on channel conditions. In literature, delay models have been used to estimate the delays and different models have been proposed. In these investigations, delays have been modeled to be constant [19], time-varying yet deterministic [20], or even random [21–23]. Constant and deterministic time-varying delay models result in deterministic NCSs, thereby making it possible to use deterministic control methods. However, such models result in loss-of-performance.

The available stochastic delay models can be divided into two categories: delays are either mutually independent (empirical distribution model) or dependent (Markov Chain, Hidden Markov Chain models). Fitting an empirical distribution may not always be possible considering the variations in sharing, channel quality, and congestion. The Markov Chain model in [21], and the Hidden Markov Model in [23] usually result in stochastic controllers. Markov Chain Monte Carlo model for delay, based on online acceptance-rejection sampling, proposed in [22] was used to design deterministic controller. This design methodology cannot be extended for designing adaptive controllers.

Alternatively, the paper [17] proposed to use the tools of computational intelligence such as Q-learning in designing stochastic optimal controller for NCSs subjected to random communication delays. Usually delays in communication channel depend on various parameters that are inherently random and therefore, the estimator needs also learning capabilities from previous scenarios. To our best knowledge there are no delay models in NCSs that use learning from previous network conditions. To accommodate learning capabilities, this investigation proposes to estimate the random delays using a special case of RBF neural network, called Minimal Resource Allocation Network (MRAN). Since the delays are random and it is difficult to model them using an analytical model, a model based estimation scheme has been avoided here and instead a data driven estimation scheme has been used. For this kind of estimation, a neural network scheme is preferred and here we specifically use MRAN for a fast estimation of the random delay as it is ideal for on-line learning and real-time implementation.

The goal of this paper is to propose a computationally efficient adaptive controller for tracking the reference signal in the presence of random communication delays that additionally assures optimality without introducing significant computation delays. The delay samples collected from experimentation on Modbus over TCP/IP network are used to generate the estimate using a MRAN. We use MRAN due to its simplicity, and ability to generate real-time estimates of the delay considering various network conditions such as network loading, channel length, number of connected nodes, and contention ratio after being trained from experimental data. Existing random delay models in literature consider only channel loading, and extending them to include other network conditions is complex. The proposed controller uses the LQR approach to update the gains. The choice of the objective function guarantees optimal control inputs and reduced tracking errors. The controller iteratively solves the discrete-time algebraic Ricatti equation using backward iteration to compute the adaptive gains, and uses the estimate of delay available from MRAN for computing the gains. Finally, to illustrate the effectiveness of the proposed approach, we use the delay model obtained from studies conducted on Modbus over TCP/IP communication channels in our simulations. Our results show that the adaptive controller tracks the reference signal in the presence of time-varying delays and also shows good regulatory performance. Unlike [24,25] where the adaptive LQR controller is based on Gaussian delay model with restrictive assumptions on steady-state

convergence of tuning matrices, this paper uses the MRAN delay estimates and backward iteration algorithm. Note that this according to our knowledge the first extension of MRAN to NCSs.

The paper is organized as six sections. In Sect. 2, we study the NCS model and give the problem statement. Section 3 describes the methodology to estimate the time-varying gains using MRAN and Sect. 4 presents the controller design. Numerical simulation along with experimental results are discussed in Sect. 5. Conclusions and future directions of research are presented in Sect. 6.

2 Problem Formulation

Consider the linear time-invariant (LTI) system

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bu^*(t) \\ y(t) &= Cx(t) \end{aligned} \tag{1}$$

with controller

$$u^*(t) \;=\; u_k \qquad \forall t \in [kh + \tau_k, (k+1)h + \tau_{k+1}) \tag{2}$$

where $x(t) \in \mathbb{R}^n$, $u^*(t) \in \mathbb{R}^m$, $y(t) \in \mathbb{R}^p$ are the state, input and output vectors respectively and A, B, C are constant matrices with appropriate dimensions. Networks N_1 and N_2 are used to connect sensor and controller output to the controller and plant respectively. NCS with dynamics (1) and (2) is shown in Fig. 1.

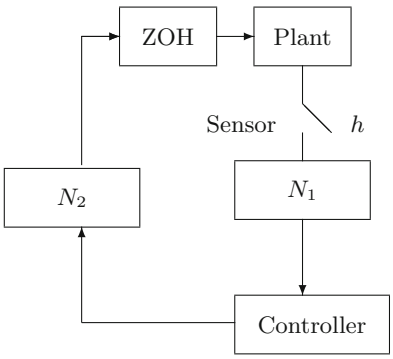


Fig. 1. Networked control system (NCS)

Total delay in the control input τ_k is the sum of networked-induced delays in the channels N_1 and N_2, denoted by τ_{sc} and τ_{ca}, respectively and computation delay is assumed to be negligible, then the total delay in control signal τ_k is given by (Eq. (3))

$$\tau_k = \tau_{sc} + \tau_{ca} \quad \text{for} \quad k = 1, 2, 3, \ldots \tag{3}$$

Assumption 1. We assume that the total delay in the channel is less than sampling time h,

$$\tau_k \leq h \tag{4}$$

Assumption 2. Re-ordering of packets in control and sensor signals is not permitted.

Assumptions 1 and 2 ensure proper working of NCS by eliminating out-of-sequence control and sensor information. Sampling the system (Eq. (1)) with time h under our assumptions leads to discrete-time NCS model [26]:

$$x_{k+1} = \phi x_k + \Gamma_0(\tau_k)u_k + \Gamma_1(\tau_k)u_{k-1} \tag{5}$$

where $\phi = e^{Ah}$, $\Gamma_0(\tau_k) = \int_0^{h-\tau_k} e^{A\lambda}\ B\ d\lambda$ and $\Gamma_1(\tau_k) = \int_{h-\tau_k}^{h} e^{A\lambda}\ B\ d\lambda$.

The NCS in (Eq. (5)) can be reformulated by letting $Z_k = [x_k, u_{k-1}]^T$ as the augmented state vector, this leads to Eq. (6)

$$Z_{k+1} = \tilde{\Phi}_k Z_k + \Gamma u_k \tag{6}$$

where

$$\tilde{\Phi}(k) = \begin{bmatrix} \phi & \Gamma_1(\tau_k) \\ 0 & 0 \end{bmatrix}$$

and

$$\Gamma = \begin{bmatrix} \Gamma_0(\tau_k) \\ I \end{bmatrix}$$

It can be noted that NCS in (Eq. (5)) is time-varying due to random delays τ_k in the channels. Note that the controller, based on the knowledge of upper bound of delays, can ensure the stability of the system but not the desired performance [2,27]. To meet the performance specification, the controller is needed that adapts its gains in accordance with the time-varying delays in the network. In particular, the controller should

1. Reduce the deviations of the state x_k from the reference signal r_k in the presence of time-varying delays
2. Stabilize the system
3. Ensure optimality of the control input.

As the design requirements are conflicting, the controller design becomes challenging.

3 Estimation of Delay τ_k

During experimentation with the industrial network, it was observed that the length, number of inputs/outputs, contention ratio for the channels (more slave

PLCs), and loading of the controller (number of lines-of-code/rungs in a ladder) affected the network delays. These factors are inherently time-varying and random, and it is unrealistic to develop a mathematical model that incorporates all these. To estimate the random delays, we use MRAN due to its capability to incorporate various factors that affect delays. Moreover, MRAN has learning capabilities that can be used to generate delay estimates after initial training with delay data based on network conditions. Note that this is the first application of MRAN in NCSs.

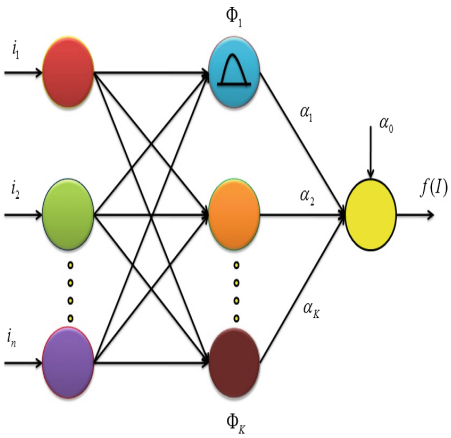


Fig. 2. Structure of RBF cf [28]

Figure 2 shows a typical Radial Basis Function network, with n inputs $I = [i_1 ... i_n]$ and a single output $f(I)$. The hidden layer consists of K computing units, Φ_1 to Φ_K, that are connected to the output by K connection weights α_1 to α_K. The output of the network has the form (where for simplicity a scalar input is assumed) in Eq. (7)

$$f(I) = \alpha_0 + \sum_{k=1}^{K} \alpha_k \Phi_k(I) \tag{7}$$

where $\Phi_k(I)$ is the response of the kth hidden neuron to the input I, and α_k is the weight connecting the kth-hidden unit to the output unit. By α_0 is denoted the bias term and K represents the number of hidden neurons in the network. In (Eq. (7)), $\Phi_k(I)$ is a Gaussian function given by Eq. (8)

$$\Phi_k(I) = exp(-\|I - \underline{\mu_k}\|_2^2 / \sigma_k^2) \tag{8}$$

where μ_k is the mean and σ_k^2 is the variance of the Gaussian function. The Euclidean norm is denoted by the symbol $\| \ \|$.

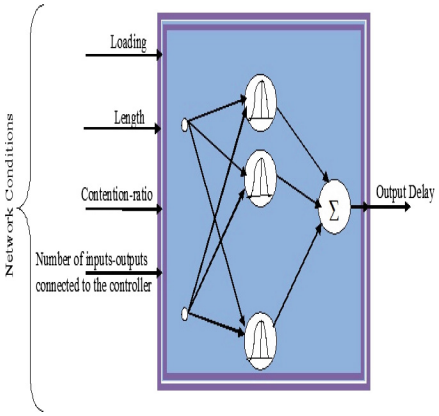


Fig. 3. MRAN-based delay estimation

The inputs to MRAN in our framework are network conditions and the output is delay estimate, see Fig. 3.

In the MRAN algorithm training begins with no hidden units. At each time-epoch M in off-line learning (from delay data) as training data (I_M, Y_M) is received, the network is built based on some growth and pruning criteria. The algorithm adds/prunes hidden neurons, as well as adjusts the existing network, according to the data received. For an outline of the various steps in the algorithm, see [28–32].

4 Adaptive Controller Design

This section presents the controller that adapts its gains based on the estimates of the time-varying delays in the channel with the ultimate goal to track the reference signal $r_k \in \Re^n$. The delay estimates are obtained from MRAN the inputs of which are the features of network conditions such as loading, length, contention-ratio, number of inputs-outputs connected to the controller. MRAN is trained using delay samples from high and low loading conditions. After initial learning MRAN generates at its output the estimate of delay. After training, MRAN generates estimate of the random delays when presented with samples of delays during the testing phase.

The control objective is to obtain good regulation, and tracking performance with minimal control effort and reasonable control complexity. For that purpose we employ LQR based approach to modify the controller gains that guarantee optimality (Fig. 4).

Theorem 1. Consider the NCS in Eq. (5) with time-varying delays τ_k. The controller is given by (Eq. (9))

$$u_k = L_x(\tau_k)x_k + L_u(\tau_k)u_{k-1} + L_r(\tau_k) \tag{9}$$

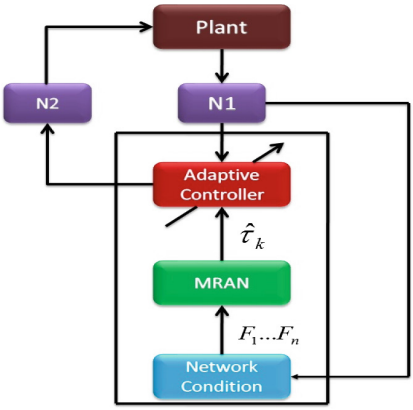


Fig. 4. Adaptive controller for NCSs

with adaptive gains

$$\begin{aligned} L_x(\tau_k) &= -(R + \Gamma_0^T P_{k+1} \Gamma_0)^{-1} \Gamma_0^T P_{k+1} \phi \\ L_u(\tau_t) &= -(R + \Gamma_0^T P_{k+1} \Gamma_0)^{-1} \Gamma_0^T P_{k+1} \Gamma_1 \\ L_r(\tau_t) &= -(R + \Gamma_0^T P_{k+1} \Gamma_0)^{-1} (\Gamma_0^T q_{k+1}) \end{aligned} \tag{10}$$

minimizes the cost function in Eq. (11)

$$\begin{aligned} J = \sum_{k=0}^{N-1} &((x_k - r_k)^T Q (x_k - r_k) + u_k^T R u_k) \\ &+ (x_N - r_N)^T Q_N (x_N - r_N) \end{aligned} \tag{11}$$

where r_k is the desired reference signal at time instant k, x_N is the terminal state, $Q \geq 0$ and $R > 0$ are weighing matrices of appropriate dimensions, and $Q_N \geq 0$ is the terminal weighing matrix.

Proof: Define the cost-to-go function at time k as,

$$\begin{aligned} J_k(x_k) = \min_{u_k, \ldots u_{N-1}} &\sum_{\chi=k}^{N-1} ((x_\chi - r_\chi)^T Q (x_\chi - r_\chi) + u_\chi^T R u_\chi) \\ &+ (x_N - r_N)^T Q_N (x_N - r_N) \\ &\textit{subject to relations} \quad (Eq.(5)) \end{aligned} \tag{12}$$

Note that $J_k(x_k)$ gives the minimum cost-to-go starting from the state x_k at time instant k. As $J_k(x_k)$ is quadratic, the cost-to-go has the form (Eq. (13))

$$J_k(x_k) = x_k^T P_k x_k + 2 q_k x_k + r_k^T Q r_k \tag{13}$$

where $q_k = -Q^T r_k$, and $P_k = P_k^T \geq 0$ that can be found recursively using backward iteration algorithm by taking $P_N = Q_N$.

Assuming that J_{k+1} has the given form, from dynamic programming principle, suppose we know $J_{k+1}(x_{k+1})$ and look for optimal input u_k the Bellman equation is given by (Eq. (14))

$$\begin{aligned} J_k(x_k) &= \min_u (x_k - r_k)^T Q(x_k - r_k) + u_k^T R u_k \\ &\quad + J_{k+1}(\phi x_k + \Gamma_0 u_k + \Gamma_1 u_{k-1}) \\ &= (x_k - r_k)^T Q(x_k - r_k) \\ &\quad + \min_u (u_k^T \; R \; u_k + (\phi x_k + \Gamma_0 u_k + \Gamma_1 u_{k-1})^T \\ &\quad P_{k+1}(\phi x_k + \Gamma_0 u_k + \Gamma_1 u_{k-1}) \\ &\quad + 2q_{k+1}^T(\phi x_k + \Gamma_0 u_k + \Gamma_1 u_{k-1}) \end{aligned} \tag{14}$$

From the necessary conditions of optimality, we obtain

$$R u_k + \Gamma_0^T P_{k+1}[\phi x_k + \Gamma_0 u_k + \Gamma_1 u_{k-1}] + \Gamma_0^T q_{k+1} = 0 \tag{15}$$

Little manipulation of Eq. (15) leads to Eq. (9) with the adaptive gains given by Eq. (10). □

Obviously, the computation of adaptive controller gains requires the knowledge of P_{k+1} and the estimate of time-delay τ_k to compute $\Gamma_0(\tau_k)$ and $\Gamma_1(\tau_k)$. We compute P_{k+1} recursively working backward from $k = N$ by letting $P_N = Q_N$, where Q_N is the terminal weighing matrix. The estimate of τ_k is obtained as described in Sect. 3. The Gains are computed from Eq. (10). Having obtained the results required for computing the gains, we now proceed to illustrate the controller by combining experiments with simulation.

5 Results and Discussions

This section presents the results obtained from experiments and simulations to illustrate the performance of the proposed adaptive controller.

5.1 Delay Estimation from Experiments

To estimate the time-varying delays, experimentation was conducted on Modbus over TCP/IP by connecting several controllers, and then loading the network (by increasing the number of inputs/outputs and number of rungs in PLC ladder diagram). Further, the length of the communication channel was varied by connecting controllers such as ABB AC 500 at different locations. The prototype of the experiment is shown in Fig. 5 wherein there is a Master controller (also called the Modbus Master), and Slave controllers.

The delays for variations of the parameters during experimentation is recorded both for training and validation phases of MRAN. During the network training phase the weights were adjusted to match the samples, whereas in the validation phase the delay estimates were compared with the estimates

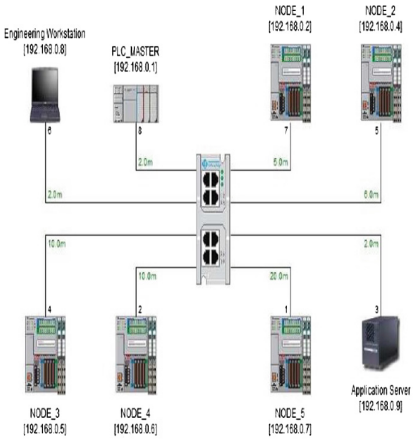


Fig. 5. Experiment pilot for delay estimation

generated by MRAN. The first 400 samples of delay data were used in learning, and the next 100 samples of data were used in validation of the estimates. After validation, in the test phase MRAN can be used to generate the delay estimates for designing the adaptive controller on-line.

The number of hidden neurons in MRAN, delay samples from the network, estimated delays, and the error (difference between the actual and estimated network delay) are shown in Fig. 6. It was found that MRAN can estimate the delay with a maximum error of 2.2 ms, and an average error of less than 1 ms over 10 samples.

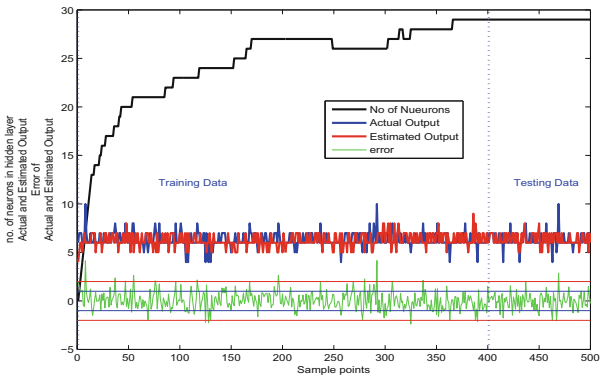


Fig. 6. Combined plot: number of Hidden neuron, actual, estimated output and error

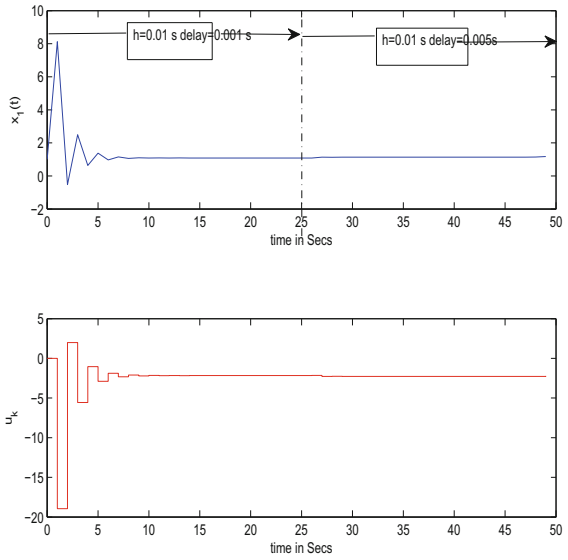


Fig. 7. Response of the system (Eq. (16)) to the proposed adaptive controller and fixed delays

5.2 Example 1

To illustrate MRAN-based adaptive controller, we consider the scalar example

$$\dot{x}(t) = 2x(t) + u(t) \tag{16}$$

with $r_k = 1$ and initial condition $x_0 = 2$ as in [12]. The poor performance of the static state feedback controller for this example has been illustrated by the authors of [14]. Our investigation uses a sampling time of 10 ms, and the estimated delay varies from 4 ms to 10 ms. Figures 7 and 8 show the response along with the control input u_k with delays being changed at $t = 25$ s and delay varying at every 10 s using delay estimates from MRAN, respectively. This result illustrates that MRAN based controller adapts the gain to track the controller input even in the case of fast variations in delay.

5.3 Example 2

The investigation considers the double integrator system

$$\begin{aligned} \dot{x}(t) &= \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} x(t) + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u(t) \\ y(t) &= \begin{bmatrix} 1 & 0 \end{bmatrix} x(t) \end{aligned} \tag{17}$$

to illustrate the effectiveness of the proposed adaptive regulator. The continuous-time dynamics (Eq. (17)) is discretized with sampling period h under the assumption that $\tau_k \leq h$, resulting in Eq. (18)

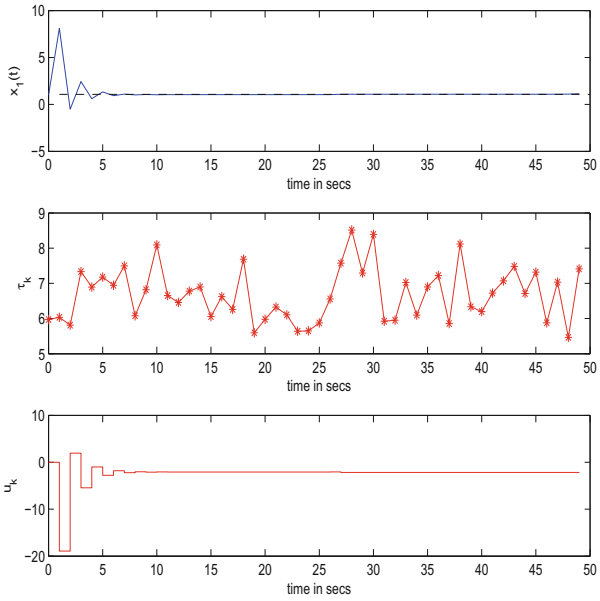


Fig. 8. Response of system (Eq. (16)) to the proposed adaptive controller and variable delays, estimated by MRAN

$$x_{k+1} = \begin{bmatrix} 1 & h \\ 0 & 1 \end{bmatrix} x_k + \begin{bmatrix} \frac{(h-\tau_k)^2}{2} \\ h - \tau_k \end{bmatrix} u_k + \begin{bmatrix} \tau_k(h - \frac{\tau_k}{2}) \\ \tau_k \end{bmatrix} u_{k-1} \tag{18}$$

The sampled-data representation of linear NCSs with total delay (at each sampling instant k) less than the sampling time, i.e. $\tau_k \leq h$, can be obtained using MATLAB routine NCSd developed by the authors of the paper, see [33].

Figure 9 shows the response of the system (Eq. 18) under the MRAN-based adaptive controller ((Eq. 5) using the constant reference signal $r_k = [0 \quad 0]^T$ for each k. The Figure additionally shows the control values u_k and the delay estimates from MRAN. The result shows the regulatory performance of the controller in the presence of delays.

Variations in the gains L_x, L_r, and L_u with time-varying delays τ_k are shown in Fig. 10. The gains are adjusted according to channel delays τ_k to track the reference state, and these variations are reflected mostly in the gain L_u.

The states of the double integrator system in Eq. (18) with the adaptive controller (Eq. (5)) and desired states $r_k = [1 \quad 0]^T$ are shown in Fig. 11. This response illustrates the tracking performance of the adaptive controller.

Variations in gain L_x, L_r, and L_u with time-varying delays τ_k are shown in Fig. 12.

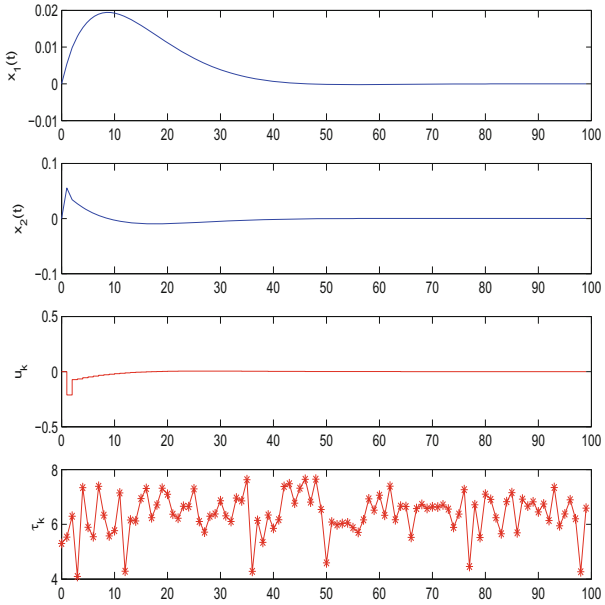


Fig. 9. States of the system and the control input with time-varying delay

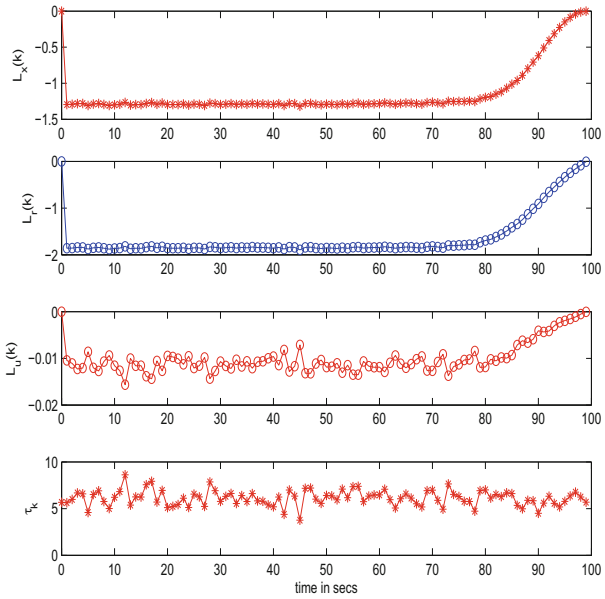


Fig. 10. Variation in gains L_x, L_r, and L_u with delays for regulatory performance

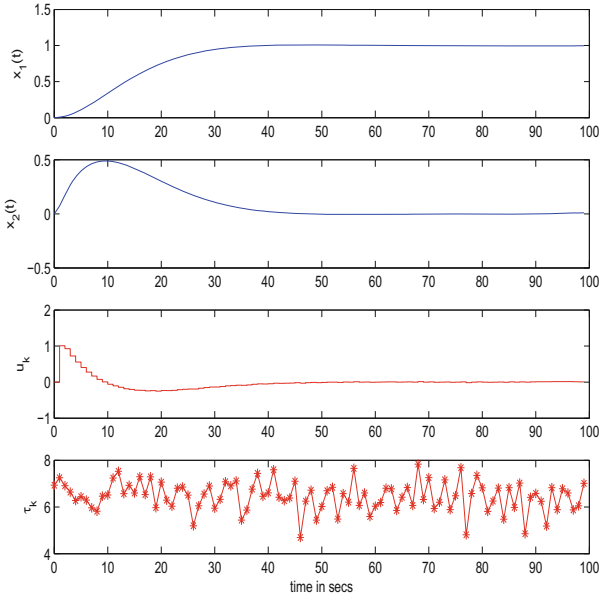


Fig. 11. States of the system and the control input for tracking the reference with time-varying delays ($x_0 = [0 \quad 0]^T$)

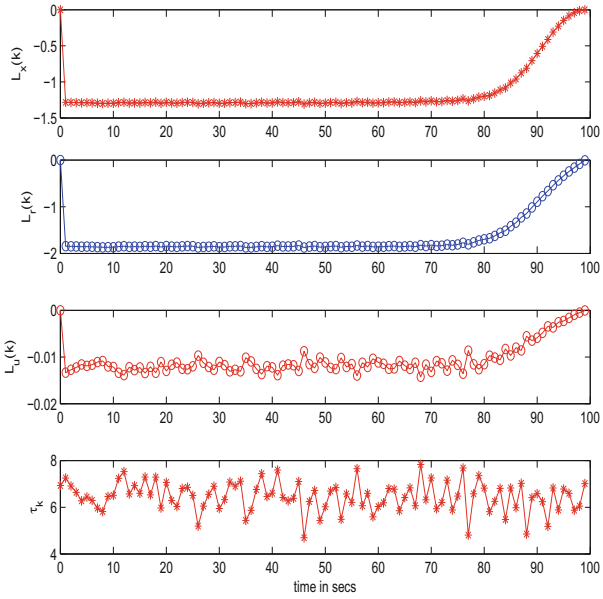


Fig. 12. Variation in gains L_x,L_r, and L_u with delay τ for tracking performance

6 Conclusion

In this investigation time varying delays in communication channels were modeled using MRAN, based on experiments conducted with Modbus over TCP/IP network. Information on network conditions such as loading, communication length, contention ratio, and number of connected nodes along with experimental delays was used to train the MRAN. Once trained using network conditions, MRAN generates delay estimates that were used to design adaptive controller for NCSs subjected to time-varying delays. MRAN-based estimation yields a learning based framework for estimating delays and, therefore, is more suitable for real-time control. The adaptive controller uses LQR approach to compute the gains and as a result, optimality is guaranteed by construction. Then, conditions for stability of the proposed controller were derived. The results were illustrated using two simple simulation examples along with delay estimates from MRAN. Results indicate that the adaptive controller modifies its gain based on time-varying delays in communication channels to meet the performance requirements. Adaptive controller by addressing packet-loss and implementing the adaptive control in real-time industrial pilot are the future prospects of this investigation.

Acknowledgments. This work was supported by European Union through European Regional Development Fund and the Estonian Research Council grant PUT481.

References

1. Hespanha, J.P., Naghshtabrizi, P., Xu, Y.: A survey of recent results in networked control systems. Proc. IEEE **95**(1), 138 (2007)
2. Seshadhri, S.: Estimation and design methodologies for networked control systems with communication constraints. Ph.D. thesis, NIT-Trichy (2003)
3. Zhang, L., Gao, H., Kaynak, O.: Network-induced constraints in networked control systems survey. IEEE Trans. Ind. Inform. **9**(1), 403–416 (2013)
4. Seshadhri, S., Ayyagari, R.: Platooning over packet-dropping links. Int. J. Veh. Auton. Syst. **9**(1), 46–62 (2011)
5. Qu, F.-L., Guan, Z.-H., Li, T., Yuan, F.-S.: Stabilisation of wireless networked control systems with packet loss. IET Control Theory Appl. **6**(15), 2362–2366 (2012)
6. Srinivasan, S., Buonopane, F., Vain, J., Ramaswamy, S.: Model checking response times in networked automation systems using jitter bounds. Comput. Ind. **74**, 186–200 (2015)
7. Balasubramaniyan, S., Srinivasan, S., Buonopane, F., Subathra, B., Vain, J., Ramaswamy, S.: Design and verification of cyber-physical systems using truetime, evolutionary optimization and uppaal. Microprocess. Microsyst. **42**, 37–48 (2016)
8. Tzes, A., Nikolakopoulos, G.: LQR-output feedback gain scheduling of mobile networked controlled systems. In: Proceedings of the 2004 American Control Conference, 2004, vol. 5, pp. 4325–4329. IEEE (2004)
9. Godoy, E.P., Porto, A.J., Inamasu, R.Y.: Sampling time adaptive control methodology for can-based networked control systems. In: 2010 9th IEEE/IAS International Conference on Industry Applications (INDUSCON), pp. 1–6. IEEE (2010)

10. Piltz, S., Bjorkbom, M., Eriksson, L.M., Koivo, H.N.: Step adaptive controller for networked MIMO control systems. In: 2010 International Conference on Networking, Sensing and Control (ICNSC), pp. 464–469. IEEE (2010)
11. Björkbom, M., et al.: Wireless control system simulation and network adaptive control. Ph.D. Dissertation, Dept. of automation science and technology, School of Science and Technology
12. Voit, H., Annaswamy, A.: Adaptive control of a networked control system with hierarchical scheduling. In: American Control Conference (ACC), 2011, pp. 4189–4194. IEEE (2011)
13. Marti, P., Yepez, J., Velasco, M., Villa, R., Fuertes, J.: Managing quality-of-control in network-based control systems by controller and message scheduling co-design. IEEE Trans. Ind. Electron. **51**(6), 1159–1167 (2004). doi:10.1109/TIE.2004.837873
14. Loden, N.B., Hung, J.Y.: An adaptive PID controller for network based control systems. In: 31st Annual Conference of IEEE Industrial Electronics Society, 2005, IECON 2005, pp. 6–pp. IEEE (2005)
15. Tahoun, A., Hua-Jing, F.: Adaptive stabilization of networked control systems. J. Appl. Sci. **7**(22), 3547–3551 (2007)
16. Chunmao, L., Jian, X.: Adaptive delay estimation and control of networked control systems. In: International Symposium on Communications and Information Technologies, 2006, ISCIT 2006, pp. 707–710. IEEE (2006)
17. Xu, H., Jagannathan, S., Lewis, F.L.: Stochastic optimal control of unknown linear networked control system in the presence of random delays and packet losses. Automatica **48**(6), 1017–1030 (2012)
18. Watkins, C.J.C.H.: Learning from delayed rewards. Ph.D. thesis, University of Cambridge (1989)
19. Luck, R., Ray, A.: Delay compensation in integrated communication and control systems: part II-implementation and verification. In: American Control Conference, 1990, pp. 2051–2055. IEEE (1990)
20. Chan, H., Özgüner, Ü.: Closed-loop control of systems over a communications network with queues. Int. J. Control **62**(3), 493–510 (1995)
21. Nilsson, J., et al.: Real-time control systems with delays. Ph.D. thesis, Lund Institute of Technology Lund, Sweden (1998)
22. Seshadhri, S., Ayyagari, R.: Dynamic controller for network control systems with random communication delay. Int. J. Syst. Control Commun. **3**(2), 178–193 (2011)
23. Ghanaim, A., Frey, G.: Modeling and control of closed-loop networked PLC-systems. In: American Control Conference (ACC), 2011, pp. 502–508. IEEE (2011)
24. Srinivasan, S., Vallabhan, M., Ramaswamy, S., Kotta, U.: Adaptive LQR controller for networked control systems subjected to random communication delays. In: American Control Conference (ACC), 2013, pp. 783–787. IEEE (2013)
25. Srinivasan, S., Vallabhan, M., Ramaswamy, S., Kotta, U.: Adaptive regulator for networked control systems: Matlab and true time implementation. In: 2013 25th Chinese Control and Decision Conference (CCDC), pp. 2551–2555. IEEE (2013)
26. Vallabhan, M., Seshadhri, S., Ashok, S., Ramaswmay, S., Ayyagari, R.: An analytical framework for analysis and design of networked control systems with random delays and packet losses. In: Proceedings of the 25th Canadian Conference on Electrical and Computer Engineering (CCECE) (2012)
27. Lian, F.-L., Moyne, W., Tilbury, D.: Optimal controller design and evaluation for a class of networked control systems with distributed constant delays. In: Proceedings of the 2002 American Control Conference, 2002, vol. 4, pp. 3009–3014. IEEE (2002)

28. Yingwei, L., Sundararajan, N., Saratchandran, P.: Identification of time-varying nonlinear systems using minimal radial basis function neural networks. In: IEE Proceedings of the Control Theory and Applications, vol. 144, pp. 202–208. IET (1997)
29. Platt, J.: A resource-allocating network for function interpolation. Neural Comput. **3**(2), 213–225 (1991)
30. Kadirkamanathan, V., Niranjan, M.: A function estimation approach to sequential learning with neural networks. Neural Comput. **5**(6), 954–975 (1993)
31. Yingwei, L., Sundararajan, N., Saratchandran, P.: A sequential learning scheme for function approximation using minimal radial basis function neural networks. Neural Comput. **9**(2), 461–478 (1997)
32. Yingwei, L., Sundararajan, N., Saratchandran, P.: Performance evaluation of a sequential minimal radial basis function (rbf) neural network learning algorithm. IEEE Trans. Neural Netw. **9**(2), 308–318 (1998)
33. Seshadhri, S.: Sampled-data model for ncss with delay less than sampling time (2012). http://www.mathworks.com/matlabcentral/fileexchange/37875/. Accessed 21 July 2013

Aspects Regarding Risk Assessment of Human Body Exposure in Electric and Magnetic Fields

Marius Lolea and Simona Dzitac(✉)

Department of Energy Engineering, University of Oradea, Oradea, Romania
mlolea@uoradea.ro, simona.dzitac@gmail.com

Abstract. The paper gives an idea of attaching the risk of electromagnetic field exposure in the general population and employees with potential health hazards that are related to the different sensitivity of the human body. The difference between the two categories of people, is that in the first case, the intersection with the devices which generating electromagnetic field is optional both in duration, number or distance, and in the second case this option is canceled due to duties of services that give the obligation of employees to stationed in areas with such types of devices. Even if there are thresholds exposure limit values for electric and magnetic field quantities considered hazardous to health, referred to in legal settlements can not say with certainty that human sensitivity and is given the same thresholds for different age and sex categories. These fit rather with probability theory. Therefore the work is appreciated and that the risk to health or human behaviors generated by electromagnetic field exposure related to probability theory and treatment of this risk is therefore using a probabilistic model.

1 Introduction

The multiple types of specific diseases of modern civilization and the large number of reported cases led to the idea of search for new cause increased risk of their generation. Among the causes were included and electromagnetic field exposure. Fears are amplified by the existence of numerous electromagnetic devices that help to increase life comfort and technical progress.

Epidemiological and laboratory studies, conducted at worldwide lavel have highlighted several problems attributed field with different degrees of certainty. The results of these studies are controversial [5, 9, 11]. However, national rules limiting the exposure of persons, the maximum limit values are provided for parameters such as electric and magnetic field and the exposure so that the effects are not harmful to health [6, 7]. The interplay between electromagnetic fields and living matter are very different and depend on the frequency of electromagnetic waves, the amplitude of characteristic parameters of the field and proportion of energy transmitted to the body [5, 9–12]. Biological implications of exposure to magnetic fields against human body, is can study methodological with a cybernetic or an analytical method [5]. Cybernetic method consider the biological system as a black box characterized input quantities, output sizes and connections between them. Input values are given by electromagnetic field parameters. Output sizes are effects on the living organism.

Connections between the sizes of input and output are formed from interaction mechanisms which are obtained on experimental basis. The analytical method is based

V.E. Balas et al. (eds.), *Soft Computing Applications*, Advances in Intelligent Systems and Computing 633, DOI 10.1007/978-3-319-62521-8_9

on mathematical modeling of biological systems. The phenomena that occur at the molecular level to the application of an electromagnetic field are: electronic excitation, the polarization, the appearance of electrical forces, thermal effects, dipolar interactions, etc. Exist, circuit diagrams of the living cell developed and mathematical models to study the electromagnetic interaction at anatomic level [5]. In a few situations electromagnetic field impact with human body it is a risk factor occupational diseases [4].

The purpose of the paper is to assess the level of human exposure through measurements values for the electromagnetic field generated by electrical power networks equipments and risk of exceeding normal levels considered dangerous of these sizes. For this second case apply a probabilistic model [1–3].

2 Probabilistic Methods from Risks Assessment

The risk associated with occurrence of a certain event comes down, in the simplest form, as the product of the probability of that event and the consequences of that event. For example, for a given level of risk desired, as the likelihood of an event decreases the potential consequences products may increase. All the events (states), through which a physical system during the experience is called field events. By associating a field event of an experiment may be obtained depth information on knowledge of the phenomenon studied [1, 3]. The probability transferred in quantitative and intuitive plan of numbers, a concrete situation pattern of events, allows some comparison of events. In this way, the probability is a measure of risk.

The concepts of probability and probability finite field can be presented as axiomatic form.

By Kolmogorov, probability is defined by the following axioms:

- each event $A \in K$, of the field events K, is associated with a positive number P {A}, called the probability of the event;
- P {E} = 1 is the probability of the event is safe and has a maximum value;
- if two events A, $B \in K$, are incompatible, then: $\{P\} = P\ A \cup B\ \{A\} + P\ \{B\}$.

The concepts of probability of appearance p, and non-appearance probability r of an event for which p + r = 1, allow to judge between safety and risk that there is a complementary relationship. On the basis of the axioms of Kolmogorov, it follows that R(p) + S(r) = 1, where R(p) is the probability of risk corresponding to an event, and S(r) - additional safety risk. The relation: R(p) + S(r) = 1 reveal that an increased risk involving a decrease of safety and a risk reduction involve an increases of the safety, the two sizes having values in the range [0, 1].

In the general case it can write:
$0 < R(p) < 1$ sau $0 < S(r) < 1$, after which, in particular:
If $R(p) \rightarrow o$, it can write $S(r) \rightarrow 1$ and inverse.
If danger determines the risk, then the risk can be defined as a measure of danger P (p) and its extent permitted limit value:

$$R(p) = \lim_{n \to \infty} P(n) = P(p)$$

For the particular case of the action of the electromagnetic field on the human body either by direct coupling or by electromagnetic radiation, the danger is delimited by limit values standardized of field quantities considered harmful. Thus, assessing the probability of exceeding the normal levels that are considered generating negative health effects, is a step in assessing risk from exposure. The risk here can be defined as the product of probability of exceeding the exposure limit values for E and B and that the consequences of these exceedances, the latter being far damage human life or health. Following graphs below of normal distributions, we can estimate the risk from producing an unwanted event by leaving the chosen interval of significance [8] (Fig. 1).

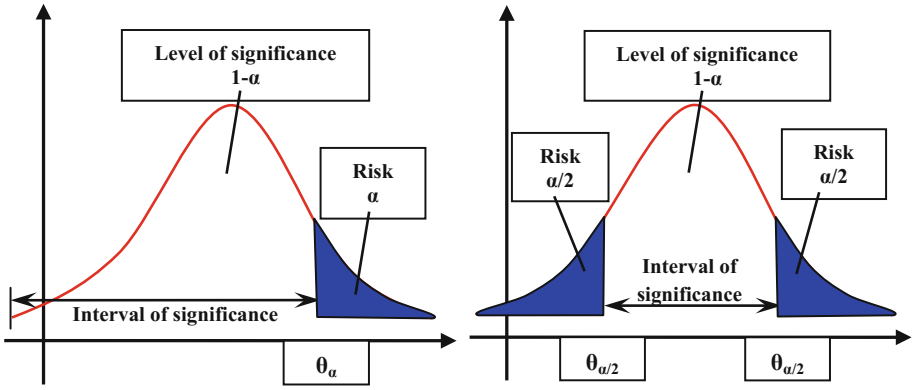


Fig. 1. The link between the meaning range and of probability: a - to a one-sided risk; b - for a bilateral risk (adapted from [8])

To a certain repartition function, meaning the connection between interval and level of significance, it is given by relation:

$$P(\theta < \theta_\alpha) = 1 - \alpha, \tag{1}$$

for a unilateral risk, or by relation:

$$P(\theta_{1-\frac{\alpha}{2}} < \theta < \theta_{\frac{\alpha}{2}}), \tag{2}$$

For a bilateral risk,

where: $1 - \alpha$ is the level of significance;

$\theta_a - \theta_b$ = interval of significance; α – the risk.

Assessed risk the consequences of exposure to electromagnetic field is generated type as unilateral effects considered dangerous to humans is only upside quantities of electric and magnetic field above their normal intake. A typical case of bilateral risk is generated by electromagnetic disturbances such as voltage fluctuations. Whether the voltage drops below a certain threshold, as in dips or rises above the maximum value that this service became Surge, risks in product supply electricity to consumers are double effect.

3 Risk Assessment Generated to Humans Exposured in Electromagnetic Field Generated by Power Grids

Assessing the risk of exceeding the permissible values for magnetic induction (B) and electric field (E) is performed by comparing the density function of considered size distribution (f(B) and/or (E)) with the permissible limit value of these sizes set of regulations in the field of workers' exposure to electromagnetic fields and/or the general population [2, 3].

For calculations follow the next steps:

(a) shall be taken from measurements the values of magnetic induction B and electric field strength E at various locations;
(b) is calculated for each location, the arithmetic mean and standard deviation of the values taken at the measuring points;
(c) Considering that values have a normal distribution, determine the probability density function expression (f(B)/f(E)). The form of this function is [3]:

$$f(M) = \frac{1}{\sqrt{2\pi} \cdot \sigma} \cdot e^{-\frac{1}{2} \cdot \left(\frac{M-m}{\sigma}\right)^2} \tag{3}$$

where:

m – the mean of measured values;
σ – standard deviation of measured values;
M = field sizes considered: {E, B}.

Measured values of the magnetic induction and electric field strength is considered that follow a normal distribution. Calculation was developed in Mathcad application, version 14.0.

The calculation algorithm initiated by the authors has the following form:

- Defines the Matrix of the maximum measurements of magnetic induction B and electric field strength, E.

It has the following tabular form for B, in the first case:

B1 :=

	Measuring data 1	2	3
1	11.4		
2	12.6		
3	13.5		
4	…		

- Number of datasets from matrix B1: → NCol := 1

The limit value for B: → Blim := 100

The computing vector, has the form:

$$
\begin{aligned}
DD := \Bigg| \; & NLinii \leftarrow length\left(B1^{\langle 1 \rangle}\right) \\
& NDate_{NCol} \leftarrow 0 \\
& \text{for } i \in 1..NCol \\
& \quad \text{for } j \in 1..NLinii \\
& \quad\quad NDate_i \leftarrow NDate_i + 1 \text{ if } B1_{j,i} \neq 0 \\
& \text{for } i \in 1..NCol \\
& \quad Media_i \leftarrow \frac{\sum_{j=1}^{NDate_i} B1_{j,i}}{NDate_i} \\
& \quad AS_i \leftarrow \sqrt{\frac{\sum_{j=1}^{NDate_i} \left(B1_{j,i} - Media_i\right)^2}{NDate_i - 1}} \\
& \text{for } i \in 1..NCol \\
& \quad Risc_i \leftarrow 1 - pnorm\left(BLim, Media_i, AS_i\right) \\
& \begin{pmatrix} Media \\ AS \\ Risc \end{pmatrix}
\end{aligned}
$$

The elements of the vector above are: DD_1 = mean, DD_2 = standard deviation, DD_3 = the risk;

No. of column inside matrix B1, k: → k := 1

$$f(B): = dnorm\left[B,(DD_1)_k,\ (DD_2)_k\right] \rightarrow \text{Density distribution function, for B;}$$

$$f1(B): = dnorm\left[B,(DD_1)_k,(DD_2)_k\right](B > BLim) \rightarrow \text{Marks the risk;}$$

- Drawing graph to distribution function, finally.

The algorithm applies to workers of Hydropower plants(HPP) and for the general population. Through the activities undertaken and functional characteristics of the electricity networks they serve workers are exposed to electromagnetic fields generated by them. The electromagnetic environment is formed due to the operation of all sources of a given site or a particular territory. Therefore for each location we need to consider

all sources of electromagnetic field. Measuring devices used were: CA 42 Field meter and SPECTRAN 5035, both equipped with internal probes of magnetic field and external for electric field. Displaying values is in V/m for E and in μT for B. Measuring step is noted by d_m and is chosen as equal to 2 m, to detect with high accuracy, the maximum values of quantities of electric and magnetic field. With measured values will be calculated the averages, standard deviations and risk probability [1–3].

For the general population analyzes a portion of a high voltage overhead power lines (PL), which crosses the city of Oradea through a crowded area. In the case of power lines(PL) measurements have been performed under conductors, in line axis.

Technical characteristics of the PL for the aperture considered, are: circuit type d.c., U_n = 110 kV, connection between substations Oradea Sud – Oradea Vest, the portion of Auchan supermarket, on Radu Enescu street, type poles: STC, aperture lenght = 126 m, composite insulators, conductors type: OL-Al 185/32 mm^2, gauges: cond.1 = 13.7 m, cond.2 = 13.4 m, cond.3 = 14.05 m, cond.4 = 13.17 m, cond.5 = 14.25 m, cond.6 = 13.76 m; number of measurements: 126/2 = 63 at measure step considered. On the HPP Tileagd case, is exemplified with the values of magnetic induction B, taken around of the generators excitation, where operational staff have access, expressed in μT. In the case of generators the measurements were carried out on a circular path around them. Threshold values within normal limits, are [6, 7]: E_{lim} = 5000 V/m for the general population, E_{lim} = 10000 V/m for employees; B_{lim} = 100 μT for the general population; B_{lim} = 500 μT for employees.

Allowable limit values were periodically adjusted by their decreasing, due to the fact that research into exposure domain have been perfected over time. Therefore, to consider the Gaussian variations of their, is appropriate.

With considered fixed limit values, Fig. 2 shows the risk graph for the aperture analyzed and in Fig. 3 shows the risk graph for operators from generators hall of HPP Tileagd, belonging to Bihor Power System. With blue color was marked risk area.

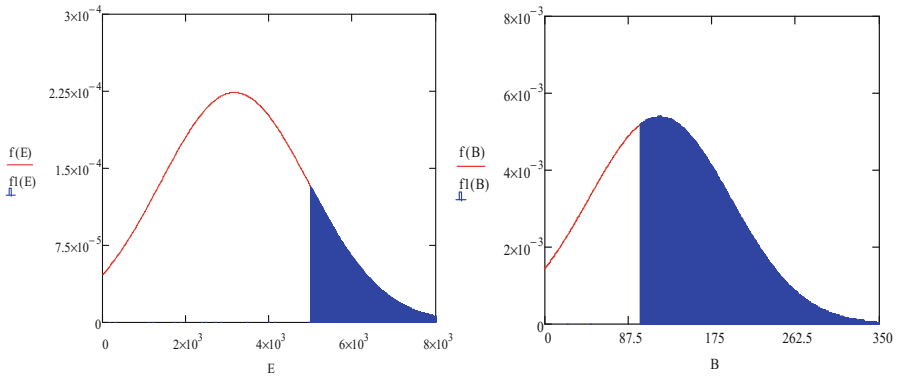


Fig. 2. The graph of risk: a - with values of E for PL 1/110 kV, aperture 1; b - with values of B for considered case inside HPP Tileagd

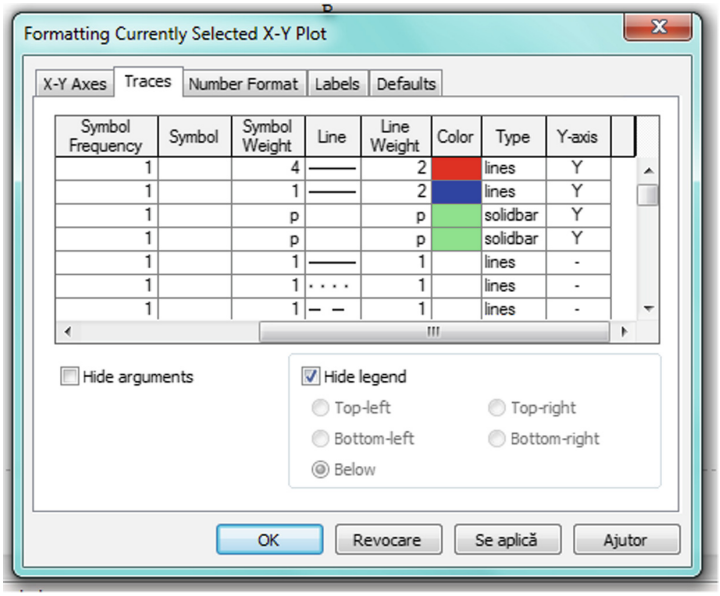


Fig. 3. Setting data for considered case in MathCad software

For evidence of common risk corresponding of the area under the intersection of the two curves are written equations corresponding to that point unknown p. For identification of solution is necessary to equating the two equations. How the two points are at the same level on the chart at the intersection of the two curves density distribution, they must have the same coordinates so $p_1(x_1, y_1) = p_2(x_1, y_1)$ and can write:

$$\text{for } f(p_1) = \frac{1}{\sqrt{2}\pi\sigma_1} \cdot e^{-\frac{1}{2}\left(\frac{p_1 - m_1}{\sigma_1}\right)^2} \tag{4}$$

$$\text{for } f(p_2) = \frac{1}{\sqrt{2}\pi\sigma_{adm}} \cdot e^{-\frac{1}{2}\left(\frac{p_2 - m_2}{\sigma_{adm}}\right)^2} \tag{5}$$

For the case in which the standard deviations for the two sets of data, ranging to Gaussian distribution are equal, result:

$$\sigma_1 = \sigma_{adm} = \sigma, \tag{6}$$

and because the two points coincide on chart, namely $p_1 = p_2 = p$, by equalizing the two expressions follows:

$$\frac{1}{\sqrt{2}\pi\sigma} \cdot e^{-\frac{1}{2}\left(\frac{p - m_1}{\sigma}\right)^2} = \frac{1}{\sqrt{2}\pi\sigma} \cdot e^{-\frac{1}{2}\left(\frac{p - m_2}{\sigma}\right)^2} \tag{7}$$

By logarithm, will get:

$$-\frac{1}{2}\left(\frac{p-m_1}{\sigma}\right)^2=-\frac{1}{2}\left(\frac{p-m_2}{\sigma}\right)^2 \tag{8}$$

And after simplifying and solving the resulting equation, the unknown being selected value will be calculated with the following expression:

$$p=\frac{m_1+m_2}{2}, \tag{9}$$

then, in case of the means of the written program running in MathCad will get: $p = 70.206$, with $m_1 = 40.412$ and $m_2 = 100$.

For the other calculation hypotheses, in which $\sigma_{adm}=\frac{1}{2}\sigma$ and $\sigma_{adm}=\frac{1}{4}\sigma$ resulting the equalities:

$$\text{For case 2}: \frac{1}{\sqrt{2}\pi\sigma}\cdot e^{-\frac{1}{2}\left(\frac{p-m_1}{\sigma}\right)^2}=\frac{1}{\sqrt{2}\pi\cdot\frac{\sigma}{2}}\cdot e^{-\frac{1}{2}\left(\frac{p-m_2}{\sigma/2}\right)^2} \tag{10}$$

$$\text{For case 3}: \frac{1}{\sqrt{2}\pi\sigma}\cdot e^{-\frac{1}{2}\left(\frac{p-m_1}{\sigma}\right)^2}=\frac{1}{\sqrt{2}\pi\cdot\frac{\sigma}{4}}\cdot e^{-\frac{1}{2}\left(\frac{p-m_2}{\sigma/4}\right)^2} \tag{11}$$

Which have the next solutions:

For Eq. (10):

$$p_{1,2}=\frac{4m_2-m_1\pm 2\sqrt{4(m_1-m_2)^2+6\sigma\ln 2}}{3} \tag{12}$$

and in case of Eq. (11):

$$p_{1,2}=\frac{16m_2-m_1\pm 2\sqrt{16(m_1-m_2)^2+30\sigma^2\ln 4}}{15} \tag{13}$$

To generate common area beneath the two curves, the left and right of the point of intersection called risk area, perform the necessary settings in Mathcad. These are indicated in Fig. 3.

In Fig. 4 is shown in green color, the resulting risk in case the two standard deviations are equal.

In case of the measured magnetic flux density values exemplified for synchronous generators hall of o HPP Tileagd, they were chosen radial directions for further circular and concentric routes around excitation systems. Since the operators do not have to stand throughout the service duration near excitation systems, the acceptable value exposure to magnetic induction is considered the value of $B_{lim} = 100\ \mu T$. In Figs. 5 and 6, σ, σ_2 and σ_3, are the standard deviations for another two considered cases.

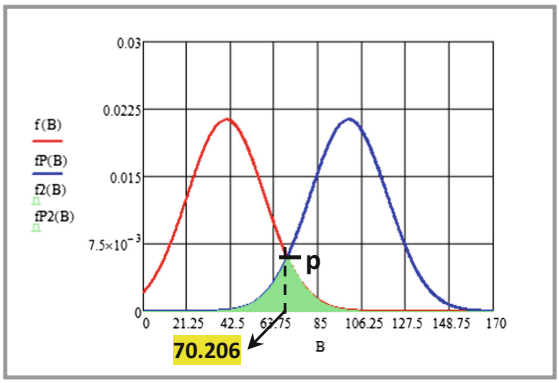


Fig. 4. Risk graph in the second case considered

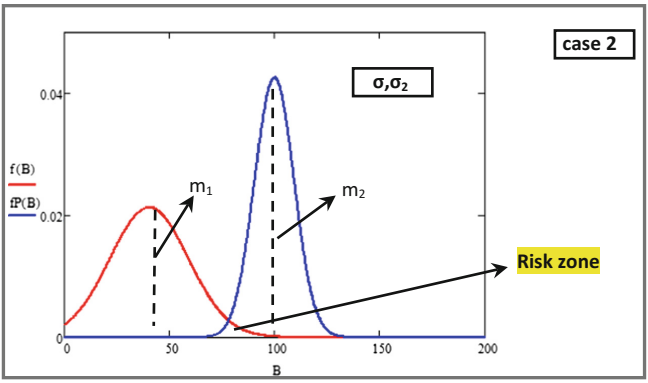


Fig. 5. The distribution of magnetic induction admissible with standard deviation in portion of 0.25% from σ of a real values measured

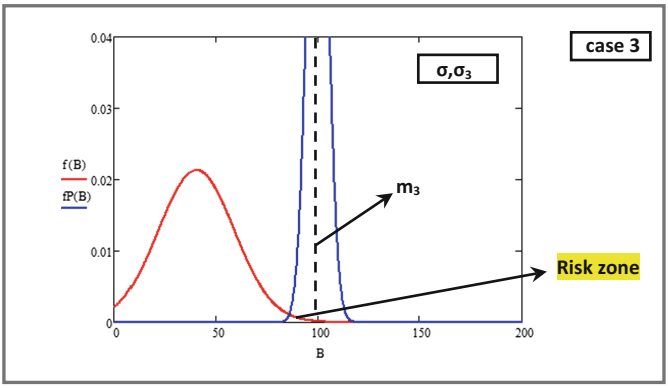


Fig. 6. The distribution of magnetic induction admissible with standard deviation in portion of 0.50% from σ of a real values measured

Algorithms presented in the paper, apply for other power plants in Bihor County. Algorithms presented in the paper, apply for other power plants in Bihor County. For the first scenario considered when considered dangerous occupational exposure limit values are considered fixed, calculated parameters are summarized in Tables 1 and 2. Probability of exceeding the admissible limits, allows ranking of these facilities in terms of the risk from exposure to the population and employees in the energy sector. This is shown in Fig. 7.

Table 1. Calculated parameters values for B, in case of HPP

Crt. No.	Objective of the measurements	Measurements number	Mean (M)	Standard deviation (σ)	Overtaking probability (risk/P_r)
1	HG hall -HPP Fughiu	246	22.5	0.094554	**0.047421**
2	HG hall - HPPSăcădat	256	37.5	0.128673	0.023425
3	HG hall - HPP Tileagd	560	25.7	0.137553	**0.039424**
4	HG hall - HPP Lugașu	580	35.8	0.154553	0.023423
5	HG hall - HPP Munteni	380	22.1	0.188553	0.015976
6	HG hallo – HPP Remeți	380	28.3	0.174553	**0.024436**

Table 2. Calculated parameters values for E, in case of PL

Crt. No.	Objective of the measurements	Measurements number	Mean (M)	Standard deviation (σ)	Overtaking probability (risk/P_r)
1	PL 1	1246	22.5	0.134554	0.037421
2	PL 2	1756	37.5	0.124673	**0.057423**
3	PL 3	2560	25.7	0.134553	**0.038424**
4	PL 4	2180	35.8	0.184553	0.033423
5	PL 5	980	22.1	0.184553	**0.046475**
6	PL 6	1380	28.3	0.154553	0.033436

Figure 7 shows the ranking of categories of people according to risk exposure caused by the electric and magnetic field in the two objectives analyzed.

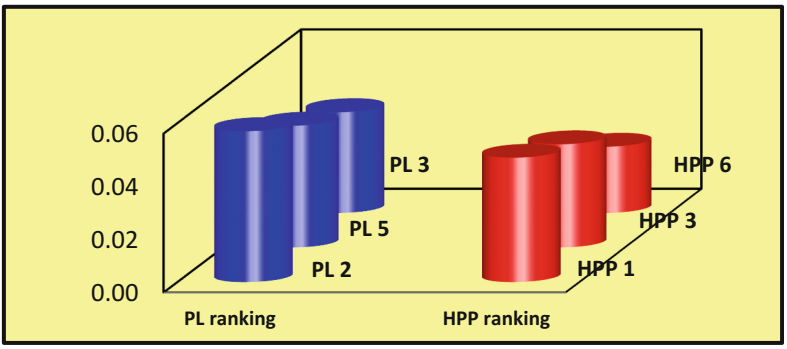


Fig. 7. The Graph of risk for exposure in electric field. PL and HPP ranking

4 Conclusions

In the context of scientific uncertainty regarding the effects of exposure to electromagnetic fields advisable to adopt the precautionary approach by implementing administrative measures, inform and train the population, especially those at risk of exposure sources and not least support from decision makers, on the development of an adequate logistics structure from monitoring electromagnetic fields and their effects in the long term.

The calculation model which is adopted, not give information about risk consequences but, based on the distribution of the measured values enable the assessment of the likelihood of exceeding the normal threshold for electric field intensity electric and magnetic field induction on exposure to electromagnetic field of workers from the power system or the general population. Exposure Risk assessment is done by comparing the results of calculation of probability values exceeded the normal values of field sizes for all staff categories analyzed. Standard deviations and default the dispersions of measured values, are more pronounced where the measuring points were distributed over large areas with a great variety of equipment and devices that work with large differences between voltages and currents service. However, if one considers that there are sensitivities and different reactions bodies in terms of exposure to electromagnetic field as suggested in the literature, along with the assumption accepted the job as the limit values allowable quantities Field pursue distribution normal, then the risk of negative effects on the human body increases. This is evident because the overlap of the two distributions is more pronounced in the lower levels of the two categories of values, measured and admissible. Then probabilistic model adopted is justified especially if it is found that the values of magnetic induction and electric field strength, results of measurements, follow a normal distribution.

References

1. Felea, I., Coroiu, N., Stoica, I., Dubău, C.: Evaluarea unor elemente de risc pentru personalul de exploatare al rețelelor electrice. Analele Universității din Oradea, fascicula de energetic, vol. I, pp. 37–50. Secțiunea II/Electroenergetică (2001)
2. Felea, I., Secui, C., Muşet, A.: Evaluation of risk affecting the exposed human organism in operational electromagnetic field. In: The 6th International workshop of Electromagnetic Compatibility CEM, Constanţa, Romania, 12–14 November 2009
3. Felea, I.,Coroiu, N., Secui, C.: Asupra riscului de expunere în câmp electromagnetic. Simpozionul Național Siguranţa în funcţionare a Sistemului Energetic SIG, SC Electrica Deva, 26–28 September, vol II, pp. 402–410 (2001)
4. Virginia, M.: Câmpuri electromagnetice de joasă frecvenţă factori de risc profesional, Rev. Acta Medica Transilvania, **II**(1), 27–30 (2011)
5. Mihaela, M.: Bioelectromagnetism, Editura Matrix Rom, Bucureşti (1999)
6. Guvernul României, Hotarârea Nr. 1136 din 30.08.2006 privind cerintele minime de securitate si sanatate referitoare la expunerea lucratorilor la riscuri generate de câmpuri electromagnetice
7. Guvernul României, Hotarârea Nr. 1193 din 29.09.2006 privind limitarea expunerii populatiei generale la câmpuri electromagnetice de la 0 la 300 GHz
8. Munteanu, T., Gurguiatu, G., Bălănuţă, C.: Fiabilitate şi Calitate în Inginerie electrică. Aplicaţii (ed.) Galati University Press (2009). ISBN 978-606-8008-25-7
9. Milham, S.: Dirty Electricity, Electrification and the Diseases of Civilization. Universe Star, New York (2012). ISBN 13:9781938908187, ISBN 10:193890818X
10. Aliyu, O., Maina, I., Ali, H.: Analysis of electromagnetic field pollution due to high voltage transmission lines. J. Energy Technol. Policy **2**(7), 1–10 (2012). ISSN 2224-3232 (paper), ISSN 2225-0573 (online)
11. Otto, M., Muhlendahl, K.E.: Electromagnetic fields (EMF): do they play a role in children's environmental health (CEH)? Int. J. Hyg. Environ. Health **210**, 635–644 (2007). www.elseiver.de/ijheh
12. Hardell, L., Sage, C.: Biological effects from electromagnetic field exposure and public exposure standards. Biomed. Pharmacother. Rev. (2), 104–109 (2008). www.elseiver.com/locate/biopha

A Communication Viewpoints of Distributed Energy Resource

Ravish Kumar[1], Seshadhri Srinivasan[2(✉)], G. Indumathi[3], and Simona Dzitac[4]

[1] ABB Corporate Research, Bangalore, India
ravish.kumar@in.abb.com
[2] International Research Center, Kalasalingam University, Virudhunagar, India
cpscourse@klu.ac.in
[3] Cambridge Institute of Technology, Bangalore, India
indumathi.ece@citech.edu.in
[4] Universtatea Din Oradea, Oradea, Romania
simmona@dzitac.ro

Abstract. Smart Grid provides a flexible and powerful framework to integrate Distributed Energy Resources (DER). DER consisting of energy sources (such as renewable, conventional, storage) and loads. Based on application, there can be different modes of operating. It can operate in parallel with, or independently from, the main grid. Because of social, economic, and environmental benefits, the demands of DER system is increasing gradually. Combination of DERs and Loads with control units forms a miniature grid which is known as Microgrid. Integrating of DERs to the Microgrid system has two aspects: electrical and communication. Electrical integration has their own challenges which are addressed in IEEE 1547 standard. However, challenges related to communication and information integration of multi vendor DERs system is still open. Currently, IEC 61850-7-420 is the only standard which defines information model for DERs, but it is not completely accepted by DER manufacturers. On the other side, IEC 61850 mapped communication protocols like MMS, GOSSE, DNP3 etc. are substation protocols and may not suitable to fulfill DERs to DERs communication demands. In this paper, we provide an overview of Smart Grid conceptual framework and highlight DER scope. We discuss DER system components, various applications and required communication features. We review IEC 61850 mapped communication protocol MMS and identify the limitation for using in DERs communication.

1 Introduction

The traditional electrical grid is an interconnected network for delivering electricity from suppliers to consumers. It consists of generating stations that produces electrical power, high-voltage transmission lines that carry power from distant sources to demand centers, and distribution lines that connect individual customers. The growing demand of electricity led to increase numbers of electrical

V.E. Balas et al. (eds.), *Soft Computing Applications*, Advances in Intelligent Systems and Computing 633, DOI 10.1007/978-3-319-62521-8_10

grids deployment located at different geographical location. The electrical grid is a broadcast grid networks, where a few central power generator provides all the electricity production to the industrial and residential customer via a large network of cables and transformers. Electrical grid structure can summarized as follows: one-way energy flow from central station generators, over a transmission network, through substations onto distribution systems, over radial distribution circuits to end-use customers. To address the increasing energy demands, the traditional electrical grids being enhanced by adding more bulk generators to produce more energy. Bulk generators mostly use fossils fuels for power generation. On this earth, there are limited fossils fuels. Rapid consumption of limited natural resources must be avoided for future sustainable energy generation.

To meet the future demands of electricity it is being focused to shift traditional grid generating power from fossils fuels to renewable resources [1]. The renewable resources are solar, wind, small hydro, tidal, biogas etc. The future of electric grid is Smart Grid (SG) [1]. The objective of SG is to improve energy efficiency and reduce carbon footprint at all the stages of energy flow: Bulk Generation, Transmission, Distribution and Consumption. SG vision is to provide an open framework and protocol for information and power exchange seamlessly. Unlike traditional electric grid, which is centralized grid, the SG is decentralized grid. It allows to integrate power sources at distribution level. The end consumer can be Prosumers, which acts as electricity consumer and producers at the same time. Prosumer have their own small Distributed Energy Resources (DERs) for electricity generation. Based on local usage, the additional generated power can be supply to utility grid. Presently, most of the prosumers have limitation, they can not supply surplus power to utility grid because of insufficient electrical and communication infrastructure [2]. The surplus overproduced electricity is just thrown away, if prosumer's energy storage are full. Lack of infrastructure support causes waste of energy. The benefits of the DER integration in smart grids has been discussed in [12–15].

Because of social, economical and environment benefits, DER technology is evolving, new DER systems are getting deployed across the globe. However, research are still lacking for providing standard communication interface for DERs [3] system. Real time information exchange is very crucial to ensure reliable power supply. Presently, most of the DER system communicates with other system either on proprietary protocols or basic protocols like Modbus. There is no standard communication protocol and interface are recommended for DER system by International Standard organization. DER system is considered as an alternative of main power grid. New applications are developed to connect with DER system to enable new features. Lack of standard interfaces and protocols limitation to provide interoperable among different manufacture devices and system.

International standard IEC 61850 [4] is known and worldwide accepted standard for Substation Automation. It decouples the application data model from communication protocols and provide flexibility for extending this standard for any application. Because of this flexibility, addition of new domain is always

possible with IEC 61850. It is also being consider as a prime mover of SG framework. Since, it brings flexible and open framework for monitor, control, and measurement applications. The IEC 61850 7-420 [5] standard defines information model for DER, that describe what information and semantics to exchange with other DER system. DER communication with existing IEC 61850 mapped protocols such as MMS, DNP3, GOOSE can be implemented. However, from DER functionality point of view, many of the communication needs may not be fulfilled by these protocols. For example, auto device registration and plug and play, session less communication among DERs, etc. Authors in [11] investigated the role of OPC UA and IEC 61850 integration for enabling smart grids and demonstrated how OPC UA can provide communication medium for smart grid elements.

Interoperability is the key aspects DER system. One DER system must be interoperable with other DER system. Interoperability is an ability of two systems to communicate and share information with each other. The communication aspect of devices can be divided into four aspects [5]: Information modeling (the type of data to be exchanged-nouns), Service modeling (the read, write, or other actions to take on data-verbs), Communication protocols (mapping the noun and verb models to actual bits and bytes), and telecommunication media (fiber optics, radio system, wire, and other equipment). Currently, IEC 61850-7-420 is the only standard which defines information model and data model for DER system. However, the existing IEC 61850 mapped protocol seems not to be suitable for DERs communication. Most of the DER system vendors does not support IEC 61850 communication protocols in the DER system. They use their proprietary for simplicity and ease of development. Manufacturers of DER system would like to continue DER communication with their proprietary protocols. As we mentioned earlier, IEC 61850 standard is designed for Substation Automation, where system configurations is static and preconfigured. Therefore, existing IEC 61850 protocols may not be suitable for DER system communication. Yoo et al. [6] has shown IEC 61850 communication architecture based Microgrid communication and only considered static DER system configuration. Sučić et al. [7] highlighted the IEC 61850 communication protocol limitations for DER system and suggested Service Oriented Architecture for DER system. In this paper, we discuss an overview of Smart Grid and DER. Further, we discuss different applications of DER system and provide the required communication viewpoints.

The rest of the paper is organize as follows. Section 2 provide an brief overview of Smart Grid Architectural Model Framework and discuss different domains and zones. In Sect. 3, we discuss the Distributed Energy Resources system, its application, and communication architecture. Section 4, we have evaluated IEC 61850 communication protocols for DER communication. Section 4 discuss challenges for implementing IEC 61850 communication protocols in DER system. And, in Sect. 5 conclude the whole paper.

2 Smart Grid Conceptual Framework

Smart Grid (SG) is a combination of different technologies and can be seen as complex system. The key enabler of SG framework is an open communication and integration framework, which allows bidirectional information and electricity flow. Figure 1 shows a conceptual Smart Grid Architecture Model(SGAM) proposed by National Institute of Standard and Technology [1].

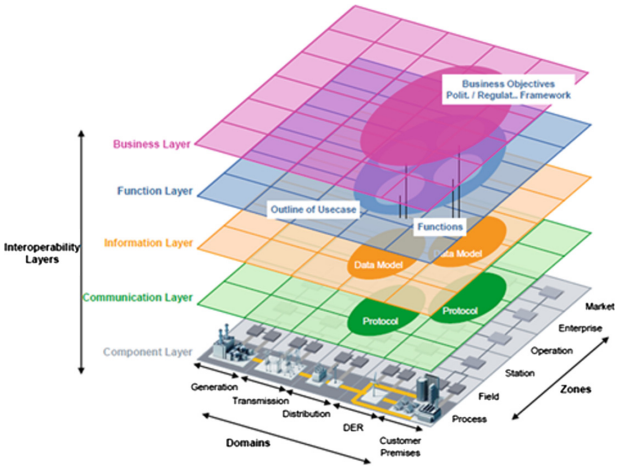


Fig. 1. SGAM conceptual framework model [1]

SGAM framework consists of Business layer, Function layer, Information layer, Communication layer, and Component layer. An interoperability aspect must be satisfied among the layer for successful Smart Grid operation. The Smart Grid system is divided into two aspects Electrical process and Information management. In the SGAM framework, domains represent electrical process and Zones represent the information management. Domains physically related to the electrical grid components: Bulk Generation, Transmission, Distribution, Distributed Energy Resources (DERs), and Customer premises. It is arranged according to the electrical energy conversion flow. Zones are related to power system management. Zone reflects the hierarchical level based on aggregation and functions separation of each hierarchical level. DER is considered as basic building block for Smart Grid system. Since, it provides ancillary distributed energy resources at distribution level, close of end consumers. The importance of DER are being recognized due to its social, economic, and environmental benefits. Growing number of DER demand is estimated to reach USD 34.94 Billion by 2022, at a CAGR of 10.9% between 2016 and 2022 [8].

3 An Overview of Distributed Energy Resources System

DER plan is consisting of different energy sources (e.g. renewable, conventional, storage) and loads. It is capable of operating in parallel with, or independently from, the main grid. DER system is more than back-up power source. Figure 2 shows an example of DER system.

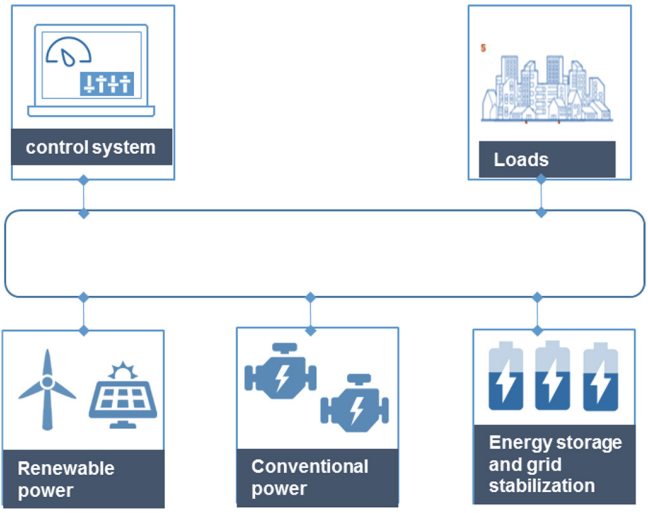


Fig. 2. Distributed Energy Resources system example

DER plant operates in two modes: Grid Connected (grid mode) and Grid Isolated (islanding mode). In the grid mode, DER plan is connected to utility grid Common Coupling Point (CCP). It is being controlled based on utility grid power supply and cost factor behavior. For example, if DER is generating energy using renewable sources, then the load will use energy from and optimize the utility grid power in order to save overall cost. While in the case of islanding mode, DER is not connected to utility grid. It acts to remote power sources. Load is completely relied on DER. Based on the condition the transition takes place between, grid to islanding mode, and islanding to grid mode. The fundamental operational requirements of DER plant to provide stability of frequency and voltage for ensuring optimal power flow to the load. Grid to islanding mode requires transition and stabilization for minimal load shedding and distribution. Islanding to grid mode requires resynchronization and minimum impact for sensitive loads during transient periods. Each of these operations are highly dependent on underlying communication infrastructure. The growing demand of DER system increased the complexity of multivendor DER elements interconnection, since most of the DER plant vendors use their own proprietary communication protocols.

Typically, DER plant may have three level of controller. Primary, secondary and tertiary controller [9]. Primary controller is machine's (such as diesel generator) local controller. The secondary controller monitors and controls the primary controller based on the real time to operate and control generating machine behavior. Tertiary controller is used to controller power flow among DER systems located at different location. In the below subsection DER applications example are discussed to explain DER operating modes.

3.1 DER Applications

DER provides an alternative local power source in parallel or independently from main grid. DER system can be simple or complex as per the application needs and use case. The following are the application scenarios of DER application.

3.1.1 Back up for Poor Power Quality of Main Grid

Poor power quality refers inability of main grid to provide stable power supply with any fluctuations. Power quality of grid strongly depends on the load characteristic and transmission and distribution grid infrastructure. Long transmission line with asymmetric loads can easily influence power quality, which can results into unbalanced voltage, harmonic, in power supply. Poor power quality may lead to power outage, if demands lead to power generation capacity. In such cases, DER system at the consumer site plays an important role. Depending on the field of application, military, industrial, commercial, or residential, power quality requirement of different DER system can be deployed to provide back up when there is poor power quality supply from main grid.

3.1.2 Natural Disasters

Natural disasters such as earthquake, tornados, hurricanes, and tsunamis may damage power grid transmission and distribution network and the area which are not directly affected by natural disasters may suffers from power outage for a week to months if damage is severe. Since DER system is not dependent on power supply of the power grid, the immediate construction of DER system can provide an alternate power supply.

3.1.3 Power Grid Failure

Large power grid, chances of disturbances due to cyber-attack, terrorist attack, human error, natural incidents, etc. are very high. The cascading effect of one unit failure can end up power outage in larger region. A grid isolated DER systems can be an alternative of Power grid failure in order to maintain power supply for crucial domains.

3.1.4 Addressing of Growing Power Demand

The growing power demand results in unstable power supply if generation capacity lags the power demand. Upgrading of power grid generation is not an easy

process it requires huge investment and effort. Alternately, DER system is good candidate for satisfying growing demand of energy by deploying DERs near the high power demand zone.

3.1.5 Continuous Power Supply

DER is the primary source of electricity and provides power supply to the load round the clock.

3.1.6 Uninterrupted Power Supply

For this application, the crucial task of DER is very short start-up and ramp time to meet load demands. DER system should be stable and efficient. Typically, the operating hours of such DER system is less than two hours.

3.2 Der Communication Architecture

From the communication point of view, DER system can be divided into three layers. Layer 1 - Machine/Primary controller to Secondary controller communication, Layer 2, Secondary controller to Central supervisory control, and Layer 3, central supervisory controller to other DER systems. These three layers are depicted in Fig. 3.

Every generation unit such as PV array, wind turbine, diesel generator has its own primary controller unit, which locally controls the generator unit. Primary controller can start, stop and regulate the power generation. Primary controller interacts with secondary controller mostly using serial to analog interface. The job of secondary control is very crucial; it monitors the status of generator units and controls the physical aspects of generator based on vital statistics such voltage, frequency, circuitry health, etc. and broadcast the status information horizontally to other Secondary controller and vertically to Central Supervisory

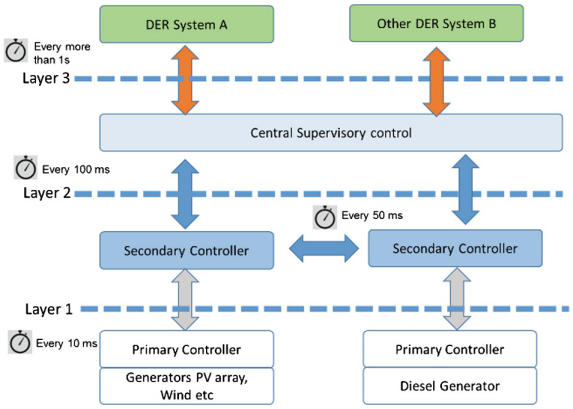


Fig. 3. Distributed Energy Resources communication architecture

Control system. Communication between primary and secondary controller has highest frequency of message. High frequency of message of message is necessary for capturing and address transient behavior of the generators such as wind turbine, diesel generator etc. The message update cycle between primary controller and secondary controller is 10 ms, Secondary controller to secondary controller is 50 ms, Secondary controller to central supervisory controller is 100 ms, and, DER system to another DER system is more than 1 s. DER system will have advance features for ensuring autonomous operations. Some of these advance features are given below.

3.2.1 Automatic Discovery of DER System

This feature allows DER system to discover new DER system in the network and synchronize their activities automatically. For example, if new wind turbine is connected to network, the existing DER system should be able to discover without having any prior inform of DER.

3.2.2 Dynamic System Configuration

This feature allows DER system element to reconfigure themselves automatically based on the overall plant status in order to provide reliable, efficient and good quality of power supply.

3.2.3 Plug and Play

This feature allow new DER element to be part of network automatically. DER system will be configured automatically by Central Supervisory Control system.

3.2.4 Semantic Data Interpretation

It is expected that dataset of DER system should be semantically interpreted by other DER system. This can be achieved if all the DER system support IEC 61850-7-420 information model.

4 Analysis of IEC 61850 Communication Protocols for DER System

The IEC 61850 standard is specially designed for Substation Automation domain. The communication model is designed in way the application specific data exchange is independent from communication model. Therefore, IEC 61850 can be extended to other domains as well. IEC61850 standard allows the integration of all protection, control, measurement and monitoring functions by one common protocol. It provides the means of high-speed substation applications, station wide interlocking and other functions which needs intercommunication between IEDs. The advantage of using IEC 61850 standard is that it ensures the interoperability and information interchangeability among multivendor devices. Device from any vendor compliant to IEC 61850 standard can communicates

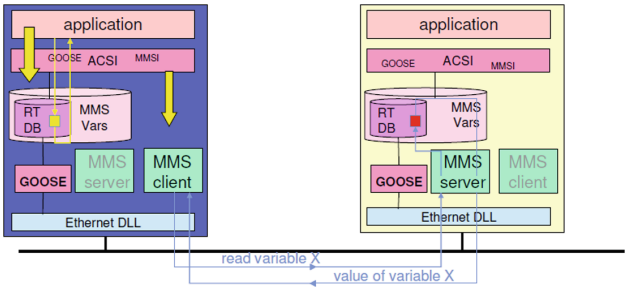


Fig. 4. IEC 61850 based message exchange

Table 1. Analysis of IEC 16850 mapped MMS protocol for DER

DER required features	IEC 61850 support
Data modeling	Yes
Interoperability	Yes
Scalability	Yes
10–100 ms message transfer	Yes
Secure communication	No
Easy configuration	No
Dynamic configuration	No
DER auto discovery	No
Plug and play integration	No
Stable for low resource device	No
Point to point communication	Yes
Point to multi-point communication	Yes
Plug and play support	No
Less engineering effort	No
Time synchronization	Yes
TCP/IP	Yes
Wireless communication	No
Web service support	No
Protocol stack extensibility	No
File transfer	Yes
Publisher/subscriber communication	No

with each other seamlessly. IEC 61850 standard defines an Abstract Communication Service Interface (ACSI) for information exchange and control among devices. The ASCI defined services are used for polling of data model information, reporting and logging of events, control of switches and functions. Peer to peer communication for fast data exchange between the feeder level devices

(protection devices and bay controller) is supported with GOOSE protocol. Voltage and current readings which are time synchronized are supported with Sampled Value communication model. Figure 4 shows the IEC 61850 based communication between two applications.

Disturbance recordings are sent using File Transfer model. A common formal description code, which allows a standardized representation of a systems data model and its links to communication services is presented by the standard and its called as SCL (Substation Configuration Language) [10]. It covers all communication aspects according to IEC 61850. Based on XML schema, this code is an ideal electronic interchange format for configuration data, an equivalent of EDDL in Process Automation applications. The data model and the communication services (defined through ACSI) are decoupled from specific communication technology. This technology independence guarantees long term stability for the standard and opens up possibility to switch over to future communication technologies.

Table 1 shows our analysis of IEC 61850 mapped MMS communication protocol for DER communication. We found MMS protocol may not be suitable for DER communication.

5 Conclusion

Distributed Energy Resources are an important Smart Grid building block. Combining multiple DERs unit forms a Microgrid which can be considered as miniature Smart Grid. Due to social, economic, and environment benefits, the demand of DERs deployment is increasing gradually. Currently DER system having interoperability issue. DER system of one manufacturer may not able to operate together because of their proprietary protocol use. In this paper, we have provided details of DER systems and it communication architecture. And provided communication view of point of DER system, which should be supported by standard protocols interface in coming future.

References

1. NIST Special Publication 1108R2. NIST Framework and Roadmap for Smart Grid Interoperability Standards, Release 2.0
2. Final report, Standards for Smart Grids. CEN/CENELEC/ETSI Joint Working Group
3. Cleveland, F.M.: IEC 61850-7-420 communications standard for distributed energy resource (DER). In: Proceedings of IEEE PES General Meeting, pp. 1–4 (2008)
4. International Standard IEC 61850: Communication networks and systems for power utility automation
5. IEC 61850-7-420 Ed. 1.0. Communication networks and system in power utility automation - Part 7-420: Basic communication structure - distributed energy resources logical nodes. March 2009. http://www.iec.ch
6. Yoo, B.-K., Yang, S.-H., Yang, H.-S., Kim, W.-Y., Jeong, Y.-S., Han, B.-M., Jang, K.-S.: Communication architecture of the IEC 61850-based micro grid system. J. Electr. Eng. Technol. **5**, 605–612 (2011)

7. Sučić, S., Havelka, J.G., Dragičević, T.: A device-level service-oriented middleware platform for self-manageable DC microgrid applications utilizing semantic-enabled distributed energy resources. Int. J. Electr. Power Energy Syst. **54**, 576–588 (2014)
8. http://www.marketsandmarkets.com/Market-Reports/micro-grid-electronics-market-917.html. Accessed Aug 2016
9. Kounev, V., Tipper, D., Grainger, B.M., Reed, G.: Analysis of an offshore medium voltage DC microgrid environment– part II: communication network architecture. In: IEEE PES T&D Conference and Exposition, pp. 1–5, Chicago, IL, USA (2014)
10. IEC 61850-6: Communication networks and systems for power utility automation – Part 6: Configuration description language for communication in electrical substations related to IEDs
11. Srinivasan, S., Kumar, R., Vain, J.: Integration of IEC 61850 and OPC UA for smart grid automation. In: 2013 IEEE Innovative Smart Grid Technologies-Asia (ISGT Asia). IEEE (2013)
12. Srinivasan, S., Kotta, U., Ramaswamy, S.: A layered architecture for control functionality implementation in smart grids. In: 2013 10th IEEE International Conference on Networking, Sensing and Control (ICNSC). IEEE (2013)
13. Verrilli, F., Gambino, G., Srinivasan, S., Palmieri, G., Del Vecchio, C., Glielmo, L.: Demand side management for heating controls in microgrids. IFAC-PapersOnLine **49**(1), 611–616 (2016)
14. Gambino, G., Verrilli, F., Canelli, M., Russo, A., Himanka, M., Sasso, M., Srinivasan, S., Del vecchio, C., Glielmo, L.: Optimal operation of a district heating power plant with thermal energy storage. In: 2016 American Control Conference (ACC), pp. 2334–2339, July 2016. IEEE (2016)
15. Maffei, A., Srinivasan, S., Iannelli, L., Glielmo, L.: A receding horizon approach for the power flow management with renewable energy and energ storage systems. In: 2015 AEIT International Annual Conference (AEIT), pp. 1–6, October 2015. IEEE (2015)

Cyber-Physical Energy Systems Approach for Engineering Power System State Estimation in Smart Grids

Seshadhri Srinivasan[1(✉)], Øystein Hov Holhjem[2], Giancarlo Marafioti[2], Geir Mathisen[2], Alessio Maffei[3], Giovanni Palmieri[3], Luigi Iannelli[3], and Luigi Glielmo[3]

[1] Berkeley Education Alliance for Research in Singapore, Singapore 138602, Singapore
seshadhri.srinivasan@bears-berkeley.sg
[2] SINTEF ICT-Applied Cybernetics, Trondheim, Norway
{Oystein.Hov.Holhjem, fgiancarlo.marfioti, geir.mathiseng}@sintef.no
[3] Department of Engineering, University of Sannio, Benevento, Italy
{amaffei, palmieri, luiannel, glielmo}@unisannio.it

Abstract. This investigation proposes a CPES architecture and model for engineering energy management application for smart grids. In particular, the investigation considers the implementation of the power systems state estimator (PSSE). The CPES architecture has three layers: physical, monitoring and applications. The physical layer consists of the grid and the various components. Since, the grid is usually engineered with various devices from multiple vendors that have different protocols and standards; data aggregation becomes a problem. The second layer of the CPES architecture overcomes this problem by proposing a middleware that aggregates data from the physical layer. The topmost layer is the applications layer, where the energy management system applications are implemented. These applications require the model, topology and information from the grid. This requires combining the physical aspects of the grid with the cyber ones. This investigation uses the common information model to model the grid and information exchanges. Then the model is combined with measurement and optimization models of the application to realize the PSSE. The proposed approach is illustrated on a Norwegian distribution grid in Steinkjer. Our results show that the CPES approach provides an easier way to engineer future smart grid applications.

Keywords: Cyber-physical energy systems (CPES) · Power systems state estimator (PSSE) · Common information models (CIM) · Service oriented architecture (SOA) · Middleware

1 Introduction

Traditionally, core responsibility of power engineers has been to operate power grids and manage the transfer of energy. With the advent of smart grids there is an unprecedented volume of data available that needs to be harmonized with the energy

V.E. Balas et al. (eds.), *Soft Computing Applications*, Advances in Intelligent Systems and Computing 633, DOI 10.1007/978-3-319-62521-8_11

flow. Consequently, communication and energy infrastructure have to be designed together for achieving cost benefits from smart grids. The communication infrastructure includes middleware, communication networks, data-models, and software platforms. Furthermore, since smart grids are massively distributed systems scalability plays a vital role within various applications such as state estimation, stability analysis and contingency analysis. These applications rely on measurements from heterogeneous sensors that are manufactured by various vendors. Engineering such applications requires a framework for handling both the cyber and physical parts of the grid simultaneously.

Cyber Physical Energy Systems (CPES) Approach is touted as a futuristic approach for engineering smart grids. They consider the tight coupling between the physical and cyber domains in their design. CPES is a recent research topic that has attracted significant attention [1]. Palensky et al. [2] proposed to study the modern energy systems using the CPES approach with the following categories: physical models, information technology, roles and individual behavior, and aggregated as well as stochastic models. Furthermore, proposed the use of CPES approach as a tool and method for the emerging complexities of the energy grids. While, specialized and useful tools for individual domains of the energy systems exist, there is no methodology to combine the different aspects of the energy and information. Motivated by this CPES approach has been studied for modelling and enabling various aspects of the energy grids. To our best knowledge, Ilic et al. [3] first proposed a dynamic model using the CPES approach. The proposed model greatly dependent on the cyber technologies supporting the physical system. Major contributors to the CPES modelling and design are Widl et al. [4–7]. In [4], the author studied the role of the continuous time and discrete-event CPES models for control applications such as demand response, load balancing, and energy storage management. The investigation [7] presented a co-simulation platform for components, controls, and power systems based on software applications such as (GridLAB-D, OpenModelica, PSAT, 4DIAC) for smart charging of electric vehicles. The use of CPES approach for energy optimization in buildings has been studied in [8–10]. The use of model based design methodology for validating control algorithms in a residential microgrid has been studied in [11]. The role of CPES system for electric charging applications has been studied in [12]. Similarly, reducing energy consumption by combining CPES and industry 4.0 was proposed in [13]. More recently, McCalley et al. [14, 15] surveyed the design techniques and applications of CPES.

In spite of these developments, the role of CPES approach for handling grid level applications has been minimum. Furthermore, CPES approach encapsulating both the physical and cyber domain has not been investigated. Among the various problems that can be tackled using CPES approach, we focus on the power system state estimation problem that has been studied in energy grids. The PSSE determines the most likely state of the power system from a set of measurements that are captured using supervisory control and data-acquisition systems. The role of PSSE is crucial for many energy management applications such as optimal power flow, monitoring, and contingency analysis. In addition, PSSE also provide snapshots of the network to energy management system applications developed by many third party applications. The role of PSSE in enabling optimal power flow [16, 33], bad-data detections, market

operations etc. [17, 32] has been studied in literature. Currently, there are various tools, standards and models used for capturing the various aspects of PSSE. The physical models used for PSSE have been reviewed in [18]. Among the communication standards used for modelling PSSE, IEC 61850 [19] and Common Information Models (CIM) standardized by the International Electromechanical Commission have been studied for implementing the grid. The investigation in [20] viewed PSSE as one of the applications enabled by IEC 61850 without much description about the implementation. The role of IEC 61850 in implementing PSSE for a distribution grid has been proposed in [21]. Similarly, CIM models for state estimation have been proposed in [22–24]. The recent research are converging more towards the use of CIM as a standard for data communication in smart grids (see, [25–30] and references therein).

The main contribution of the paper is a CPES architecture and model that can be used to engineer smart grids. The proposed architecture has three layers: physical, middleware and applications. The physical layer is the grid with the various sensors and devices. Typically the grid is composed of various components and devices that have various communication protocols, standards and manufactured by different vendors. Aggregating information from them is difficult. To overcome this, a middleware is used. The top layer of the CPES architecture is the application layer that needs to obtain the data from the middleware map it into the grid model. To model the grid, we use the common information model as it is easily extendable to model various aspects of the grid. The CIM model is tailored to the needs of the PSSE and is used for information exchange to the applications layer as well as among the applications. The application layer in addition contains the optimization and measurement models for mapping the physical variables of the grid. Furthermore, the implementation aspects of the PSSE in the application layer is also presented. The CIM based model maps the cyber aspects with the network topology and to the physical entities directly. Furthermore, the PSSE is also simultaneously realized in the application layer. This leads to a new approach for modelling and realizing future energy management applications. It is worth to mention here that the middleware part of the architecture is beyond the scope of this investigation and has not been treated in detail.

The paper is organized as follows. Section 2 presents the implementation aspects of the PSSE. Section 3 presents the CPES architecture and model. Section 4 presents the case study and implementation in the distribution network in Steinkjer. Conclusions and future course of investigation are discussed in Sect. 5.

2 Power System State Estimation

The main objective of a power system state estimator is to estimate the state of each bus, namely voltage magnitude and phase angle, using the set of redundant measurements, usually affected by errors. The set of phasors representing the complex bus voltages are called static state of the system. The State estimator, initially developed as an engineering tool, is slowly transforming into a decision support system that assists control and monitoring tasks. As with any estimator, the SE outcome is an approximation of the current network state. Moreover, from a high level supervisory control and data acquisition module, sensors measurements arrive with different time-granularities so

that measurements and state estimate can differ in phase. The physical component of the PSSE can be broken down into two parts: hardware and software. The hardware part consists of the SCADA system that collects the real-time information from the remote terminal units installed in various substations. RTU measurements include active and reactive power flows, power injections, current magnitude, voltage magnitude and phase angles. Figure 1 shows the hardware components of PSSE. Till recently, a direct measurement of voltage phase angles was considered impossible. Then, the introduction of phase measurement units (PMUs) has been made that measurement possible. In recent times, the PMUs have significantly improved the measurements accuracy by using the GPS to synchronize the time signals with an accuracy close to 1 μs.

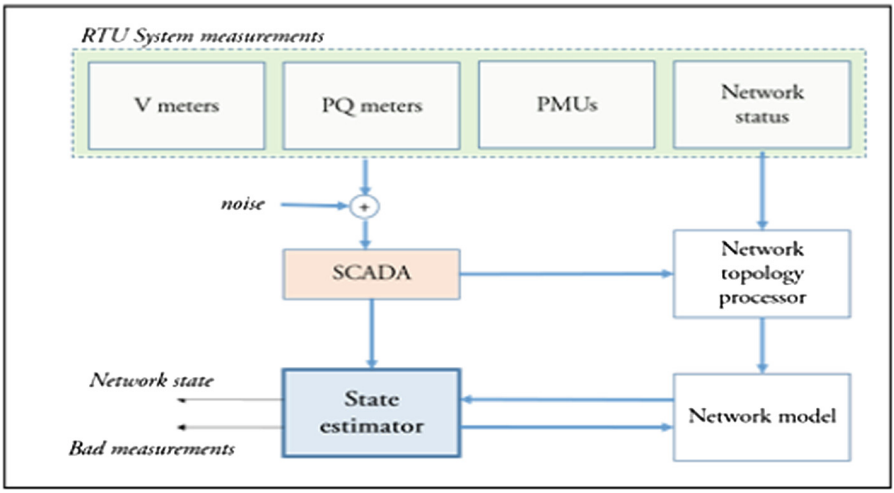


Fig. 1. Hardware components of the PSSE

The software components of the PSSE are: measurement pre-filtering, topology processor, observability analysis, state estimator, bad-data detection and bad-data suppression. The pre-filtering block checks for error measurements and unusual data. The topology processor builds the network topology required for the PSSE application using the knowledge of the physical topology. The SE algorithm then computes the state estimates. Finally, the bad data processor checks for the bad data and eliminates the error values. The state estimator block uses iterative algorithms that have finite convergence time for building the PSSE. In summary, the SE operation serves as a large-scale filter between remote measurements and high-level applications that constitute the Energy Management System (EMS).

3 CPES Model

Energy management in smart grids require orchestrating several hardware platforms, protocols and devices from multiple vendors. The EMS applications should get access to different data from these various devices. A good solution to solve the heterogeneity

issue is to design a middleware. The architecture of such a middleware is shown in Fig. 2. The bottom most layer is the physical layer that models the energy grid. The middleware aggregates the data from the grid and provides it to the PSSE application. The application layer houses the third party and other applications for EMS such as the PSSE. Energy grids typically consider the energy flows in the physical grids, however to orchestrate the smart grid, the information from heterogeneous sensors needs to be aggregated by the middleware and provided to the applications. The physical layer models the energy flows and transfers. This model needs to be augmented with the data model for implementing the EMS. The network topology and other physical aspects needs to be captured in the application layer. The flow of information among the different layers is defined by the CIM model. Here, we do not consider the implementation aspects of the middleware and the services are not elaborated. To propose the CPES model, we propose a top-down approach wherein the application layer has the PSSE application. The PSSE application requires measurements from the grid that is aggregated using the middleware. To process the information, the topology and model of the grid needs to be known. For this purpose the CIM models are used.

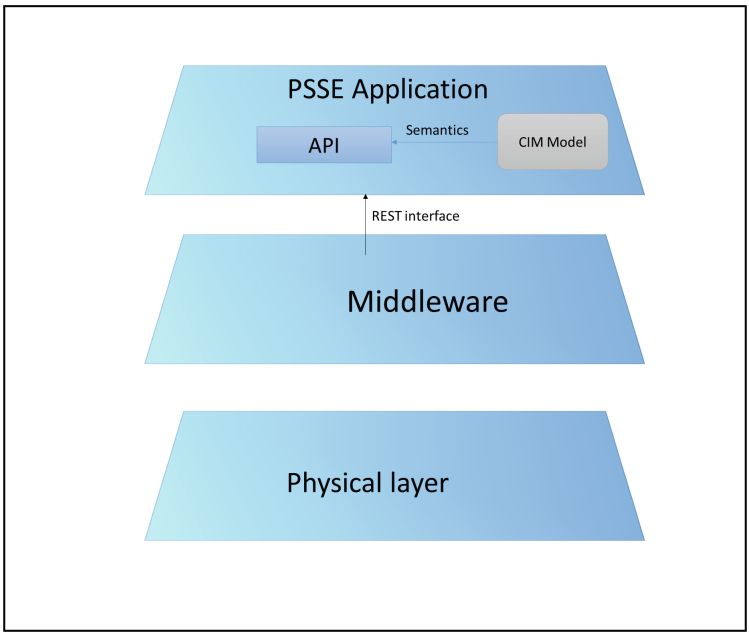


Fig. 2. CPES layered architecture

The PSSE EMS application is developed in JAVA. In order to facilitate information interoperability an application program interface (API) is defined. This API can enable different applications and services to access data and exchange information independently, alleviating difficulties with intrinsic data representation. The CIM specifies the semantics of the API. This also eliminates the difficulty of interfacing lower level

components with the physical application. The CIM model can be used to build semantics into the middleware as well. The lower level components transmit the information using dedicated communication infrastructure. There are different information models and data formats for the data. As the CIM provides interface to the applications, they are oblivious to the physical entities connected to the grid. This leads to a service oriented architecture for the middleware. In what follows we describe the measurement, optimization and grid model captured using CIM. The implementation aspects of the application layer and different aspects considered in the design of PSSE.

A. Measurement Model

The State Estimator routine is used to monitor the power network status during normal operation, where the system is in quasi-steady state responding to slowly varying load demand and power generation. It is also assumed that, in the three-phase transmission system, all loads are balanced and all transmission lines are fully transposed so that the network can be represented by its single phase equivalent diagram. Given the single phase diagram, the network is mathematically modelled and all the measurements, previously described Sect. 2, are written as function of the network state variables (i.e. voltage phasors). The equations modelling the power network are nonlinear and do not take into account all possible errors due to the uncertainties in the network parameters, e.g., metering errors and noise that may be introduced through the telecommunication systems.

Consider a vector z as the set of all available measurements. This vector can be expressed in terms of power network state variables x as follows:

$$z = \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_m \end{bmatrix} = \begin{bmatrix} h_1(x_1, x_2, \ldots, x_n) \\ h_2(x_1, x_2, \ldots, x_n) \\ \vdots \\ h_n(x_1, x_2, \ldots, x_n) \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_m \end{bmatrix} = h(x) + e \tag{1}$$

where h(x) is the measurement model, x is the state vector, and e is the error in measurements. The measurement model h(x) is a nonlinear model that maps the state vector to the measurement.

B. Optimization Model

The mathematical formulation of the presented state estimation problem is based on the maximum likelihood concept. Maximum likelihood estimator (MLE) of a random variable maximizes the likelihood function, which is defined based on assumptions of the problem formulation. The first assumption, as previously mentioned, is that the errors are distributed according to a Gaussian (or normal) distribution, with the expected value equal to zero. Thus, a random variable z is said to have a normal distribution if its probability density function $f(z)$ is given as follows:

$$f(z) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}(z-\mu)/\sigma^2} \tag{2}$$

where μ is the expected value (or mean) of $z = E(z)$ and σ is the standard deviation of z. The previous property holds for the i.i.d. assumption.

The second assumption implies that the joint probability density function (pdf) of a set of m measurements can be obtained by taking the product of individual pdfs corresponding to each measurement. The result product function is called *likelihood function* for the set of m measurements:

$$f_m(z) = f(z_1)f(z_2)\ldots f(z_m) \tag{3}$$

Essentially the likelihood function is a measure of the probability of observing a given set of measurements in vector z. For this reason we are interested in finding the parameter vector that maximize this function. In order to simplify the procedure of determining the optimum parameters, the function is commonly replaced by its logarithm, obtaining the so called *log-likelihood function*. Hence, the MLE of the state x can be found by maximizing the log-likelihood function for a given set of observations, $z_1, z_2, \ldots, z_m$. Thus, the following optimization problem is formulated:

$$\textit{minimize } J(x) = \sum_{i=1}^{m} \left(\frac{z_i - \mu_i}{\sigma_i} \right)^2 \tag{4}$$

Let us define, $r_i = z_i - \mu_i$ is the residual of measurement i. The expected value $E(z_i)$ of measurement z_i, can be expressed as $\mu_i = E(z_i) = h_i(x)$, where $h_i(x)$ is a nonlinear function relating the system state vector x to the i th measurement. The minimization problem in (4) can be modelled as an optimization problem for the state vector x:

$$\begin{aligned} \min_{x} \quad & \sum_{i=1}^{m} W_{ii} r_i^2 \\ \text{s.t.} \quad & \mathrm{z_i} = h_i(x) + r_i \quad \forall i = 1, \ldots, m \end{aligned} \tag{5}$$

where $W_{ii} = \sigma_i^{-2}$ is the inverse of the assumed error variance for the measurements. The reciprocal of the measurement variances can be thought of weights assigned to individual measurements. High values are assigned for accurate measurements with small variance and low weight for measurements with large uncertainties. A little manipulation leads to

$$\begin{aligned} \min_{x} \quad & J(x) = r^T \mathbf{W} r \\ \text{s.t.} \quad & \mathrm{z} = h(x) + r \end{aligned} \tag{6}$$

where $x, r, z, h(x)$ are vectors of respective quantities and W is the diagonal weighted matrix of the WLS problem equals to the inverse of covariance matrix of the measurements.

C. Smart Grid Model

As stated earlier CIM has been chosen as a representative mechanism for the smart grid model. The CIM models provide an abstract model for the power network using unified modelling language (UML). The CIM represents the power system entities using an object oriented approach as classes, attributes, methods and association as defined in IEC 61970 [25] and IEC 61968 [26] standards. The standard IEC 61970-301 provides as semantic model for the power system components at an electrical level and their interrelationships. The IEC 61968-11 extends the semantic models to include the data exchanges for scheduling, asset management and other market operations. Although, CIM contains most classes and their associations to represent the power system, still object models need to be adapted for implementing specific application. CIM models can be adapted by defining new classes, subclasses, methods and attributes.

CIM Profile: A profile is a delimitation of CIM which consists of a subset of classes and attributes that specify information conceptually and the relationships among the different objects. A subset of IEC 61970-456 defines the CIM necessary to describe the PSSE results. A modular approach is used in the development of CIM profile that four profiles: Equipment, Measurement, Topology, and State variable Profile. The equipment profile models physical elements of the network such as network, loads, generators, and switches. The measurement profile contains the measurement information such as active and reactive power, voltages, load angles etc. Topology profile defines the classes needed to describe the network topology considering the switching status. State variable profile contains the model for defining the state variables in the network. The relationship between the different profiles is shown in Fig. 3. The profile connected at the "from" end of the arrow depends on the profile at "to" end of the arrow.

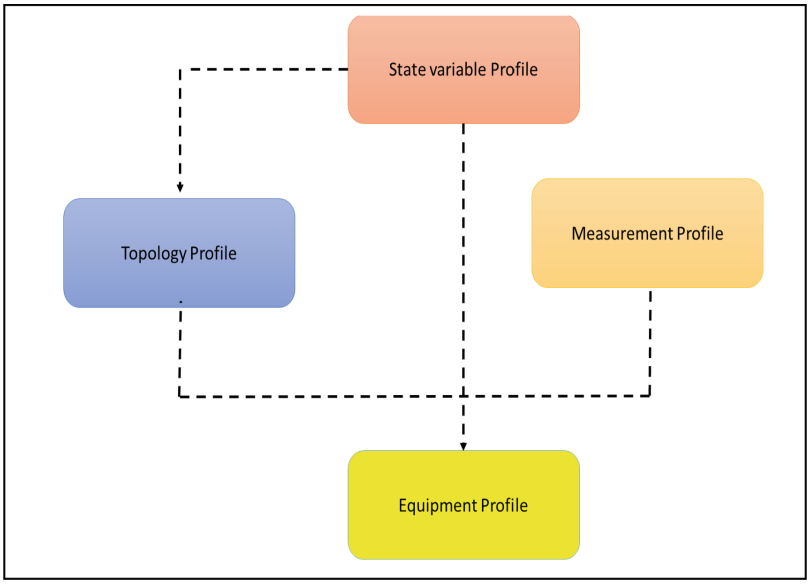


Fig. 3. CIM profiles for implementing the PSSE

As a next step the abstract and concrete classes of the various profiles are defined. As the treatment of all the profiles is beyond the scope of this investigation. Here we provide an example of the topology profile (see, Fig. 4). The topology profile defines the classes needed to describe how each of the equipment in the network is connected to each other. Topology is given by the association of the buses with the corresponding association of the terminals of the equipment. This way the network model is built as a branch flow model and can be directly used by the PSSE application. The state variable

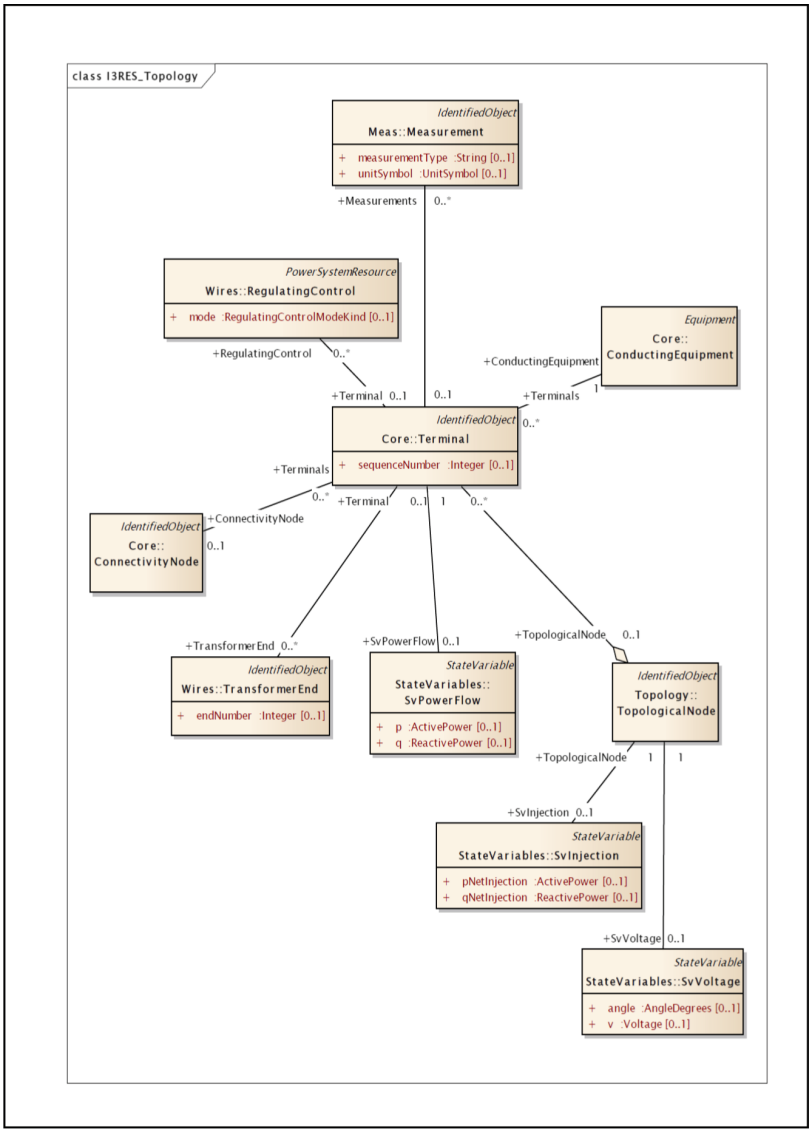


Fig. 4. Topology profile

profile of the CIM model is shown in Fig. 5. Similarly, the other CIM profiles are modelled to describe the network. In addition the abstract and concrete models required of the CIM profiles are described. The CIM model maps the network topology with the measurements and state variables. The CIM model defines the semantics for the PSSE API which is implemented in JAVA. The application uses the middleware services to query the sensors and update the network state using the PSSE application. The middleware services required can be broken down into low-level services and common services. The description of the middleware services are not considered in this investigation. REST interfaces are used to communicate between the middleware and application layer.

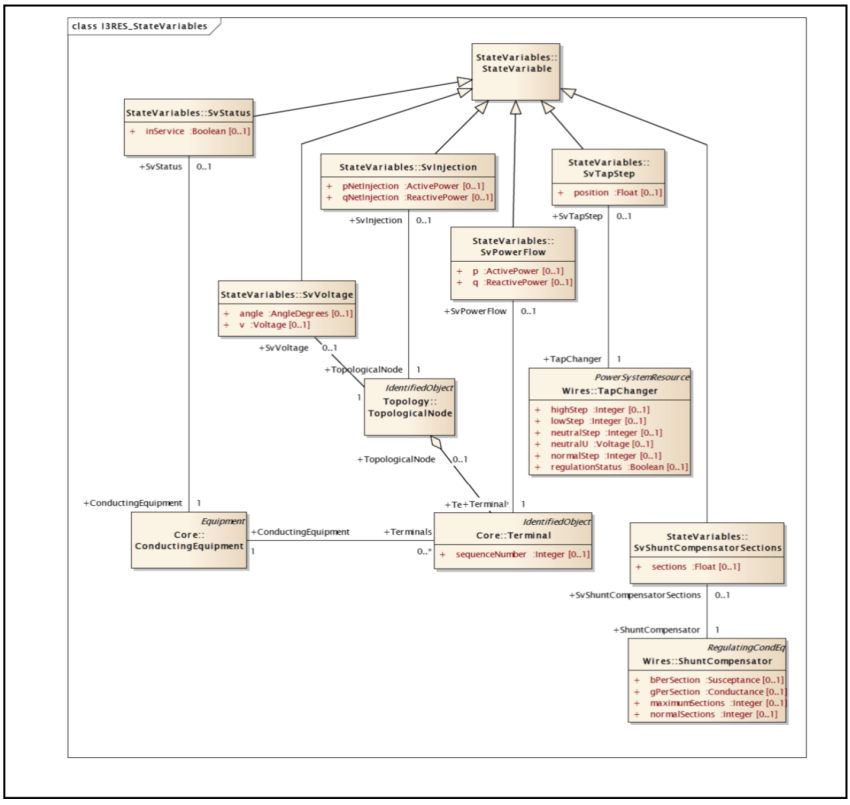


Fig. 5. State variable profile

D. Software Model of the PSSE Application

The PSSE application needs to map the data aggregated from the middleware to the network topology. To this extent the application uses the InterPSS that defines the network data in the IEEE format for processing. The application implementing the PSSE is written in JAVA and to solve the optimization model in (6) GAMS solver is

used due to its symbolic processing capability. The APIs implementing the CIM, and libraries of the InterPSS, GAMS are integrated with the application written in JAVA to realize the PSSE for the EMS.

The CPES model described above integrates the physical equations describing the energy flows in the network with the CIM to model the cyber part of the network. The resulting model encapsulates both the cyber and physical aspects of the network. Furthermore, the application software model proposed also uses the InterPSS to map the topology and uses the GAMS for realizing the PSSE. This model along with the CPES description of the network can be used to engineer energy management system application in typical smart grid. The main advantage of the proposed approach is that the method provides a comprehensive framework for co-designing both the physical and cyber aspects of the network, thereby reduces the engineering efforts. Furthermore, provides an interface between the different domain engineers involved in smart grids.

The CPES model described above integrates the physical equations describing the energy flows in the network with the CIM to model the cyber part of the network. The resulting model encapsulates both the cyber and physical aspects of the network. Furthermore, the application software model proposed also uses the InterPSS to map the topology and uses the GAMS for realizing the PSSE. This model along with the CPES description of the network can be used to engineer energy management system application in typical smart grid. The main advantage of the proposed approach is that the method provides a comprehensive framework for co-designing both the physical and cyber aspects of the network, thereby reduces the engineering efforts. Furthermore, provides an interface between the different domain engineers involved in smart grids.

4 Case Study

The CPES approach proposed in Sect. 3 was used to engineer PSSE in the energy management system of the distribution grid in Steinkjer, Norway. It is a radial distribution network that consists of a: hydro power plant with 2 generators, 32 aggregating loads, 50 link busses, and 84 transmission lines. The CIM defines the basic ontology of the set of attribute value pairs. Ontologies are often written in XML or in a Resource Document Framework (RDF), which is a suitable format for middleware information exchange through the common communication bus. The CIM standard IEC 61970-52 defines the procedure for description of the network model as a serialized RDF schema. The main concept of the RDF is called the tiple and consists of subject-predicate-object expression. An RDF document contains element that are identified by unique ID attribute and that can be referenced from other elements using that ID in a resource attribute. The CIM profile for implementing the PSSE can be represented using a RDF schema document. In details, after having selected all the classes, attributes and relationship among the classes, tools such as CIMTool can be used for generating the RDF schema. As stated earlier the PSSE application uses the InterPSS and GAMS to provide the state estimates for the different applications. The CIM model also defines the information exchange among the EMS applications. The CPES engineered PSSE application is tested in the pilot distribution network. The residuals of the PSSE are used to evaluate the accuracy of the state estimates. Figure 7 shows the residuals

computed for the real and reactive power in p.u. values. Our results show that the PSSE provided an accurate estimates with error between 2–4% for the 85 buses in the distribution network (Fig. 6).

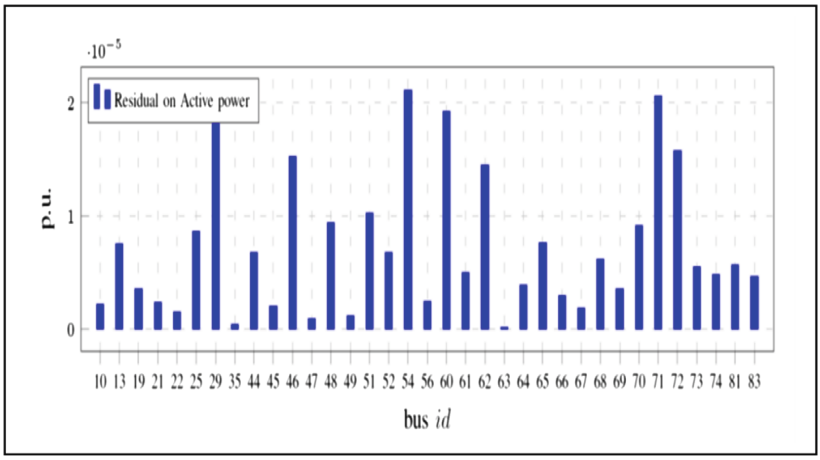


Fig. 6. Residuals of the real power in PSSE

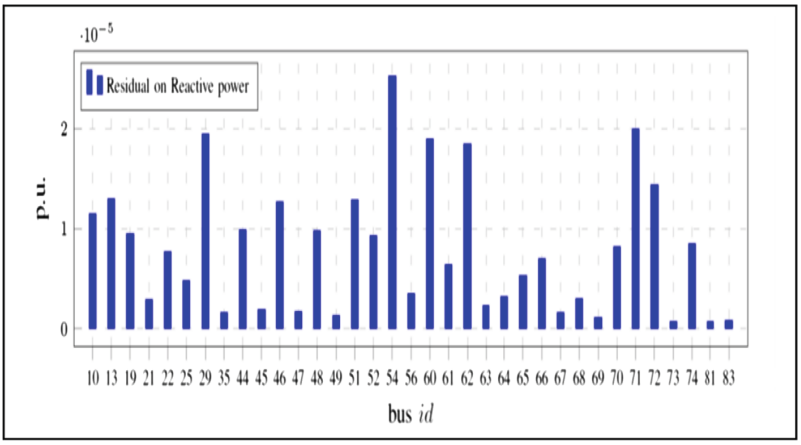


Fig. 7. Residuals of the reactive power in PSSE

5 Conclusions

This investigation proposed a cyber-physical energy systems approach for engineering PSSE in smart grids. The proposed approach modelled both the physical and cyber aspects of the PSSE simultaneously. The physical model was represented by the

measurement and optimization model. While, the common information model was used to describe the cyber part of the CPES model. The proposed approach can be used to engineer PSSE applications in smart grids with heterogeneous sensors from various vendors and considering interoperability constraints. Furthermore, the proposed CPES approach can also be used for building a service oriented architecture (SOA) based middleware. The details of the implementation of the middleware are not discussed in the investigation. Extending the CPES approach to model optimal power flow with PSSE is the future course of this investigation.

Acknowledgement. This research is supported by the European Union's Seventh Framework program (FP7 2012-2015) for the ICT based Intelligent management of Integrated RES for the smart optimal operation under the grant agreement n. 318184, project name: I3RES.

References

1. Khaitan, S.K., McCalley, J.D.: Cyber physical system approach for design of power grids: a survey. In: 2013 IEEE Power and Energy Society General Meeting (PES). IEEE (2013)
2. Palensky, P., Widl, E., Elsheikh, A.: Simulating cyber-physical energy systems: challenges, tools and methods. IEEE Trans. Syst. Man Cybern.: Syst. **44**(3), 318–326 (2014)
3. Ilic, M.D., Xie, L., Khan, U.A., Moura, J.M.F.: Modeling future cyber-physical energy systems. In: IEEE Power and Energy Society General Meeting-Conversion and Delivery of Electrical Energy in the 21st Century, pp. 1–9. IEEE (2008)
4. Palensky, P., Widl, E., Elsheikh, A.: Simulating cyber-physical energy systems: challenges, tools and methods. IEEE Trans. Syst. Man Cybern.: Syst. **99**, 1–10 (2012)
5. Widl, E., Palensky, P., Elsheikh, A.: Evaluation of two approaches for simulating cyber-physical energy systems. In: Proceedings of 38th Annual Conference on Industrial Electronics, IECON 2012, pp. 3582–3587 (2012)
6. Stifter, M., Widl, E., Filip, A., Elsheikh, A., Strasser, T., Palensky, P.: Modelica-enabled rapid prototyping of cyber-physical energy systems. In: IEEE Workshop on Modeling and Simulation of Cyber Physical Energy Systems, pp. 1–6 (2013)
7. Elsheikh, A., Awais, M.U., Widl, E., Palensky, P.: Co-simulation of components, controls, and power systems based on open source software. In: Proceedings of 2013 IEEE Power and Energy Society General Meeting, pp. 1–5 (2013)
8. Kleissl, J., Agarwal, Y.: Cyber-physical energy systems: focus on smart buildings. In: Proceedings of the 47th ACM Design Automation Conference, pp. 749–754 (2010)
9. Zhao, P., Godoy, S.M., Siddharth, S.: A conceptual scheme for cyberphysical systems based energy management in building structures. In: 9th IEEE/IAS International Conference on Industry Applications (INDUSCON), pp. 1–6 (2010)
10. Shein, W.W., Tan, Y., Lin, A.O.: PID controller for temperature control with multiple actuators in cyber-physical home system. In: 15th IEEE International Conference on Network-Based Information Systems (NBiS), pp. 423–428 (2012)
11. Ge, Y., Dong, Y., Zhao, H.: A cyber-physical energy system architecture for electric vehicles charging application. In: 2012 12th International Conference on Quality Software (QSIC). IEEE (2012)
12. Al Faruque, M.A., Ahourai, F.: A model-based design of cyber-physical energy systems. In: Design Automation Conference (ASP-DAC), 2014 19th Asia and South Pacific. IEEE (2014)

13. Martin, B., et al.: A new approach to increasing energy efficiency by utilizing cyber-physical energy systems. In: 2013 Proceedings of the 11th Workshop on Intelligent Solutions in Embedded Systems (WISES). IEEE (2013)
14. Khaitan, S.K., McCalley, J.D., Liu, C.C. (eds.): Cyber Physical Systems Approach to Smart Electric Power Grid. Springer, Heidelberg (2015)
15. Khaitan, S.K., McCalley, J.D.: Design techniques and applications of cyberphysical systems: a survey. IEEE Syst. J. **9**(2), 350–365 (2014)
16. Maffei, A., Srinivasan, S., Iannelli, L., Glielmo, L.: A receding horizon approach for the power flow management with renewable energy and energy storage systems. In: 2015 AEIT International Annual Conference (AEIT), pp. 1–6. IEEE, October 2015
17. Huang, Y.-F., et al.: State estimation in electric power grids: meeting new challenges presented by the requirements of the future grid. Signal Process. Mag. **29**(5), 33–43 (2012). IEEE
18. Kashyap, N.: Novel resource-efficient algorithms for state estimation in the future grid. Ph.D. dissertation, Department of Electrical Engineering, Aalto Unieristiy, Finalnd (2012)
19. Srinivasan, S., Kumar, R., Vain, J.: Integration of IEC 61850 and OPC UA for Smart Grid automation. In: 2013 IEEE Innovative Smart Grid Technologies-Asia (ISGT Asia), pp. 1–5. IEEE, November 2013
20. Mohagheghi, S., et al.: Applications of IEC 61850 in distribution automation. In: 2011 IEEE/PES Power Systems Conference and Exposition (PSCE). IEEE (2011)
21. Choi, S., Kim, B., Cokkinides, G.J.: Feasibility study: autonomous state estimation in distribution systems. IEEE Trans. Power Syst. **26**(4), 2109–2117 (2011)
22. Xu, K.-N., Xin-Gong, C., Xi-Ji, X., Yan-Shi, C.: Model design of electric system state estimation based on CIM. In: Power and Energy Engineering Conference, 2009. APPEEC 2009. Asia-Pacific. pp. 1–4. IEEE (2009)
23. LiJun, Q., Meng, L., Ying, W., CuiJuan, H., HuaWei, J.: Cimbased three-phase state estimation of distribution network. In: 2011 International Conference on Advanced Power System Automation and Protection (APAP), vol. 1, pp. 667–672. IEEE (2011)
24. Sharma, A., Srivastava, S., Chakrabarti, S.: An extension of common information model for power system multiarea state estimation. Syst. J. **PP**(99), 1–10 (2014). IEEE
25. Chen, Y.: Industrial information integration—a literature review 2006–2015. J. Ind. Inf. Integr. **2**, 30–64 (2016)
26. Cintuglu, M.H., Martin, H., Mohammed, O.A.: An intelligent multi agent framework for active distribution networks based on IEC 61850 and FIPA standards. In: 2015 18th International Conference on Intelligent System Application to Power Systems (ISAP), pp. 1–6. IEEE, September 2015
27. Belikov, J., Kotta, Ü., Srinivasan, S., Kaldmäe, A., Halturina, K.: On exact feedback linearization of HVAC systems. In: 2013 International Conference on Process Control (PC), pp. 353–358. IEEE, June 2013
28. Soudari, M., Srinivasan, S., Balasubramanian, S., Vain, J., Kotta, U.: Learning based personalized energy management systems for residential buildings. Energy Build. **127**, 953–968 (2016)
29. Srinivasan, S., Kotta, U., Ramaswamy, S.: A layered architecture for control functionality implementation in smart grids. In: 2013 10th IEEE International Conference on Networking, Sensing and Control (ICNSC), pp. 100–105. IEEE, April 2013
30. Landa-Torres, I., et al.: The application of the data mining in the integration of RES in the smart grid: consumption and generation forecast in the I3RES project. In: 2015 IEEE 5th International Conference on Power Engineering, Energy and Electrical Drives (POWERENG). IEEE (2015)

31. Cavalieri, S., Regalbuto, A.: Integration of IEC 61850 SCL and OPC UA to improve interoperability in Smart Grid environment. Comput. Stand. Interfaces **47**, 77–99 (2016)
32. Verrilli, F., et al.: Model predictive control-based optimal operations of district heating system with thermal energy storage and flexible loads. IEEE Trans. Autom. Sci. Eng. **14**(2), 547–557 (2017)
33. Maffei, A., Srinivasan, S., Castillejo, P., Martinez, J.F., Iannelli, L., Bjerkan, E., Glielmo, L.: A semantic middleware supported receding horizon optimal power flow in energy grids. IEEE Trans. Ind. Inform. **PP**(99), 1. doi:10.1109/TII.2017.2655047. http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7822985&isnumber=4389054

Knowledge-Based Technologies for Web Applications, Cloud Computing, Security Algorithms and Computer Networks

About the Applications of the Similarity of Websites Regarding HTML-Based Webpages

Doru Anastasiu Popescu[1(✉)], Ovidiu Domșa[2], and Nicolae Bold[3]

[1] Faculty of Mathematics and Computer Science, University of Pitesti, Pitesti, Romania
dopopan@yahoo.com
[2] "1 Decembrie 1918" University of Alba Iulia, Alba Iulia, Romania
domsaddd@yahoo.com
[3] Faculty of Management, Economic Engineering in Agriculture and Rural Development, University of Agronomic Sciences and Veterinary Medicine Bucharest, Slatina Branch, Bucharest, Romania
bold_nicolae@yahoo.com

Abstract. The study of the similarity between web applications has extended alongside with the informational explosion resulted from the fast communication means through Internet. The copyright of web applications is difficult to be appreciated in this domain and this is the reason for the development of novel web technologies and mechanisms of measuring the similarity between two webpages. In this paper, we will present a modality of measurement of the similarity degree between two webpages regarding the HTML tag-based webpages. The degree of similarity will be determined approximately, being dependent of the webpages used from the both websites and the tags set used in the comparison of the webpages. The selection of webpages in order to determine the degree of similarity between two webpages will be made using genetic algorithms. In the final part of the paper there are presented the results obtained with the implementation of the algorithm presented in the paper.

Keywords: Webpage · Tag · Algorithm · Chromosome · Gene · Similarity

1 Introduction

The similarity is a notion which appears frequently in many domains: science, linguistics, copyright, arts, transportation etc. In the last period, different mechanisms of measuring the similarity between different objects were developed, using particularly complex algorithms and mathematical results.

A variant of measuring the similarity between two objects presumes the identification of certain character sequences which characterize adequately these objects and then these sequences are analyzed by means of specific algorithms to obtain a number which reflects the similarity. This type of algorithms is used in [4, 5]. We will use this idea in this paper for measuring the similarity of two webpages. The number that will

V.E. Balas et al. (eds.), *Soft Computing Applications*, Advances in Intelligent Systems and Computing 633, DOI 10.1007/978-3-319-62521-8_12

be obtained to define to what extent two webpages are similar will be determined by examining the tag sequences which are used at the web pages created using HTML.

The large number of files used for creating the webpages is an impediment for creating algorithms which calculate the similarity between two webpages and this is the reason for which we will use a fixed number of files for this calculus. The generation of these files will be made using genetic algorithms, which will provide several variants of these files and for each variant we will use the editing distance method for the calculus of similarity, presented in detail in [7].

The main motivation for our work is the finding of the degree of similarity between two web applications. Another motivation is related to finding different methods of calculating this degree, in this paper being presented a method using genetic algorithms.

In Sect. 2 we will present some definitions and examples necessary for the approximate calculus of the degree of similarity between two webpages.

Section 3 will contain the presentation of the selecting modality required for the measurement of the similarity between two web pages, using genetic algorithms.

In Sect. 4, we will present the modality of calculating the similarity of webpages using the results from Sect. 3.

Section 5 is dedicated to the presentation of several results obtained after the algorithms presented in Sects. 3 and 4 were implemented.

2 Preliminary Notions

In order to determine the similarity between two webpages which will be denoted by WA (web applications): WA1 and WA2, we will use only the files formed of HTML tags. These files have the .html and .htm extensions. To simplify the future notations, we will denote by $P = \{p_1, p_2, \ldots, p_m\}$, respectively with $Q = \{q_1, q_2, \ldots, q_n\}$ the sets of web pages associated with the applications WA1, respectively WA2.

Next, we will interpret by p_i or q_j the file which contains the tags of the corresponding web page ($1 \leq i \leq m$, $1 \leq j \leq n$).

In the approximate calculus of the similarity between two web pages we will not take into consideration a fixed set of tags, this set being denoted by TG.

For the next definitions, it is useful to denote by $T1_i$ the sequence of tags from p_i which are not found in TG and by $T2_j$ the sequence of tags from q_j which are not found in TG, where $1 \leq i \leq m$, $1 \leq j \leq n$.

In the calculus of the similarity between two sequences of tags we will use the editing distance (or the Levenshtein distance), presented in detail in [3, 7, 11]. In the classic definition, the character strings are used, in this paper the characters being replaced by tags.

In the paper [7] we have done a similar study. We will use, in this matter, some notations used in the referred paper:

- DL(s, t) is the minimum number of operations to transform the string s into the string t;
- $d(p_i, q_j)$ is the degree of similarity between two webpages p_i and q_j;

- $d(p_i, WA_2)$ is the degree of similarity between the webpage p_i and the web application WA_2;
- $d(WA_1, WA_2)$ is the degree of similarity between two web applications WA_1 and WA_2;
- $ad(WA_1, WA_2)$ is the approximate degree of similarity between two web applications WA_1 and WA_2.

3 The Choice of Files for the Calculus of the Similarity Between Two Webpages

At the notations used in the previous sections we add two natural numbers k and h with $1 \leq k \leq m$, $1 \leq h \leq n$. k and h are chosen so that the runtime of the algorithm of the calculus of the similarity between two web pages to be enclosed in limits closed to the ne calculated for all the web pages.

In Fig. 1 the general scheme of the algorithm as in [3, 8] is presented. The algorithm will be used twice for generating subsets with k web pages from WA1, respectively with h web pages from WA2.

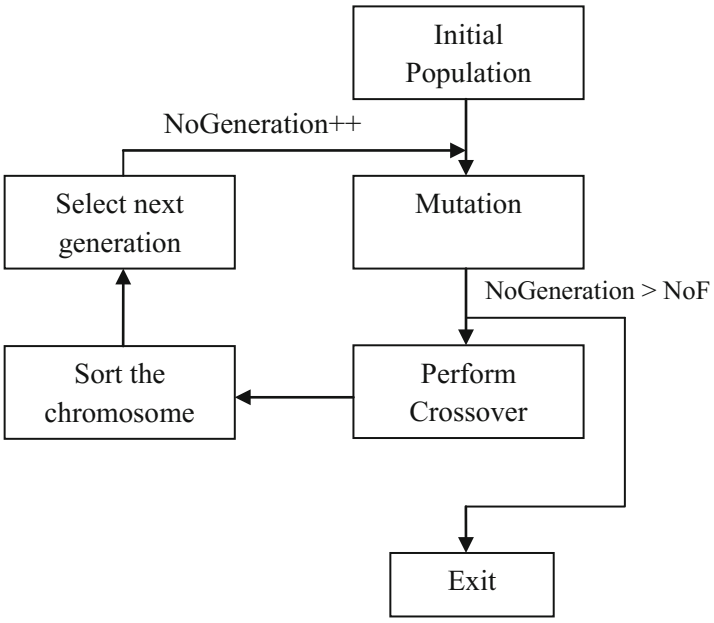


Fig. 1. Scheme of a genetic algorithm

The algorithm presented next has as input data a string of characters which represents the path where a web application can be found and a natural number k, less or equal with the number of web pages built using HTML tags, denoted by n. The web

pages will be codified with numbers from 1, 2, … to n. The output data of the algorithm will be represented by a bi-dimensional array with k columns, each line representing a chromosome, namely a subset with k distinct elements from {1, 2, …, n}. The subsets obtained from the array will be distinct.

The steps of this algorithm are:

Step 1. The character sequence s and the number k is read. This sequence represents the path of the web application.
Step 2. The files with the extension .html or .htm from the tree with the subfolders that are rooted in s are determined. The names of these files will be memorized alongside their path in the string p[1], p[2], …, p[n].
Step 3. The number of generations NoF which will be used in the genetic algorithm is read.
Step 4. Using the genetic algorithm presented above, it is built a bi-dimensional array with the elements Pop[i][j], i = 1, Nocrom and j = 1, k, where Nocrom represents the number of chromosomes which will be obtained. Nocrom can be read.
Step 5. The chromosome obtained at step 4 are sorted ascendingly depending on the fitness functions, but these chromosomes are not distinct. This is why a new bi-dimensional array will be built. This array contains the elements PopDist[i][j], i = 1, NoCromDist and j = 1, k.

Observations

1. In the genetic algorithm from the step 4, the initial population is randomly generated using the numbers from the set {1, 2, …, n}, these numbers being the genes that form the chromosomes, n being the number of web pages from WA. The chromosome is formed from k genes. The initial population used in the implementation presented in Sect. 6 has 1200 chromosomes.
2. For the i^{th} chromosome from the population which contains the genes Pop[i][1], … Pop[i][j], …, Pop[i][k], the fitness is calculated as being the sum of the degrees of similarity of between the pairs of web pages p[Pop[i][j]] and p[Pop[i][r]], $1 \leq j$ $r \leq k$ and it is being denoted by f(i). Thus:

$$f(i) = \sum_{j=1;r=j+1}^{k-1;k} d(Pop[i][j], Pop[i][r])$$

3. The i^{th} chromosome is more performant than the j^{th} chromosome if f(i) < f(j), which means that the web pages from the i^{th} chromosome are more different to each other than the ones from the j^{th} chromosome.
4. The mutation operation is made by identifying a number h (randomly generated) from {1, 2, …, n} which is not found in a chromosome and randomly determining a position poz in the chromosome. After that, the gene from the position poz is replaced with h.
5. For the crossover operation, chromosomes which contain at least one different number are randomly generated. The crossover of the two chromosomes with the indices i and j which have this property will be made in this way:

- it is built a string of numbers with all the genes from the two arrays, denoted by $x = (x_1, x_2, \ldots, x_{2k})$.
- the array x is sorted ascending and a string of numbers $y = (y_1, y_2, \ldots, y_w)$ containing the distinct numbers form x. $w > k$, because of the fact that the two chromosomes has at least one gene (number) different.
- two new chromosomes are built from the array y, the first chromosome containing the first k components from y and the second with the latter k components from y.

4 The Approximate Calculus of the Degree of Similarity Between Two Webpages

In this section, we will keep the notations from the previous sections. The algorithm presented in the next lines will determine the approximate degree of similarity between two webpages WA1 and WA2, using k web pages form WA1 and h web pages from WA2.

The steps of this algorithm are:

Step 1. Two strings of characters s and t and the natural numbers k and h are read. These two strings represent the paths where the two webpages WA1 and WA2 can be found.
Step 2. The files with the extension .html or.htm from the tree with the folders which are rooted in s are determined. The names of these files will be memorized alongside with their paths in the string p[1], p[2], …, p[m].
Step 3. The files with the extension .html or .htm from the tree with the folders which are rooted in t are determined. The names of these files will be memorized alongside with their paths in the string q[1], q[2], …, q[m].
Step 4. Using the algorithm from the Sect. 4 twice, once for WA1 and once for WA2, we determine two bi-dimensional arrays with chromosomes (web pages indexes):

$$(\text{PopDist1}[i][j])_{i=1,\text{NoCromDist1};j=1,k} \text{for WA1}$$

respectively

$$(\text{PopDist2}[i][j])_{i=1,\text{NoCromDist2};j=1,h}, \text{ for WA2.}$$

The applications that contains the web pages with the indexes from the chromosomes from the lines of the previous bi-dimensional arrays will be denoted by $(WA1_i)_{i = 1, \text{NoCromDist1}}$ respectively $(WA2_j)_{j = 1, \text{NoCromDist2}}$.
Step 5. We determine $ad(WA1_i, WA2_j)$ for $i = 1$, NoCromDist1and $j = 1$, NoCromDist2. The average of these numbers will be the number which represents the approximation of the similarity between the webpages WA1 and WA2.

Observations

1. The presented algorithm complexity is O(Nocrom•(max(m, n)^2 + Nocrom •log (Nocrom))).
2. The algorithm presented in this section and the one from the Sect. 4 can be modified in order for other definitions of the similarity between web pages, as the well presented in [6] or [8] or other modalities of measuring the similarity between the character sequences as the ones from [1, 10, 11] to be used.

5 Implementation

The implementation made in Java led to results closed to the ones obtained for the similarity between two webpages using the definition 5 from the Sect. 2, meaning the fact that all the web pages from the both webpages were included.

Table 1 presents some information about the webpages used at the test of the Java program which implements the algorithms from Sects. 5 and 6. The set of tags used at the implementation of the program was made from the tags:

```
<html>
</html>
<head>
</head>
<a>
</a>
<div>
```

Table 2 presents some results obtained for the webpages presented in Table 1: the degree of similarity (using the algorithm from [7]) and the approximation of the degree of similarity (using the algorithm presented in this paper) (Fig. 2).

Table 1. Webpages using for Java program

Notation	Number of files	Number of HTML files
E1	125	23
E2	85	15
E3	90	16
E4	104	24
E5	97	17
E6	2431	476
E7	344	35

The results obtained using the implementation of the algorithms presented in this paper shows the fact that they can be used to calculate the degree of similarity between two webpages with quite an acceptable error (generally, under 0.01). An important aspect is the fact that the presented algorithms can also be used for other modalities of defining the degree of similarity between two files.

Table 2. Results to implement algorithms in the Java language

k	WA1	h	WA2	Similarity degree	Approximate similarity degree
12	E1	10	E2	0.6527269615774745	0.6151740849736574
12	E1	10	E3	0.6528103923698071	0.5965780161873493
12	E1	15	E4	0.6390043190028488	0.6247214290536568
12	E1	10	E5	0.636510512472533	0.6347343901389451
15	E1	100	E6	0.34485409963776364	0.318678530256979
15	E1	20	E7	0.319443831408161	0.291547650419385
100	E6	20	E7	0.568263602360978	0.536853457204536

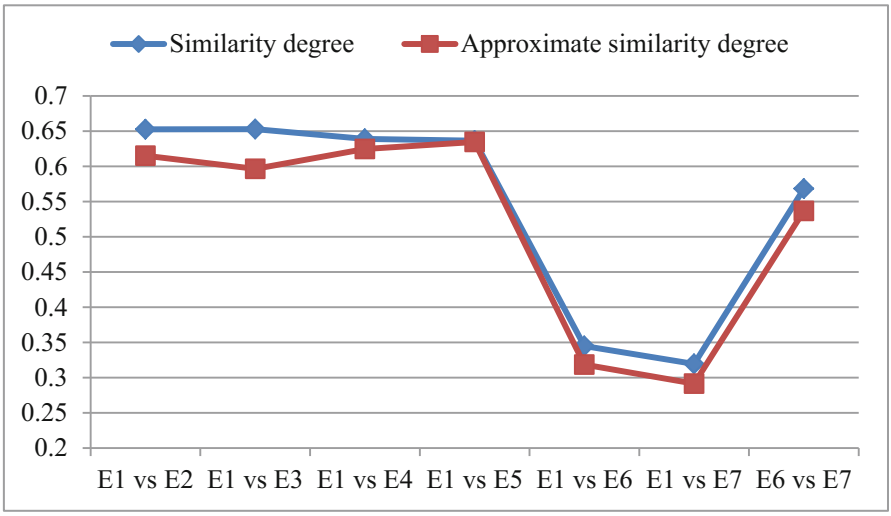


Fig. 2. Results to implement algorithms in the Java language (similarity degree versus approximate similarity degree)

6 Conclusions

This paper showed a model for the approximate calculus of the degree of similarity between two webpages, each of them being made from several other components. Particularizing, the applications are webpages and the components are web pages with the source code built from tags. A central thing of this model is the genetic algorithm, which helps us at the generation of component subsets with certain restrictions. This type of algorithms is used in many situations, in [2, 3, 9] being presented some applications of these algorithms.

The presented model can also be adapted for other modalities of defining the similarity between the files that compose a web application. For a future work, we propose to create a tool which will calculate the approximate similarity between two webpages using several formulas to calculate the similarities between the files.

References

1. Okundaye, B., Ewert, S., Sanders, I.: Determining image similarity from pattern matching of abstract syntax trees of tree picture grammars. PRASA Johannesburg, pp. 83–90 (2013)
2. Caldas, L.G., Norford, L.K.: A design optimization tool based on a genetic algorithm. In: Automation in Construction, ACADIA 1999, vol. 11, no. 2, pp. 173–184 (2002)
3. Darby, S., Mortimer-Jones, T.V., Johnston, R.L., Roberts, C.: Theoretical study of Cu–Au nanoalloy clusters using a genetic algorithm. J. Chem. Phys. **116**(4), 1536 (2002)
4. Kim, H.-S., Cho, S.-B.: Application of interactive genetic algorithm to fashion design. Eng. Appl. Artif. Intell. **13**(6), 635–644 (2000)
5. Remani, N.V.J.M., Rachakonda, S.R., Kurra, R.S.R.: Similarity of inference face matching on angle oriented face recognition. J. Inf. Eng. Appl. **1**(1) (2011)
6. Popescu, D.A., Maria, D.C.: Similarity measurement of web sites using sink web pages. In: 34th International Conference on Telecommunications and Signal Processing, TSP 2011, 18–20 August, Budapest, Hungary, pp. 24–26. IEEE Xplore (2011)
7. Popescu, D.A., Nicolae, D.: Determining the similarity of two web applications using the edit distance. In: 6th International Workshop on Soft Computing Applications, 24–26 July 2014, Timisoara, Romania, 6th IEEE SOFA. LNCS, pp. 681–690 (2014)
8. Popescu, D.A., Dan, R.: Approximately similarity measurement of web sites. In: ICONIP, Neural Information Processing. Proceedings LNCS. Springer, 9–12 November 2015
9. Rahmani, S., Mousavi, S.M., Kamali, M.J.: Modeling of road-traffic noise with the use of genetic algorithm. Appl. Soft Comput. **11**(1), 1008–1013 (2011)
10. Dietterich, T.G.: An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization. Mach. Learn. **40**(2), 139–157 (2000)
11. Cormen, T.H., Leiserson, C.E., Rivest, R.R.: Introduction to Algorithms. MIT Press, Cambridge (1990)

Energy Efficient Cache Node Placement Using Genetic Algorithm with Probabilistic Delta Consistency Using Flexible Combination of Push and Pull Algorithm for Wireless Sensor Networks

Juhi R. Srivastava(✉) and T.S.B. Sudarshan

Department of Computer Science and Engineering,
Amrita School of Engineering, Bengaluru Campus,
Amrita Vishwa Vidyapeetham, Amrita University, Kasavanahalli,
Carmelaram P.O., Bangalore 560 035, Karnataka, India
juhi80in@yahoo.co.in, sudarshan.tsb@gmail.com

Abstract. Minimum energy consumption and minimum service delay with maximizing quality of service is key to a worthy WSN. Network lifetime and network connectivity lead to optimization. However, to handle the uncertainties present in WSNs, we need powerful heuristics to solve the optimization problem. A possible way to minimize power consumption is by the use of caching popular data by optimally selecting and placing cache nodes in the network. We propose the use of a powerful searching tool and a well-known soft computing technique; a multi-objective genetic algorithm to achieve optimization goals in the presented work. The proposed work proceeds in two steps to give a complete optimized network system with increased network lifetime. In the first step we use genetic algorithm to carefully select and place cache nodes in the network with aims of maximizing field coverage and minimizing network energy usage. In the second step we perform cache consistency on the cache nodes for valid data retrieval. We use the Probabilistic Delta Consistency (PDC) with Flexible Combination of Push and Pull (FCPP) algorithm. The cache consistency algorithm is implemented on the MAC layer and comparisons are made with 802.11 MAC layer without cache consistency using metrics; routing load, packet delivery ratio, end-to-end delay, normalized load and energy consumed. The network overhead is reduced considerably leading to speedy data access. The algorithm is designed for a clustered environment and is an extension of our previous work where we have successfully proved that genetic algorithm gives better results for node deployment when compared to the state of the art node placement algorithm ScaPCICC. The results are obtained by running the experiments on Matlab and ns2.

Keywords: Genetic algorithm · Cache consistency · Wireless Sensor Network · Optimization · Clustering

V.E. Balas et al. (eds.), *Soft Computing Applications*, Advances in Intelligent Systems and Computing 633, DOI 10.1007/978-3-319-62521-8_13

1 Introduction

A Wireless Sensor Network (WSN) is comprised of a set of several identical sensor nodes with limited CPU and storage capacity and is a special class of ad hoc networks. In a typical example, nodes are capable of sensing environmental attributes, such as temperature, light and humidity, allowing researchers to monitor an area of interest remotely, inconspicuously and continuously. Some of the unique characteristics of a Wireless Sensor Networks like ability to withstand harsh environmental conditions, ability to cope with node failures, mobility, dynamic network topology, communication failures, heterogeneity of nodes, large scale of deployment, unattended operation makes their use in certain applications indispensable. Many applications for WSN have been discussed in the literature. A few include environmental/habitat monitoring, acoustic detection, seismic detection, military surveillance, inventory tracking, medical monitoring, smart spaces, process monitoring and traffic management. However, to work seamlessly and achieve application and user level satisfaction, a WSN faces many challenges. The goal of achieving application-level QoS, with minimum energy consumption and minimum service delay in WSNs becomes a very challenging task.

Minimum energy consumption and minimum service delay with maximizing quality of service is key to a worthy WSN. A sensor node failure due to energy depletion would lead to network partitions, path breaks, new route setup and therefore, latency in service and data provisioning. It is observed that most of the energy spent by a sensor node is during listening to the channel, data routing and sensing. Routing or data transmission is a foremost reason for high energy dissipation by sensor nodes. When compared; the amount of energy spent during computation with the energy spent during communication, it is observed that transmitting a single bit through a transceiver expends more energy than computing a single instruction on a processor. Routing in WSNs is data centric than address centric. All sensors detecting and sensing the event report to the sink generating redundancy of messages. However, one disadvantage associated with routing is every time an information is required, it has to be fetched from the source which might be several hops away from the sink. This also invokes participation of all the sensor nodes present in the path thus leading to energy drainage. Literature show tremendous contributions in the design of exceedingly competent routing algorithms for WSNs, however, even with a highly energy efficient data routing protocol, a WSN that is designed for continuous monitoring application can achieve more energy efficiency by other means. Battery life time of a sensor node can be extended if amount of communication and computation done by a node is reduced.

A possible way to minimize power consumption is by the use of caching the data. Caching of popular data can save a sensor network from exploitation of many of its scarce resources. To mention a few, caching helps in reducing amount of communication between nodes in the network, reduces network wide transmission, hence, reduces interference because of nodes, overcomes variable channel conditions, reduces network congestion, speeds data access by nodes and provides QoS with minimum energy consumption. However, caching can also considerably impact the system energy expenditure if not implemented efficiently; there are two vital design issues to be addressed in caching; first selection of appropriate nodes to cache data and their

optimal positioning, i.e., optimal cache node placement and second, is cache consistency. Because of the mobility associated with nodes, caching strategy also requires cache consistency schemes to ensure that the cached objects are consistent with those stored in the server. Our study focuses on the data access algorithms using cache for WSNs. In literature, many algorithms have been proposed to address these issues to the best possible way but very few work reflect the use of any soft computing techniques to achieve the same.

Soft Computing is the fusion of methodologies that were designed to model and enable solutions to real world problems, which are not modeled, or too difficult to model, mathematically. These problems are typically associated with fuzzy, complex, and dynamical systems, with uncertain parameters. The uncertainty associated with the network state information is inevitable in terms of, both, fuzziness and randomness. A powerful advantage of soft computing is the complementary nature of the techniques. Used together they can produce solutions to problems that are too complex or inherently noisy to tackle with conventional mathematical methods.

In our work, we have used genetic algorithm as an optimization technique to place the Caching Nodes (CN) optimally. We also perform a Flexible Combination of Push and Pull (FCPP) [4] algorithm on the cached nodes to provide user-specified consistency prerequisites using the Probabilistic Delta Consistency (PDC) model. We have used a clustered WSN network environment with cluster head and cluster members to implement our algorithms. We use MATLAB and ns2 simulators for performing our experiments.

2 Related Works

2.1 Genetic Algorithm

Optimization helps to enhance system performance by selecting optimal point within a given set of strictures or circumstances. The key concept in optimization is to be able to make assessment of the system resources and accordingly make decisions that can minimize the exploitation of these resources to achieve the desired objective with maximized desired output. Considering WSNs, we understand that battery life time of a sensor node can be extended if amount of communication and computation done by a node is reduced. In other words, network lifetime and network connectivity lead to optimization. However, to handle the uncertainties present in WSNs, due distance between nodes, communication channel, we need powerful heuristics to solve the optimization problem and therefore soft computing has been involved to achieve the desired optimization goals in the presented work.

The aim of the presented work is to optimize the placement of cache nodes in WSNs. We try to select finest number of nodes in a given WSN to serve as cache nodes for complete network coverage. This optimal number of nodes to be selected as cache nodes is a tricky task as selecting large number of CNs may create redundant messages, caching new information in their small memories will require more enhanced techniques of cache replacement. All this would lead to depletion of vital resources like sensor and network energy, overload network bandwidth and also create network

congestion. While a smaller number of CNs may not be able to cover the entire application and may lead to network partitions hence leaving areas un-sensed while reporting an event.

It has been observed that for many applications, the search space for optimization is excessively large or excessively expensive in terms of computation leading to a NP-hard search space. In such instances it is advised to use random search algorithms that also form the basis for studying non-linear programming techniques in research. Algorithms for optimization under this category are robust to noise, non-linearities and discontinuities. These random search algorithms can adapt to new environmental situations and yield optimized values even when the system parameters cannot be predicted at priori. Random search algorithms are also called as multi-objective programming methods as they can address more than one objective function for a given problem statement while maintaining coherency with the boundary conditions. Algorithms that mimic human brain and behavior have revolutionized the current available optimization methods. Some of these well-known algorithms are Memetic Algorithm (MA), Genetic Algorithm (GA), Shuffled Frog Leaping (SFL), Particle Swarm Optimization (PSO), Ant Colony Systems and Simulated Annealing (SA) [2]. All these techniques try to solve a problem in a comparable way; however, they still differ marginally in employing their optimization procedure. Each of these algorithms employs a depiction to suit the technique and then arrive at an optimal solution iteratively.

Genetic Algorithms are powerful techniques in solving optimization problems with huge search spaces with large number of variables where usage of procedural algorithms is complicated. They can solve optimization problems which are non-linear, stochastic, non-differentiable, discontinuous or combinatorial in nature. Genetic algorithms employ evolution process observed in biological populations for selection of best off springs that lead to optimal solutions in successive iterations. This soft computing method generates chromosomes which are the optimization parameters represented in the form of bit strings that allows encoding of both discrete and continuous parameters contained in the same problem statement. An advantage of using GA is that the variables in GA take values within their described constraints and boundaries and therefore, no solutions fall out of the upper and lower limits defined in the algorithm. If at all leak or violations occur it is due to functional restrictions of the design [2].

Authors Khanna et al. [5] have addressed creation of clustering with optimal cluster heads and cluster members in WSNs using GA. Their work is limited in identifying limits in large networks thereby applying GA and hence finding shortest convergence time becomes challenging.

In his work, author Yang [11] uses GA for placing sink nodes optimally. Their algorithm helps in fast data delivery without much use of energy during data transmission. However, a major limitation of the work is that sink nodes can be located only at certain sites known as feasible sites.

Author Youssef et al. [13], has used GA for energy efficient gateway positioning to reduce data access delay. However, their work is not scalable and therefore does not allow large number of gateways in the network.

Authors Yong et al. [12], have used GA in a cluster based WSN to find optimal transmission power of cluster heads and cluster members for data transmission with

reduced energy consumption. Their work does not consider nodes which are out of range for transmissions.

Authors Mollanejad et al. [6], focus on repositioning of base stations using GA for faster data transmission. They consider a small error bound along with network parameters like residual energy and distance for optimized location allotment. However, they consider a static network with each sensor having a GPS system thereby increasing the cost.

Wang et al. [10] have used GA for placing sensors which lead to an optimal path planning. However, they have not considered energy parameter in their algorithm.

Author Hoyingcharoen et al. [3], have used GA for selecting optimal number of sensors to guarantee minimum detection probability. However, his work does not consider energy and connectivity of nodes.

Sun et al. [9], use GA for deploying relay nodes for complete network coverage. They have modeled their system as a graph and use only a single objective function with a penalty constraint that removes any leaks in the solution from the constraint.

The proposed work is an extension of our previous work [7], where we use genetic algorithm against state of the art algorithm ScaPCICC for optimal cache node placement in WSNs. We have been successful in proving that genetic algorithm performs better in terms of reducing data access delay and number of messages in the network, while achieving energy efficient optimal cache node deployment in WSNs.

2.2 Cache Consistency

It is crucial that the data fetched from a cache is not old and represents a fresh version. We can achieve this by using cache consistency schemes that maintain the cached data consistent with those existing in the server (Fig. 1). Cache updates can be initiated by the server, CN or can be a cooperative process. In the server based approach also called push method, invalidation messages are broadcasted by the server to all the CNs

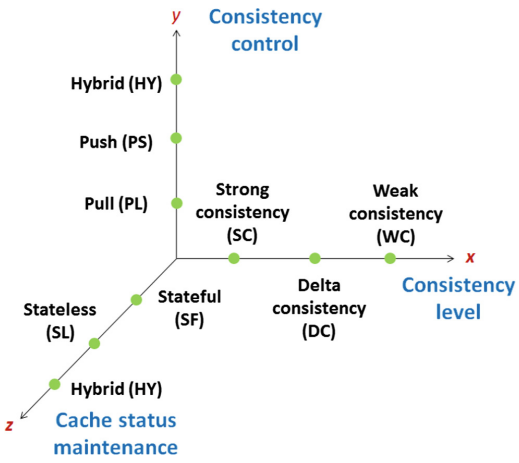


Fig. 1. Consistency models

indicating the update status of data items; whereas in the second case also known as pull mechanism, it is the CN that polls the server to determine the version of the cached item. However, for some applications a combination of push-pull method can be used. This cooperative cache consistency scheme allows for fresh retrievals of cached data at all points of time, however, cooperative schemes combine the advantages and the disadvantages of both the methods.

Depending on the application, the server may or may not keep the record of CNs and their associated data items. If records are maintained then it is called the Stateful (SF) approach otherwise it is known as the Stateless (SL) system. Though SF approach is not scalable but yet it is more efficient as server performs multicast of invalidation reports thus reducing on the number of messages in the network. SL keeps track of the

Table 1. Summary of cache consistency schemes

Cache consistency schemes	Consistency level	Cache status maintenance	Limitations
Push model	Strong consistency	STATEFUL	Overhead caused as server has to maintain state of all cache nodes
Pull model	Weak consistency	STATELESS	Since cache initiated pull, cache node miss updated data
LEASE	Strong consistency	STATEFUL	If lease period is more, critical data might not be updated by server
TTR mechanism	Strong consistency	STATELESS	TTR value has to be adaptive based on hot and cold data items in order to maintain stringent consistency requirements
TTL mechanism	Strong consistency	STATELESS	No assurance from the server that the datum will not be modified within TTL expires
Client Poll/Pull each read mechanism	Strong consistency	STATELESS	Every read at the cache node adds to round trip delay & network overload due to the generated control messages
Hybrid mechanism	Strong consistency	STATELESS	Disadvantages of both TTL and client poll mechanisms
Server invalidation mechanism	Strong consistency	STATEFUL	State has to be maintained indefinitely causing wastage of space & many invalidations have to be send when an object is modified causing network overload
IR based invalidation mechanism	Strong consistency	STATELESS	Long query latency, nodes having same copies query server separately thus increasing overheads & clients in long dose mode may miss IRs
UIR based invalidation mechanism	Strong consistency	STATELESS	Overhead in network caused due to UIR and IR messages floated

update history and occasionally floods the updates through invalidation messages generating extensive communication overhead. However, according to application constraints the consistency requirements can be stringent (strong consistency) leading to all time fresh version delivery otherwise can tolerate data which does not guarantee consistency also called weak consistency. If the tolerance of serving a data is never more than more than Δ time with the data at the server, then it satisfies delta consistency. In MANETS a more common technique used for consistency is called the probabilistic consistency model. Here a query is answered when the data item received back is consistent with the server with a probability of at least Γ, where Γ; $(0 \leq \Gamma \leq 1)$ is a design pre-defined consistency level. Table 1 gives a summary of various cache consistency schemes [14].

We employ Flexible Combination Of Push and Pull (FCPP) algorithm for cache consistency [4]. The main idea behind using FCPP is that it allows the data source node to delay the update till all the ACKs are received or until the time-out values for all the silent caching nodes expire. FCPP is a strong consistency algorithm that considers Lease protocol as a special case. FCPP benefits users by providing user defined consistency requirements, thus allowing invocation of Pull or Push mechanism whenever necessary. We now proceed to discuss the presented work in detail.

3 Proposed Algorithm

3.1 Cache Node Placement Using Genetic Algorithm - Methodology and Network Model

We propose a multi-objective GA based node placement algorithm for WSNs which optimizes the placement of cache node in WSN by optimizing the network functional parameters like field coverage and number of sensors per CN, while minimizing network constraints like network energy usage, the number of out of coverage or unconnected sensor nodes and number of overlapping CNs. The technique is designed for a clustered WSN used for monitoring applications and focuses on positioning of three types of sensor nodes in a 2-dimensional $L \times L$ square grid configuration with predefined Euclidean distance between the units. It identifies nodes as CNs, n1 (high signal range, HSR), n2 (low signal range; LSR). The algorithm also considers inactive sensor nodes which have no role and are silent though not dead in the network. GA in the proposed algorithm tries to situate these nodes optimally so that the field coverage and number of sensors per CN is maximized. The design assumes placing of each sensing node on the intersection points of the square grid for complete network coverage (Fig. 2). The nodes categorized as CN, LSR, HSR, and inactive nodes sensor nodes are encoded in a row by row style with each individual represented by a bit string of size $2L^2$ (Fig. 3) [7].

For cache node discovery we use Scaled Power Community Index Cooperative Caching (ScaPCICC) [1]. For a given graph G = (V, E), it is defined as

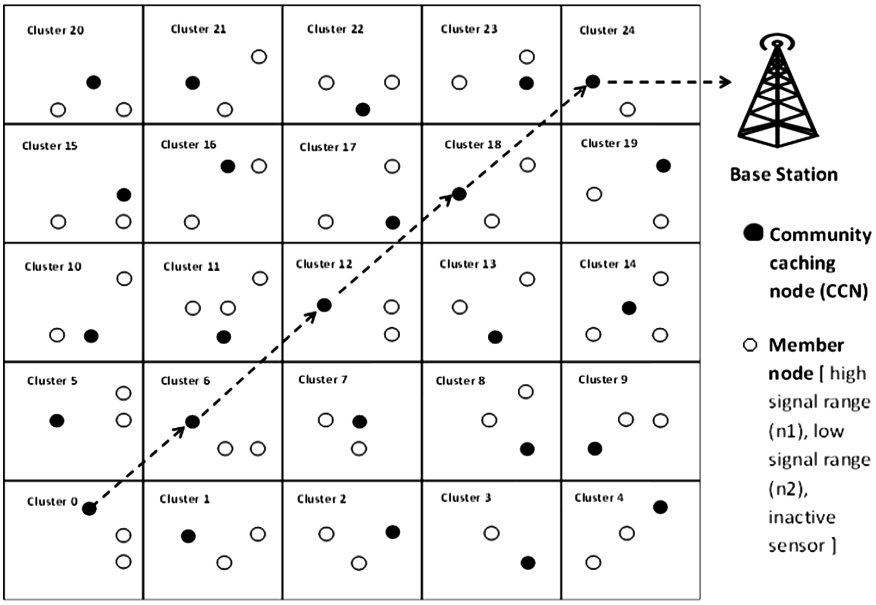


Fig. 2. Network diagram

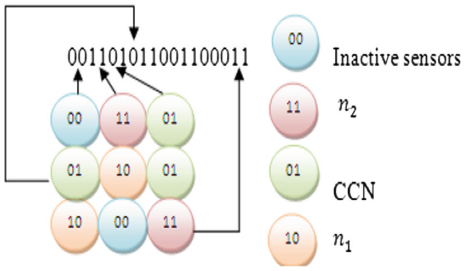


Fig. 3. Node representation in WSN

$$\text{ScaPCICC} = \text{PCI(v)}/\text{C(v)}, \tag{1}$$

where,
PCI(v) = power community index
C(v) = clustering coefficient

PCI is defined as one hop direct connection or neighbors of node v which is represented as vertex in the graph. Also known as degree of node v and C(v) is described as

$$c(v) = (2 * Lv)/(dv * (dv - 1)) \quad (2)$$

where,
dv = Degree of node v
Lv = # of links among dv neighbor of node v

In other words, PCI value of a node indicates its control over other nodes whereas, Scaled PCI gives the minimum cardinality of all the nodes connected in the subgraph and the adjacent nodes in the dominating set of graph G. ScaPCICC displays high accuracy in selection of nodes and can work well for isolated nodes. Nodes selected using ScaPCI become the cache nodes which along with other nodes are encoded in bit string to form chromosomes that can proceed to the next level of optimization process using GA.

In the proposed algorithm we use a weighted sum approach to organize all the desired objectives in a single fitness function.

$$F = \min\left(\sum_{i=1}^{5} k_i x_i\right) \quad (3)$$

where, F is defined as

$$F = -\alpha_1 FC + \alpha_2 SpCN + \alpha_3 NetE + \alpha_4 OpCnE - \alpha_5 SORE \quad (4)$$

and values of coefficients α_i = 1, 2, 3, .. are decided based on design constraints and experimentations. We define field coverage consisting of components like N_{cn} (number of CNs sensors), number of nodes with HSR value; Nn1, number of LSR nodes; Nn2, number of inactive sensors; Ninact, number of out of range sensors; Nor and total number of sensing points; Ntotal. We define SpCN as Sensors-per-Caching Node, NetE as network energy, OpCnE is defined as Overlaps-per-Caching node- Error and SORE as Sensors-Out-of-Range Error [7].

3.2 Cache Consistency Using Flexible Combination of Push and Pull (FCPP)

Once cached nodes are created and popular or most frequently accessed data are cached then the next step is to maintain cache consistency. Cache consistency is necessary to ensure valid data access between the source node that detects the data first and its cached copies. Maintaining cache consistency incurs heavy overhead in WSNs as different users have different consistency requirement. For example some applications require frequent updates like stock market whereas; few applications allow some tolerance level in accessing the data like weather forecast. It therefore, becomes essential to allow users to tune in to their consistency requirements leading to a flexible adaptation to data update rate in the cached nodes. This also results in a trade-off between maintaining user-defined consistency requirements and consistency maintenance cost. The consistency model selected in the presented work is called the Probabilistic Delta Consistency (PDC) method. To further maintain user- specified consistency condition

at minimum cost we use the Flexible Combination of Push and Pull (FCPP) algorithm. We will discuss each technique in detail.

Probabilistic Delta Consistency allows users the flexibility to define their consistency requirements in two dimensions orthogonal to each other. The x-axis indicates the maximum tolerable deviation in δ time or value that can exist between the source data and the cached copies. Whereas, the y-axis denotes the probability p, which accounts for the minimum ratio of queries answered by consistent cached copies. PDC model is versatile in setting the consistency requirements in a network. PDC represented as (p, δ) helps identify the consistency level of a system. For example PDC (0, 100) indicates strong consistency whereas PDC (*, 0) indicates weak consistency (Fig. 4) [4].

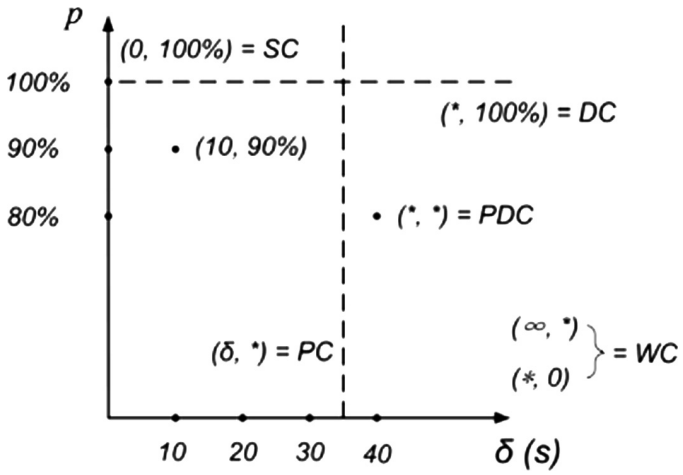


Fig. 4. Probabilistic delta consistency representation

To satisfy the consistency requirement with minimum cost we use the FCPP algorithm. FCPP associates a time-out value with every cached copy based on the δ and p values. Cache updates are made by the data source node by sending an (INV) message for which it receives an acknowledgement (INV_ACK) from cached nodes for valid time-out values of the cached copies. However, data updates can be postponed until it receives all the ACKs or until the time-out values of the entire un-responding nodes end or the upper limit of the acceptable delay of data update is reached. The key issue in implementing FCPP is to be able to quantify the tradeoff between consistency and cost. FCPP converts to Pull option when the time-out value 1 is shifts to zero, whereas, it transforms to Push format. FCPP works on both cache node and the data source node. On cache node the algorithm is described below:

FCPP on caching node

Step 1: Receive data request query

(a) IF (l > 0) serve the request with the consistent cache copy;
(b) ELSE // the time-out value l is decreased to zero

Step 2: generate a RENEW message for a new time-out value and update copy if needed;
Step3: Time-out value l is renewed; serve the user request with the updated cache copy;

FCPP on the data source node

Case 1: The source data is ready to be updated
Step1: Send invalidation message (INV) to caching node with positive l;
Step2: IF (INV_ACK is received from all cached nodes or D seconds have expired)
(2.1) Update the source data;
Case 2: On arrival of a RENEW message
Step3: IF (source data update is complete)
(3.1) Send updates to the caching node;
Step4: IF (no pending updates)
(4.1) Allow cache nodes to cache copies with time-out values 1;

FCPP is implemented on to the MAC layer of the cache node. We move to the simulations and results to see the advantages of the proposed algorithms.

4 Simulation and Results

4.1 GA Simulation and Results

GA Simulation parameters are provided in (Table 2). At the start of the GA simulations, each coefficient were assigned a unity value, however, with course of the experimentation the optimal values of the variables were determined. Figure 5 shows

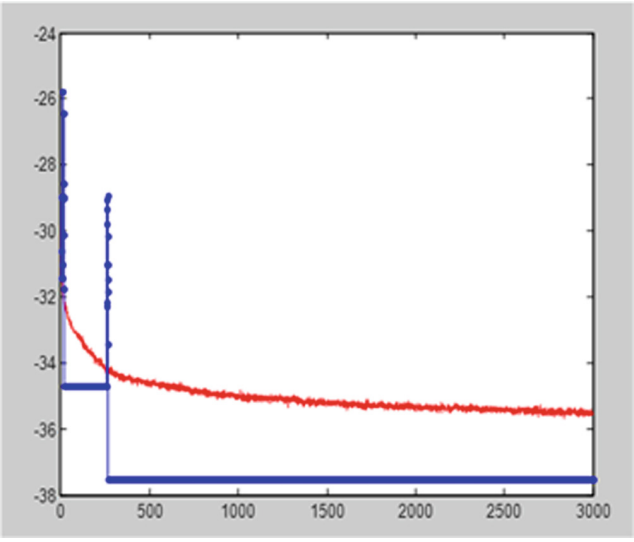


Fig. 5. Fitness values for the best individuals against average fitness value of the population

Table 2. GA simulation parameters

Parameter	Value
Area	Grid size of 700 cm × 1200 cm
Crossover	Two-point (p = 0.8)
Mutation	Nonlinear (p = 0.1). elitism
Selection	Roulette wheel
Number of GA runs	3000
Number of nodes	100 (results shown for 34 nodes)
Stability seen (termination criteria – no change in value)	300 runs
Base station location	Node numbered 0, placed at the centre

the fitness values for the best individuals against average fitness value of the population during the GA run. The value stabilizes to 37.8 from 300 runs improving the fitness value to 9.2% at 300 runs and 5.7% at 3000 generations (Fig. 5).

In spite of initial randomness due to search for best individuals, the network energy stabilizes to optimum value of 1.73 (Fig. 6).

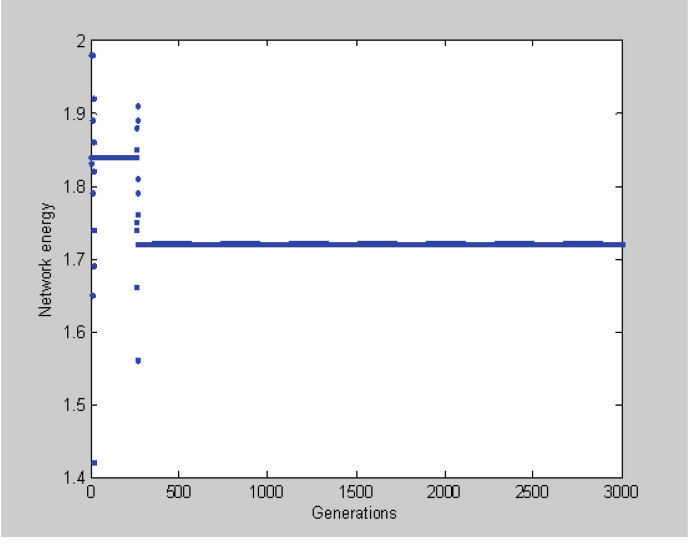


Fig. 6. Network Energy (NetE)

Figure 7 shows that the OpCnE parameter which is the number of overlaps between the cache nodes successfully reduces as the GA runs proceed towards stabilization. The OpCnE value becomes zero as the experiment progresses from 300 GA iterations.

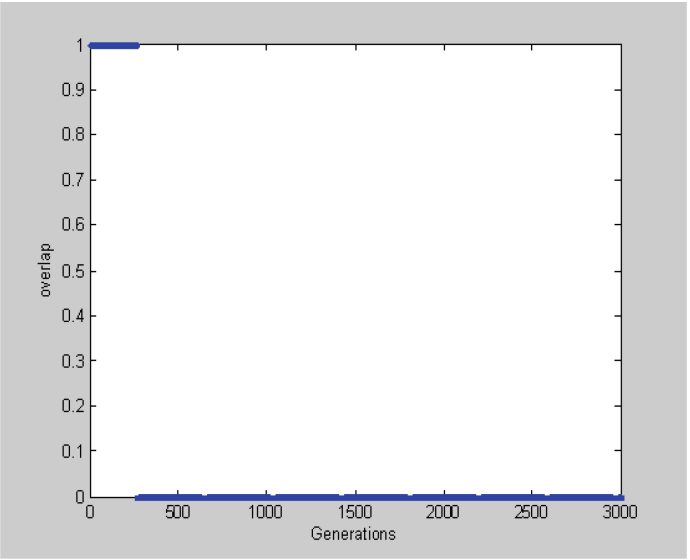


Fig. 7. Overlaps-per-Caching node-Error (OpCnE)

From Fig. 8, it is evident that after the experiment stabilizes the sensors out of range for a cache node is reduced by 11.2% (11.2 mark) and finally diminishes to zero from 300 generations onwards.

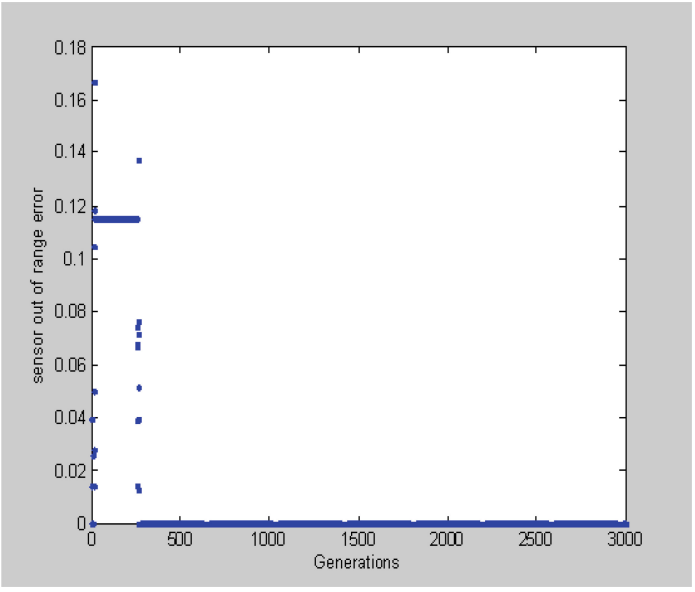


Fig. 8. Sensor Out of Range Error (SORE)

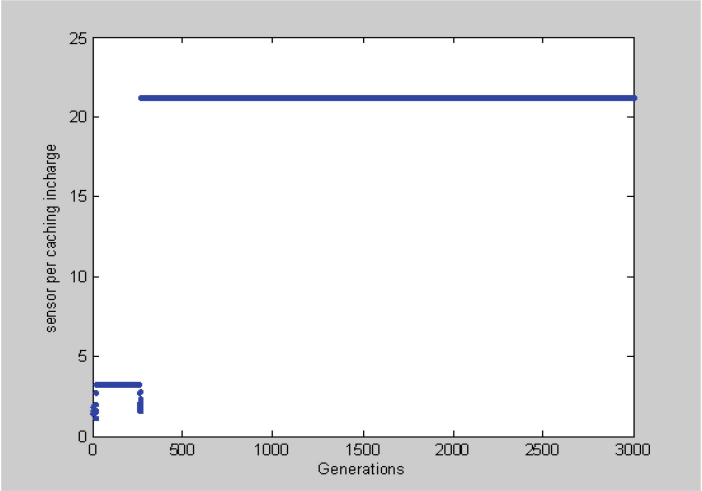


Fig. 9. Sensor-per-Cache Node (SpCN)

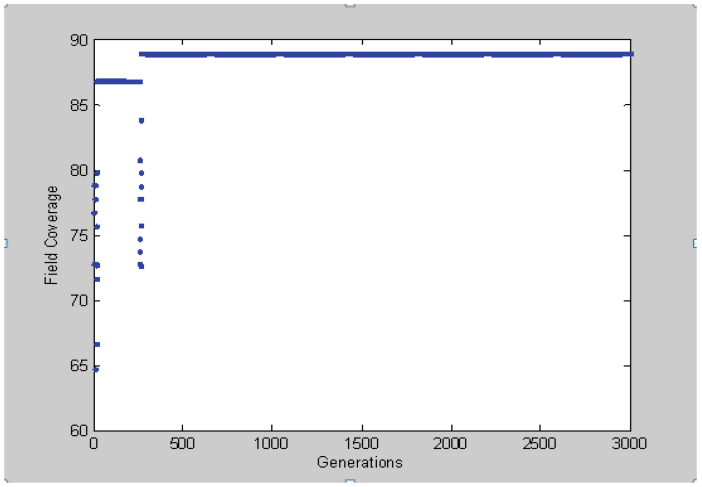


Fig. 10. Field coverage (FC)

Figure 9, is aimed at optimizing the fitness function parameter SpCN. The number of Sensors- per- Cache Node graph indicates an increased value of 22 from its previous value of 3.6 mark. This experiment leads to graph in Fig. 10 which shows 89% increase in the filed coverage. Maximizing filed coverage and reduced data access delay with minimum energy usage achieved through optimal cache node selection and placement is the key concept in the presented work.

The optimized network obtained from the MATLAB simulator after running GA gives a text file with location information of all the nodes in the WSN. The cluster

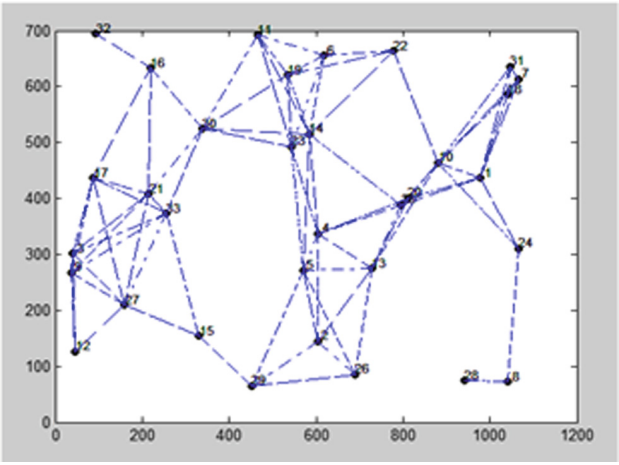


Fig. 11. Connectivity graph of the optimized network obtained after running GA

heads, cluster members and the identified cache nodes locations coordinates are now fed into the ns2 simulator for maintaining cache consistency requirements. Table 3 shows the text file which is the optimized network obtained after using GA for cache placement in WSN for PCI > 70. The PCI value helps to decide the optimal number of CNs required to cover the entire network.

Figure 11 shows the connectivity graph of the optimized network (degree of each node) obtained after running GA over WSN for optimal selection and placement of CNs. From Fig. 11 we can understand the roles of each sensor node in the network for one complete cycle of data transmission. A useful cluster configuration is obtained only if we have a good CH (Cluster head) to CM (Cluster member) ratio so that, clustering covers the entire network using minimum network energy enhancing network lifetime. The optimal ratio of CH to CM for N nodes organized in a clustered architecture is given by CH fraction [8].

$$\mathrm{CH_{Frac}} = \mathrm{CH}/\mathrm{N} \tag{5}$$

where, N is total number of nodes given by following equation

$$\mathrm{N} = |\mathrm{CH}| + |\mathrm{CM}| \tag{6}$$

4.2 FCPP Simulation and Results

FCPP simulation parameters are provided in (Table 4). In the presented, work Flexible Combination of Push and Pull (FCPP) algorithm has been deployed on to the MAC Layer. Comparison of FCPP implemented network with 802.11 MAC layer is done in terms of Packet Delivery Ratio, End-to-end delay, normalized routing load, Routing overhead (in terms of number of packets) and Energy consumed (Tables 5 and 6).

Table 3. Optimized network generated by using the genetic algorithm

Node ID	X coordinate	Y coordinate	Role
1	978	437	Head
2	604	145	Head
3	41	301	Head
4	605	337	Member
5	571	272	Head
6	618	655	Member
7	1065	614	Member
8	1042	72	Head
9	36	268	Member
10	882	464	Head
11	465	692	Head
12	47	126	Cache node
13	728	275	Head
14	585	515	Cache node
15	331	154	Member
16	219	633	Head
17	88	436	Head
18	1040	586	Member
19	534	621	Member
20	811	401	Cache node
21	215	408	Member
22	777	664	Head
23	543	493	Member
24	1065	311	Cache node
25	796	388	Head
26	689	85	Head
27	160	209	Member
28	941	74	Member
29	452	65	Member
30	338	524	Cache node
31	1046	636	Head
32	94	694	Member
33	255	373	Member

Figures 12, 13, 14, 15 and 16 shows that FCPP performs better over 802.11 MAC layer. Tables 5 and 6 shows a comparative analysis between FCPP implemented MAC layer with 802.11 MAC layer.

Table 4. FCPP simulation parameters

Parameter	Value
Number of nodes	30–100
Area	700 cm × 1200 cm
Base station/sink	Center of the area (node ID = 0), assumed stationary
Maximum node velocity (V_{max})	20 m/s
Minimum node velocity (V_{min})	0 m/s
Pause time (T_p)	0 m/s
Simulation time	100 s
MAC	IEEE 802.11
Consistency requirement p	60–90%
Consistency requirement δ	2 s–20 s
Routing protocol	AODV
Maximum delay before data update	2 s
Average update interval	20 s–80 s
Mobility model	Random way point

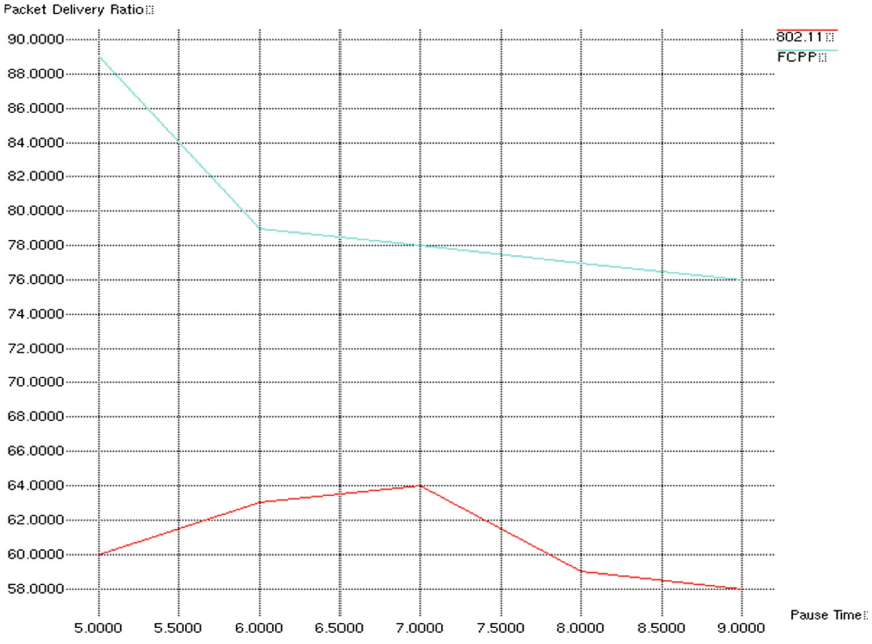


Fig. 12. Packet delivery ratio 802.11 MAC layer versus FCPP implemented MAC layer

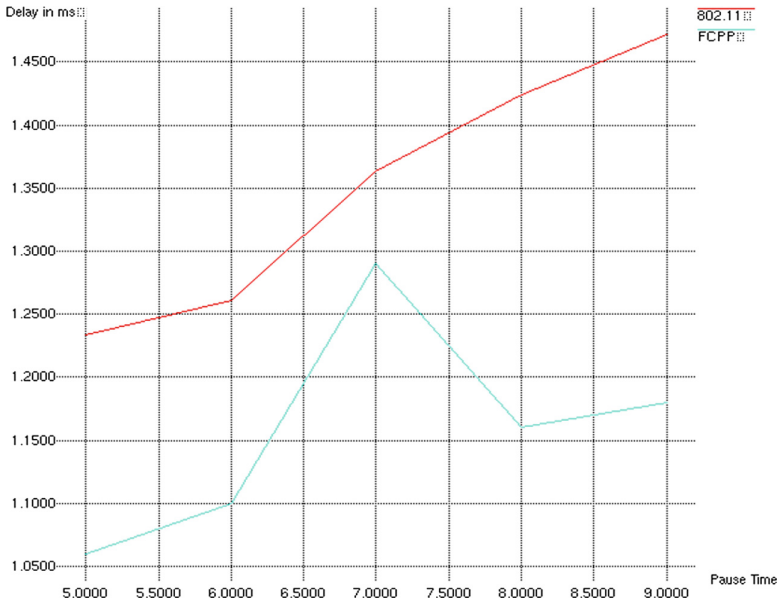


Fig. 13. End-to-end delay 802.11 MAC layer versus FCPP implemented MAC layer

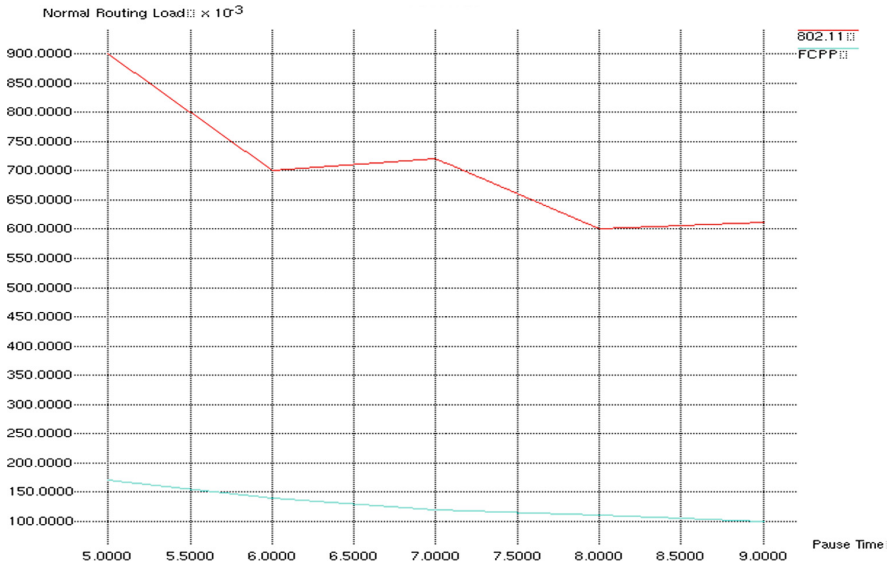


Fig. 14. Normalized routing load 802.11 MAC layer versus FCPP implemented MAC layer

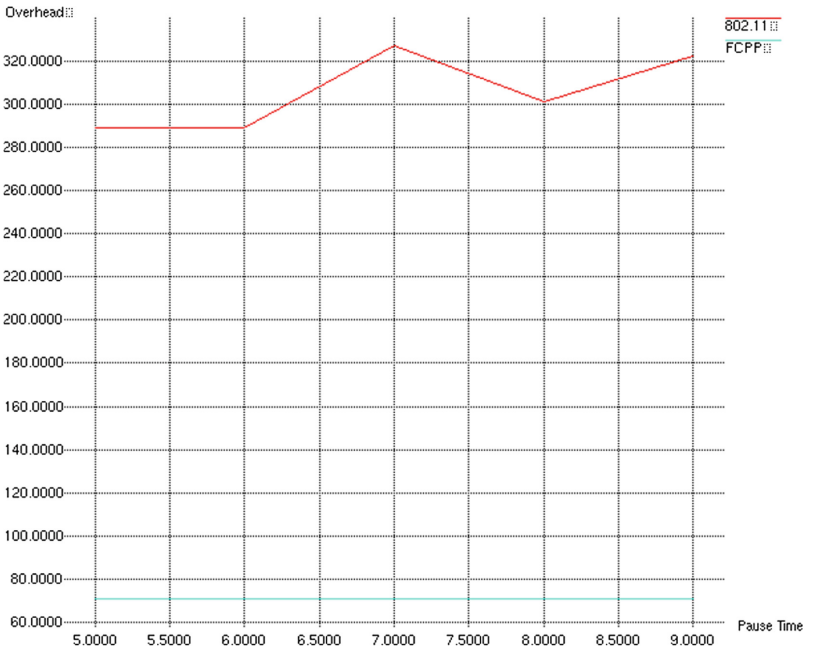


Fig. 15. Routing overhead 802.11 MAC layer versus FCPP implemented MAC layer

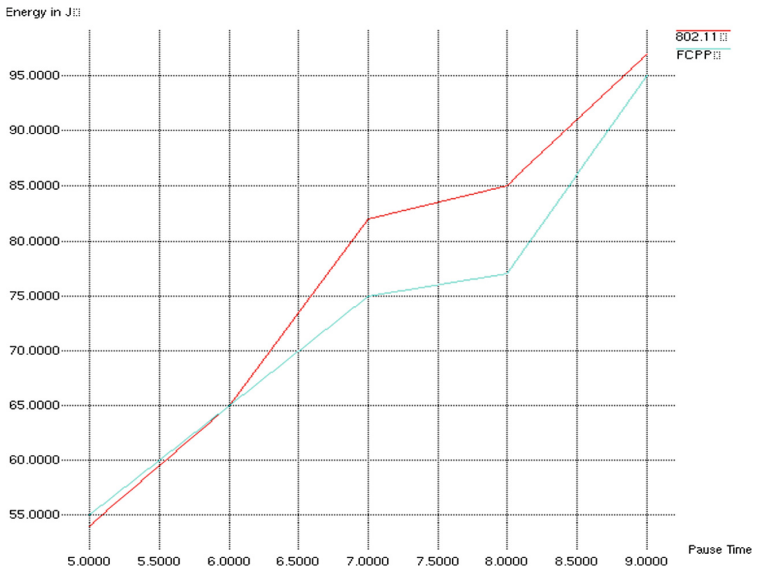


Fig. 16. Energy consumption 802.11 MAC layer versus FCPP implemented MAC layer

Table 5. 802.11 MAC layer

For pause time (in seconds)	Packet delivery ratio	End-to-end delay	Normalized routing load	Routing overhead	Energy consumed
5	60.48	1.234	0.954	289	54.48
6	63.56	1.260	0.757	289	65.05
7	64.48	1.36	0.72	327	82.41
8	59.05	1.42	0.64	301	85.29
9	58.49	1.47	0.611	322	97.42

Table 6. FCPP implemented MAC layer

For pause time (in seconds)	Packet delivery ratio	End-to-end delay	Normalized routing load	Routing overhead	Energy consumed
5	81.64	1.06	0.174	71	55.53
6	79.37	1.10	0.149	71	65.32
7	78.46	1.13	0.129	71	75.62
8	77.90	1.16	0.114	71	85.89
9	76.69	1.18	0.103	71	95.55

5 Conclusions and Future Work

In the presented work we propose a scheme for cache node selection and optimal placement of selected cache nodes in WSNs using genetic algorithm. The proposed algorithm uses a multi objective fitness function and aims at minimizing the network energy usage during data transmission, thus resulting in an increased network lifetime. The technique is designed for a clustered WSN used for monitoring applications. The work further progresses by performing cache consistency on the cached nodes to guarantee valid data retrieval by users. To achieve this we use the Flexible Combination of Push and Pull algorithm which is implemented over the 802.11 MAC layer. The algorithm FCPP is run on cached nodes and data source node for valid data access by users. The cache consistency algorithm is implemented on the MAC layer and comparisons are made with 802.11 MAC layer without cache consistency using metrics like routing load, packet delivery ratio, end-to-end delay, normalized load and energy consumed. The results obtained clearly show that cache placement using GA is effective in achieving maximum field coverage of 89% in WSN while reducing network energy consumption thus granting an increase in network lifetime.

The cache consistency algorithm FCPP further helps the network in faster data retrieval without message overheads generated in the network. It proves to be a very convenient choice for users to tune their consistency requirements with ease and at the same time avoid too many transmissions in the network thus leading to an energy efficient network system. However, the algorithm FCPP is limited when frequency of data updates are high and the network is unstable. FCPP algorithm performs best when

the user requirements are more severe. The experiments were completed using MATLAB and ns2 simulator.

For future work, we propose enhancement in the MAC layer to further improvise on the results. We would also like to consider varied needs of dissimilar users for data retrieval using FCPP. We finally propose to deal with situations with mobile base stations.

Conflict of interest. Authors declare that they have no conflict of interest

References

1. Dimokas, N., Katsaros, D., Tassiulas, L., Manolopoulos, Y.: High performance, low complexity cooperative caching for wireless sensor networks. Wirel. Netw. **17**, 717–737 (2011). doi:10.1007/s11276-010-0311-x
2. Elbeltagi, E., Hegazy, T., Grierson, D.: Comparison among five evolutionary-based optimization algorithms. Adv. Eng. Inf. **19**(1), 43–53 (2005). doi:10.1016/j.aei.2005.01.004
3. Hoyingcharoen, P., Teerapabkajorndet, W.: Fault tolerant sensor placement optimization with minimum detection probability guaranteed. In: Eighth International Workshop on the Design of Reliable Communication Networks (DRCN) (2011)
4. Huang, Yu., Cao, J., Jin, B., Tao, X., Lu, J., Feng, Y.: Flexible cache consistency maintenance over wireless adhoc networks. IEEE Trans. Parallel Distrib. Syst. **21**(8), 1150–1161 (2010)
5. Khanna, R., Liu, H., Hwa Chen, H.: Self-organization of sensor networks using genetic algorithms. In: Proceeding of International Conference on Communications, vol. 8 (2006)
6. Mollanejad, A., Khanli, L.M., Zeynali, M.: DBSR: dynamic base station repositioning using genetic algorithm in wireless sensor network. In: Second International Conference on Computer Engineering and Applications, pp. 521–525 (2010)
7. Srivastava, J.R., Sudarshan, T.S.B.: Energy-efficient cache node placement using genetic algorithm in wireless sensor networks, "Soft Computing" Journal, Springer (2014). doi:10.1007/s00500-014-1473-8
8. Srivastava, J.R., Sudarshan, T.S.B.: A genetic fuzzy system based optimized zone based energy efficient routing protocol for mobile sensor networks (OZEEP), Appl. Soft Comput. J., Elsevier, vol. 37, pp. 863–886, December, 2015. doi:10.1016/j.asoc.2015.09.025
9. Sun, P., Ma, J., Ni, K.: A genetic simulated annealing hybrid algorithm for relay nodes deployment optimization in industrial wireless sensor networks. In: IEEE International Conference on Computational Intelligence for Measurement Systems and Applications (CIMSA) (2012)
10. Wang, Y., Li, X., Tian, J.: Blackboard mechanism based genetic algorithm for dynamic deployment of mobile sensor networks. In: International Conference on Electronic & Mechanical Engineering and Information Technology (2011)
11. Yang, L.: Determining sink node locations in wireless sensor network. In: IEEE International Conference on Systems, Man, and Cybernetics, vol. 4, Taiwan (2006)
12. Yong, M., Huludao, L., Yu, Y., Yan, W.: Optimization design of coal mine body sensor network based on Genetic Algorithm. In: International Conference on Network Security, Wireless Communications and Trusted Computing (2009). doi:10.1109/NSWCTC.2009.347
13. Youssef, W., Younis, M.: Intelligent gateways placement for reduced data latency in wireless sensor networks. In: Proceedings of the 32nd IEEE International Conference on Communications (ICC 2007). Glasgow, Scotland, UK (2007)
14. Cao, J., Zhang, Y., Xie, L., Cao, G.: Data consistency for cooperative caching in mobile environments. Computer **40**(4), 60–67 (2007)

A Fast JPEG2000 Based Crypto-Compression Algorithm: Application to the Security for Transmission of Medical Images

Med Karim Abdmouleh(✉), Hedi Amri, Ali Khalfallah, and Med Salim Bouhlel

Research Unit: Sciences and Technologies of Image and Telecommunications, Higher Institute of Biotechnology, University of Sfax, Sfax, Tunisia
medkarim.abdmouleh@isggb.rnu.tn, hedi.amri@setit.rnu.tn, ali.khalfallah@enetcom.rnu.tn, medsalim.bouhlel@enis.rnu.tn

Abstract. Over the past years, the use of telecommunications and information technologies in medicine is evolving. This involves the development of the applications bound to the telemedicine and based on a network medical image transmission. Therefore, the optimization of medical application performances remains a necessity. In this paper, we propose a novel and efficient crypto-compression algorithm. This novel scheme concerning the application of a partial encryption to the JPEG2000 file format. Our algorithm is rapid, efficient, secure and it perfectly preserves the performances of the JPEG2000 compression algorithm. In addition, the proposed transmission scheme is adapted to the Telediagnostic sector and can be easily integrated in JPEG2000 coder.

Keywords: Crypto-compression · JPEG2000 · RSA · Medical images · Transmission · Telemedicine

1 Introduction

The extraordinary evolution of information technologies in the medical sector is motivated by political, professional and industrial wills. Currently, Telediagnostic is considered as the most potential sector in telemedicine [11–15]. It allows two or several medical teams to exchange medical information and to comment on it in a context of diagnosis help. This sector is the result of the creation of "proficiency poles" that is involved by the medical hyper specialization. Given the importance of this sector in the improvement of the care quality, the reduction of treatment costs and the universalization of the medical practices and knowledge, it is necessary to optimise this class of applications. To do this, both encryption and compression must be performed [4].

Various encryption schemes for JPEG2000 images have been proposed [16,17,19,26,30,35,36]. Based on the order to mix compression and encryption

V.E. Balas et al. (eds.), *Soft Computing Applications*, Advances in Intelligent Systems and Computing 633, DOI 10.1007/978-3-319-62521-8_14

together to reduce the overall processing time [6,7,20,25,28,34], but they are either insecure or too computationally intensive and may induce degradation of coding efficiency.

In fact, network communication, especially on the World Wide Web, can be intercepted. Therefore, there is a risk of placing at the disposal of eavesdroppers the secrecies of patients' state during the transmission of medical data through a non-protected network. Thus, the encryption of the medical image is imposed to assure the deontology and the safeguard of patient medical secrecy. In particular, the asymmetric encryption can also be used to assure the authentication of medical image transmission.

In addition, because of the important size of medical images after the digitalisation, we must compress them in order, first, to improve the capacity of storage, and second to optimize theirs transmission through networks (reduction in the transmission time and in the obstruction of networks).

To satisfy these conditions, the classic approach consists in applying a compression algorithm to the medical image [4,27]. The output is then encrypted with an independent cryptographic algorithm. Unfortunately, the processing time in classical cryptographic algorithms is too long to satisfy the Telediagnostic required speed because this type of application is mainly used in a case of emergency.

The rest of this paper is organized as follows. In the Sect. 2, the encryption algorithm RSA and the JPEG2000 still image compression algorithm are briefly described to facilitate the development of our proposed algorithm. Section 3 presents the principle of our approach. Then, the proposed encryption and decryption procedure is detailed. After that, the robustness of our algorithm against the different types of attacks is discussed. Finally, the advantages of our algorithm are enumerated. Section 4 concludes the paper.

2 Background

2.1 The Asymmetric Encryption Algorithm: RSA

The RSA is a public-key cryptosystem that was invented by Rivest et al. [31]. It may be used to provide both confidentiality and authentication, and its security is based on the intractability of the integer factorization. The cryptosystem RSA utilizes two distinct keys: the public key and the private key. The first type can be communicated freely, through an insecure channel, to all the correspondents susceptible to exchange data with the possessor of the private key. On the other hand, the private key must remain confidential [2] (Fig. 1).

2.1.1 Key Generation

For the creation of an RSA public key and a corresponding private key, each person X should, first, generate two large random (and distinct) primes p and q, that have roughly the same size and then compute [29]:

$$n = p \times q \; and \; \phi = (p-1)(q-1)$$

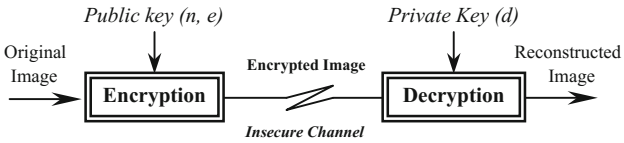


Fig. 1. Principle of asymmetric encryption.

Afterwards, he should select a random integer e:

$$1 < e < \phi, \text{ such that } \gcd(e, \phi) = 1$$

Next, the extended Euclidean algorithm is used to compute the unique integer d:

$$1 < d < \phi, \text{ such that } e \cdot d \equiv 1 (\text{mod} \phi)$$

As result, X's public key is (n, e) and X's private key is d [29].

2.1.2 Algorithm

Let's suppose that a person Y wants to send encrypted data to X who must be able to decrypt these encrypted data.

If the size of n in bits is t, the encryption processing requires, firstly, the break of data into k bits of blocks and the conversion of these bocks into their decimal representation *(m)*. To encrypt each integer m, which is in the interval $[0, n-1]$, Y should obtain X's public key (n, e) and compute the corresponding c:

$$c = m^e \bmod n$$

Convert this c into binary representation, redo this operation with all the k bits of blocs, and finally send the encrypted data to X.

After doing all the necessary conversions and breaks, X should use the private key d to recover m from c [29]:

$$m = c^d \bmod n$$

The RSA adds a very significant benefit: it can serve to authenticate the sender of the data (e.g. a digital signature) [9,10,21]. In fact, the RSA keys are complementary: if one encrypt with a public key, it must use the private key for recovering the encrypted data (Fig. 2). Inversely, if one encrypt with a private

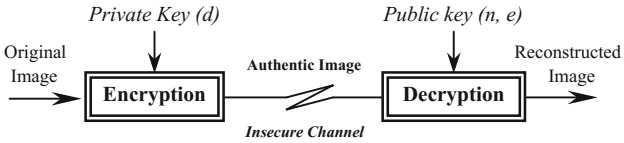


Fig. 2. Principle of authentication with RSA algorithm.

key, it must use the public key to recover the encrypted data. Therefore, when X encrypts his own image with his own secret key, Y will be able to use the X's public key to decrypt the image, and in this case, Y is certain that it is X who sent him the image because he is the only person who possess the secret key.

2.1.3 Security

The security of the RSA algorithm is generally equivalent to the hardness of the factoring problem. In fact, to find the private key d, the only possibility is the factorization of n [33]. However, we need to be careful in the choice of the key size for RSA because of the increase in computing power and the continuing refinements of the factoring algorithms [33].

The last factoring record was achieved on December 12, 2009 when a six-institution research team led by T. Kleinjung threw down a challenge to decrypt RSA-768 number. The effort took almost 2000 2.2 GHz-Opteron-CPU years according to the submitters, just short of three years of calendar time [22].

2.2 JPEG2000

The JPEG2000 standard was created by the Joint Photographic Experts Group (JPEG) [1], also denominated as ISO/IEC 15444. It's an image compression algorithm that allows great flexibility, not only for the compression of images, but also for the access to the compressed data, such as the several progressive decoding modes: by resolution, by quality, by position or by region. Currently, this standard is divided into fourteen distinct parts, where Part 1 [18] describes the core coding system and Part 2 its extensions. However, only the first part reached the International Standard (IS). This part describes mainly the decoding algorithm and the compressed data format: the codestream is the result of a juxtaposition of headers containing coding parameters and bit streams (a compressed form of image data).

Since our approach is about the application of a partial encryption in the JPEG2000 file format, only a short preview about the codestream building will be presented in the next paragraph.

The JPEG2000 file format (JP2 format): This format provides a foundation for the storing application of the specific data (metadata). A JP2 File represents a collection of boxes (Fig. 3) which design a building block defined by a unique box type and length [1].

As shown in Fig. 4, four fields compose a box: the LBox, TBox, XLBox, and DBox fields.

The LBox field specifies the length of the box in bytes. The TBox field indicates the type of box (i.e., the nature of the information contained in the box).

The XLBox field is an extended length indicator that provides a mechanism for specifying the length of a box the size of which is too large to be encoded in the length field alone. If the LBox field is 1, then the XLBox field is present and contains the true length of the box. Otherwise, the XLBox field is not present [1].

The DBox field contains data specific to the particular box type. Some of the JP2 File boxes are independent, and some of those boxes contain other boxes.

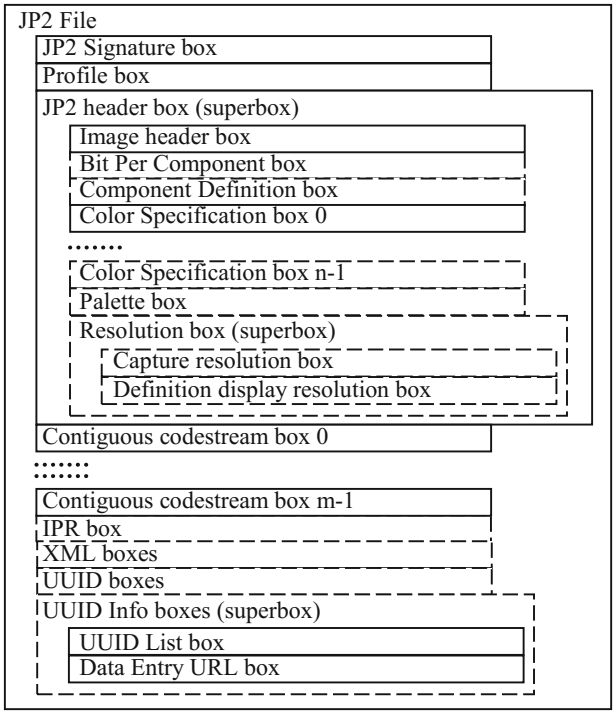


Fig. 3. Conceptual structure of JP2 File.

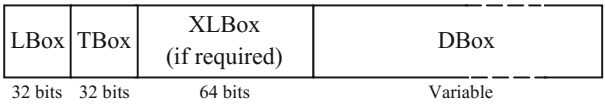


Fig. 4. Box structure.

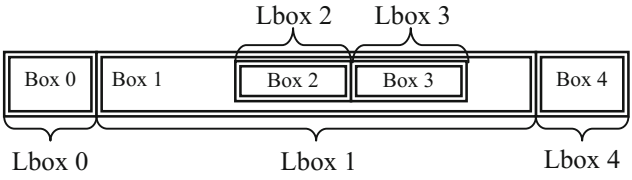


Fig. 5. Illustration of box length.

As a matter of terminology, a box that contains other boxes in its DBox field is referred to as a superbox. The binary structure of a file is a contiguous sequence of boxes (Fig. 5).

3 The Fast Crypto-Compression

In this section, we will develop the proposed JPEG2000 based crypto-compression algorithm. This novel scheme concerning the application of a partial encryption to the JPEG2000 file format (Fig. 6). Firstly, the principle of the novel approach, on which our algorithm is based, is presented. Second, the proposed robustness of our algorithm to the different type of attacks is proved. Finally, the advantages of our crypto-compression technique are enumerated.

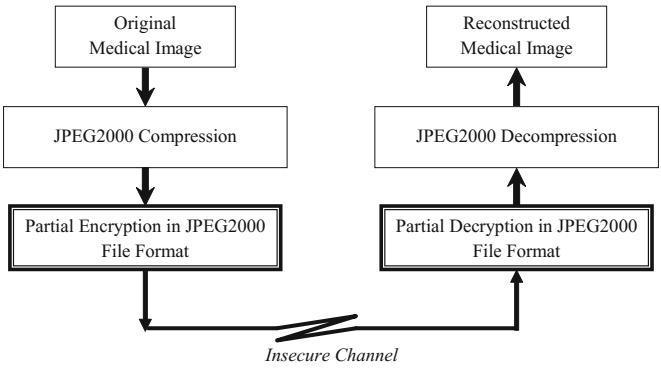


Fig. 6. Scheme of our approach.

3.1 Principle of Our Approach

If we examine the JP2 File structure attentively, we conclude that most of the information in this type of file is independent (Structure in packet). Therefore, to design an efficient crypto-compression algorithm, we must touch the totality of the file structure.

Since the JP2 File format is composed of boxes, and the first 32 bits (Lbox) of every box contain the box length information. Actually, the basic idea of our approach consists of encrypting only the Lbox part of every box in the JP2 File format excepting the boxes included in a superbox. Consequently, it will be impossible to recover the original data because the correct box length is unknown. As a result, the image is protected.

3.2 Encryption Algorithm

In Sect. 2, it is noted that the principle of the RSA algorithm consists of encrypting a blocs of K bits (where K is the length in bits of the RSA key). Hence, if one uses the RSA algorithm for encrypting, the length L of the data that can be encrypted, in bits must be a multiple of K:

$$L = n \times K \ (n \text{ is an integer})$$

Since it is possible that the Lbox bit sequence is not a multiple of K bits, we propose the following encryption procedure:

- Regrouping all the Lboxes in a structure of a contiguous bit sequence without changing their order (Fig. 7a). This (32 × B) bit sequence must be placed at the beginning of the JP2 File (where B represents the number of JP2 File boxes excepting those that are included in a superbox).
- Adding 32 bits in order to indicate the length of the Lbox sequence length (32 × B bits) because the recovering of the original data requires the transmission of the Lbox sequence length. This few added bits must also be placed at the beginning of the JP2 File but before the Lbox sequence (Fig. 7b).
- Determine an integer N that verifies:

$$N < K \text{ and } (32 + 32 \times B + N) \text{ is a multiple of } K$$

Then, we encrypt the $(32+32\times B+N)$ bits with the RSA encryption algorithm (Fig. 7c).

For decryption the encrypted data in the reception, we must:

- Decrypt the K bits from the beginning of JP2 File. The first 32 decrypted bits allow the determination of the Lbox sequence length (32 × B bits) (Fig. 7d).
- Search the integer N, with the same way that is employed at the encryption stage. Once N is determined, we must decrypt the remaining encrypted data: $[(32 + 32 \times B + N) - K]$ bits.
- Delete the first 32 bits that indicate the Lbox sequence length.
- Place the new first 32 bits (corresponding to the Lboxes of the first Box in JP2 File) just after the Lbox bit sequence. Consequently, the length of the first box is known. The second 32 bits are placed just after the end of the first box. Since the length of the second box is known, we can place the next 32 bits at the end of the second box and so forth until all the Lboxes in their original places (Fig. 7e).

After the decryption stage, the original image is recuperated by decoding the resulting JP2 File with any classical JPEG2000 decoder.

3.3 Security

In this part, we discuss the robustness of our crypto-compression algorithm against different attacks through the cryptanalysis that is the study of breaking encryption algorithms. In this discussion, we assume that the cryptanalyst have full access to the description of the algorithms, as well as full access to the insecure channel through which the image is transmitted [33].

The security of our image cryptosystem will be analysed for the following four types of attacks, such as exhaustive key search attack, cipher image-only attack, known-plain image attack, and chosen-plain image attack.

In the Exhaustive key search, the cryptanalyst tests each of the possible keys one at a time until the correct plain image is recognized [8]. Since we use

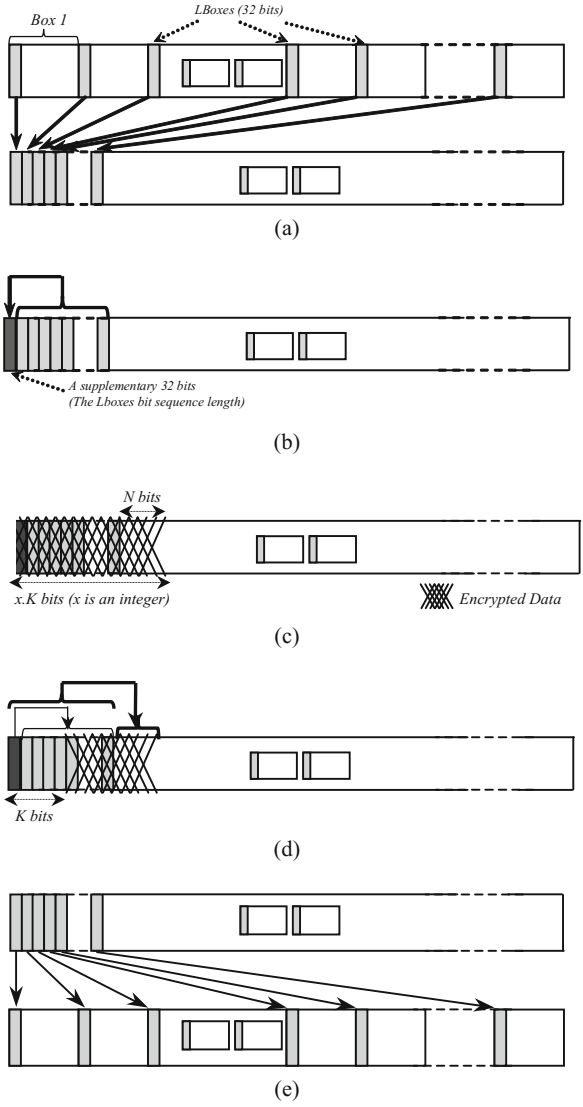


Fig. 7. The proposed encryption and decryption stages.

the RSA keys for the encryption and decryption of the box bit sequence in our algorithm, the break of our algorithm with this type of attack is equivalent to the test of all the possible RSA Keys.

In the cipher image-only attack, the cryptanalyst has access only to several images that are encrypted with the same key. The cryptanalyst attempts to recover the corresponding plain image or the decryption key [8].

Since it is impossible to recovering the RSA decryption key, and it is impossible to recovering any information about the original data without the acquaintance of the length of the JP2 File boxes (Lbox), our algorithm is robust against this type of attacks.

In the Known-plain image attack, the cryptanalyst has access to the cipher image and the corresponding plain image of several images encrypted with the same key. The cryptanalyst attempts to recover the key or to design an algorithm to decrypt any image that is encrypted with the same key [33]. In the case of our algorithm, the encryption is made on a bloc of K bits (RSA principle). The result of the RSA encryption of two original K bits blocs that contain few different bits is two K bits blocs that are completely different. Since the original bit sequence of two similar images contain certainly few different bits, it is impossible to find a relation between the original and the encrypted blocs. Consequently, it is useless to cryptanalyse our algorithm with a know-plain image attack.

In the Chosen-plain image attack, the cryptanalyst is allowed to choose the plain image that is encrypted, and observe the corresponding cipher image. The cryptanalyst's goal is the same as that in a known-plain image attack. The cryptanalyse of our algorithm is equivalent to that of the bit sequence that is encrypted with RSA encryption algorithm. As a consequence, the break of our crypto-compression algorithm with the chosen plain image attack is equivalent to that of the RSA technique with this type of attack.

3.4 Advantages of Our Scheme

- Our algorithm is rapid. In fact, we encrypt only the Lbox part (32 bits) from the whole information that can be contained in the JP2 File. Therefore, our approach allows an enormous reduction of the encrypting-decrypting processing time.
- Our algorithm perfectly preserves the performances of the JPEG2000 compression algorithm [32] because the encryption is made after all the compression stages, and therefore, it does not tamper with the compression ratio. The 32 bits, which are added to communicate the Lbox bit sequence length, are too negligible, therefore, one cannot consider this fact as an inconvenient.
- Moreover, our algorithm is secure. In fact, it is demonstrated in Sect. 3 that the security of our crypto-compression system is equivalent to that of the RSA algorithm. This security depends essentially on the key length. According to the RSA factoring records, if we use a key size in the range of 1024 to 2048 bits, we can affirm that our algorithm is robust to the different attacks for many years.
- Our algorithm is adapted to the medical image transmission in Telediagnostic [3,5,23,24]. In addition, since RSA permits the authentication of the sender of the medical image [9,10,21], we can assure a secure and an authentic medical image communication with our algorithm.

4 Conclusion

In this paper, we proposed a new crypto-compression algorithm that significantly reduces the encryption and decryption time and can assure a secure and an authentic transmission in medical image communication. Based on the structure of the JP2 File, which is based on a succession of box and superbox, we developed, in this work, an original approach concerning the encrypting in JP2 File only the part that indicates the length of the boxes (the Lbox).

Given the originality and the efficiency of this approach, the proposed algorithm, which is derived from, is rapid, secure, and preserves the excellent performance of JPEG2000 algorithm and adapted to the medical image transmission in Telediagnostic sector. In addition, our algorithm can be applied in many other telecommunication sectors like Tele-expertise, secure broadcasting, and distributed image retrieval.

References

1. ISO/IEC 1.29.15444-1, "JPEG2000 part i final committee version 1.0", September 2004
2. PKCS#1 v2.2, RSA cryptography standard, RSA Laboratories (2012)
3. Amri, H., Hanna, F., Lapayre, J.C., Khalfallah, A., Bouhlel, M.S.: Repro: a new reduction/expansion protocol to increase the performance of image transmission in medical telediagnosis platforms. Biomed. Eng.: Appl. Basis Commun. **27**(06), 1550054 (2015)
4. Bouhlel, M.S., Kammoun, F., Garcia, E.: An efficient DCT-based crypto-compression scheme for a secure and authentic medical image transmission. J. Test. Eval. Appl. Sci. Eng. **34**(6), 459–463 (2006)
5. Chaabouni, I., Fourati, W., Bouhlel, M.S.: Using ROI with ISOM compression to medical image. IJCVR **6**(1/2), 65–76 (2016)
6. Chang, C.C., Hwang, M.S., Chen, T.S.: A new encryption algorithm for image cryptosystems. J. Syst. Softw. **58**(2), 83–91 (2001)
7. Chang, H.K.C., Liu, J.L.: A linear quadtree compression scheme for image encryption. Signal Process.: Image Commun. **10**(4), 279–290 (1997)
8. Cheng, H., Li, X.: Partial encryption of compressed images and videos. IEEE Trans. Signal Process. **48**(8), 2439–2451 (2000)
9. Cramer, R., Damgard, I.: New generation of secure and practical RSA-based signatures. In: 16th Annual International Cryptology Conference, pp. 173–185. Springer Berlin Heidelberg, Berlin, Heidelberg (1996)
10. Cramer, R., Shoup, V.: Signature schemes based on the strong RSA assumption. ACM Trans. Inf. Syst. Secur. **3**(3), 161–185 (2000)
11. Dey, N., Bose, S., Das, A., Chaudhuri, S.S., Saba, L., Shafique, S., Nicolaides, A., Suri, J.S.: Effect of watermarking on diagnostic preservation of atherosclerotic ultrasound video in stroke telemedicine. J. Med. Syst. **40**(4), 1–14 (2016)
12. Dey, N., Mukhopadhyay, S., Das, A., Chaudhuri, S.S.: Analysis of P-QRS-T components modified by blind watermarking technique within the electrocardiogram signal for authentication in wireless telecardiology using DWT. Int. J. Image, Graphics Signal Process. (IJIGSP) **4**(7), 33–46 (2012)

13. Dey, N., Nandi, B., Das, P., Das, A., Chaudhuri, S.S.: Retention of electrocardiogram features insignificantly devalorized as an effect of watermarking for a multimodal biometric authentication system. In: Advances in Biometrics for Secure Human Authentication and Recognition, pp. 175–212 (2013)
14. Dey, N., Pal, M., Das, A.: A session based blind watermarking technique within the NROI of retinal fundus images for authentication using DWT, spread spectrum and harris corner detection. Int. J. Mod. Eng. Res. (IJMER) **2**(3), 749–757 (2012)
15. Dey, N., Samanta, S., Yang, X.S., Das, A., Chaudhuri, S.S.: Optimisation of scaling factors in electrocardiogram signal watermarking using cuckoo search. Int. J. Bio-Inspired Comput. **5**(5), 315–326 (2013)
16. Fang, J., Sun, J.: Compliant encryption scheme for JPEG2000 image code streams. J. Electron Imaging **15**(4), 043,013–043,013–4 (2006)
17. Grangetto, M., Grosso, A., Magli, E.: Selective encryption of JPEG2000 images by means of randomized arithmetic coding. In: IEEE 6th Workshop on Multimedia Signal Processing, pp. 347–350 (2004)
18. Grosbois, R., Santa-Cruz, D., Ebrahimi, T.: New approach to JPEG2000 compliant region of interest coding. In: SPIE's 45th Annual Meeting, Applications of Digital Image Processing XXIV, vol. 4472, pp. 267–275 (2001)
19. Gu, G., Ling, J., Xie, G., Li, Z.: A chaotic-cipher-based packet body encryption algorithm for JPEG2000 images. Signal Process.: Image Commun. **40**, 52–64 (2016)
20. Jones, D.: Application of splay trees to data compression. Commun. ACM **31**(8), 996–1007 (1988)
21. Katz, J.: Digital Signatures. Springer, US (2014)
22. Kleinjung, T., Aoki, K., Franke, J., Lenstra, A.K., Thomé, E., Bos, J.W., Gaudry, P., Kruppa, A., Montgomery, P.L., Osvik, D.A., Te Riele, H., Timofeev, A., Zimmermann, P.: Factorization of a 768-bit RSA modulus. In: Proceedings of the 30th Annual Conference on Advances in Cryptology, pp. 333–350 (2010)
23. Lazrag, H., Naceur, M.S.: Despeckling of intravascular ultrasound images using curvelet transform. In: 6th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT), pp. 365–369 (2012)
24. Lazrag, H., Naceur, M.S.: Wavelet filters analysis for speckle reduction in intravascular ultrasound images. In: 6th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT), pp. 375–379 (2012)
25. Li, X., Knipe, J., Cheng, H.: Image compression and encryption using tree structures. Pattern Recognit. Lett. **18**(11–13), 1253–1259 (1997)
26. Lian, S., Sun, J., Zhang, D., Wang, Z.: A selective image encryption scheme based on JPEG2000 codec. In: 5th Pacific-Rim Conference on Multimedia PCM, pp. 65–72 (2004)
27. Lima, J., Madeiro, F., Sales, F.: Encryption of medical images based on the cosine number transform. Signal Process.: Image Commun. **35**, 1–8 (2015)
28. Matias, Y., Shamir, A.: A Video Scrambling Technique Based On Space Filling Curves. In: Proceedings of Advances in Cryptology (CRYPTO), pp. 398–417. Springer Berlin Heidelberg, Berlin, Heidelberg (1988)
29. Menezes, A.J., Van Oorschot, P.C., Vanstone, S.A.: Handbook of Applied Cryptography. Taylor & Francis (1997)
30. Modrzyk, D., Staworko, M.: A high-performance architecture of JPEG2000 encoder. In: Proceedings of the 19th European Signal Processing Conference (EUSIPCO), pp. 569–573 (2011)
31. Rivest, R.L., Shamir, A., Adleman, L.: A method for obtaining digital signatures and public-key cryptosystems. Commun. ACM **21**(2), 120–126 (1978)

32. Saidani, T., Atri, M., Said, Y., Tourki, R.: Real time FPGA acceleration for discrete wavelet transform of the 5/3 filter for JPEG 2000. In: 6th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT), pp. 393–399 (2012)
33. Schneier, B.: Applied Cryptography: Protocols, Algorithms, and Source Code in C. Wiley, New York (1996)
34. Wang, C., Ni, J., Huang, Q.: A new encryption-then-compression algorithm using the rate-distortion optimization. Signal Process.: Image Commun. **39**, 141–150 (2015)
35. Wu, H., Ma, D.: Efficient and secure encryption schemes for JPEG2000. In: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol. 5, pp. 869–872 (2004)
36. Yan, S., Lin, Q.: Partial encryption of JPEG2000 images based on EBCOT. In: International Conference on Intelligent Control and Information Processing (ICICIP), pp. 472–476 (2010)

IoThings: A Platform for Building up the Internet of Things

Andrei Gal(✉), Ioan Filip, and Florin Dragan

Department of Automation and Applied Informatics,
University Politehnica of Timisoara,
Bvd. V. Parvan, no. 2, 300223 Timisoara, Romania
andrei.g88@gmail.com,
{ioan.filip,florin.dragan}@aut.upt.ro

Abstract. Given the multitude and diversity of smart devices surrounding us, along with the variety of applications distributed on these smart devices, this paper proposes the architecture for a platform that will be used to interconnect different smart devices in terms of hardware capabilities and software features, making feasible the concept of Internet of Things (IoT). The paper defines the architecture of the platform, expanding the functionality and correlation between modules that make up the platform, and proposes a software representation of a *Thing* and also a software representation of a *Context* as used in the IoT network. Using both software representations, platform modules manage tasks that discover all available devices in a surrounding area and create proper context for devices to collaborate, schedule smart device communication and establish communication between available devices in the network. The great advantage of the proposed solution is that it provides an open platform, extensible from an architectural point of view with the possibility to integrate with new smart devices in order to enlarge the IoT network.

Keywords: Internet of things · Smart devices · Context-aware · Platform · Software architecture

1 Introduction

A mobile device can be described as being a computational device having a hardware structure similar to that of a personal computer (PC) and also a processing power compared to it. The main design feature of all mobile devices is that they are that small that a person can easily move a device from one place to another. Usually their dimensions are basically compared to that of a human hand.

The evolution of mobile devices to what we now define as being a mobile device took place in quite a small period of time. The impact of the technological development in this direction was great on society affecting it on different levels such as economics, social media, commerce, press, culture, and nevertheless having the greatest impact on the communication level in the society. Below there are defined 5 periods in the evolution of mobile devices:

V.E. Balas et al. (eds.), *Soft Computing Applications*, Advances in Intelligent Systems and Computing 633, DOI 10.1007/978-3-319-62521-8_15

- *The Brick (1973–1988)* - this is the first known era in the mobile telephony as in this period the fundamentals of future cell phone communication were being put in place along with the appearance on the market of the first mobile phone (cell phone) invented by Martin Cooper of Motorola company in 1983 [1].
- *The Candy Bar (1988–1998)* - this era represents one of the biggest leaps in the mobile technology. During this period the design of the cell phones changed drastically both because the usage of second-generation (2G) technology and the increased density of cellular sites [1].
- *The Feature Phone (1998–2008)* - Taking in consideration the technological advancement to the mobile devices, this era is considered having the lowest impact. Most of the changes occurred to the mobile devices were at a software level as software features were developed to make use of the available hardware such as music listening applications, photo taking applications, game applications. The most groundbreaking change was the introduction of the internet and the addition of GPRS modules along with the usage of the new 2.5G network [1].
- *The Smartphone (2002–present)* - This era overlapped the one before as there is no clear difference between devices developed during the previous era and this one. Though introducing the internet on a mobile phone might be the event that triggered the so called smartphone era, the real route was made when introducing a common operating system (OS) on each of the devices developed. This addition led to the possibility of developing OS based application which gave birth to the concept of smartphone (a phone that is as smart as a PC). Alongside the development on a large scale of smartphones companies started to extend the area of development creating other smart devices such as smart TV's, smart watches, smart cards, smart boards, smart pads, smart fridges, smart locks etc. [1].
- *The Touch (2009–present)* - This era overlaps the one before and is related to the introduction of the touchscreen technology on smart mobile devices but also on simple smart devices. The first touchscreen technology was implemented by Apple Inc. on their first iPhone and since then every smartphone uses it. As with smartphone technology, touchscreen technology began to be used increasingly more and applied to a number of existing smart devices, nowadays having touchscreen smart TV's, touchscreen smart watches, touchscreen fridges touchscreen laptops etc. [1].

2 An Overview of the Used Concepts

2.1 Internet of Things (IoT)

During the mobile device development evolution most of the device could communicate between them only by using the GSM network, therefore the devices were mainly used as mobile phones with no possible way to interconnect them. This was mostly related to the fact that there was a limitation of the communication environments that the mobile devices possessed.

Along with the development of smartphones with Wi-Fi and Bluetooth modules applications that use these modules to exchange data were more often developed. Most applications exchanged address book information such as contact entries, text messages, and call history but there were also applications developed to exchange files such as music or photo files. With the development of these devices and applications the need of having a live interconnection for exchanging relevant data between the existing devices has increased to a point that the concept of Internet of Things appeared.

The Internet of Things can be defined as being a network on a global scale for interconnecting smart devices (*Things*) around the world. Making a comparison with the internet as we know it where people are the agents of communication and data transferred between people is mostly referring general social information, the IoT uses *Things* as agents of communication, while the data transferred between the *Things* represent relevant data for the *Things* engaged in the process of communication. As the concept is being still in early stages there are several possible interpretations of the *Things* that make up the IoT. Considering the current development of smart devices focused mostly on the home automation services, a possible basic IoT infrastructure can be seen in Fig. 1. All smart devices in Fig. 1 play the role of *Things* in the IoT infrastructure [2, 3].

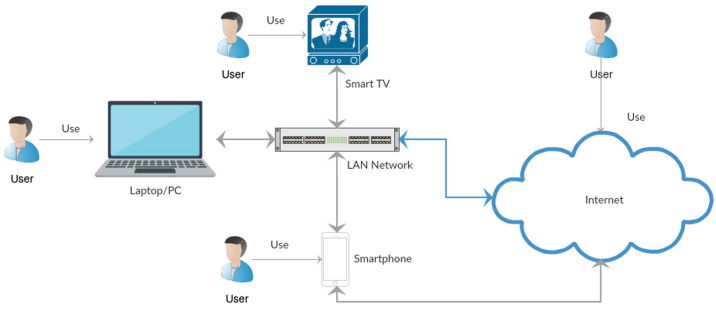


Fig. 1. Basic IoT infrastructure

In particular, the simplest use of the IoT, by referring to Fig. 1 shown above, could be that of a user being able to just power on/off the smart TV from a simple TV Manager application the user has installed on the smartphone he owns.

When referring to the term home automation a more complex infrastructure of the IoT could be represented by involving a lot more *Things* in the process of automating the home. *Things* like temperature sensors, pressure sensors, humidity sensors or light sensors embedded in equivalent smart devices can give feedback upon action that could be taken to adjust the ambient comfort in the home. The actions taken to adjust the environment could be done by *Things* like windows', air conditioners', dehumidifiers', electric heaters', louvers' all developed as smart devices that could be controlled remotely like in Fig. 2 [3].

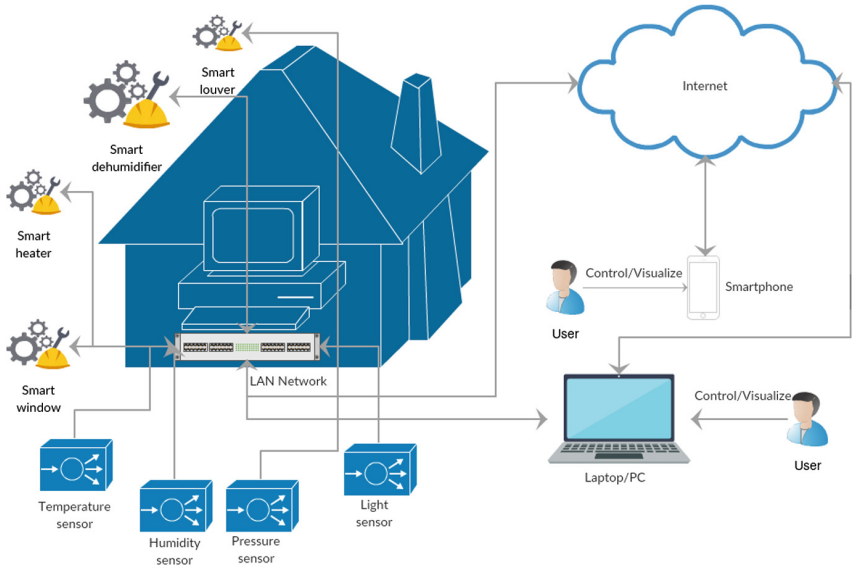


Fig. 2. Complex IoT infrastructure

By taking in consideration various points of interest, the IoT infrastructure can be extended by adding *Things* relevant in fields of study like health care, smart buildings, waste management, smart roads, social media and a lot more fields where it can be possible to build a level of abstraction so that Things can be developed to serve purposes like people's security, replacement of human actions, increased comfort level in an environment, reducing the possibility of human error, etc. In an ideal IoT world, smart devices will serve users by doing autonomous work behind the scenes, accomplishing the ultimate purpose for which such an enlarged IoT infrastructure is created, that being interconnectivity between devices and services that work together to do something useful for mankind [3, 4].

2.2 Context-Aware Computing

To present in a clear light the essential aspects of the proposed platform it can be mentioned that like many current architectures proposed for implementation of IoT infrastructure, *IoThings* was also designed taking in consideration the paradigms of *context-aware computing*, especially emphasizing on the main attribute of any context-aware system namely the *adaptation of the system to its location, to the people and objects in the system or connected to the system, and also to the changes of the objects over time* [5, 6].

The concept was developed based on the ubiquitous computing paradigm proposed in the early 1990's and it refers to a suite of mobile systems that can take information from their environment and adapt to it by taking appropriate decisions. The core

elements of these context-aware systems are represented by the context information and the actions (decisions) taken to adapt the mobile systems to the context at a certain point in time and space.

The platform presented in this paper follow the lines drawn by both the context-aware computing concept and IoT concept, proposing a modular platform architecture that encapsulated core aspects from both concepts presented above with the purpose to create a context-aware IoT network where devices can interconnect forming the Internet of Things [7, 8].

3 *IoThings* - Software Level Concepts

IoThings is a platform designed to support a series of smart devices that are different in terms of hardware functions and could run under different or same operating systems. The platform is designed, to have independent modules overlapping each of the functions of the platform, altogether coordinating the devices to adapt to the environment they reside in, for creating the proposed context. Due to the focus on the context-aware computing concept, the platform is structured around the term of *context information.* As the platform is being designed for implementing an IoT solution platform, all the smart devices that use the platform are represented as entities belonging to a certain *context*, and all relevant information about a smart device is translated in a *context entity information* [9, 10].

3.1 Software Representation of a Thing

As defined by the IoT concept, a Thing is represented in the real world by a smart device and its main characteristic is that it can interconnect with other Things for exchanging data, this way creating a large network of Things which eventually can create the Internet of Things.

Having a multitude of various Things that are being different in terms of hardware structure, the platform defines a set of hardware functionalities that a Thing can have. For example, Things can get measures from the environment, can record sounds/videos, can move, can carry other Things, can sense proximity, can start/stop engines etc. A Thing being represented as a class at a software level, each of the hardware functionalities will be defined as an interface, so for the abstraction of each function a Thing has, it will have to implement that particular interface. In Fig. 3 is presented a class diagram for the interfaces the platform supports. It should be mentioned here that the class diagram can be extended to include any other possible hardware features a Thing can own but also adapt existing features to requirements of an accepted Thing for the platform [11].

As seen in Fig. 3, a Thing represented at a software level in the *IoThings* platform contains the sum of all functions that that Thing is supporting at a hardware level. Going onwards the platform also reduces to a level of software abstraction the general information a Thing can have like: name, physical form, physical dimension, operating system installed on, operating system version, software version etc. As depicted in Fig. 4 the platform exposes a generic interface that the Thing will have to implement in order to work with the platform.

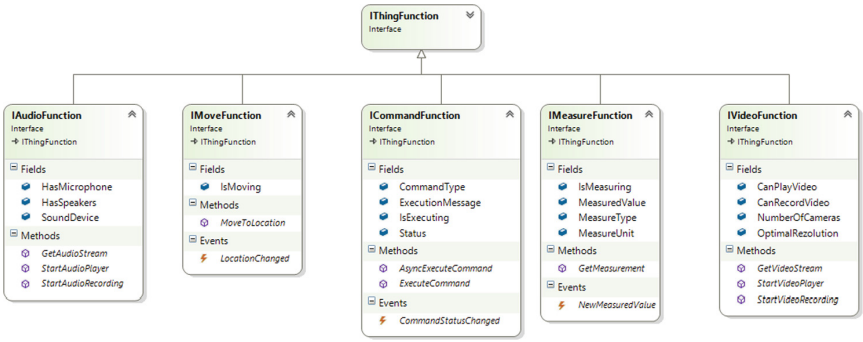


Fig. 3. Defined interfaces for the software representation of a *Thing's* functions

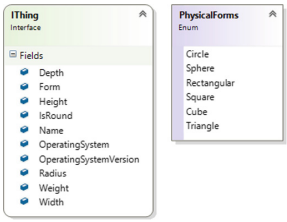


Fig. 4. Defined interfaces for the software representation of a Thing's general information

The general information interface contains nine properties as below:

- *Name* – property that defines the name of the Thing.
- *OperatingSystem (OS)* – property that defines the operating system under which the software embedded in the Thing runs. This property is relevant in terms knowing how to handle the communication between Things that have different operating systems installed.
- *OperatingSystemVersion* – property that gives the platform the information regarding the version of the OS and as like the one above it's important in communication between Things running different OS/OS versions.
- *Form* – property that represents the physical form of the Thing. As it can be seen in Fig. 4 this can have several values like: *Circle, Sphere, Rectangular, Square, Cube, Triangle.*
- *Height* – property that is set to the real height of the Thing.
- *Width* – property that is set to the real width of the Thing.
- *Depth* – property that is set to the real depth of the Thing.
- *Weight* – property that is set to the real weight of the Thing.
- *IsRound* – this property is set if the Thing represented has a round shape.
- *Radius* – this property comes with the one above and it represents the actual radius of the Thing in case it has a round shape.

The function of a Thing that can move is represented by the *IMoveFunction* interface and it contains the following:

- *IsMoving* – property that is set whenever the Thing is in motion. This is relevant in the context of having to give the Thing a command to move knowing it is already in motion.
- *MoveToLocation* – this method is used to give the Thing the command to move to a specified location in the available space.
- *LocationChanged* – this event is triggered by the Thing whenever its location has been changed. This way the platform is able to locate each of the Things at a specified moment in time.

To represent the capacity of a Thing to take measurements from the environment it resides in the *IMeasureFunction* interface propose below members:

- *IsMeasuring* – property specifying the fact that the Thing is currently taking measurement from the environment. The property is useful when using Things that work with big data values or have time consuming sensors.
- *MeasuredValue* – property that retains the measured value from the environment at a certain point in time.
- *MeasureType* – this property is set to the relevant type of measurement the Thing deals with. As seen in Fig. 5 the property can have several values like:
- *MeasureUnit* – property is set to the relevant unit of measurement used by the *MeasureType* property, and it can take values like:

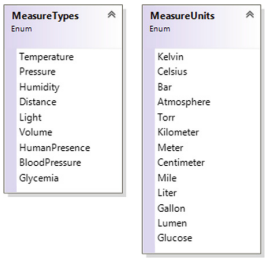


Fig. 5. Measure types and measure units used by *IMeasureFunction* interface

- *GetMeasurement* – the method is used to retrieve the last value read by the Thing from the environment. This will also update the *MeasuredValue* property.
- *NewMeasuredValue* – event is used to be triggered whenever a new value has been measure from the environment.

The ability of a Thing to execute commands that have an effect on the environment is represented by the *ICommandFunction* and it has the following composition:

- *CommandType* – property that can be set with the following values as in Fig. 6: *Start, Stop, Open, Close, Lock, Unlock, Drag, Move, Push, Pull, Cut*.

- *Status* – property that is set with the result of the command's execution, as for to know whether the command ended successfully has entered some state relevant for the context in which the Thing is. The values possible for this property can also be seen in Fig. 6.
- *IsExecuting* – this property is set whenever a command starts its execution and is reset when the command end the execution.
- *ExecutionMessage* – property will retain the message thrown by the command after finishing the execution.
- *AsyncExecuteCommand* – the method is used to start the asynchronous execution of the implemented command. As the execution of a command can be time consuming this method offers the possibility to run the execution on a separate thread in case necessary in the context.
- *ExecuteCommand* – the use of this method implies the execution of the command awaiting for it to finish, blocking the thread from where it is called.
- *CommandStatusChanged* – this event is fired up by the Thing when the status of the command in execution is changed.

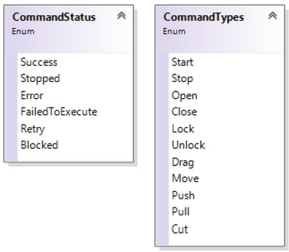


Fig. 6. Command types and command status used by *ICommandFunction* interface

For the video ability of a Thing, the *IVideoFunction* interface has to be implemented having the structure below:

- *CanPlayVideo* – property that defines whether the Thing can play or not videos.
- *CanRecordVideo* – property that defines whether the Thing can record or not videos.
- *NumberOfCameras* – property set to the number of available cameras that the Thing has to offer.
- *OptimalResolution* – property set with the optimal resolution for the video. This will be used for the Thing to record/play video by setting the resolution to the one in the property.
- *GetVideoStream* – this method is used to get the video stream from a certain camera and use it as output for any of the applications using the platform and using live streaming feature.
- *StartVideoPlayer* – this method is used to tell the Thing to start playing a video in case it has the *CanPlayVideo* property set.
- *StartVideoRecording* – this method is used to tell the Thing to start recording video in case it has the *CanRecordVideo* property set.

The structure of the *IAudioFunction*, making available the audio representation at the software level for a Thing consists of:

- *HasMicrophone* – property set when a Thing has a microphone incorporated for the possibility to record audio.
- *HasSpeakers* – property set if a Thing has speakers available to play audio sounds.
- *SoundDevice* – property specifying the sound device used by the Thing.
- *GetAudioStream* – this method is used to get the audio stream from a certain device and use it as output for any of the applications using the platform and using live audio feature or voice recording feature.
- *StartAudioPlayer* – this is used to start playing and audio file in case a Thing has *HasSpeakers* property set.
- *StartAudioRecording* – this method is used to start recording audio is a Thing has *HasMicrophone* property set.

Taking in consideration the *context information* core element of the *IoThings* platform, all Things known by the platform have to be included in a context so that the context-awareness paradigm is respected. To include all *Things* in a certain *context* the platform integrates the *Thing* concept into the *context information* concept by creating a *context entity* that encapsulates any *Thing* known by the platform. In order to do a better drawing of the *context* concept any *context entity* will consist in two major parts as seen below in Fig. 7.

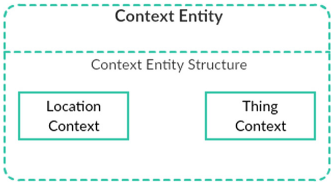


Fig. 7. Context Entity structure

As it can be seen in Fig. 7 a *Context Entity* is composed of a *Location Context* and a *Thing Context*. The *Thing Context* is represented by the software representation of a Thing which is described above, as the *Location Context* represents the location of the *Context Entity* in the context space bordered by the IoT network that is built with the use of *IoThings* platform. Particularly speaking the *Location Context* consist of relevant information regarding the position in time and space where a specific entity is. From a structural point of view, the *Location Context* consists of:

- *GPS coordinates* – if possible this is to be provided by the Thing in the Context Entity, or could be provided by entities from the same context.
- *Context reference* – a reference to the context that the entity belongs to.
- *Time reference* – a time reference based on where the context resides in terms of geo-location. It can contain data as local time, local day, local month [12].

3.2 Software Representation of a Context

To be able to have an idea of how *IoThings* platform uses the *Context* concept to create an abstraction at the software level we'll highlight some definitions of the concept.

A *Context* could be seen as environmental information that is part of an application's operating environment and that can be sensed by the application. This typically includes the location, identity, activity and state of people, groups and objects that are part of the environment.

Another possible definition of a *Context* could be that of the interaction between humans and computers in socio-technical systems that takes place referring to the physical and social situation in which computational devices and environments are embedded. The context is determined by the people involved (including their background knowledge and their intentions), the objective of the interaction (including the tasks to be carried out), and the time and place where the interactions occur [13].

Given the two definitions above, *IoThings* platform defines the *Context* as being a self-aware environment, where entities different in terms of functional abilities can interact between them or with people by giving feedback with a certain status of the environment at a point in time, or accepting commands if necessary, to adjust the environment for serving a predefined scope.

Having this in consideration the software representation of a *Context* from the *IoThings* platform point of view consists in elements like in Fig. 8:

- *Name* – property set with the name of the Context.
- *Location* – a property that defines the geographical coordinates where the Context is settled.
- *Scope* – property set with the predefined scope of the Context.
- *Status* – property giving the current status of the Context. Possible status of the Context can be determined by a full analysis of the Scope of the Context.
- *NetworkID* – property set with the ID of a certain network (if any) where the Context establishes communication between entities owned.
- *AvailableContextEntities* – a list of *Context Entity* objects that are residing in the specified Context. These entities make up the Context itself by adapting and adjusting Context variables to achieve the predefined Scope.
- *SelfAwarePercent* – property which set the percentage of self-awareness of the Context. The percentage is to be calculated depending on several factors like number of autonomous entities in the Context, number of human interactions in achieving the Scope, number of errors in executing a command in the Context etc.

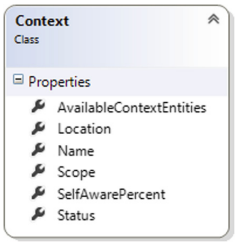


Fig. 8. Software representation of a Context

4 *IoThings* Architecture

IoThings platform being designed for working with *Things* in terms of *Context Entities*, and respecting the used concepts, its architecture is composed of modules that provide functionalities like communication with any known *Context Entity,* defining a *Context* using the available entities, managing contexts and context entities, all modules making use of the software representations defined by the platform in order to interfere with the smart devices inhabiting a real context environment. The platform's architecture, represented at the modules level is depicted in Fig. 9.

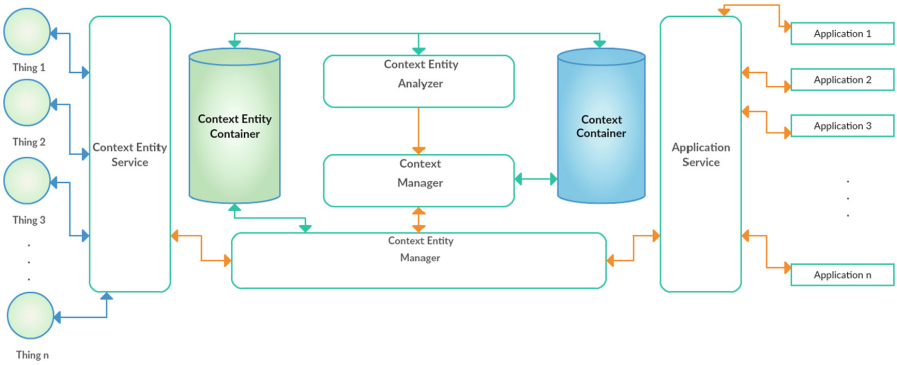


Fig. 9. *IoThings* architecture

a. *Context Entity Service* - this module represents a service exposed by the platform that gives the possibility for any *Context Entity* to communicate with the platform. Through this module a smart device that is defined as a possible *Context Entity* will be known by the platform as an available *Context Entity* and can be used in a proper *Context* based on the location and the hardware functions that the device is capable of. Also the service can be used for communication between different devices that cannot communicate due to hardware/software interoperability problems.
b. *Context Entity Analyzer* – this module deals with the analysis of all entities in terms of both location context and functions that the entity is capable of, resulting in the creation of a new *Location Context* where the analysed entity will be included or in the attachment of the analysed entity to an already existing *Location Context*.
c. *Context Entity Manager* – this module represents the management level of the platform related to the entities communicating with the platform, and it performs three major functions: extracts all possible entities that can communicate with the platform into a *Contact Entity Container*, establishes communication between the external applications available to interact with a certain context via *IoThings* and relates any defined *Context* with an external application or with any of the entities owned by the *Context.*
d. *Context Manager* – this module represents the management level of the platform related to the contexts that are defined by it. The module communicates with the *Context Entity Analyzer* and based on the result of the analysis manages the

resulting *Location Context* in order to create a new *Context* or manage existing ones in a *Context Container*. Also this module provides information for any of the external applications regarding any context created and managed by the platform.

e. *Application Service* – this module acts as a service exposed by the platform in order to allow external applications to communicate with it. The applications are used by people in certain environments to either get feedback from the Things in the environment, or to command Things that can be triggered actions on.

5 Conclusions

This paper presents a platform which is able to manage smart devices that have different hardware functions and software installed components, by putting together two abstractions at a software level for concepts defined in the *Internet of Things* and *Context-Aware Computing* fields of interests.

A *Thing* is represented by the platform as a class that implements an interface related to general information that a *Thing* has, and also interfaces related to different functions that a *Thing* possesses.

A *Context* is represented by the platform as a class that puts altogether relevant information that defines a context-aware system like: location, scope, status, context entities belonging to the context and self-awareness.

The platform architecture is designed to communicate with all *Things* that are defined respecting the proposed software representation of a *Thing*, in order to create a container with all available *Things* for the platform. The platform will extract *Things* from the container, and by subjecting them to an analysis based on the location of the *Thing* and its capabilities, will create *Contexts* that will serve a predefined scope.

The platform will manage both *Things* and *Contexts* in order to create an interconnection between the two terms by attaching to a *Context* as many *Things* as needed for serving the scope for which the *Context* was defined. The connection will be translated in real situations, where smart devices forming a real environment, will adapt to be able to maintain/achieve a certain desired state of the environment. Also the platform will tend to increase the degree the self-awareness for each of the *Contexts* defined by emphasizing the use of the Things in an autonomous mode.

The platform makes available an application level where external applications can communicate with the platform for getting feedback from a certain *Context* or for sending commands to *Things* that act as actuators in the *Context*. In particular, the application level enables the human factor the possibility to intervene by just visualizing the state of the *Context* or by acting on some *Things* in that certain *Context.*

Acknowledgments. This work was developed through the Partnerships in Priority Areas - PN II, with the support of ANCS, UEFISCDI, Project No. 36/2012, Code: PN-II-PT-PCCA-2011-3.2-1519.

References

1. Fling, B.: Mobile Design and Development. O'Reilly Media Inc., Sebastopol (2009)
2. Stankovic, J.A.: Research directions for the internet of things. IEEE Internet Things J. **1**, 3–9 (2014)
3. Jie, Y., Pei, J.Y., Jun, L., Yun, G., Wei, X.: Smart home system based on IOT technologies. In: International Conference on Computational and Information Sciences (ICCIS), pp. 1789–1791 (2013)
4. Moser, K., Harder, J., Koo, S.G.M.: Internet of things in home automation and energy efficient smart home technologies. In: IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 1260–1265 (2014)
5. Schilit, B., Adams, N., Want, R.: Context-Aware Computing Applications, pp. 85–90 (1994)
6. Yürür, O., Liu, C.H., Sheng, Z., Leung, V.C.M.: Context-awareness for mobile sensing: a survey and future directions, In: IEEE Communications Surveys & Tutorials, pp. 68–93 (2014)
7. da Costa, C.A., Yamin, A.C., Resin, C.F.R.: Toward a general software infrastructure for ubiquitous computing. IEEE Pervasive Comput. **7**, 64–73 (2008)
8. Kindberg, T., Fox, A.: System software for ubiquitous computing. IEEE Pervasive Comput. **1**, 70–81 (2002)
9. Salber, D., Dey, A.K., Abowd, G.D.: The context toolkit: aiding the development of context-enabled applications. In: SIGCHI Conference on Human Factors in Computing Systems, pp. 434–441 (1999)
10. Hristova, A., Bernardos, A.M., Casar, J.R.: Developing an ambient home care system: context toolkit-based design and implementation. In: Ninth International Conference on Parallel and Distributed Computing, Applications and Technologies, pp. 455–461 (2008)
11. Wang, W., De, S., Toenjes, R., Reetz, E., Moessner, K.: A comprehensive ontology for knowledge representation in the internet of things. In: IEEE International Conference on Trust, Security and Privacy in Computing and Communications, pp. 1793–1798 (2012)
12. Zafari, F., Papapanagiotou, I., Christidis, K.: Microlocation for internet-of-things-equipped smart buildings. IEEE Internet Things J. **3**, 96–112 (2015)
13. Gerhard, F.: Context-aware systems: the 'right' information, at the 'right' time, in the 'right' place, in the 'right' way, to the 'right' person. In: International Working Conference on Advanced Visual Interfaces, 22–25 May, pp. 287–294 (2012)

Imperialist Competition Based Clustering Algorithm to Improve the Lifetime of Wireless Sensor Network

Ali Shokouhi Rostami[1], Marzieh Badkoobe[1], Farahnaz Mohanna[1], Ali Asghar Rahmani Hosseinabadi[2(✉)], and Valentina Emilia Balas[3]

[1] Department of Communication Engineering, University of Sistan and Baluchestan, Zahedan, Iran
A. Shokouhi@b-iust.ac.ir, Mbadkoobe@gmail.com, F_Mohanna@ece.usb.ac.ir
[2] Young Researchers and Elite Club, Ayatollah Amoli Branch, Islamic Azad University, Amol, Iran
A.R.Hosseinabadi@iaubeh.ac.ir
[3] Aurel Vlaicu University, Bd. Revolutiei 77, 310130 Arad, Romania
Valentina.balas@uav.ro

Abstract. In Wireless Sensor Network (WSN) nodes have limited energy and cannot be recharged. Clustering is one of the major approaches to optimize consumption of energy and data gathering. In these networks, clustering must be special to prolong network lifetime. In WSN, clustering has heuristic nature and belongs to NP-hard problems. In complex problems, search space is too big and grows exponentially. Because it takes too much time and cost, finding a deterministic optimized solution is difficult in such a short time. In this situation population-based algorithms are beneficial in finding optimum solutions. In this paper, a clustering algorithm is investigated and a novel idea, in line with the population-based algorithm, is presented. The proposed algorithm uses Imperialist Competition Algorithm (ICA) for the clustering of nodes. The results show that this algorithm postpones the dead time of nodes and prolongs network lifetime, compared to other discussed clustering algorithms.

Keywords: Clustering · Imperialist Competition Algorithm · Lifetime · WSN

1 Introduction

Wireless sensor network (WSN) consists of many tiny sensor nodes. These nodes are typically equipped with battery, processor, memory and radio to send and receive data, which will be deployed compactly in the area for a specific purpose. Having gathered the required information from the environment, sensor nodes that are connected with each other via wireless links process them, and then deliver them to Base Station (BS) that is responsible for gathering information [1, 2].

A wireless sensor node (node) is a small electronic device which has a limited energy resource, and generally the battery cannot be changed or recharged. Therefore lifetime of a sensor node is absolutely dependent on battery life [3, 4]. In WSN, most of

V.E. Balas et al. (eds.), *Soft Computing Applications*, Advances in Intelligent Systems and Computing 633, DOI 10.1007/978-3-319-62521-8_16

the energy is consumed in data transmission, therefore approaches that reduce transmission greatly save energy and increase network lifetime. Clustering is one of the most popular approaches used in reducing energy consumption. Clustering protocols, in comparison to direct protocols, not only decrease data transmission, but also, due to load balancing between nodes, reduce energy consumption [5, 6].

Selecting Cluster Heads (*CH*) and cluster formation procedures must make the clusters balanced, reduce the exchange of messages and keep the time complexity constant which is independent of the growth of the network. This selection process is very challenging [1, 7].

Clustering protocols proposed in the literature are, in general, classified into two groups: clustering protocols in homogeneous networks and clustering protocols in heterogeneous networks. The important clustering protocols in homogeneous networks are LEACH [8], extensions of LEACH [9, 10], PEGASIS [11], extensions of PEGASIS [12], TCCA [13], EEHC [14], HEED [15] and TEEN [16]. The important clustering protocols in heterogeneous networks are SEP [17], BSIDR [18], DEEC [19] and extensions of DEEC [20].

Clustering has a heuristic nature and belongs to NP-hard problems [9]. In heuristic problems, search space is too big and grows exponentially. In such cases, finding a deterministic answer needs too much time and cost, therefore it is difficult to find deterministic optimized answers in a short time. In this situation population-based algorithms are beneficial to determine optimum answers. Genetic Algorithm (GA) [21], Particle Swarm Optimization (PSO) [22], Gravitational Emulation Local Search Algorithm (GELS) [23–25], Gravitational Search Algorithm (GSA) [26–30] and Imperialism Competition Algorithm (ICA) are the popular population-based algorithms.

These optimization protocols are applied to WSN to decrease the amount of data transmission and therefore prolong the network lifetime by finding the most suitable route in transmission and selecting optimal *CH*s [31]. These protocols are effective in many aspects like: optimal formation of clusters [32], selecting optimal nodes as cluster heads [10, 33], finding optimal number of clusters, and finding the optimal paths [34, 35].

In this paper, ICA is used for clustering. This protocol considers the conditions and limitations of WSN. In clustering topic, a protocol is needed that reduce data transmission and subsequently reduces energy consumption. This leads to an increase in network lifetime. In order to achieve these objectives, a new clustering protocol is proposed which first clusters all nodes using ICA and then in each cluster chooses the appropriate nodes as cluster heads. By using this protocol, more energy is remained in the nodes, and the lifetime of the network increases significantly.

2 Related Work

Many algorithms in the field of energy efficiency have been proposed in the past and clustering is one of the most effective methods in this field. Following are a few of them.

2.1 Low-Energy Adaptive Clustering Hierarchy (LEACH)

LEACH [8] is one of the first and most popular energy efficient hierarchical clustering algorithms for WSN that is designed for reducing energy consumption. As the role of *CH* is rotated randomly and *CH*s directly forward their data to the BS with a one-hop route, this algorithm provides energy balancing. In LEACH formation of clusters and selection cluster heads, the *CH*s are completely distributed and there is no central coordination. *CH*s are selected randomly and all nodes have the same chance to be a cluster head. LEACH provides the following key areas of energy saving:

- No overhead is wasted in deciding which node becomes cluster head as each node decides this independent of other nodes.
- CDMA allows clusters to operate independently, as each cluster is assigned a different code.
- Each node calculated the minimum transmission energy required to communicate with its cluster head and transmits with this optimized power level.

2.2 Stable Election Protocol (SEP)

SEP [17] is an extension of the LEACH algorithm. It is represented for heterogeneous networks. There are two types of node bases on initial energy and equipment: normal node and advance node. Advance nodes have more chances to be the cluster head. The other stages are similar to LEACH. This protocol improves lifetime and provides better energy balancing, compared to LEACH.

2.3 Distributed Energy Efficiency Clustering (DEEC)

DEEC [19] is used for both homogeneous and heterogeneous networks. Unlike LEACH and SEP, in this protocol cluster heads are not chosen randomly, but they are selected based on residual energy. According to this, in each round, nodes that have more energy, have a better chance of becoming the cluster head. All nodes must have global knowledge of the network to calculate the residual energy in each round. This adds a large overhead on the network.

3 Imperialist Competition Algorithm (ICA)

To solve the optimization problems, different methods have been proposed such as Genetic Algorithm (GA) [36], Particle Swarm Optimization (PSO), and the Ant Colony Optimization (ACO). Imperialism Competition Algorithm (ICA) is a new evolutionary optimization method inspired by imperialism competition [37]. Like other evolutionary algorithms, ICA starts with an initial population called country [38]. There are two types of countries, colonies and imperialists. Some of the best countries in the population will be imperialists and the rest become colonies, which together forms the empires. Imperialistic competition takes place between these empires. During this competition, weak empires collapse and powerful ones take possession of their colonies. The power of an empire which is the counterpart of fitness function in *GA*, is inversely proportional to its cost.

After competition, the imperialists are selected and colonies are divided, then colonies move toward their imperialist. After dividing all colonies among imperialists and creating the initial empires, the competitions are started. Colonies start moving toward their relevant imperialist based on assimilation policy. Empires that cannot succeed in this competition and cannot increase their power (or at least prevent a power loss), are eliminated from the competition.

This competition leads to a gradual increase in powerful empires and decrease in the weak ones. Eventually weak empires collapse and lose their power. Imperialistic competition converges to a state in which there exists only one empire and colonies have the same cost function value as the imperialist.

In an N dimensional optimization problem, a country is a $1 \times N$ array. This array is defined as below:

$$Country = [P_1, P_2 \ldots P_N]$$

The cost of a country is found by evaluating the cost function f at the variables (P_1, $P_2 \ldots P_N$), then

$$C_I = f(country_i) = \mathrm{f}(P_{i1}, P_{i2} \ldots P_{iN})$$

The algorithm starts with N initial countries called N_{pop}, and N_{imp} best of them are chosen for imperialists. N_{col} remaining countries are colonies that each belongs to an empire. Powerful empires have greater number of colonies and the weaker one have less number of colonies. To distribute the colonies among imperialists proportionally, the normalized cost of an imperialist is defined as follows:

$$C_n = max_i c_i - c_n \tag{1}$$

where c_n- is the cost of n^{th} imperialist and C_n is its normalized cost. Imperialists began to improve their colonies and colonies move towards imperialists. This motion is shown in Fig. 1. In this movementθ and x are random numbers with uniform distribution as illustrated in formula (2) and d is the distance between colony and the imperialist.

$$x - U(\circ\ ,\ \beta\ \times d)\,\theta \sim U\,(-\gamma, \gamma) \tag{2}$$

where β is a number greater than 1. If $\beta > 1$, colonies and imperialist are closing together. γ is an adjustment of deviation from the original direction. The values of β and γ are arbitrary.

The total power of each empire is determined by the sum of its power and average power of its colonies.

$$TC_n = cost(imperialist) + \xi \times mean\{cost(colonies\ of\ empires)\}$$

Where TCn is the total cost of n-th Empire and ζ is a positive number which is considered less than one.

The main steps of ICA are as follows:

- Select some random point and initialize the empires.
- Move colonies toward their relevant imperialist (Assimilation).
- Randomly change the position of some colonies (revolution), If there is a colony with the imperialist.
- Exchange the position of the colony and imperialist.
- Until reaching similar empires.
- Compute the total cost of all empires.
- Pick the weakest colony from the weakest empire and hand it over to one of the empires (Imperialist competition).
- Eliminate the powerless empires.
- Exit if stop condition are satisfied, otherwise do further assimilation and continue.

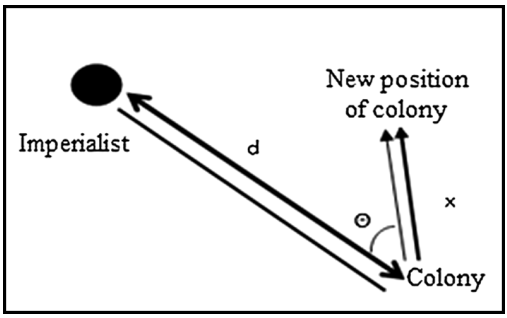


Fig. 1. Colony moving toward imperialist [38]

4 The Proposed Algorithm

In this work a WSN is considered as composed of n nodes which are the same and have two modes: Cluster Head Nodes (*CH*) and normal nodes. Since the energy consumption of *CH*s are greater than normal nodes; appropriate *CH*s are chosen in each cluster that needs less energy for transferring data to the base station and for receiving data from cluster members.

This protocol is part of a fixed clustering algorithm and done with ICA. ICA-Algorithm consists of three phases: the set up phase, the cluster head selection phase, and the steady up phase. Set up phase is executed just once in the base station, whereas other phases are repeated in each round. In *CH* selection phase, appropriated *CH*s are chosen for each cluster. In the steady up phase, nodes gather data and then each node transfers its data to the corresponding *CH*'s. The *CH*s then transfer this received data to the base station.

4.1 Setup Phase

In this phase, nodes are clustered into k_{opt} cluster using ICA. After nodes are located in their place, they send their located information to the base station. Then the BS, by using this information, calculates k_{opt}. Nodes are clustered using ICA and the clusters are formed. BS broadcasts this information to all nodes.

Location of nodes is given to the ICA as input parameters. In this case, finding the k_{opt} point as the center of cluster is the objective. In so doing, sum of square distance of each cluster members to center of cluster, plus sum of square distance of center of clusters to BS, are minimized.

4.2 *CH* Selection Phase

Unlike the pervious phase which is performed only once in the BS, this phase is repeated in each round. After clusters are formed in the BS and cluster information is broadcasted to all the nodes, each node will realize who belongs to which cluster. Each node sends a message to its neighbors which are in the range of its transmission and thereby will inform its ID and cluster number. Each node receives its neighbor's message, saves it in the table information of nodes which have similar cluster number to it. Then this table is sent to other cluster members. Using this table, every member node, can update its tables, thus all cluster members are identified.

Most energy is consumed in transmission and transmission distance has direct and exponential direct in energy consumption. Center of cluster is center of gravity, too, for this reason if *CH* located around center of gravity, distance between all cluster members to *CH* will be approximately equal, hence energy consumption in transfer of data, will be almost same for all nodes. This causes better balance energy consumption of nodes and prevents preterm death.

If *CH* is located around the center of gravity, it is optimized, however this condition is not sufficient for cluster head selection. As the numbers of nodes that are located around center, are limited, and if it was the only criterion, the energy of nodes adjacent to center is quickly finished and nodes will die soon. As a result, load balancing and energy balancing is destroyed. Hence, there are other criteria for selecting cluster heads.

*CH*s consume more energy than member nodes, therefore the remaining energy of nodes must be noted in *CH* selection. Nodes which have more residual energy have higher priority than those which have less residual energy. Accordingly, in choosing *CH*, remained energy should be considered.

Other criteria that are involved in selecting *CH*, is the distance between node and BS. The closer to the BS the CHs are placed, the lesser the energy used to transfer data. If this does not happen, the node that play role of *CH*, is remained as its role until its energy is finished end and node dies. This causes *CH*s to die sooner than members and the network balancing energy to be destroyed. The optimal status is when energy of all nodes is finished at the same time. Though, if the role of *CH* don't replace, it will not happen.

4.3 Steady up Phase

This phase starts after *CH* selection phase and repeat in each round. To avoid collision and energy efficiency, TDMA scheduling is run and a time slot is allocated for each member node so that it can send its data to *CH*. To avoid wasting energy, the rest of time, they are inactive.

CH receives data from all cluster members, then aggregate received data and send them to BS.

When all *CH*s send their aggregated data to BS, the round is finished.

Model of energy consumption is the same energy model proposed in [8] in order to achieve an acceptable. Signal-to-noise ratio (SNR) in transmitting an L bit message over a distance d, energy expanded by the radio is given by:

$$E_{TX}(L,d) = \begin{cases} L \times E_{elec} + L \times \varepsilon_{fs} \times d^2 & if\ d \le d_0 \\ L \times E_{elec} + L \times \varepsilon_{mp} \times d^4 & if\ d \ge d_0 \end{cases} \tag{3}$$

Where E_{elec} is the energy dissipated per bit to run the transmitter or the receiver circuit, ε_{fs} and ε_{mp} depend on the transmitter amplifier model we use, and d the distance between the sender and the receiver. By equating the two expressions at $d = d_0$, we have

$$d = \sqrt{\frac{\varepsilon_{fs}}{\varepsilon_{mp}}} \tag{4}$$

To receive an L bit message the radio expends

$$E_{RX}(L) = L \times E_{elec} \tag{5}$$

5 Simulation Result

For simulation of this algorithm, we have selected a rectangle area where 100 nodes are randomly distributed in the 100 × 100 area. The base station is located at the center (50, 50). The simulation is done with MATLAB. The parameters of simulation are listed in Table 1.

Table 1. Simulation parameters

Value	Parameters
0.5 J	Initial energy
10pJ/bit/m2	E_{fs}
0.0013pJ/bit/m4	E_{mp}
50 nJ/bit	E_{DA}
4000 bit	Packet length
50 nJ/bit	E_{elec}
87.7 m	D_0
3000	Round number

In this simulation we suppose a network with assumptions as follows:

- All nodes are located randomly.
- All nodes are static and have limited energy.
- At the beginning energy of all nodes are the same.
- All nodes are aware of their position and their amount of remaining energy.
- All nodes can be both Cluster head and normal node.

For clustering of nodes, the first step is the calculating the optimal number of clusters. According to [8] and simulation parameters, it will be $75\ m < d_{toBS} < 185\ m$.

Therefore it is expected that the optimal number of clusters be in the range 1–6.

Figures 2 and 3 show that by increasing the number of clusters, the average residual energy of the network is increased, death of nodes is delayed and network lifetime is

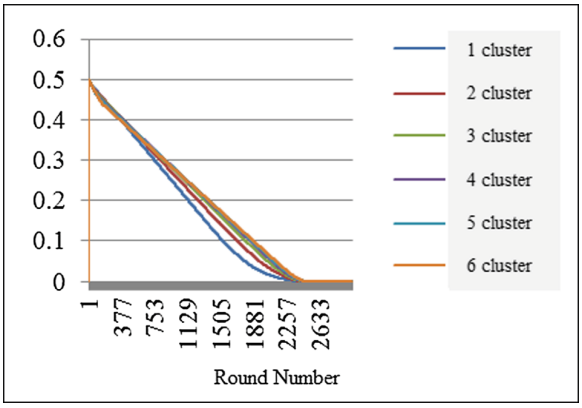


Fig. 2. Average residual energy in the various number of clusters

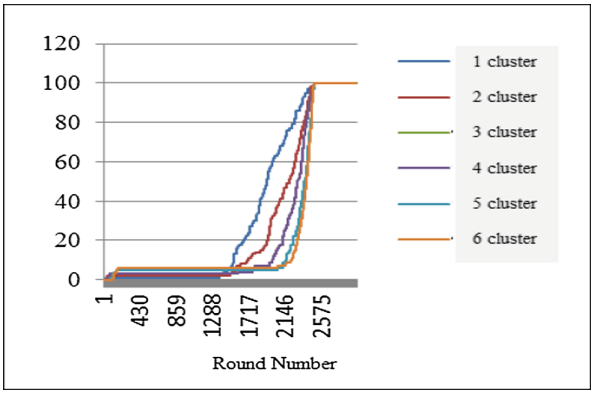


Fig. 3. Number of dead nodes in the various number of clusters

increased. In this experiment the optimal number of clusters is considered 6. ICA divides the network into 6 clusters so that the total distance of each cluster member to cluster center and the total distance from all cluster heads to the base station, is minimized. Figure 4 shows the formation of nodes in 6 clusters. Circles same in color are same cluster members and squares indicate center of each cluster (Fig. 5).

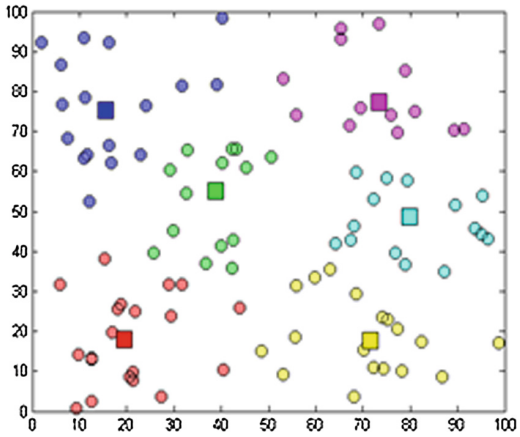


Fig. 4. Clustering nodes into 6 cluster using ICA

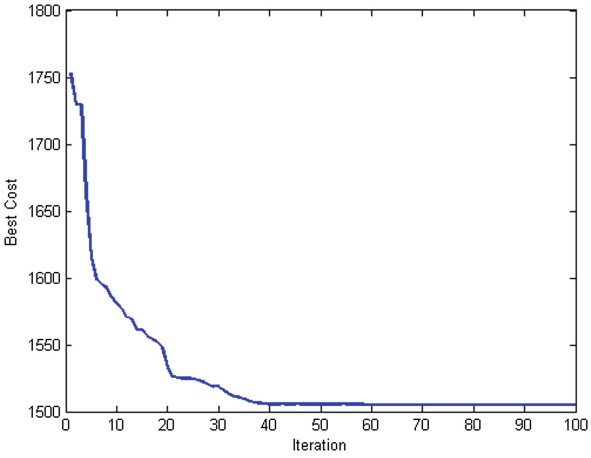


Fig. 5. ICA convergence rate

Simulation parameters for ICA are shown in Table 2.

The result of this protocol is compared to discuss clustering algorithm. The results of these comparisons are shown in Figs. 6 and 7.

Table 2. ICA simulation parameters

Value	Parameter
100	Number of countries
10	Number of imperialist
0.1	Revolution probability
100	Iteration
6	Number of cluster

In this protocol, one of the cluster head selection parameter is proximity to the cluster center. The closer the nodes are to the center, the more their chance of being a cluster head is. Therefore the nearest nodes to center are the first priority to become cluster heads.

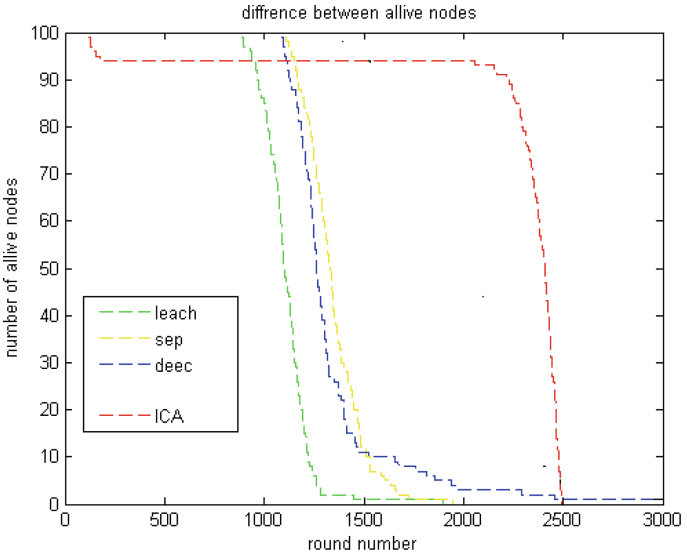


Fig. 6. Comparing numbers of alive nodes in each round

Figures 6 and 7 show in ICA, regardless of the percentage of nodes that are dead in early rounds, the remaining nodes almost die at the same time.

As shown in Fig. 8, in early rounds ICA-Clustering have lesser average remaining energy than other algorithms. The reason is that some percentage of nodes die very quickly, but in other rounds ICA-Clustering have more remaining energy in each round and therefore the lifetime is improved than the others. ICA-Clustering is compared to

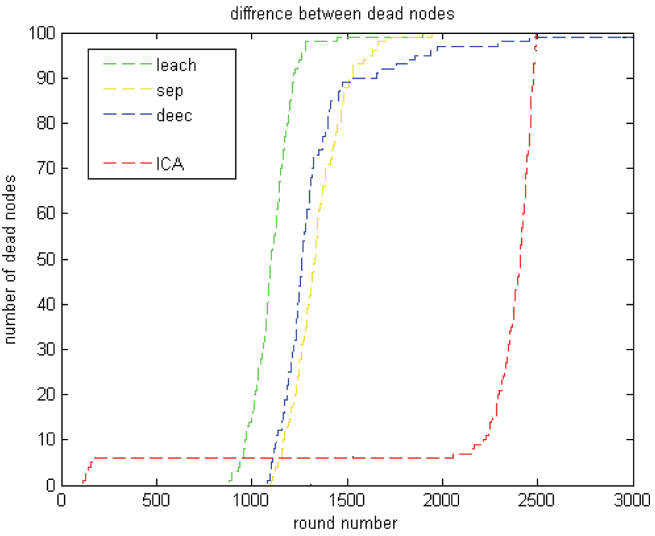


Fig. 7. Comparing number of dead nodes in each round

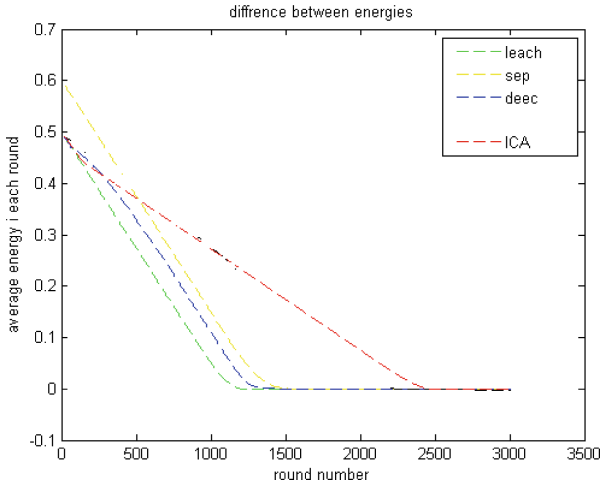


Fig. 8. Comparing number of average residual energy in each round

LEACH, SEP, DEEC and could obtain lifetime improvement about 40%, 36% and 26% respectively than them.

The lifetime of this protocol is compared to some clustering algorithm shown in Fig. 9.

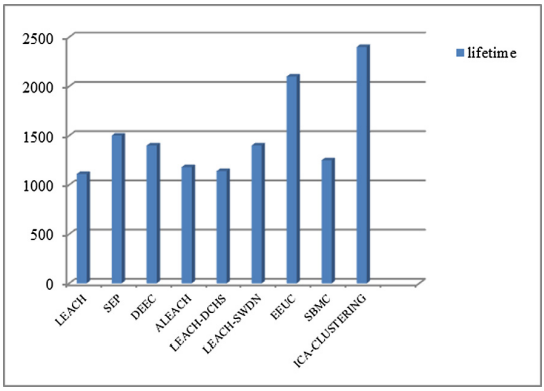


Fig. 9. Comparing lifetime of proposed algorithm and some clustering algorithm

6 Conclusion

Wireless Sensor Network consists of a large number of tiny nodes that are distributed in the environment for specific purposes. Sensor nodes have limited energy and cannot be recharged. Algorithms used in WSN must both be energy efficient and improve network lifetime. Clustering is one of the approaches that is used for energy efficiency. In this paper we introduce a new fixed-clustering algorithm named ICA-Clustering. In this algorithm nodes are clustered with ICA in base station and then in those cluster nodes which are closer to the center of the cluster and have more residual energy. This improves its chance of being a cluster head. This algorithm, compared with other proposed clustering algorithms, leads to more energy efficiency and improves network lifetime.

References

1. Rabeay, J.M., Ammer, M.J., da Silva, J.L., Patel, D., Roundry, S.: PicoRadio supports ad hoc ultra-low power wireless networking. IEEE Comput. Mag. **33**, 42–48 (2000)
2. Elahi, A., Hosseinabadi, A.R., Rostami, A.S.: Multi-hop fuzzy routing for wireless sensor network with mobile sink. Int. J. Sci. Eng. Res. **4**(7), 2431–2439 (2013)
3. Kumarawadu, P., Dechene, D.J., Luccini, M., Sauer, A.: Algorithms for node clustering in wireless sensor networks: a survey, pp. 295–300, December 2008
4. Tavakkolai, H., Yadollahi, N., Yadollahi, M., Hosseinabadi, A.R., Rezaei, P., Kardgar, M.: Sensor selection wireless multimedia sensor network using gravitational search algorithm. Indian J. Sci. Technol. **8**(14), 1–6 (2015)
5. Karenos, K., Kalogeraki, V., Krishnamurthy, S.: Cluster-based congestion control for sensor networks. ACM Trans. Sensor Netw. **4**, 1–39 (2008)
6. Rostami, A.S., Bernety, H.M., Hosseinabadi, A.R.: A novel and optimized algorithm to select monitoring sensors by GSA. In: International Conference on Control, Instrumentation and Automation (ICCIA), pp. 829–834 (2011)

7. Elahi, A., Hosseinabadi, A.R., Rostami, A.S.: Improving news document clustering based on a hybrid similarity measurement. In: International Conference on Intelligent Computing and Intelligent Systems (ICIS), pp. 1–6 (2011)
8. Heinzelman, W.R., Chandrakasan, A., Balakrishnan, H.: An application-specific protocol architecture for wireless microsensor networks. In: IEEE Tmns. Wireless Commun. pp. 660–670, October 2002
9. Heinzelman, W.R., Chandrakasan, A., Balakrishnan, H.: An application-specific protocol architecture for wireless microsensor networks. In: IEEE Transactions on Wireless Communications, pp. 660–670, October 2002
10. Wang, A., Yang, D., Sun, D.: Clustering algorithm based on energy information and cluster heads expectation for wireless sensor networks. Comput. Electr. Eng. **38**(3), 662–671 (2012)
11. Lindsey, S., Raghavendra, C.S.: PEGASIS: Powere-fficient Gathering in Sensor Information System. In: Proceedings IEEE Aerospace Conference, pp. 1125–1130, March 2002
12. Yueyang, L., Hong, J., Guangxin, Y.: An energy-efficient PEGASIS-based enhanced algorithm in wireless sensor networks, China Commun. Technol. Forum (2006)
13. Selvakennedy, S., Sinnappan, S.: An adaptive data dissemination strategy for wireless sensor networks, Int. J. Distrib. Sens. Netw., 3(1), 23–40 (2007)
14. Bandopadhya, S., Coyle, E.: An energy efficient hierarchical clustering algorithm for wireless sensor networks. In: Proceeding of IEEE INFOCOM, vol. 3, pp. 1713–1723, April 2003
15. Fahmy, S., Younis, O.: HEED: a hybrid energy-efficient distributed clustering approach for ad hoc sensor networks. IEEE Trans. Mobile Comput. **3**(4), 366–379 (2004)
16. Manjeshwar, A., Agrawal, D.P.: TEEN: a protocol for enhanced efficiency in wireless sensor networks. In: The Proceedings of the 1st International Workshop on Parallel and Distributed Computing Issues in Wireless Networks and Mobile Computing, April 2001
17. Smaragdakis, G., Matta, I., Bestavros, A.: SEP: a Stable Election Protocol for clustered heterogeneous wireless sensor networks. In: Proceedings of the International Workshop on SANPA, pp. 251–261 (2004)
18. Varma, S., Nigam, N.: U.S. Tiwary, Base Station Heterogeneous Wireless Sensor Network using clustering, pp. 1–6 (2008)
19. Qing, L., Zhu, Q., Wang, M.: Design of a distributed energy-efficient clustering algorithm for heterogeneous wireless sensor networks. Comput. Commun. **29**(12), 2230–2237 (2006)
20. Liu, Z., Zheng, Q., Xue, L., Guan, X.: A distributed energy-efficient clustering algorithm with improved coverage in wireless sensor networks. Future Gener. Comput. Syst. **28**(05), 780–790 (2012)
21. Haupt, R.L., Haupt, S.E.: Practical Genetic Algorithms, 2nd edn. Wiley, Hoboken (2004)
22. Kennedy, J., Eberhart, R.: Particle swarm optimization. Proc. IEEE Int. **4**, 1942–1948 (1995)
23. Hosseinabadi, A.R., Siar, H., Shamshirband, S., Shojafar, M., Nizam, M.H., Nasir, M.: Using the gravitational emulation local search algorithm to solve the multi-objective flexible dynamic job shop scheduling problem in Small and Medium Enterprises. Ann. Oper. Res. **229**(1), 451–474 (2015). Springer
24. Rostami, A.S., Mohanna, F., Keshavarz, H., Hosseinabadi, A.R.: Solving multiple traveling salesman problem using the gravitational emulation local search algorithm. Appl. Math. Inf. Sci. **9**(2), 699–709 (2015)
25. Hosseinabadi, A.R., Kardgar, M., Shojafar, M., Shamshirband, S., Abraham, A.: GELS-GA: hybrid metaheuristic algorithm for solving multiple travelling salesman problem. In: International Conference on Intelligent Systems Design and Applications (ISDA), pp. 76–81 (2014)

26. Hosseinabadi, A.R., Yazdanpanah, M., Rostami, A.S.: A new search algorithm for solving symmetric traveling salesman problem based on gravity. World Appl. Sci. J. **16**(10), 1387–1392 (2012)
27. Hosseinabadi, A.R., Farahabadi, A.B., Rostami, M.S., Lateran, A.F.: Presentation of a new and beneficial method through problem solving timing of open shop by random algorithm gravitational emulation local search. Int. J. Comput. Sci. **10**(1), 745–752 (2013)
28. Hosseinabadi, A.R., Ghaleh, M.R., Hashemi, S.E.: Application of modified gravitational search algorithm to solve the problem of teaching hidden Markov model. Int. J. Comput. Sci. **10**(3), 1–8 (2013)
29. H. Tavakkolai, A. R. Hosseinabadi, M. Yadollahi, T. Mohammadpour, "Using Gravitational Search Algorithm for in Advance Reservation of Resources in Solving the Scheduling Problem of Works in Workflow Workshop Environment", Indian Journal of Science and Technology, Vol. 8(11), 1–16, June 2015
30. Hosseinabadi, A.R., Kardgar, M., Shojafar, M., Shamshirband, S., Abraham, A.: Gravitational search algorithm to solve open vehicle routing problem. In: 6th International Conference on Innovations in Bio-Inspired Computing and Applications (IBICA 2015), Chapter Advances in Intelligent Systems and Computing, Kochi, India, pp. 93–103. Springer (2016)
31. Shijun, H., Yanyan, D., Zhou, R., Zhao, S.: A clustering routing for energy balance of WSN based on genetic algorithm. In: International Conference on Future Computer Support Education, IERI Procedia, vol. 2, pp. 788–793 (2012)
32. Shahvandi, L.K., Teshnehlab, M., Haroonabadi, A.: A novel clustering in wireless sensor networks used by imperialist competitive algorithm. Int. J. Adv. Eng. Sci. Technol. **8**(2), 276–280 (2011)
33. Bayraklı, S., Zafer Erdogan, S.: Genetic algorithm based energy efficient clusters (GABEEC) in wireless sensor networks. In: The 3rd International Conference on Ambient Systems, Networks and Technologies, vol. 10, pp. 247–254 (2012)
34. MurtalaZungeru, A., MinnAng, L., PhooiSeng, K.: Classical and swarm intelligence based routing protocols for wireless sensor networks: a survey and comparison. J. Netw. Comput. Appl. **35**, 1508–1536 (2012)
35. Bore Gowda, S.B., Puttamadappa, C., Mruthyunjaya, H.S., Babu, N.V.: Sector based multi-hop clustering protocol for wireless sensor networks. Int. J. Comput. Appl. **43**(13), 33–38 (2012)
36. Shojafar, M., Kardgar, M., Hosseinabadi, A.R., Shamshirband, S., Abraham, A.: TETS: a genetic-based scheduler in cloud computing to decrease energy and makespan. In: 15th International Conference on Hybrid Intelligent Systems (HIS 2015), Chapter Advances in Intelligent Systems and Computing 420, Seoul, South Korea, vol. 420, pp. 103–115. Springer (2016)
37. Shamshirband, S., Shojafar, M., Hosseinabadi, A.R., Abraham, A.: OVRP_ICA: an imperialist-based optimization algorithm for the open vehicle routing problem. In: International Conference on Hybrid Artificial Intelligence Systems (HAIS), vol. 9121, pp. 221–233. Springer, LNCS (2015)
38. Atashpaz-Gargari, E., Lucas, C.: Imperialist competitive algorithm: an algorithm for optimization inspired by imperialistic competition. IEEE Congr. Evol. Comput. **7**, 4661–4666 (2007)

Wireless Sensor Networks Relay Node Deployment for Oil Tanks Monitoring

Ola E. Elnaggar, Rabie A. Ramadan(✉), and Magda B. Fayek

Cairo University, Giza, Egypt
ola_alnaggar@yahoo.com, rabie@rabieramadan.org, magdafayek@gmail.com

Abstract. Oil tanks monitoring is an important problem in the field of oil production. Tanks might be distributed in large areas and dedicated people might be assigned for each tank to monitor it in terms of its temperature, the surrounding area, and even the level of oil inside. Wireless Sensor Networks (WSNs) could be one of the best solutions to this problem. The problem of these tanks is that they are far from each other and they might have obstacles among them that block the sensors' signals. At the same time, WSN consists of many low-cost nodes with limited power and communication. In addition, these sensors are either randomly deployed in the monitored field or deterministically installed in specific places. In both cases, gaps may appear in the network leaving uncovered areas/tanks. In this paper it is improved the connectivity of the network used for oil tanks monitoring by adding minimum number of relay nodes to it. To solve the problem of oil tanks monitoring by using WSN are used the following algorithms: the Divided Network Area Algorithm (DNAA) by dividing the network area to small squares and connecting theses small areas; the Best Path Algorithm (BPA) by making the network connected when selecting the best path for all sensors and guarantee that all nodes are connected; and the Adjustable Communication Range with Best Path Algorithm (ACR-BPA) working with both the communication ranges and the network paths. With different problem settings, the proposed algorithms are examined and compared to a state-of-art greedy algorithm.

1 Introduction

Wireless Sensor Networks (WSNs) are taking huge attention due to its wide range of application in the last few years. Some of these applications are air pollution monitoring, health care monitoring, forest fire detection, Oil & Gas Remote Monitoring applications for example pipeline monitoring [9] and tank level monitoring [10], etc. WSN consists of large number of wireless sensor nodes, which are deployed in particular area to measure certain phenomenon such as temperature, sound, pressure, etc. These sensors send their measured data to a central processing unit, which collects all the data and develops the decision. Each sensor has its sensing rang, communication rang and limited power.

Sensing rank is the monitored area around the sensor. Communication rank is the minimum distance between two sensors to send the data between them.

V.E. Balas et al. (eds.), *Soft Computing Applications*, Advances in Intelligent Systems and Computing 633, DOI 10.1007/978-3-319-62521-8_17

Due to the limitations of the wireless sensors prolonging their lifetime becomes the important attribute in WSN. Another problem in WSN is the gaps between sensors due to the random deployment of the sensors. Random deployment of sensor nodes most of the time makes WSN disconnected. To solve this problem we need to add some relay nodes to the network to make it connected. In order to achieve better communication and maximize the network lifetime, relay nodes are deployed in the network using an efficient deployment algorithm.

One of the most important applications of WSNs is monitoring oil tanks. Oil tanks are containers available in many shapes: vertical and horizontal cylinder; open top, close top and flat bottom. Sensors could be deployed in/on the tanks to measure certain features such as tank's temperature, oil level, and surrounding environment. However, with large number of tanks, the connectivity of these sensors might be a problem. To solve this problem we need to deploy a minimum number of relay nodes on the disconnected areas in the network to make the network connected. At the same time, WSN lifetime is another issue that has to be considered during the relay deployment which makes the problem more complex.

The paper presents three algorithms to solve the monitoring oil tanks problem. The first algorithm, entitled Divided Network Area Algorithm (DNAA), in which it divides the network area into grid of cells, could be in a form of squares. Relay nodes will be added to the center of these cells for their purpose is to enhance sensors connectivity. The second algorithm, entitled Best Path Algorithm (BPA), exploits the concept of best path to the centralized node (sink node/base station). It computes the best path for each sensor node and tries to add the relay nodes on these paths. The last algorithm is named Adjustable Communication range with Best Path Algorithm (ACR-BPA). This algorithm is a modification to BPA with the use of one of the sensors features which is the adjustable communications ranges.

The remainder of the paper is organized as follows. Section 2 states some of the most related work to the work done in this paper. Section 3 presents the problem definition. The algorithms involved to solve the oil monitoring problem are shown in Sect. 4. Section 5 presents the simulation and the comparison of the algorithms. Conclusion of the paper and the future work is presented in the final section.

2 Related Work

The scientific literature includes papers that consider the connectivity and the energy consumption in WSNs [1–6]. In [1], the authors considered deploying of few additional nodes possible to reconnect a disconnected network. In [2,7], the main goal is to fully enhance the connectivity of the wireless sensor networks by adding an available set of relay nodes to the network after going through some stages called levels.

In the first level, the authors divided the network area into certain numbers of equal regions and represented each region by a relay in its center. In the

second level, the best K relays' locations are chosen by solving a semi-definite programming (SDP) optimization. Through the third level, it iteratively refines the solution by dividing each obtained relay's region into a number of smaller regions and repeating the same procedure. In [3], a relay sensor placement algorithm to maintain the connectivity is proposed. They formulated this problem into a network optimization problem, named Steiner Minimum Tree with Minimum Number of Steiner Point (SMT-MSP). This study restricts the transmission power of each sensor to smallest value, then add relay nodes to guarantee the connectivity.

In [4], three heuristic algorithms are proposed for achieving connectivity of a randomly deployment ad hoc wireless networks. This work connects the network with minimum number of relay nodes with maximum utility from a given number of the additional nodes for the disconnected network. Another work proposed in [6] where the authors studied single-tiered constrained relay node placement problems, under both the connectivity requirement and the survivability requirement.

In this paper, we study relay node deployment problem in WSN. Oil tanks monitoring is considered as a realistic case study application. In oil tank problem we need to make all sensors on each tank connected to the base station by adding minimum number of relay nodes in the optimal places to prolong the network lifetime.

3 Oil Tanks Monitoring Problem

The Oil & Gas industry is one of the most prevalent industries for the application of Wireless Sensor Technology. WSN solutions for the oil, gas and power generation reducing maintenance costs, lowering the energy consumption, minimizing downtime, improving equipment performance, enhancing safety and centralize controls. Wireless technology has the potential to be beneficial in many regards. Eliminating the need for cables can contribute to reduced installation and operating costs; it enables installations in remote areas and it allows for cost efficient. Examples on Oil & Gas Remote Monitoring applications are Pipeline Integrity Monitoring and Tank Level Monitoring.

In this paper, we focus on Tank Level Monitoring problem using WSN. Oil tanks are containers holding oil with open top, close top and *Oil Pumping Units (OPU).* Most of the Oil Pumping Units (OPU) are manually monitored (Fig. 1).

This oil-pumping system use a high power-consuming process and is incapable of OPU's structural health monitoring. In Tank Level Monitoring with WSN, the condition of the oil storage tanks can be monitored using sensors: level sensor, temperature sensor and gas sensor. These sensors are fixed inside the oil storage tanks. The sensor output is given to the sink node. Based on the condition of the oil storage tanks and the oil pumping motor is controlled.

The motivations of using WSN in Tank Level Monitoring are: the special nature of oil exploration (the majority of oil pumping units are spread over barren hills, mountains and deserts) and the existing oil-pumping systems with manual control.

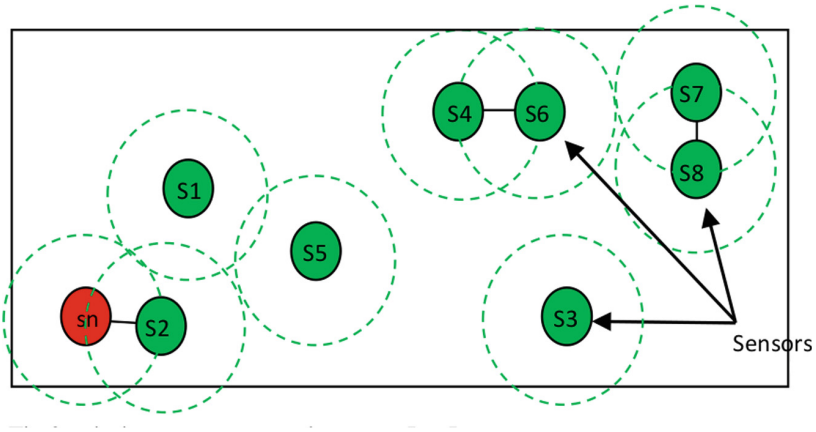


Fig. 1. Wireless sensor network on area $L \times L$.

Manual control systems have three disadvantages: (1) The OPU administrator must go to the oil field frequently to see the OPU status regardless the weather and the place. (2) Power consumption for OPU is huge during the oil-pumping process. (3) It is hard to fix the monitoring problem because it depends on people. For these reasons, WSN is used to monitor oil tank levels.

Let consider a wireless sensor network in a two dimension space. All sensor nodes (terminal nodes) need to send all information about oil level on a certain tank to a sink node. The connectivity of these sensors depends on the position of the sensors and their communication ranges.

As can be seen in Fig. 2, as an example, all sensors need to send all data to the Sink Node (SN). Sensor S2 can send to SN without a problem because S2 is directly connected to SN but for sensor S1, some relay nodes need to be added for the purpose of connectivity. Therefore, the work in this paper is step towards solving this problem using minimum number of relay nodes and saving the overall network lifetime.

4 Algorithms for Solving the Oil Tanks Monitoring Problem

In this section, three algorithms are proposed to solve the oil tanks monitoring problem. The algorithms try to add some relay nodes to the WSN to make it connected with minimum number of relay nodes and try to prolong the network lifetime as well. The algorithms are: Divided Network Area Algorithm (DNAA), Best Path Algorithm (BPA) and Best Path Algorithm with Ant Colony Optimization (BPA-ACO). The following subsection introduces the Ant Colony Optimization algorithm that will be further used in the BPA-ACO algorithm.

4.1 Ant Colony Optimization

Ant Colony Optimization (ACO) is a class of algorithms. The first algorithm, Ant System, was initially proposed by Colorni, Dorigo and Maniezzo [11–13]. The base of ACO is to simulate the real behavior of ants in nature. The ant colony provides indirect communication with the help of ant's pheromone. Pheromones are chemical substances which attract other ants searching for food.

The attractiveness of a given path depends on the quantity of pheromones detected by an ant. The quantity of pheromone is governed by some rules and depends on the attractiveness of the route. The use of attractive route ensures that the ant exudes more pheromones on its way back and so that the path is much attractive for other ants. The evaporation of pheromones is time dependent. When the way is no longer used, pheromones are evaporated and the ants begin to use other paths. ACO steps are as follows.

First Step: For first sensor, *m*artificial ant start with random number from 1 to maximum number of communication levels (in this paper there are 6 levels).

Second Step: Each ant builds a solution by adding one communication level after the other until it reaches the last sensor. The selection of the next communication range depends on certain probability. The probability p_k^i of transition of a virtual ant from the node i to the node k is given by formula 1; τ_i - indicates the attractiveness of transition in the past, η_i - adds to transition attractiveness for ants, ni- set of nodes connected to point i, without the last visited point before i, β, α- system dependent parameters.

$$p_k^i = \frac{(\eta_i^\alpha + \tau_i^\beta)}{\sum({\eta_{ni}}^\alpha + \tau_{ni}^\beta)} \tag{1}$$

Third Step: Pheromone update, virtual ant is using the same reverse path as the path to the food source based on its internal memory, but in opposite order and without cycles. After elimination of the cycles, the ant puts the pheromone on the edges of reverse path according to formula (2); $\tau_{ij}(t)$ is the value of pheromone in step t, $\Delta\tau$ is the value by ants saved pheromones in step t. Values $\Delta\tau$ can be constant or they can be changed depends on solution quality.

$$\tau_{ij}(t+1) = \rho\tau_{ij}(t) + \Delta\tau(t) \tag{2}$$

Fourth Step: At last, the pheromones on the edges are evaporated. The evaporation helps to find the shortest path and provides that no other path will be assessed as the shortest as given in Eq. (3); ρ is a user-defined parameter called *evaporation coefficient.*

$$\tau_{ij}(t+1) = (1-\rho)\tau_{ij}(t) \tag{3}$$

In our problem, a group of artificial ants searches solutions by starting each ant with random communication level for first sensor and completing the path until last one finishes depending on certain probability. This operation is inspired from the Theorem 1 given in [8]:

Theorem 1. To maximize the lifetime of the WSN, for the sensor nodes Ni and Nj, the power level assignment should satisfy $x_i \geq x_j$ *for* $i < j$.

The high communication range is given high probability when the sensor is far away from the sink node and low probability is given to the nearest sensor to the sink node. When all ants reach to the end of the path, the pheromone gets updated depending on the path length and the consumed power. Therefore, the path covering the pipe with law power consumption has high pheromone value. ACO will stop after reaches the maximum number of iterations and returns the best path.

4.2 Greedy Algorithm

In [8] is solved the oil tanks monitoring problem using the Greedy algorithm. The greedy algorithm starts with certain number of sensor nodes are deployed with certain position (the position of sensor nodes are on the top of oil tanks). The greedy algorithm includes the following three phases:

Initial Graph Phase: Find the initial graph G(N_t, E(L_t, R_t)) where N_t is the terminal nodes with location L_t and communication range R_t. When the distance between two sensors is less than the communication range, the two sensors are connecting. In this phase, no relay nodes are added.

Constructing Delaunay Phase: Construct Delaunay by using the terminal nodes. The construction of the Delaunay is illustrated as follows.

Let S be the set of the points in the two dimension space. The Voronoi diagram of S, denoted as Vol(S) which is decomposed into Voronoi cells {Va: a ∈ S} defined as: $Va = \{x \in R2 : |x - a| \leq |x - b|, \forall b \in S\}$.

The dual of the Voronoi diagram is the Delaunay triangulation $Del(S)$. $Del(S)$ is geometrically realized as a triangulation of the convex hull of S. After constructing Delaunay, the algorithm calculates the length of the three edges of each triangle. If the length of the edge is not larger than the transmission range Rt, then connect it.

Triangle Type Phase: After Phase 1, the algorithm divides the Delaunay triangles into three types. In Type 1, the length of all edges of the triangle is larger than Rt and smaller than $2Rt$.

In Type 2, the longest edge of the triangle is at most $4Rt$, while the shortest edge is larger than Rt and at most $2Rt$. The properties of triangles different from Types 1 and 2 are defined as Type 3. For the triangles of Type 1.

The algorithm places one relay node to connect five nodes that are formed by three adjacent triangles. Second, it places one relay node to connect four nodes that are formed by two adjacent triangles. Third, it adds one relay node to connect three nodes of one triangle. For the triangles of Type 2, it tries to place two relay nodes to connect three nodes of one triangle. For the triangles of Type 3, it adds relay nodes to connect the nearest disconnected nodes pair along the edge of the triangle. See [8] for other details on the greedy algorithm.

4.3 Divided Network Area Algorithm

The Divided Network Area Algorithm (DNAA) starts with a certain number of sensor nodes (N) deployed on the top of the oil tanks. All sensor nodes start with the same communication level R_t and the same power level. Sensor nodes assume to send its own data about the oil level, temperature, pressure, and any other features from the oil tank to the ink node (SN).

DNAA includes the following phases:

Dividing Area Phase: Divide the network area to squares with edges equal to the communication rang R_t. On each square which include at least one sensor, add one relay node at the center of it. DNAA didn't add any relay nodes in squares without sensor nodes. Each two sensors with distance less than R_t are connected. In Fig. 2 are ten sensors. Assume sensor number 1, $s1$, is the sink node. All sensors need to send the data to $s1$. After DNAA had divided the area to squares, it adds relay nodes at the center of each square that has at least one sensor in it, see $s12, s13, s15$, etc. then, it draws a line (line with dark blue) between the two sensors that have distance d between them where $d \leq R_t$ ($s12$ and $s3$).

Connecting Sensor Nodes Phase: Make the disconnected node connected with the nearest node with adding relay nodes like sensor number 6 in Fig. 3 or without relay node if the distance between the two sensors is less than R_t. For instance, from Fig. 2, to connect $s1$ with $s2$ DNAA adds a relay node $s18$ and the same for $s6$ and $s15$, it adds $s19$.

Relay Nodes Connected Phase: Make the main relay nodes (relay nodes added in the center of the square) connected by adding other relay node(s) on the shortest path between the two of them as shown in Fig. 3.

All of the sub-graphs need to be connected (Fig. 3). DNAA connects sub-graphs: $s1, s2, s11, s18$ with sub-graphs: $s13, s14, s4, s5$ and some relay nodes $s20$:s23. Repeat this phase until all of the sub-graphs are connected.

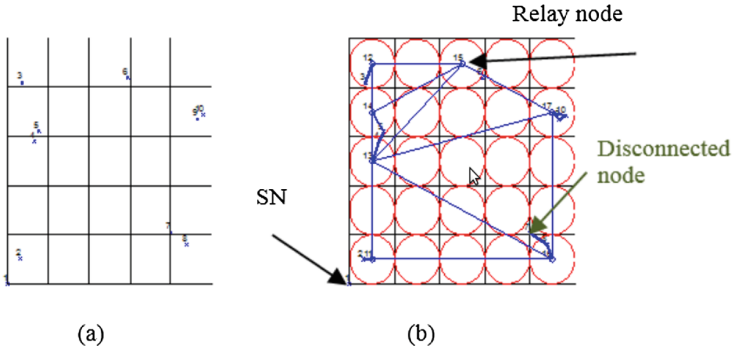


Fig. 2. (a) Before deploying the relay nodes, (b) after adding relay nodes at the center of the square.

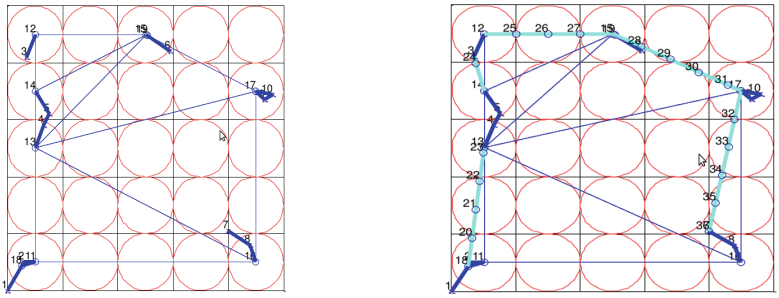


Fig. 3. Making the disconnected nodes connected (left) and the disconnected relay nodes connected (right).

4.4 Best Path Algorithm

BPA starts with the same conditions for greedy algorithm and DNAA. It starts with a certain number of sensor nodes (N) are deployed on the top of the oil tanks. All sensor nodes are assumed to have the same communication rang R_t and the same power level. Sink node collect all data from all sensor nodes in the WSN. The BPA consist of the following phases.

Connection Without Relay Node Phase: In this phase, each two sensors with distance less than R_t are assumed connected.

Connection With Relay Node Phase: In this phase BPA adds relay nodes to make the network connected. BPA first computes all paths from each sensor to the sink node. BPA calls the path as the best path if the number of relay nodes added to make the path connected was minimum number than others. To do this, BPA gives each link in the path a weight. The weight of a disconnected link is the needed number of relay nodes to make it connected. Every connected link has a weight equal to zero. The path weight is equal the sum of all links' weight, see Fig. 4.

As in Fig. 4, all sensors need to send its data the sink node ($s1$ is the sink node in this example). The links with green color mean that the two sensor are connected (the distance between them is less than R_t) with weight $w = 0$. Links with red color means that the distance between the two sensors is greater than R_t; in this case, the algorithm must add some relay nodes to make the link connected. Links with $w = 3$ means that they need three relay nodes to the link connected.

For instance, sensor 4 can send its data to the sink node through many paths. For example, Path 1 goes through nodes $4, 9, 2$, and 1 with $weight = weightlink$ (4:9) + $weightlink$ (9:2) + $weightlink$ (2:1) which is equal to $1 + 0 + 0 = 1$. Path 2 goes through nodes $4, 3$, and 1 with weight equals to 0+3 =3. Sensor 4 will send its data through path 1 because it has the minimum weight. In the next step $link$ (4:9) will be with $w = 0$ because it will be connected.

The algorithm will repeat step two until all sensors will be connected.

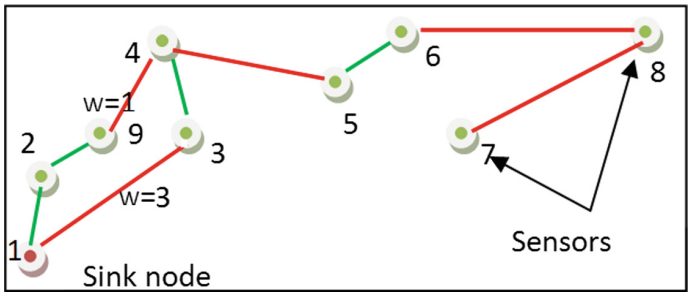


Fig. 4. Path weight in BPA. (Color figure online)

4.5 Adjustable Communication Rang with Best Path Algorithm

ACR-BPA assumes sensors with adjustable communication range $R1, R2 \ldots R6$ as in [8] instead of using one communication range. The algorithm starts with a certain number of sensor nodes deployed on the top of oil tanks with certain positions. All sensor nodes are assumed to have the same communication power and adjustable communication range.

In the *first step*, the algorithm selects the best path for all sensors while the *second step* consists of two parts:

Part A: Choose the communication range (RC) of the sensors depending on the distance between them. For example, in Fig. 4, if the distance d between two sensors is less than or equal to $R6$ (the maximum communication range), ACR-BP doesn't add any relay node. When: $R5 < d \leq R6$, RC for the two sensors is R6; $R4 < d \leq R5, RC$ for the two sensors is $R5 \ldots d \leq R1, RC$ for the two sensors is $R1$.

Part B: Add the relay nodes to the link when $d > Rmax$. To deploy relay nodes in this step ACR-BP uses Ant Colony Optimization (ACO) deployment algorithm.

5 Numerical Experiments and Discussions

In this section, the results for using WSN in Level Tank Monitoring problem will be illustrated. The next subsections show the effect of using the proposed algorithms on the network lifetime in case of different network settings.

Experiment 1: Comparison among the four algorithms in terms the required numbers of relay nodes.

In this experiment, there are certain number of oil tanks are established randomly on square area with length L, $L = 100$ m. One sensor node is deployed on the top of each tank. The experiment studies the effect of using the four algorithms to make the wireless sensor nodes connected.

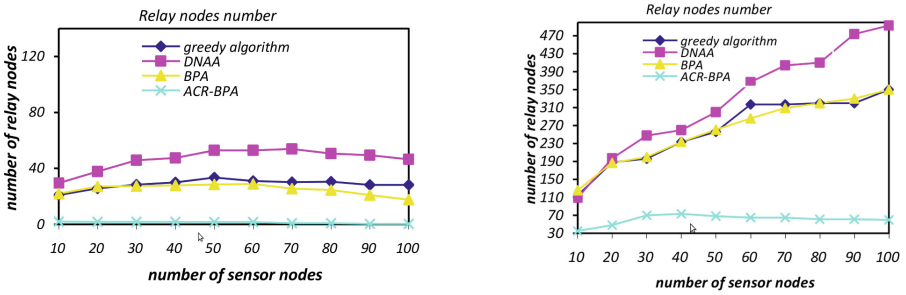


Fig. 5. The number of relay nodes added in the four algorithms in square area with L = 100 m (left) and L = 500 m (right).

The wireless sensor nodes, n, (where n is the number of oil tanks) changes from $n = 10$ to $n = 100$ sensors. All sensors as assumed to have the same power level. Sensors communication ranges as assumed the same for all algorithms except for ACR-BPA in which it utilizes different communication ranges in its operation.

From the results in Fig. 5, DNAA used the maximum number of relay nodes even more than greedy algorithm. ACR-BPA seems to be the best in terms of the required number of the relay nodes added for the network to be connected. ACR-BPA uses the minimum number of relay nodes compared with the rest of the algorithms. These results apply to all different networks with different number of nodes. Therefore, the algorithms could be ranked from the best to the worst in terms of the deployed number of relays as ACR-BPA, BPA, Greedy, and then DNAA. In fact, the implementation of DNAA on a real environment could be somehow difficult due to the special nature of oil exploration and oil drilling, the majority of oil pumping units (OPU) are spread over barren hills, mountains and deserts.

Experiment 2: Comparison among the proposed algorithms in terms of the required numbers with large monitored area.

In this experiment, again the number of oil tanks are changed from 10 to 100; however, the monitored area in this set of experiments are considered as a square of side length L = 500 m.

This large area certainly affects the required number of relay nodes. Therefore, the purpose of this set of experiments is to examine the behavior of the proposed algorithms as well as the greedy algorithm.

As can be seen in Fig. 5, DNAA still requires the maximum number of relay nodes; on the other hand, ACR-BPA used the minimum number of relay nodes. At the same time the behavior of the greedy algorithm and BPA is almost similar in terms of the number of the relay nodes they use. ACR-BPA enhances the results from Greedy algorithm with almost 23.2%, and BPA acts much better than the greedy algorithm with almost 0.009%.

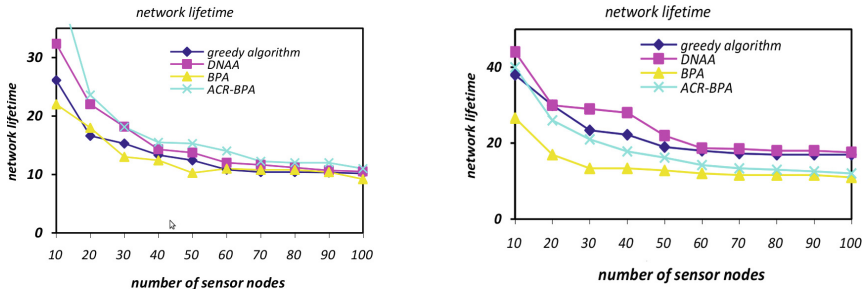


Fig. 6. The effect of adding relay nodes on the network lifetime in the four algorithms on square area with L = 100 m (left) and L = 500 m (right).

Experiment 3: Network lifetime under the operation of the proposed algorithms.

In this set of experiments, we examine the performance of the proposed algorithms compared to the greedy algorithm in terms of network lifetime. The area of oil tanks is assumed a square shape $L \times L$ where $L = 100$ m. A sensor is deployed on the top of each oil tank. The term network lifetime mean, in this context, either the first node dies or a relay nodes dies. The network lifetime is computed as the number of the round before any or both of these conditions occurs.

As shown in Fig. 6, the number of oil tanks as well as the number of sensors change from 10 to a 100 on a square area of 100×100 m. Oil tanks are deployed randomly in all cases. The Fig. 6 shows that ACR-BPA gives the best lifetime over all of the other algorithms. For instance, at 60 oil tanks, the network still operated for 26 rounds. In addition, DNAA algorithm comes in the second rank after the ACR-BPA in which at 60 oil tanks, it keeps the network operate for 24 rounds. At the same time, BPA and the greedy algorithms have the same performance for almost 12 rounds with 50 oil tanks are deployed. The results is confirmed in Fig. 6 even when the monitored area is increased to a square with side length $L = 500$ m.

Experiment 4: Comparison among the proposed algorithms in terms of the required number of relay nodes and network lifetime with monitored area of equilateral triangle shape.

Here the performance of the algorithms are examined when equilateral triangle area are used. The network topology in this case is assumed to differ from the square monitored area. There is a certain number of oil tanks are established randomly on equilateral triangle area with edge $L = 200$ m.

One sensor node is deployed on the top of each tank. The set of experiments illustrate the effect of the network shape on the network lifetime and the number of relay nodes. The number of oil tanks increases from $n = 10$ tanks to $n = 100$. DNAA can't implement in this experiment because it requires the monitored areas to be divided area into square shapes which is not valid in our case.

From the results shown in Fig. 7, the change in the shape of the network area (oil tank area) causes a change in the required number of relay nodes and

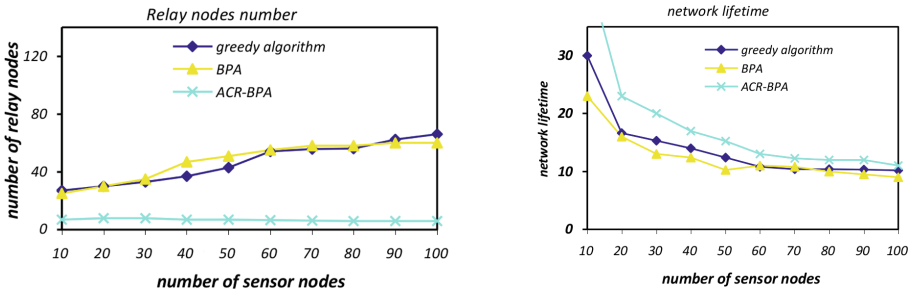


Fig. 7. The number of relay nodes added (left) and effect of adding relay nodes on the network lifetime (right) into the four algorithms in equilateral triangle area with L = 200 m

the network lifetime. However, ACR-BPA still over-performing other algorithms in terms of required number of relay nodes and network lifetime. At the same time, BPA and the greedy algorithm performing almost similar to each other especially with the large number of oil tanks. For instance after 60 oil tanks are deployed in the monitored field, their curves are overlapping as can be seen in Fig. 7.

6 Conclusion and Future Work

In this paper, the problem of oil tanks monitoring is solved using three proposed algorithms. The algorithms are compared to each other and with an existing greedy algorithm. ACR-BPA algorithm based on ACO seems to have the best performance over the compared ones. However, ACR-BPA is suitable only for nodes with multilevel communication ranges. Therefore, BPA could be suitable for nodes with fixed communication range and ACR-BPA will be the best for network with nodes having variable communication range. The conducted extensive set of experiments show that the ACR-BPA and BPA are much better than the greedy algorithm in terms of the number of deployed relays and network lifetime. For our future work, we would like to examine the performance of the proposed algorithm on network with different topologies and monitored areas with different shapes.

Acknowledgements. The study was conducted under the auspices of the IEEE-CIS Interdisciplinary Emergent Technologies task force.

References

1. Li, N., Hou, J.C.: Improving connectivity of wireless ad hoc networks. In: Proceedings of the Second Annual International Conference on Mobile and Ubiquitous Systems: Networking and Services (MobiQuitous 2005), pp. 314–324 (2005)

2. Ibrahim, A.S., Seddik, K.G., Ray, K.J.: Improving connectivity via relays deployment in wireless sensor networks. IEEE GLOBECOM, pp. 1159–1163 (2007)
3. Cheng, X., Du, D.Z., Wang, L., Xu, B.: Relay sensor placement in wireless sensor networks. Wirel. Netw. **14**(3), 347–355 (2008)
4. Koskinen, H., Karvo, J., Apilo, O.: On Improving Connectivity in Static Ad Hoc Networks by Adding Nodes. Med-Hoc-Net (2005)
5. Hou, Y.T., Shi, V.Y., Sherali, H.D., Midkiff, S.F.: On energy provisioning and relay node for wireless sensor network. IEEE Trans. Wirel. Commun. **4**, 2579–2590 (2005)
6. Misra, S., Hong, S.D., Xue, G., Tang, J.: Constrained Relay Node Placement in Wireless Sensor Networks to Meet Connectivity and Survivability Requirements (2008)
7. Chang, J.-H., Jan, R.-H.: An efficient relay sensor placing algorithm for connectivity in wireless sensor network. In: Embedded and Ubiquitous Computing. Lecture Notes in Computer Science, Vol. 4096, pp. 874–883 (2011)
8. Guo, Y., et al.: Sensor placement for lifetime maximization in monitoring oil pipelines, In: Proceedings of the 1st ACM/IEEE International Conference on Cyber-Physical Systems, pp. 61–68 (2010)
9. Elnaggar, O.E., Ramadan, R.A., Fakry, M.B.: WSN in monitoring oil pipelines using ACO and GA. In: The 5th International Conference on Sustainable Energy Information Technology (SEIT-2015), vol. 52, pp. 1198–1205 (2015)
10. Barani, R., Jeyashmi, V.: Oil well monitoring and control based on wireless sensor network using atmega 2560 controller. Int. J. Comput. Sci. Commun. Netw. **3**, 341–346 (2013)
11. Colorni, A., Dorigo, M., Maniezzo, V.: Distributed optimization by ant colonies. In: Proceedings of ECAL 1991, European Conference on Artificial Life. Elsevier Publishing, Amsterdam (1991)
12. Dorigo, M., Maniezzo, V., Colorni, A.: The ant system: an autocatalytic optimizing process, Technical report TR91-016. Politecnico di Milano (1991)
13. Dorigo, M.: Optimization, learning and natural algorithms. Ph.D. thesis. Politecnico di Milano, Milano (1992)

A Smart Way to Improve the Printing Capability of Operating System

Adeel Ahmed[1], Muhammad Arif[1(✉)], Abdul Rasheed Rizwan[2], Muhammad Jabbar[1], and Zaheer Ahmed[1]

[1] Department of Computer Science, University of Gujrat, Gujrat, Pakistan
adeel.ahmed@uog.edu.pk,
{m.arif, m.jabbar, zaheer}@uog.edu.pk
[2] Department of Computer Science, Virtual University of Pakistan, Lahore, Pakistan
liveneeo@gmail.com

Abstract. Operating system is the core of computer science and task scheduling is the major topic in this domain. Priority base scheduling is always remains hot topic in the domain of operating systems. In Priority base printing always higher priority given to those printings jobs which are processed more quickly rather than the lower priority base printings jobs. In this way low priority jobs are delayed again and again. In this research detail study is conducted regarding the scheduling algorithms in operating system. Priority base printings become bottlenecks who are assigned lowest priority. They may have to wait till their turn may be up to mid night in large departments. Now we are going to discuss a specific algorithm which can solve above issues in a smart way. A new method is proposed to solve the priority scheduling problems. This method is very handy for those users who are assigned a higher priority authority in printing mechanism. It also allows the lowest priority base printing to print data. In this solution we are not going to ignore the highest printing jobs. The highest priority base printing jobs will be continued in the same way. Our main objective is to maintain a balance between the high priority and low priority printing jobs without suffering from continues delay.

Keywords: Priority printing · Printer · Spooler · Algorithm · Smart scheduling · Printing capability

1 Introduction

Priority base scheduling is always remains hot topic in the domain of operating systems. Printing Architecture (PA) is a major component of any Operating System. PA is consists on printer driver and the application which sends the prints to printer [1]. Operating system is the core of computer science and task scheduling is the major topic in this domain. Printer drivers are attached with spool simultaneous peripheral operations on-line used to add printing job to buffer. Spooling provides best behavior to such I/O which has different data rates in receiving and sending mechanisms. Spooler is a kind of queue which takes input from rear and out at front part of the queue. The spooler sends data to target printer's I/O port which can be in serial, network, USB, wireless

V.E. Balas et al. (eds.), *Soft Computing Applications*, Advances in Intelligent Systems and Computing 633, DOI 10.1007/978-3-319-62521-8_18

communication channel and parallel [2]. In Priority base printing always higher priority given to those printings jobs which are processed more quickly rather than the lower priority base printings jobs. In this way low priority jobs are delayed again and again. In this research detail study is conducted regarding the scheduling algorithms in operating system. Priority base printings become bottlenecks who are assigned lowest priority. They may have to wait till their turn may be up to mid night in large departments.

Every printing job has some operating to complete its printing processes. For e.g., user sends printing command from an application which include direct printing, networks printing and internet base printing, after sending a command from the application it enters to the spooler (Simulates Peripheral Operation online) [3]. It also allows the lowest priority base printing to print data. In this solution we are not going to ignore the highest printing jobs. The highest priority base printing jobs will be continued in the same way. Our main objective is to maintain a balance between the high priority and low priority printing jobs without suffering from continues delay. Now we are going to discuss a specific algorithm which can solve above issues in a smart way. A new method is proposed to solve the priority scheduling problems. This method is very handy for those users who are assigned a higher priority authority in printing mechanism [4]. It also allows the lowest priority base printing to print data. In this solution we are not going to ignore the highest printing jobs. The highest priority base printing jobs will be continued in the same way. Our main objective is to maintain a balance between the high priority and low priority printing jobs without suffering from continues delay.

The spooler performs specific processing on the sent job and passes to the printer driver. The printer driver translates the job according to the bitmap of the printer. Printing job to local or remote computer is performed by printer provider. Printer provider is responsible for print queue, functions and Printing Priority Queue (PPQ) [5]. PPQ performs high priority job before the low priority for printing. We can set the priority of individual documents in the general tab, use the priority slider (Windows OS) low priority is 1 and highest priority is 99.

2 Related Works

Priority queue is different from normal queue. The normal queue data structure is base on FIFO. But priority queue is a kind of box which receives different priority jobs and sent the highest priority out to the calling function. Here is a conceptual picture of a priority queue is shown in the Fig. 1.

Figure 1 shows how that printing jobs are inserted in to the queue and the highest priority job is taken out at first. It also shown that the lowest priority jobs are remained in the queue waiting for their turn. Printing jobs are sent to priority queue to sort their positions; different kinds of method are used for sorting their priorities just as array representation, link list representation and heap sort etc. [4]. All these data structure have computation time to sort the requested prints. If a printer is already working on hundreds request and so much busy and in that position while a printer is going to print documents at this position it may receive another request and it have to compute sort on all prints, and this situation may exists for long time. This can stop the working of all prints [6]. We can see before sorting priority queue jobs, this job can enter the priority

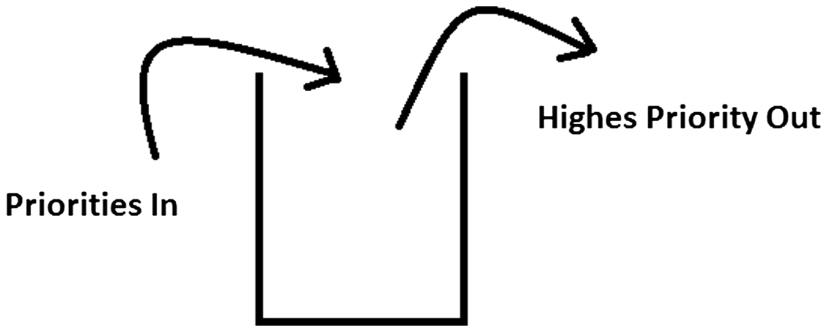


Fig. 1. Priority queue [2]

queue in different order. After entering the jobs into priority queue an algorithm is applied to sort them. Absolutely this algorithm takes time to sort them. Figure 2 illustrates the before sorting view.

Fig. 2. Before sorting overview

Starvation can be caused in PPQ. Just e.g., there 1 to 99 priorities for users according to their work in a origination. 99 have the highest priority and 1 has the lowest priority in case of Microsoft OS. From 99 to let's say 20 users are already in printing but remaining 19 to 1 may b in starvation. It can be from 1 to 19 has urgent information to give his boss but he is the victim of starvation. So just PPQ is not so better in such kinds of case for printing jobs. Figures 3 and 4 illustrates the priority Queue.

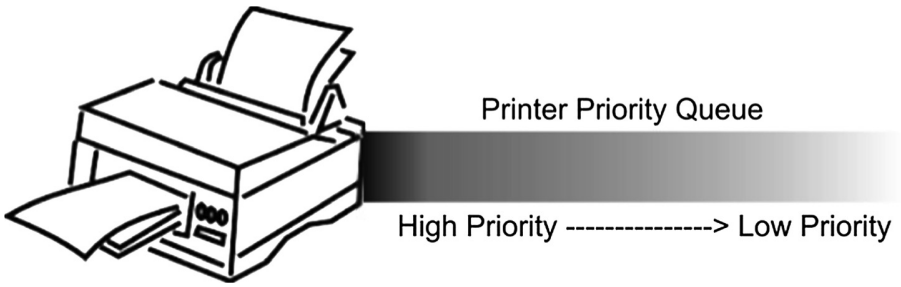


Fig. 3. Priority queue [7, 8]

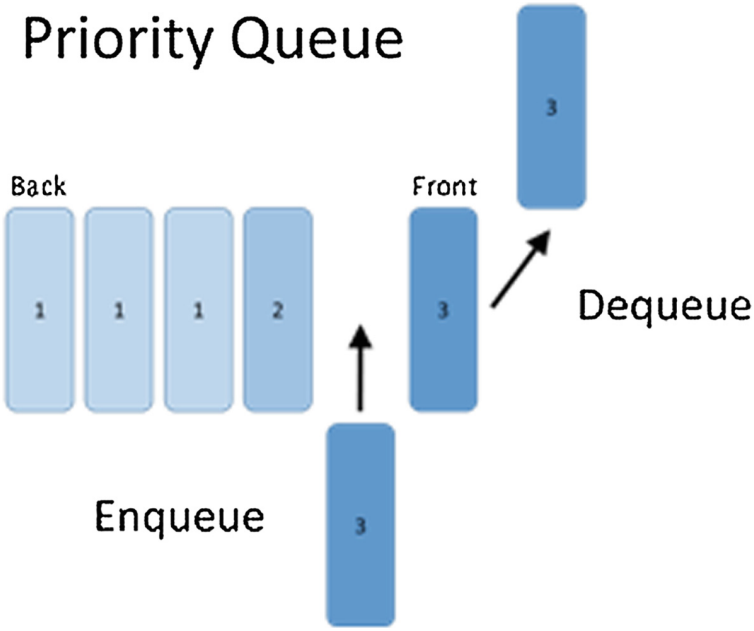


Fig. 4. Priority queue enqueue and dequeue [13]

In Fig. 4 it is shown that enque and dequqe functions are takinf the highest priority job first and then move to the lowest priority jobs. Assigning different priority level for different kinds of groups with high and low priority level of queue is applied to Microsoft Windows server edition [5]. According to Microsoft to use the best way of priority level printing, create multiple logical printer of the target printer. Each priority level should be different to other group of priority level. For example users in Group A should be able to connect with printer with highest priority just as priority number one. And Group no. 2 should access the printer with the priority level two and further so on [9, 10, 16].

The method printing not only support LAN but also the WAN in client systems [11]. This will become a huge network so we have to manage it in a smart way. The method of printing first include to identify the server method to receive and print the jobs [16]. We can customize printing priority level for others groups. Click on start button and click on Open Printers and Faxes.

Right click on target printer, click on selected printer **Properties**, and then choose **Advanced** tab. Now click on **Priority tab**, click the up or down arrows, and then click **OK**. Here you can set priority level from 1 to 99. "99" is the highest priority level and "1" is lowest priority level in printing [17]. Select the appropriate priority level and click **OK**. Now to add new logical printer so to add a new logical printer click on **Add Printer** to add another logical printer for the same existing physical printer. Now again we have to set its priority level according to our feet needs. Allow the low level users to use the first logical printer and medium level users allow them to use second logical

printer, following the way we set high level user for the highest logical printer. When we need to open Printer and Faxes dialog box, we have to click on start button and next click on Printer and Faxes [12].

3 Problem Statement

Microsoft does not allow it to manage lowest priority jobs automatically. Microsoft Server provides manually setting to overcome this issue. In Microsoft environment we have to change the priority of documents which are waiting to get printer. Even the selected job's priority changing does not affect other priority jobs, yet it is not a good solution for large business environments. It is also happen in Microsoft Server edition the lowest priority base printing is fixed up to a fix time of night. The lowest printing priority is selected at midnight for printings. The time setting is hard set so in time of free printer, the time set priority documents are not allowed to printer to use the printer resource. So this solution to the problem is not so much efficient. Higher priority should in top ascending order numeric numbers and lower priority job should rear part of ascending order in numeric numbers.

4 Results and Discussions

Queue is a linear data structure where an item is removed from top end (Front) and new item is inserted at back end (Rear). This is shown in the Fig. 5.

Fig. 5. Queue insertion and deletion

Figure 5 shows the insertion and deletion of the jobs in the queue box. This is a common fashion which is used in normal kind of queue and data structure. But we are going to use this queue data structure in a different way with the help of linked list. Linked list allow to add and remove a node in anywhere of the list. We don't have to shift up and down of other part in linked list to remove or to add a node as in array data structure. Let's suppose the following is a linked list queue.

In above figure different priority numbers are shown in Fig. 6, let's suppose these are the priority numbers of printing commands. And we need to send command of

Fig. 6. Priority numbers

priority number two, which is recently entered into queue. But due to its priority we have to send it first to printer. Here we need to remove this node from the linked list. Now we can see according to its data structure nature we should send command of top node to printer. We also not are going to sort it according to its priority, because we don't want to use CPU cycle here due to efficiency issue. After removing the target node, now see in Fig. 7 [13].

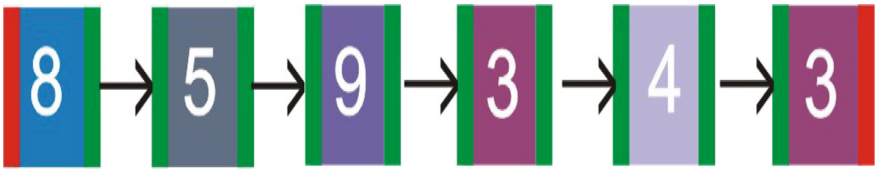


Fig. 7. Target node removal

After knowing all kinds of issues in PPQ lets solve these issues in a smart way [14]. Before going in detail of our algorithm first take overview the flow chart of our algorithm. This describes the working of algorithm in better understanding [15, 16].

4.1 Proposed Methodology (Boss Algorithm)

Boss Algorithm is operating section of our algorithm. It should receive the print command before going to spooler. In standard scenario printing command works as [Application > Spooler > Printer]. But in BA this order should [Application > BA > Spooler > Printer]. When BA will receive a printing command from an application it should check the priority of the printing job. If the priority is one, it should directly send to Ready Priority Queue (RPQ). In other case if the sent priority print is not one it should send to Printing Priority Queue (PPQ) [17]. The Boss Algorithm is depicted in the Fig. 8. The pseudo code of the algorithm is given below.

```
If (priority == 1)
        {
        Move to the RPQ→Spooler→Printing
        }
Else
{
Move to PPQ→RPQ→Spooler→Printing
}
```

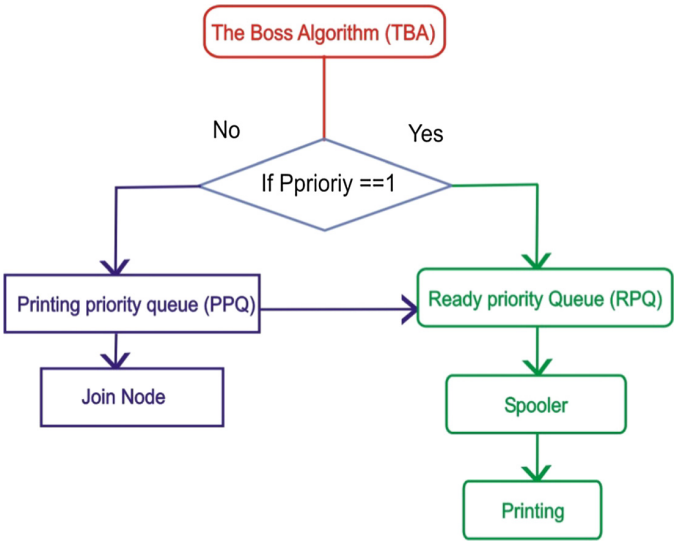


Fig. 8. Boss algorithm

4.2 Printing Priority Queue (PPQ)

PPQ should receive all kinds of prints which have not priority of one. Suppose we have following prints which are recently sent to PPQ. Before going in detail of PPQ, first let discuss its structure. In Fig. 9 we can see the different prints have arrived in PPQ, different color boxes telling us the different prints in PPQ. Above white color numeric numbers are telling us the priority numbers which are 3, 5, 13, 8, 7 up to 9. Below numbers of the figure such as 0×001, 0×002 so on up to $0 \times 00n$. These numbers are addresses of the prints which are in PPQ in a link list point of view. PPQ data structure should program with link list. It is because we need add and delete operation, we know add and delete operations are much fast in link list against any other data structure. When a print enters into PPQ, its priority should be written and it address should also append with the print. Other words we can say that two kinds of information should have with every print. The Fig. 6 has different priorities without having the highest number priority [18]. When PPQ will receive a print having priority number 2, it should send before other prints to RPQ with the same priority number. We can see in Fig. 6 there is one print which has the priority number two. It should directly send to RPQ without further processing. It is because we also need a difference in RPQ which is receiving only priority number one. If we change the priority up to one in PPQ then it will be difficult for RPQ which job should send to printer first. So we should send print from PPR to RPQ in priority number two. When RPQ send command to printer which having the priority number two, RPQ should signal to PPQ I am serving your request and now you should rescan your printing jobs again. This process should do on each process which having the priorities number two in RPQ. PPQ is responsible for following processes.

I. Send the print to RPQ which has priority number 2 without further processing.
II. Increase the priority up to two, which has low priority than 2.
III. After sending print to RPQ, PPQ is responsible to delete the node and join nodes.

Now we take an example to understand PPQ. Now again see the following picture of Fig. 9.

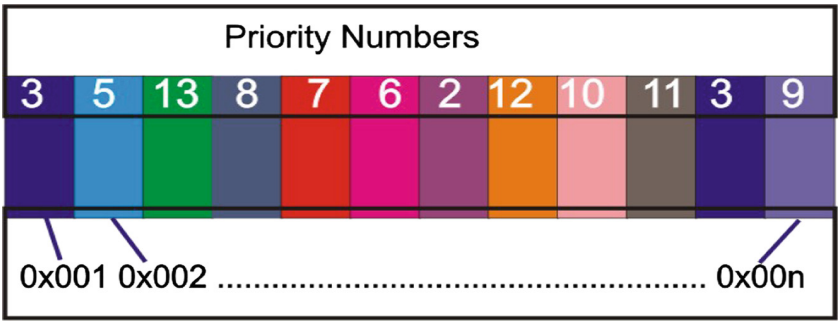


Fig. 9. Prints arrival in the queue

After finishing all jobs which have the priority number 1, RPQ should signal to PPQ to send it printing jobs. So PPQ should send the priority number two printing job to RPQ, after it PPQ should delete the node after sending RPQ.

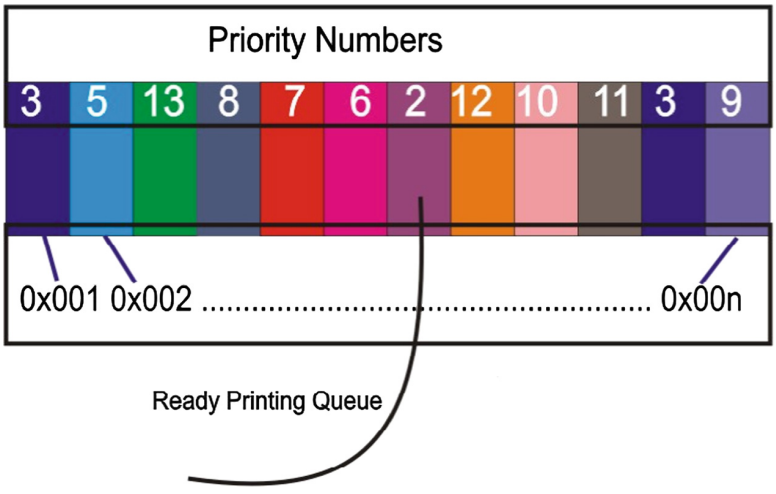


Fig. 10. Print job is sending to Ready Priority Queue

Figure 10 is showing that priority number 2 print job is sending to Ready Priority Queue (RPQ) without scanning and increasing priority number. After sending the print job, now PPQ should delete and join the sent printing job.

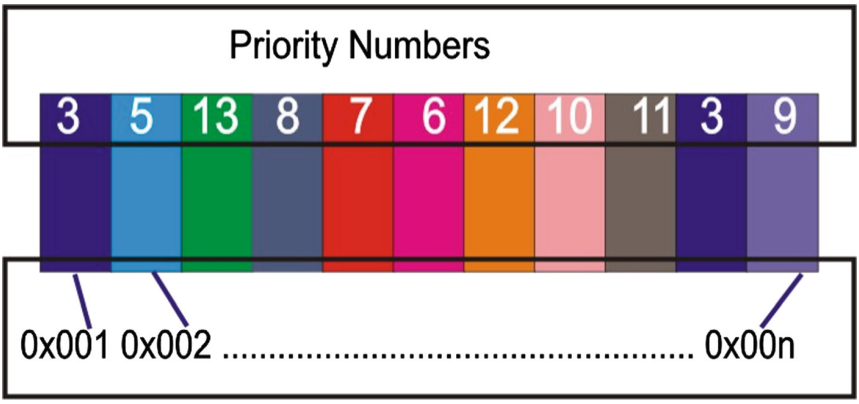


Fig. 11. Delete and join node

Figure 11 is showing after sending, deleting and joining the node process. Next PPQ should wait and receive all prints commands while PPQ does not receive the signal of scanning. After scanning the signal of scanning PPQ should scan again all prints and increase the priority with number one, if there is again a printing command which the priority number two, it should directly send to RPQ without further processing. In Fig. 11 we can see there are two print commands which have priority number 3, mean on one scanning its priority should 2.

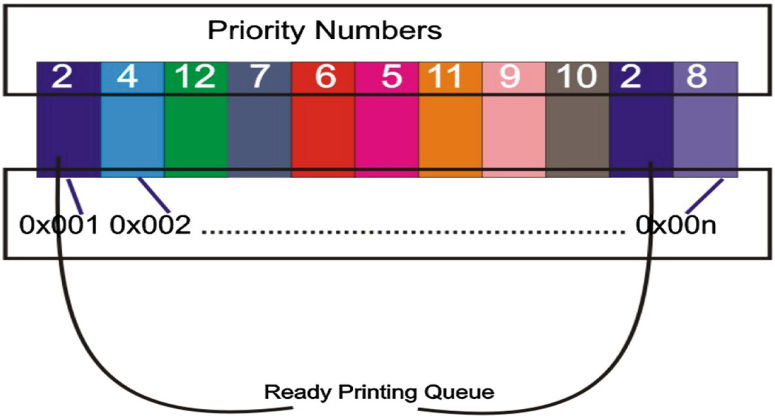


Fig. 12. Increasing the priority number

Now we can see the difference in Figs. 11 and 12 after scanning and increasing the priority number. Now we can see there are two prints commands which have priority level two. Now PPQ should send these two prints commands to RPQ, and after it PPQ should delete these nodes.

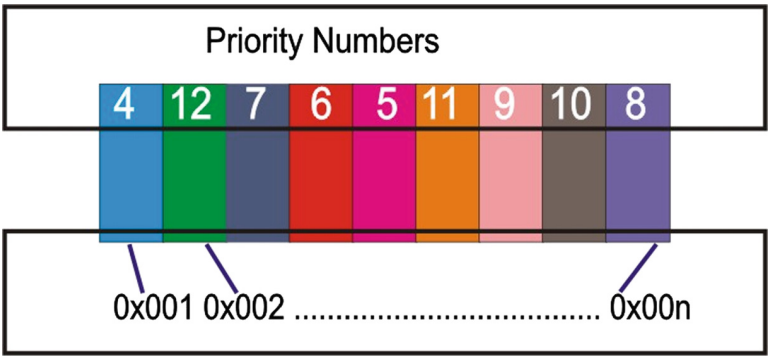


Fig. 13. PPQ after deleting and joining two nodes

Figure 13 is showing PPQ after deleting and joining two nodes. Now PPQ should again wait signal from RPQ for further scanning, increasing priority, sending print jobs, deleting and joining nodes.

4.3 Ready Priority Queue (RPQ)

RPQ should attach to spooler for sending prints commands to printer. RPQ should have two kinds of priority type prints number one, and number two. Priority prints which have priority one, should send to printer before the printing having the priority number two. RPQ can differentiate easily that which print should send to printer first. When priority having number one should not in RPQ then priority two should send to printer for printing in batch sequence [19]. If printer is busy in priority two batch sequence and a new printing command inter in RPQ, after completing the prints of the under process printing then next turn should priority number one. When RPQ is working on priority number two, it should signal to PPQ for further scanning and increasing the priority which it has in its queue. This situation may happen there is one print RPQ and PPQ have many other prints jobs. There is no such priority exists which level increase to two for RPQ in just one scan. On empty queue of RPQ is should signal again and again to send print command to it. Now let's take an example to understand RPQ.

Fig. 14. Two types of priorities are in RPQ

In Fig. 14 we can see two kinds of priorities are in RPQ, the priority number one came directly from BA after checking its priority. The priority number two is come from PPQ. This number of priority might be up to level two. In PPQ if the priority is two it should directly send to RPQ without increasing and extra processing over it.

In Fig. 14 the right side command first came into RPQ so it should send to spooler first, second from the right side second last print command came into RPQ so it should send to spooler at position number two. Now next is priority number two which is in third number from right side. Because we have more prints which have priority number one, so the number two priority should not send to printer. The next print is also the priority number one which is at fifth number from left to right in Fig. 14.

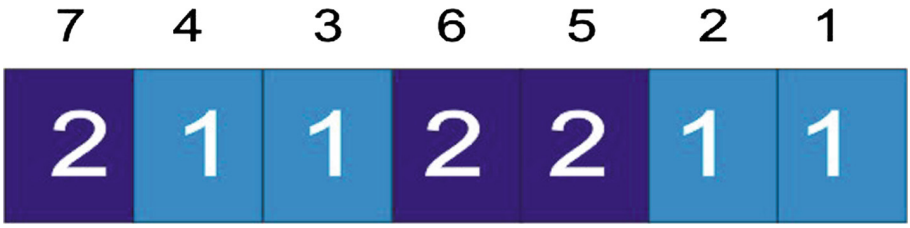


Fig. 15. Way of numbering

Figure 15 is showing us the way of numbering in which prints should send to spooler for printing. The new coming prints should treat in FCFS fashion if they have the same priority, the only difference is that priority number one and priority number two. On serving priority number two RPQ should signal to PPQ to rescan the priority numbers and increase the numbers according to its appropriate sequence.

4.4 Join Node (JN)

Join Node is sub program part of PPQ, JN should perform two kinds of work. It should first remove the current sent print job node, after it, it should join the deleted node left and right nodes.

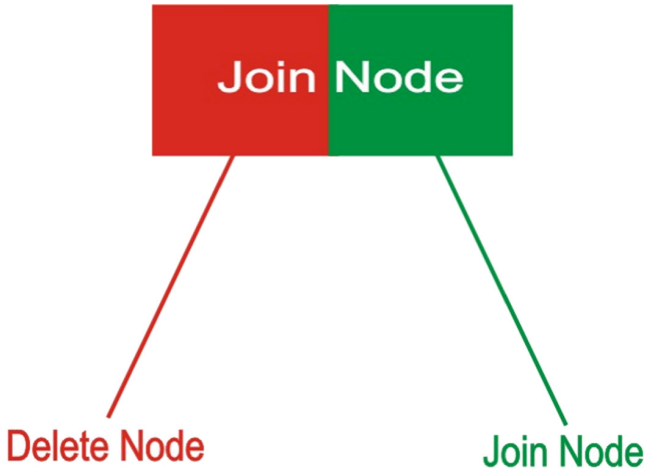


Fig. 16. Complete picture of join node

Figure 16 is showing a complete picture of JN, JN should work Delete and Join Node. When a print should send to RPQ, PPQ should call next Join Node function. The first part of the Join Node should delete the sent print command node, after deleting the node next JN should call its next function to join node [20]. While working in Deleting and joining the nodes PPQ should wait and not to scan the nodes. When JN return its control to PPQ, then PPQ should precede its next work.

5 Conclusion and Future Work

BOSS algorithm is best way for printing all kinds of priority level. It provides turn to every level of priority and it also gives first priority to the highest level priority users. It never let down or star-vat the low priority users. Low priority level users not have to wait their print in mid night. There is no need to add virtual printer and extra overhead for server and low priority level users. BA is complex in sense handling all kinds of priorities in same time. Change the priority of the queued jobs in every search setup the over head of CPU. In this way this proposed solution can be implemented in huighly demanded environment where printing capabilities are lacked.

In future we are planning to implement this scenario in well developed organization and gain the benefits of newly designed technique. This will be more beneficial for the print and press industries to groom their business. This idea can be utilized for priority based process scheduling mechanism.

References

1. Macauley, R.M., Martin, T.S., Vachon, G.C., Cosgrove, P.A., Sasson, S.J., Wilson, E.D.: Distributed Printing Social Network. Google Patents (2015)
2. Shaw, L.F., Teng, C.-C., Sykes, K.W., Endres, R.E.: Device Independent Spooling in a Print Architecture. Google Patents (1998)
3. Laurent, M.S., Onischke, M., Kuindersma, M., Krishnammagaru, D., Stairs, J., Noreikis, K.: System and Method for Releasing Print Jobs Based on Location Information. Google Patents (2015)
4. Sanders, R.E., Frazier, D.P.: System and Method for Batch Printing High-Volume Electronic Documents from a Network. Google Patents (2010)
5. Gould, K.V.W., Small, G.J., Galipeau, S.R., Kirkland, C.J.: Methods and Systems for Data Prioritization. Google Patents (2014)
6. Zhang, D., Dechev, D.: A lock-free priority queue design based on multi-dimensional linked lists. IEEE Trans. Parallel Distrib. Syst. **27**, 613–626 (2016)
7. Gruber, J., Träff, J.L., Wimmer, M.: Benchmarking Concurrent Priority Queues: Performance of k-LSM and Related Data Structures. arXiv preprint arXiv:1603.05047 (2016)
8. Lenharth, A., Nguyen, D., Pingali, K.: Priority queues are not good concurrent priority schedulers. In: European Conference on Parallel Processing, pp. 209–221 (2015)
9. Kinoshita, K., Higashizaki, Y.: Remote e-mail printing. Google Patents (2011)
10. Nan, X., He, Y., Guan, L.: Optimal resource allocation for multimedia cloud based on queuing model. In: 2011 IEEE 13th international workshop on Multimedia signal processing (MMSP), pp. 1–6 (2011)

11. Abaev, P., Gaidamaka, Y., Samouylov, K.E.: Queuing model for loss-based overload control in a SIP server using a hysteretic technique. In: Internet of Things, Smart Spaces, and Next Generation Networking. Springer, pp. 371–378 (2012)
12. Nakayama, Y., Takewa, Y., Sumikura, H., Yamanami, M., Matsui, Y., Oie, T., et al.: In-body tissue-engineered aortic valve (Biovalve type VII) architecture based on 3D printer molding. J. Biomed. Mater. Res. Part B: Appl. Biomater. **103**, 1–11 (2015)
13. Lansing, S., Pantelias, N., Vu, Y., Gomez, F.J.: System and Method for Dropping Lower Priority Packets that are Slated for Wireless Transmission. Google Patents (2011)
14. Al Hanbali, A., Alvarez, E., Heijden, M.: Approximations for the waiting time distribution in an M/G/c priority queue (2013)
15. Stanford, D.A., Taylor, P., Ziedins, I.: Waiting time distributions in the accumulating priority queue. Queueing Syst. **77**, 297–330 (2014)
16. Singh, C.J., Jain, M., Kumar, B.: MX/G/1 queuing model with state dependent arrival and second optional vacation. Int. J. Math. in Oper. Res. **4**, 78–96 (2012)
17. de Souza, R.M., Morabito, R., Chiyoshi, F.Y., Iannoni, A.P.: Incorporating priorities for waiting customers in the hypercube queuing model with application to an emergency medical service system in Brazil. Eur. J. Oper. Res. **242**, 274–285 (2015)
18. Takagi, Y., Nakagawa, M.: Printing System, Printing Method, Print Server, Control Method, and Computer-Readable Medium. Google Patents (2012)
19. Alistarh, D., Kopinsky, J., Li, J., Shavit, N.: The SprayList: a scalable relaxed priority queue. ACM SIGPLAN Not. **50**, 11–20 (2015)
20. Kim, C., Choi, S.: A Study on Efficient Join Mechanism Using Streaming-Service-Time in Mobile P2P Environment (2014)

A Comparative Study for Ontology and Software Design Patterns

Zaheer Ahmed, Muhammad Arif(✉), Muhammad Sami Ullah, Adeel Ahmed, and Muhammad Jabbar

Department of Computer Science, University of Gujrat, Gujrat, Pakistan
{zaheer, m.arif, msamiullah, adeel.ahmed, jabbar.ahmed}@uog.edu.pk

Abstract. Ontology design patterns have been extended from software design patterns for knowledge acquirement in semantic web. The main concern of ontology design patterns is to how concepts, relations and axioms are established in a ontology using ontological elements. Ontology design pattern provide the solution of quality modeling for ontologies. In user prospective the presentation of ODP play a central role for the reusability of ontologies. This research determine the improvement areas in the presentation of content ODPs. Improvement in presentation can ultimately improve the understandability of a pattern from user perspective. Our objective is to analyze the template of different software engineering patterns (SEP) and ODP. On the basis of this analysis we suggest possible changes in current template and pattern presentation. It also includes determining the most important information about patterns which can help an ontology engineer in selecting an appropriate pattern. Presentation of design patterns is related to issues such as reuse, guidance and communication. Our main goal is to evaluate the current patterns presentation. The evaluation is focused on the analysis of current patterns. The ontology design pattern templates were compared with existing templates of other patterns for determine the improvement areas. The template of an ODP consists of many parts, the first question is to identify the most important and vital information concerning the design patterns. This information would help an ontology engineer to select an appropriate design pattern for the required ontology. The second question is about the users who work with ontology design patterns. Generally, users are divided into two categories; novice and expert ontology engineers. Novice users are the end-users who use design patterns to implement in the ontologies. Expert ontology engineers are those who actually develop ontology design patterns. Each category of user has its own information requirement regarding design patterns.

Keywords: Ontology design pattern · Templates · Content ontology design pattern · Software engineering pattern

1 Introduction

ODPs are semantic patterns that provide quality modeling solutions for ontologies. ODPs play an important role in learning and teaching ontology engineering [1]. They facilitate the automatic and semi-automatic ontologies construction and provide a base

V.E. Balas et al. (eds.), *Soft Computing Applications*, Advances in Intelligent Systems and Computing 633, DOI 10.1007/978-3-319-62521-8_19

for creating ontologies in different domains [2]. In user prospective the reusability of ontology depends on presentation of ODP. So far only a small catalogue of patterns exist which is available online at the ontology design pattern portal. In this portal, ODPs are described using a template with a set of headings that should be filled out when entering a new pattern. The template defines a standard way for constructing new patterns. There are possibilities to discuss modeling issues, review and suggest changes in patterns [4].

In computer and information science ontology is defined as a "*formal, explicit specification of a shared conceptualization*" [3]. One of the main problem areas is reusability of ontologies. The existing ontologies are available at online ontology repositories which provide guidelines to ontology users. Due to unfamiliar logical structure the existing ontologies provide limited support. Must be learned the good practices form literature. This problem is solved by implementing common solution as we learn in software engineering [4]. The patterns facilitate and to some extent automate the construction of ontologies. The development of patterns in the ontology field is very popular as that in software engineering. The patterns are defined for reuse and aim at facilitating the construction process very much like the way it is done in software engineering or architectural planning of buildings [5]. The purpose of design patterns is to solve the design problems. The patterns provide a useful way for handling the problems of reusability in a development process. In SE the common practices to build software through design and architecture patterns. This practice also follow in ontology engineering [2].

Ontology design patterns provide modeling solutions of ontologies design problems. They provide a base for creating ontologies in different domains. Patterns are also used for evaluation of ontologies [4]. Ontology design patterns (ODPs) are of several types. They are divided into 6 families; Content patterns, Structural Patterns, Presentation patterns, Correspondence Patterns, Lexico-Syntactic Patterns and Reasoning Patterns [6]. This thesis deals with the presentation of content ontology design pattern. It describes the design issues of the presentation of ontology design patterns. There may be certain information that an ontology user need to understand a pattern but it is not available in the description, our next task will be to examine the missing information in the current ontology design pattern templates. Finally, based on the results of above questions, we will made suggestions for the potential improvement in the current templates of ODP presentation.

2 Related Work

Knowledge reuse is common way to improve the quality of work artifacts. Pattern is a common way to improve reusability. There are other ways to support reusability, i.e. in object oriented programming the concept of program components related to the reusability of design. To achieve reusability, number of design technique available in object oriented software development model for more reusable building block. An obstacle for reuse methodologies is the lack of motivation among developers. Before

starting the process, the developer needs to establish a reuse library which requires extra efforts. Reuse process is divided in to two steps *Design for reuse* and *Design by reuse*.

To expedite design by reuse, first the design for reuse process must be established [5, 7].

Reusability is applied at different levels. In software engineering, reusability can be applied at following three levels [5]:

Requirements Reuse: It deals with the models of the domain or the generic model of the requirement domain.

Design Reuse: It deals with the models, data structures and algorithms.

Software Component Reuse: It deals with reuse of software classes and editable source code.

The development process of ontology is time consuming and more effort required form developer. Reuse of an ontology from an ontological engineering perspective can be hard. This is even more when there are large ontologies to be reused [8].

Ontology Library: For ontologies to be grouped and organized so that they may be reused further and for ontology integration, maintenance, mapping and versioning, an important tool known as ontology library systems was developed. These systems must fulfill all ontological reuse needs and must be easily accessible [9].

Ontology Matching: Ontology matching is a process of judging correspondence between semantically related ontologies, solving the problem of semantic heterogeneity and can be used in ontology merging, query answering, data translation etc. So ontology matching facilitates interoperability between matched ontologies [10].

A pattern is something re-occurring that can be applied from one time to another and also from one application to another. These concepts are in use in our daily life and also in our professional life. We use old solution as patterns. We search patterns in our surroundings that can be useful. In reuse the pattern are best practices that provide the good design [5, 11].

Currently in the computer science field the most common and popular patterns are software patterns. Most of the software development projects, where applying functional or object oriented design are conducted using patterns. The patterns are used for increasing reusability, product quality and for managing complexity of the system development process. According to the phase of the development process where they are used these patterns are divided into different kinds [5].

The most common categories are the following:

- Analysis Patterns [5, 12].
- Architecture Patterns [5, 13].
- Design Patterns (see [11, 14–17]).
- Programming Language Idioms (see [3, 13]).

Analysis patterns are used to describe the conceptual structures of business processes for the different types of business domains like how to transform these processes into software [3, 12, 18].

An overall structuring principle is used while constructing a viable software architecture. Architectural patterns are considered as templates for solid software architectures [3].

Design pattern describe the common solution of overall design problem in certain context through the depiction object and class communication. Design pattern provide the description of participation classes, their instance, their roles and interaction, distribution of responsibility. One object oriented design problem address by the certain design pattern that also provided sample code to describe the partial solution of the problem [14].

The lowest level of patterns is represented by idioms. The implementation of particular design issue is dealt with by the idioms. The aspects of both, the design and implementation, are treated by them [13].

2.1 Problem Statement

The ontology design patterns (ODPs) have divided in grouped of six different families: Structural ODPs, Content ODPs, Reasoning ODPs, Presentation ODPs, Correspondence ODPs and Lexico Syntactic ODPs [6].

Graphically the types of patterns are represented as (Fig. 1):

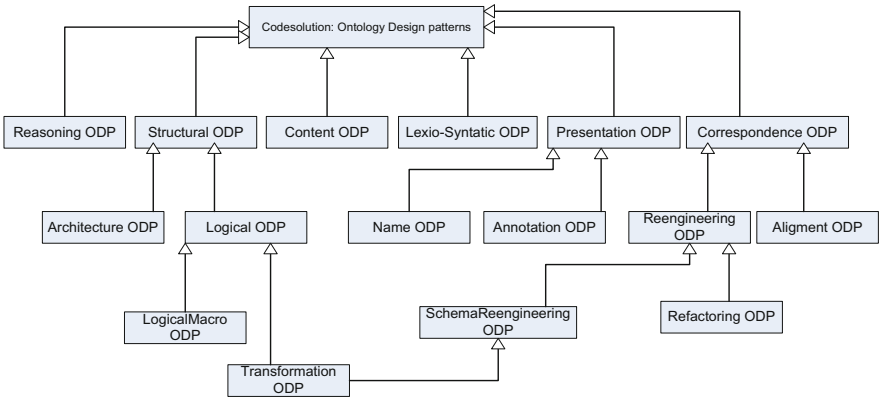


Fig. 1. Group of ontology design pattern [19]

In this paper, our focus of research will be the Content ODPs. The comparison of ODPs is limited to software patterns and data model patterns because these are the most used and well known patterns. There are many ontology languages available for development but we will only focus on OWL ontologies [19].

This research guide the improvement areas in the presentation of content ODPs. Improvement in presentation can ultimately improve the understandability of a pattern from user perspective. The focus of research is to analyze different content ODPs and provide some possible recommendation in current templates of the ODP presentation. It also includes determining the most important information about patterns which can help an ontology engineer in selecting an appropriate pattern.

Presentation of design patterns is related to issues such as reuse, guidance and communication. Our main goal is to evaluate the current patterns presentation. The evaluation is focused on the analysis of current patterns.

The template of an ontology design pattern consists of many parts, the first question is to identify the most important and vital information concerning the design patterns. This information would help an ontology engineer to select an appropriate design pattern for the required ontology. The second question is about the users who work with ontology design patterns. There may be certain information that an ontology user need to understand a pattern but it is not available in the description, our next task will be to examine the missing information in the current ontology design pattern templates.

3 Methodologies

There are many types of patterns but we have limited our comparison to those of software engineering and data models patterns, since these are the well know and common use patterns.

Software patterns have been compared to ontologies by Devedzic in one of his article [16] where the author argued that there is a significant overlap between the two concepts and that it is the aim, while generality and practical usage of these concepts differ. The concepts of patterns and ontologies have some common goals, e.g. sharing and reusing of knowledge. Both these concepts necessitate hierarchies of concepts, relationships, vocabularies and constraints. Moreover, both of them can be seen as using an object-oriented paradigm [3].

First, we analyze what current templates exist in other fields and then compare them to ontology design pattern templates to analyze the difference. In our study, evaluation was done on the results of the evaluation of different patterns. Templates of the patterns were compared to identify the difference and similarities in their presentation. Each part of the templates was studied with respect to its objective and the content provided in that part [19].

A template of a pattern is a standard way of representing a pattern. In a broad sense, a pattern template has four important elements. These elements are: Name, Problem, Solution and Consequences [3, 11, 19].

The different kinds of pattern templates are given below with their description.

3.1 Design Pattern Template

This template proposed by Guarino [11] in their book "Design pattern Element of Reusable Object-Oriented Software" [11]. Table 1 shows the different parts of the template and their description (Table 2).

Table 1. Design pattern template [11, 19]

Design pattern template	
Elements	Description
Pattern name and classification	Name is a short summary of the pattern. There are several DP, we need a way to classified them in a family. The section classification refers these families of design pattern
Intent	It is a brief description that explain the following. How the design pattern work? What is the main goal of the pattern and what are the particular design issues or problem solve by the pattern
Also known as	Another name of the pattern, If the pattern has other name
Motivation	It has a scenario that describes a design problem and explain the class and object structures in the pattern describe the problem solution. The problem solution will facilitate you to understand the more abstract description of the pattern
Applicability	This section discus the situations in which the design pattern applicable
Structure	It illustrates a detailed specification of the structural aspects of the pattern. It includes a graphical representation of the classes in the pattern using the notation of OMT (Object Modeling Technique). This section also has interaction diagrams to illustrate sequences of requests and collaboration diagram for description of collaboration between objects
Participants	This section describes the different parts of the pattern and their relation. In design pattern the participants are classes and/or objects
Collaborations	This section describes how the participants collaborate to carry out their responsibilities
Implementation	Implementation gives guidelines for implementing the pattern. It gives hints and techniques which one should be aware before implementing the pattern. For example if there are language specific issues
Sample code	Sample code is a code fragment that illustrates how you might implement the pattern in a programming language
Known uses	Known Uses is the examples of the use of the pattern in real systems. It includes a minimum of two examples from different domains
Related patterns	Related patterns are described, i.e. what are the closely related patterns to this given pattern? What are important differences? With which other patterns should this one be used?

Table 2. Builder design pattern [19]

Builder design pattern	
Elements	Description
Pattern Name and Classification	Builder, creational patterns
Intent	To split the construction of the complex object from its representation so that[same construction process can create different representations
Also Known As	
Motivation	Fig. 2
Applicability	When the builder pattern are used • The algorithm for creating a complex object should be independent of the parts that make up the object and how they're assembled • The construction process must allow different representations for the object that's constructed
Structure	Fig. 3
Participants	Builder, ConcreteBuilder, Director, Product
Collaborations	The given interaction diagram describe how Builder and Director cooperate with a client Fig. 4
Implementation	Fig. 5

Sample code	/Abstract Builder
	class abstract class TextConverter{
	abstract void convertCharacter(char c);
	abstract void convertParagraph();
	}
	.
	.
	.
	.
	public static void main(String args[]){
	Client client = new Client();
	Document doc = new Document();
	client.createASCIIText(doc);
	system.out.println("This is an example of Builder Pattern");
	}
	}

Known uses	The RTF converter application is from ET++. Its text building block uses a builder to process text stored in the RTF format
Related patterns	Abstract factory

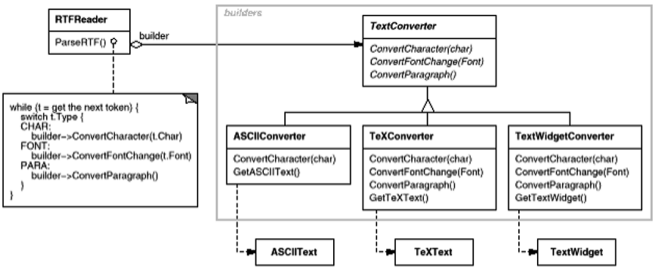


Fig. 2. Builder design pattern

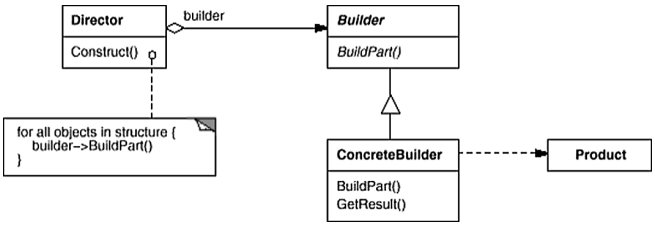


Fig. 3. Builder design pattern

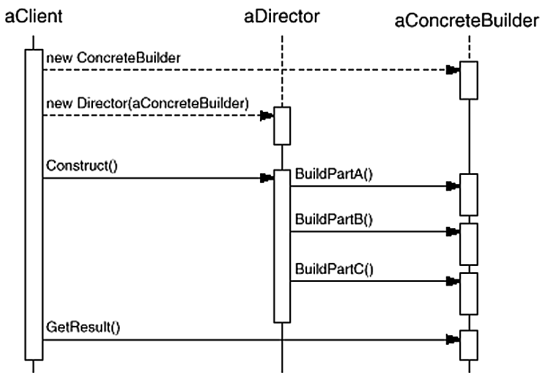


Fig. 4. Sequence diagram

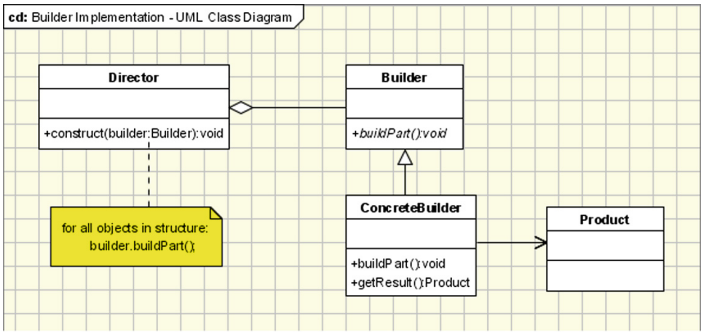


Fig. 5. Class diagram

3.2 Analysis Pattern Templates

Given below is the Analysis Pattern template described by Fernandez and Liu in their article "The Account Analysis Pattern" [20]. This template is also described in the book "Pattern-Oriented Software Architecture" [21] (Tables 3 and 4).

Table 3. Analysis pattern template [19, 21]

Analysis pattern template	
Elements	Description
Pattern name	It describes the name for referring to the pattern and also other names if the pattern has another name
Intent	It is a short statement that answers many questions like what does that design pattern do? What is the main goal of the pattern and what are the particular design issues or problems solved by the pattern
Example	Provides a real world example, which shows an existing problem and exemplifies the need of the pattern
Context	The section context is a fundamental component of a pattern. It provides an indication of the applicability of a pattern
Problem	It defines the recurring problem that is solved by the general solution. Problem is a fundamental component of a pattern because it is the reason for the pattern. The problem which is addressed by the pattern is described in this section
Solution	The solution details the participating entities in the solution, the collaborations between them and their behavior
Example resolved	Example Resolved gives the solution of the given example
Know uses	Know Uses is the example of the use of the pattern in a real system. It includes a minimum of two examples from different domains
Consequeses	It details the benefits that a pattern can offer and any possible restrictions
Related patterns	Related patterns are described what are the closely related patterns to this given pattern? What are important differences? With which other patterns should this one be used?

Table 4. Account analysis pattern [19, 22]

Account analysis pattern template	
Elements	Description
Pattern name	Account Analysis Pattern
Intent	The Account pattern keeps track of accounts of customers in institutions. These customers can perform transactions of different types against the accounts
Example	Consider a banking institution, where customers have accounts of different types, e.g., checking, savings, loan, mortgage, etc. For the convenience of their customers, the bank may have several branches or offices located in different places
Content	There are many institutions, e.g., banks, libraries, clubs, and others, that need to provide their customers or members with convenient ways to handle financial obligations, charge meals, buy articles, reserve and use materials, etc.
Problem	Without the concept of account users need to carry large amounts of cash, may have trouble reserving items to buy or borrow, and would have serious problems sending funds to remote places
Solution	Start from class Account and add relevant entities; in this case customers, cards, and transactions. Build an institution hierarchy describing the branches of the institution and relate accounts to the branches Fig. 6
Example resolved	An example for Bank accounts is shown in Figure. The classes contained in the model include Bank, BranchOffice, Account, CheckingAccount, Customer, BankCard, TransactionSet (TXSet), and Transaction, with their obvious meanings. Class TXSet collects all the transactions for a user on his account for a given period of time. There are, of course, other types of accounts Fig. 7
Know uses	The following are examples of uses of this pattern: • Banks, where customers have financial accounts of different types • Libraries, where patrons can borrow books and tapes • Manufacturing accounts, where materials are charged
Consequneses	The pattern has the following advantages: • It is clear that this model provides an effective description of the needs and can be used to drive the design and implementation of the software system. Not using a similar model would result in code that is hard to extend and probably incorrect • One can easily add other use cases: freeze account and activate/deactivate account The liabilities of this pattern come from the fact that to limit the size of the pattern and to make it more generic we have left out: Different types of customers. Each variety of customers could be handled in a special way.
Related patterns	Accountability pattern

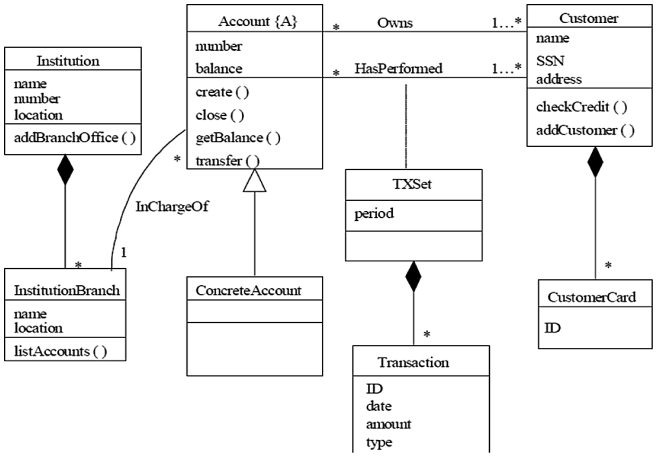


Fig. 6. Class diagram

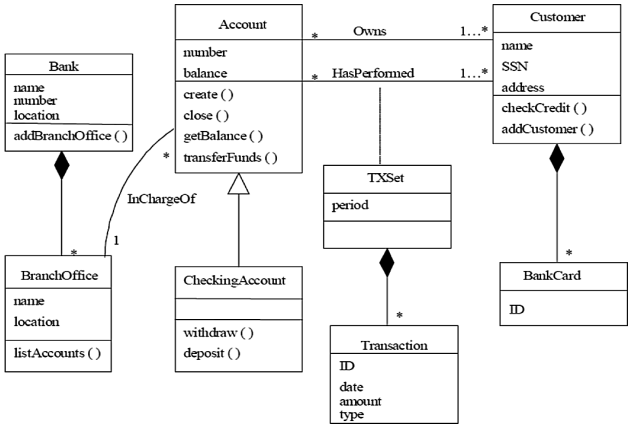


Fig. 7. Class diagram

3.3 Architecture Pattern Templates

This template was described by Ayodele Oluyomi in his article "Patterns and Protocols for Agent-Oriented Software Development" for the Agent internal Architecture-Structure Patterns [22] (Table 5).

Example

The pattern description has been slightly abbreviated for readability issues (Table 6).

Table 5. Architecture pattern template [19, 21]

Architecture Pattern Template	
Elements	Description
Name	A brief summary of the pattern
Problem	It defines the problem which a pattern can solve
Context	Different kinds the circumstances in which a pattern can be applied
Forces	It contains description of various forces and constraints that can affect the desired objectives
Solution	This section describes the different part of the pattern and their relation
Known uses	Know Uses are examples of the use of the pattern in real system. We include minimum two examples from different domains
Result context	This section description of possible effects on the initial context when the solution is applied and also the resulting advantages and disadvantages
Related pattern	Related patterns are described; What are the closely related patterns to this given pattern? What are important differences? With which other patterns should this one be used?

Table 6. Agent as delegate pattern [19, 22]

Elements	Description
Name	Agent as delegate
Classification	Multiagent system architecture-definitional
Problem	How should the role of a user be converted to an agent or agents in an agent based system while maintaining confidentiality of user information?
Context	A user role carries out activities in a system where confidentiality of user information is critical
Forces	Goals: to achieve optimum performance and maximize gains by taking decisions based on outcome of activities carried out Responsibilities: the responsibilities of this role involve carrying out both non trivial operational tasks and making concluding decisions based on the execution of the tasks carried out. User specific information is used in making the decisions. However, the user information should not be included in the execution of the operational tasks for security and confidentiality reasons
Solution	This pattern describes an approach for translating a role into agents. It prescribes translating a complex user with sensitive data into two types of agents which are User Agent and Task Agents. The pattern specifies the relationship and control that should exist between these two types of agents
Known uses	Know Uses section mentions the use of the pattern in various scenarios. It includes minimum two examples from different domains
Result context	The interaction between the assistant agent and the task agents has to be analyzed, modeled and implemented Adaptation/Integration: a user role can be translated into more than one assistant agents depending on the complexity and volume of the user information and decision making process
Related pattern	Agent as mediator

3.4 Data Model Patterns

Data Model Patterns help modelers to develop quality models by standardizing common and well-tested solutions for reuse [3]. The objective of data model pattern is to explain a starting point for data modelers [23].

A data model pattern can be implemented by adding additional attribute to any entity in a model or by adding a new entity or a relationship to an existing model. David Hay presented a Universal Data Model in [24]. It is a theoretical model which explains the basic principles of a data model pattern. See Fig. 8:

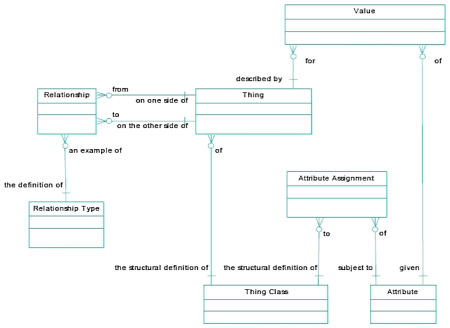


Fig. 8. Universal data model [19, 23]

In [38], Hay mentioned conventions for building data models. These conventions are guidelines for creating new patterns. They help in establishing a framework which data modelers can follow to reuse data model patterns. The modeling conventions are divided into three levels; Syntactic Conventions, Positional Conventions and Semantic Conventions [19].

Syntactic Conventions: This is the first type of conventions of modeling and deals with the symbols to be used. In the process of syntactic convention evaluation, the crucial point to remember is that there are two audiences in data modeling. The first audience is that community of users which use the models and their descriptions for the verifying whether or not the environment and requirements are actually understood by the analysts. The set of systems designers is the second audience. They make use of the rules of business as implied by the models to be the basis on which their design of computer systems is based [25].

Positional Conventions: This is the second type of Data modeling convention and dictates how entities of a model are laid out. They are concerned with the organization of elements and the overall structure of a model [25].

Semantic Conventions: Semantic conventions are those conventions that address the question of how can the meaning of a model be conveyed. These conventions help to represent common business scenarios in a standard way [25].

3.5 Template of Content Ontology Design Patterns

Ontology design patterns are similar to software design patterns. The core idea of describing software design patterns is to use a template and collect them by means of a catalogue. In order to describe ODPs we can use a similar approach as used in software engineering but the difference is that, the template used for the presentation has been optimize for the web and defined in an OWL annotation schema. It is the same used on the semantic web portal http://www.ontologydesignpatterns.org. This part contains a template of Content ODPs which is composed of the following information fields, defined in the annotation schema [2, 5, 26] (Table 7):

Table 7. Content ontology design pattern template [2, 5, 19]

Elements	Description
Name	It contains the name of the pattern. The names of patterns should be descriptive and unique names that help in identifying and referring to the patterns
Submitted by	This part of the template includes author names. In the portal it gives the link to the author page
Also known as	It gives the alternative names for the ODP, since it might be possible that the pattern has some other name but this part is not compulsory
Intent	This part of the template describes the goal of the ODP. Intent is a description of the goal behind the pattern and the reason for using it
Domain	This part of the template concerned with the area, domain and where the ODP is applicable
Competency question	It contains a list of competency questions expressed in natural language that are covered by the pattern. A competency question is a classical way of capturing a use case. A competency question is a simple query which an ontology engineer can submit to knowledge base to perform a certain task
Solution description	It describes how the given pattern provides the solution to a design problem in a certain context
Reusable OWL Building Block	It is a reusable representation of the pattern. This part is basically the implementation of the design pattern. It contains the URI of the OWL implementation of the content pattern, i.e. the reusable component available for download
Consequences:	This part of the template contains a description of the benefits and/or possible trade-offs when using the ODPs
Scenarios	Giving examples or scenarios where the given pattern implemented
Known uses	This part of the template gives examples of real ontologies where the ODP is used. This part of template is the example of real usages of the pattern
Other references	This part of the template contains references to resources (e.g. papers, theories, and blogs) that are related to the knowledge encoded in the ODPs

(*continued*)

Table 7. (*continued*)

Elements	Description
Examples	This field contains a link of an example owl file which is reusable. The example owl file presents a possible scenario which may sometime also include a UML diagram of classes and their relationships
Extracted from	Contains the URI (if any) of the ontology from which the pattern has been extracted
Reengineered from	It contains the name of the reference ontology which has been used reused in the pattern
Has components	This field refers to components of the Content ODP which are in turn ODPs themselves
Specialization of	This part of the template refers to ontology elements or ODPs. The specialization relation between ontology elements of ODPs consists of creating subclass of some ODP class and/or sub properties of some ODP properties
Related ODP	This part contains the names of the patterns which related to the current pattern based on generalization, specialization or composition. It also mentions other patterns that are used in corporation with the current pattern
Elements	This part of the template describes the elements (classes and properties) included in the ODP, and their role within the ODP
Diagram representation	This part of the template depicts a graphical representation of the ODP
Additional information	In Additional information authors provided that informatuion which is not avalible in the rest of the template

Example
(See Table 8).

Table 8. Classification pattern [4, 19]

Elements	Description
Name	Classification
Submitted by	ValentinaPresutti
Also known as	
Intent	To represent the relations between concepts (roles, task, parameters) and entities (person, events, values), which concepts can be assigned to. To formalize the application (e.g. tagging) of informal knowledge organization systems such as lexica, thesauri, subject directories, folksonomies, etc., where concepts are first-order elements
Domain	General
Competency question	• What concept is assigned to this entity? • Which category does this entity belong to?

(*continued*)

Table 8. (*continued*)

Elements	Description
Solution description	
Reusable OWL Building Block	http://www.ontologydesignpatterns.org/cp/owl/classification.owl
Consequences:	It is possible to make assertions about e.g., categories, types, roles, which are typically considered at the meta-level of an ontology. Instances of Concept reify such elements, which are therefore put in the ordinary domain of an ontology. It is not possible to parametrize the classification over different dimensions e.g., time, space, etc.
Scenarios	Mac OSX 10.5 is classified as an operating system in the Fujitsu-Siemens product catalog
Known uses	
Web references	
Other references	
Examples	
Extracted from	http://www.loa-cnr.it/ontologies/DUL.owl
Reengineered from	
Has components	
Specialization of	
Related ODP	
Elements	• **Concept** (owl:Class). A concept is a Social Object. The classifies relation relates concepts to entities at some time • **Entity** (owl:Class). Anything: real, possible, or imaginary, which some modeller wants to talk about for some purpose • **classifies** (owl:ObjectProperty). A relation between a Concept and an Entity, e.g. the Role 'student' classifies a Person 'John' • **is classified by** (owl:ObjectProperty). A relation between a Concept and an Entity, e.g. 'John is considered a typical rude man'; your last concert constitutes the achievement of a lifetime; '20-year-old means she's mature enough'
Diagram representation	Fig. 9
Additional information	

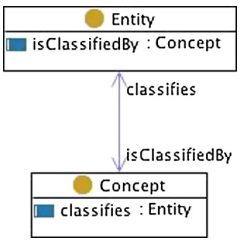


Fig. 9. Class diagram

4 Study Analysis and Results

The work starts with the literature review. Our first task was to study the current patterns that exist in other fields and compare them to ontology design patterns' templates to analyze the difference.

4.1 Comparison of Pattern Templates

The basic knowledge of problems is in software engineering modeled by conceptual models that are known as Analysis patterns. The patterns could be illustrated using a UML notation [3]. For Example the Analysis Pattern is implemented by using a UML Class diagram, Sequence Diagram and State Diagram which were described in section. Looking at an example of an analysis pattern, one can easily gauge the above relationship between ontologies and Software Patterns, and other patterns used in Computer Science. This similarity is evident even in, for example, graphical representations of an Analysis Pattern and an ontology pattern [3].

4.2 Comparison of Templates of Software Patterns and Content ODPs

This comparison is between the templates of software patterns and content ontology design patterns. We have described the elements of both software pattern templates and content ODPs in this chapter in Sects. 3.1 to 3.5, which provide us with a good starting point for comparison. There are several types of software patterns that are available and several techniques to present them but we selected Analysis Patterns, Architecture Patterns and Design Patterns. The template of a pattern is a standard way of representing a pattern. In a broad sense, a pattern template has four essential elements. These elements are: Name, Problem, Solution and Consequences as describe in Sect. 3.

This comparison is based on similarities and differences between the Content ODP template and software pattern templates. We compare each element of the ODP template with software pattern templates. Comparison is based on the names, content and the overall presentation of the template.

The parts described in Table 9 are considered as basic parts of a pattern. The description of these parts is stated. The table shows how these basic parts are described in software patterns and ODP templates.

Context: In Software Patterns, the section context describes the situations where the given pattern is applied. Every pattern has a context based on its application area. In content ODPS, context is defined in the form of domains and scenarios where the pattern is applicable.

Problem: For Analysis Patterns, Design Patterns and Architectural Patterns, the section Problem or Motivation describes the problem that can be solved by implementing these patterns. In analysis patterns, the section problem describes some generic use cases. In Design Patterns, some of the patterns use scenarios for giving more description of patterns. Such description should be able to clarify the details of the problem. In ODPs, the Problem is described by Competency Questions. Competency Questions consists of a list of competency questions expressed in natural language that

Table 9. Representing basic elements of other patterns [19]

Pattern type	Context	Problem	Solution	Consequence	Example
Ontology design pattern	Domain	Competency question	OWL Building Block, Element, Solution description and Graphical Representation	Consequence	Scenario, Example (OWL file)
Analysis pattern	Context	Problem	Solution	Consequence	Example, example resolved
Design pattern	Applicability	Motivation	Structure, Participant Collaboration, implementation and sample code	Not provided	Sample code
Architecture pattern	Context	Problem	Solution	Not provided	Not provided

are covered by the pattern. The section Competency Question describes the problem which is to be solved by the ODP. Competency questions are the requirements that an ontology should fulfill.

Solution: In Analysis Patterns, Architecture Patterns and Design Patterns, the section Solution gives a fundamental solution principle to be used in the pattern for solving a problem. In Analysis Patterns, the solution is described by using UML diagram and also a brief description of different parts of the pattern is given. In Design Pattern, the sections structure, participants and collaboration describe the solution. The section Structure is a graphical representation of the classes in the pattern which use the notation of the object modeling technique (OMT). It also uses interaction diagrams to illustrate the sequence and collaboration between the objects. The section participants give a detailed description of the classes and objects and their responsibilities. The section collaboration describes how the participants collaborate to carry out their responsibilities. In Architecture pattern, the section solution gives the description, using text and a graphical representation, as to how we can achieve the intended goals and objectives. It also describes the variants and specializations of the solution. It gives the description of people and computing actors and their collaboration. In ODPs, the section solution consists of four parts: OWL Building Block, Elements, Solution description and Graphical Representation. The section Solution description describes how a given pattern can solve the problem in the context. The section OWL Building Block contains an OWL ontology with reusable classes and reusable properties. The section Elements briefly describes Classes and Properties of the pattern implementation and the role of these classes and properties within the ODP. The Graphical representation gives a visual presentation of the pattern classes and their relations.

5Consequence: In both software patterns and content ODPs, the section consequence describes the possible benefits and limitations on the solution after using the pattern. What are the results of using the pattern? In some ODPs it is more about unexpected consequences and limitations.

Example: For Analysis Patterns, Design Patterns and Architecture Patterns, the section Example gives the real world example, which shows the existence of problems and needs for the patterns. For Analysis Patterns and Design Patterns the section example consists of two parts. In the first part the problem and in second part the implementation. In content ODPs, example is given in the form of a scenario and an example OWL file which is the OWL implementation of the scenario.

Table 10 describes the common elements of the template of content ontology design patterns and software patterns (analysis patterns, design patterns and architecture patterns). These elements are described in the background chapter with details in Sect. 2.3.2. The section motivation of design pattern is similar to the section problem of other software patterns.

In Table 10 the vertical line had different columns that shows the label of different pattern. The horizontal line shows the template heading and symbol X describes the presence of common element.

Table 10. Representing common elements of other patterns [19]

Template heading	Ontology design pattern	Analysis pattern	Design pattern	Architecture pattern
Name	×	×	×	×
Known uses	×	×	×	×
Intent	×	×	×	×
Consequences	×	×	×	
Also known as	×		×	
Classification			×	×

The template of content ODPs has developed by following the template of software design patterns. The comparison of both patterns reveals that both patterns have a lot of similarities and the current template of content ODPs includes all the elements of software patterns except 'forces'. Forces define constraints and problems that can affect the solution. In content ODPs, only the benefits and/or possible trade-offs when using the ODPs on the initial problem are mentioned. Apart from the software patterns, the content ODPs has some additional parts which have been added according to the pattern requirements.

Unique Sections: Content ODPs have some unique sections, which are not described in other Software Pattern Templates. These sections are: EXTRACTED FROM, REENGINEERING FROM and HAS COMPONENTS, as mentioned in the background chapter in Sect. 2.5.

4.3 Comparison of Data Model Patterns and Ontology Design Patterns

Data model patterns and ODPs are different to each other because data model patterns are presented only in graphical form while ODPs have much more detailed graphical and textual description. While content ODPs have an official catalogue where users can select from a list of patterns, there is no official catalogue for data model patterns.

The template of an ODP has a description of different parts of the pattern which is presented in the graphical and textual form. To implement an ODP, an ontology engineer has to study the description to understand a pattern. A data model pattern is presented in the form of an UML diagram. The diagram is built by following conventions which are a set of rules. These conventions standardize a data model hence makes it easier for a data modeler to reuse a pattern.

The reusability of both patterns depends on different factors. Content ODPs can also be used directly as they are, just like data models, although they are usually specialized but this depends on their generality. In data model patterns, it is up to the data modeler to decide whether to use a real model to make some minor adjustments or to use a completely abstract model of a problem.

The similarities of data model patterns and ODPs lie in the graphical representation of a problem. Also as we saw in Table 11, different parts of both patterns can be mapped to each other hence knowledge from data model patterns can be translated into ODPs. The use of conventions in data model patterns makes it easier to understand the diagram. These conventions and practices can be translated into guidelines for creating more uniform diagrammatic notations for ODPs.

Table 11. Mapping of elements of data model pattern and ontology design pattern [19, 22]

Data model pattern	Ontology design pattern
Entity	Concept
Attribute	Relation to "attribute"
Subtype/supertype	Subsumption hierarchy
Relationships	Relations
Mutually exclusive sets	Disjoint concepts

Overall, the comparison of software patterns and data model patterns with content ODPs shows that there are a lot of similarities in the templates of software patterns and content ODPs. The difference lies in the presentation of the content. The graphical representation of content ODPs can be made uniform by defining conventions as it is done in data model patterns.

In an experiment, the knowledge from data model patterns was translated into ontology design patterns by mapping the parts using the KAON tool [2]. Table 11 shows the mapping between the different parts. The above mapping shows that the knowledge stored in the data model patterns can be translated into ODPs. This knowledge can be reused to create new ODPs from existing data model patterns.

5 Conclusions

The purpose of this study was to improve the presentation of content ontology design patterns. This research is part of an effort to solve the larger problem of engineering high quality ontologies. One possible solution to this problem is to introduce reuse in ontology engineering which can standardize the ontology development process and

reduce the time and effort involved in it. ODPs are considered best practice to achieve this objective. To encourage the use of ODPs for ontology development, their presentation must be explicit and precise.

Suggestions are based on the comparison of the different patterns with the ontology design patterns. Based on the literature review, we propose following improvements in the current template for content ODP:

The Graphical Representation should have a more uniform diagrammatic notation. This can be done by defining standards or conventions which ontology engineers can follow while creating patterns. As guidance, the conventions used in data model patterns can be studied to define similar conventions for content ODPs. Also, namespaces provided in the Graphical Representation section should be made explicit in the Additional Information section.

Scenarios should be more explicit. While scenarios presented in the General Description section are simple, they can be more complex in general. Further, the scenario section should describe the binding from concrete elements to the pattern elements. At ODP portal, scenarios of several patterns are missing. They must be included in each pattern because they were considered as important parts. Software patterns include the implementation section with the solution of complex problem. Implementation help user to understand the pattern.

This research is part of an effort to solve the larger problem of engineering high quality ontologies. One possible solution to this problem is to introduce reuse in ontology engineering which can standardize the ontology development process and reduce the time and effort involved in it. ODPs are considered best practice to achieve this objective. To encourage the use of ODPs for ontology development, their presentation must be explicit and precise.

6 Future Works

This study will be improved through experiment. Determine how well the current ODP template supports the understanding and usage of Content ODPs. To get experts opinions on the current structure of the Ontology design pattern template. There may be certain information that an ontology user need to understand a pattern but it is not available in the description. This research will improve by examine the missing information in the current ontology design pattern templates.

References

1. Uschold, M., Gruninger, M.: Ontologies and semantics for seamless connectivity. ACM SIGMOD Record **33**(4), 58–64 (2004)
2. Blomqvist, E.: Fully automatic construction of enterprise ontologies using design patterns: initial method and first experiences. In: OTM Confederated International Conferences On the Move to Meaningful Internet Systems. Springer (2005)
3. Gruber, T.R.: A translation approach to portable ontology specifications. Knowl. Acquis. **5** (2), 199–220 (1993)

4. Presutti, V., Gangemi, A.: Content ontology design patterns as practical building blocks for web ontologies. In: International Conference on Conceptual Modeling. Springer (2008)
5. Blomqvist, E.: State of the Art: Patterns in Ontology Engineering (2004)
6. Gangemi, A., Presutti, V.: Ontology design patterns. In: Handbook on Ontologies, pp. 221–243. Springer (2009)
7. Pree, W.: Meta patterns—a means for capturing the essentials of reusable object-oriented design. In: European Conference on Object-Oriented Programming. Springer (1994)
8. Doran, P.: Ontology reuse via ontology modularisation. In: KnowledgeWeb Ph.D. Symposium. Citeseer (2006)
9. Prabhu, V., Kumara, S., Kamath, M.: Scalable Enterprise Systems: An Introduction to Recent Advances, vol. 3. Springer Science & Business Media, New York (2012)
10. Rebstock, M., Janina, F., Paulheim, H.: Ontologies-Based Business Integration. Springer Science & Business Media, Heidelberg (2008)
11. Guarino, N.: Semantic matching: formal ontological distinctions for information organization, extraction, and integration. In: Information Extraction a Multidisciplinary Approach to an Emerging Information Technology, pp. 139–170. Springer (1997)
12. Fernandez, E.B., Yuan, X.: An analysis pattern for reservation and use of reusable entities. In: Proceedings of PLoP (1999)
13. Buschmann, F., et al.: Pattern-oriented software architecture: a system of patterns (1996). Part II, 2001
14. Kampffmeyer, H., Zschaler, S.: Finding the pattern you need: the design pattern intent ontology. In: International Conference on Model Driven Engineering Languages and Systems. Springer (2007)
15. Cooper, J.W.: The Design Patterns Java Companion, vol. 218. Addison-Wesley, Upper Saddle River (1998)
16. Devedzic, V.: Ontologies: borrowing from software patterns. Intelligence **10**(3), 14–24 (1999)
17. Gamma, E.: Design Patterns: Elements of Reusable Object-Oriented Software. Pearson Education India, Delhi (1995)
18. Fernandez, E.B.: Building systems using analysis patterns. In: Proceedings of the Third International Workshop on Software Architecture. ACM (1998)
19. Lodhi, S., Ahmed, Z.: Content Ontology Design Pattern Presentation (2011)
20. Fernandez, E.B., Liu, Y.: The account analysis pattern. In: EuroPLoP (2002)
21. Buschmann, F., et al.: Pattern-Oriented System Architecture: A System of Patterns, pp. 99–122. Wiley, Chichester (1996)
22. Oluyomi, A.O.: Patterns and protocols for agent-oriented software development (2006)
23. Silverston, L., Inmon, W.H., Graziano, K.: The Data Model Resource Book: A Library of Logical Data Models and Data Warehouse Designs. Wiley, Hoboken (1997)
24. Hay, D.: Data Model Patterns Conventions of Thought. Dorset House, New York (1996)
25. Hay, D.C.: A comparison of data modeling techniques (1999)
26. Gangemi, A.: Ontology design patterns for semantic web content. In: International Semantic Web Conference. Springer (2005)

Biomedical Applications

MainIndex Sorting Algorithm

Adeel Ahmed[1], Muhammad Arif[1(✉)], Abdul Rasheed Rizwan[2], Muhammad Jabbar[1], Zaheer Ahmed[1], and Muhammad Sami Ullah[1]

[1] Department of Computer Science, University of Gujrat, Gujrat City, Gujrat, Pakistan
{m.arif, adeel.ahmed, m.jabbar, zaheer, msamiullah}@uog.edu.pk
[2] Department of Computer Science, Virtual University of Pakistan, Lahore, Pakistan
liveneeo@gmail.com

Abstract. Sorting algorithm remained hot topic in computer science from the birth of computer science to achieve maximum performance. Fortunately this achievement became possible due to good and fast sorting algorithms, such as heap sort, merge sort, radix sort and other sorting algorithms. Till this achievement is also under research to find more efficient algorithms. In sorting algorithm arrays and link list data structures are commonly used. We know arrays are efficient if we need consecutive kind of data structure and link lists are useful when we need to add and remove items in the data structure. In other word we can say both data structures have own its merits and demerits. So in our sorting algorithm we are going to use both kinds of data structure. We will use in our MainIndex sorting algorithm arrays as the MainIndex and link list as sorting cells. MainIdex sorting algorithm need some kind of information just the length of the number which is going to sort and the value of the number which is going to sort in sorting cell.

Keywords: MainIndex · Sorting cell · Selection sort · Link list · Merge sort · Bubble sort · Quick sort

1 Introduction

A sorting algorithm exchanges the numbers in a specific order according to the fit needs of the users. Any sorting algorithm has merging and searching functions. Some other exist which used the functions to sort the numbers [1]. The data is mostly taken from an array which is faster in sense of random access [2]. Link list provide a fast method when we need to add or delete node for numbers against arrays [3]. Sorting algorithms remain hot topic since the birth of computer due to time complexity. Time complexity is the total time required by an algorithm of the program to complete its given task [4]. The time complexity is measurement by big O notation of the algorithm. The numbers of elements are measured to compute the complexity time. Sorting algorithms are the fundamental concepts for a computer programmer, computer scientist and IT

V.E. Balas et al. (eds.), *Soft Computing Applications*, Advances in Intelligent Systems and Computing 633, DOI 10.1007/978-3-319-62521-8_20

professionals [5]. When we search in database mostly used ORDER BY clause, using this query actually we are using the sorting algorithm. A common example is our phone book of address, we save our contact numbers in the mobile in a different order, but when we need to find a contact they are listed into a sequence [6]. First look we should compare various sorting algorithms in dealing with large set of data, in large set of data some algorithms become very slow in practical use. Next look is about the memory usage of the algorithms, mostly faster algorithms takes large memory rather than slow one algorithm [7]. If we look about the progress in the memory usage of today, large memory is not main issue, but issue is faster algorithms [8].

But we even imagine which algorithm is fastest; the speed of the algorithm depends on the working and environment where the sorting is done. In sense which kind of data is being to sort? We cannot use the same algorithm for sorting 200 numbers on such kind of database which cannot fit into memory, so such kinds of algorithms are designed in different ways.

2 Related Works

Today many programming languages are providing built in functionalities which are usable for sorting numbers [9]. Java API, .Net framework and C++ all provides built-in sorting techniques [10]. For different kinds of data types such as floating numbers, integer number and characters need to sort according to the requirements. But sometime we need a generic kind of sorting algorithms for generic data structure [11]. This thing leads us to a complicated problem. But OOP provides us to solve such kinds of problems [12]. So many good algorithms are designed but no one algorism is Silver Bullet. At this point we see the complexity of five algorithms. As given in the Table 1.

Table 1. Complexity of the algorithms

Sorting algorithm	Worst case	Average case	Best case
Bubble sort	O(n2)	O(n2)	O(n)
Insertion sort	O(n2)	O(n2)	O(n)
Merge sort	O(n log n)	O(n log n)	O(n log n)
Heap sort	O(n log n)	O(n log n)	O(n log n)
Quick Sort	O(n2)	O(n log n)	O(n log n)

2.1 Bubble Sort

The bubble sort algorithm is a well-known sorting algorithm. Its belongs to O (n2) sorting algorithm. But bubble sort is not very efficient for sorting of large amount of data, other hand it is stable and adaptive algorithm [13].

```
Bubble Sort Algorithm
for i = 1 to n;
    swapping = false
  for j =  n: i+1;
   if a [j] < a [j-1];
      swap a[j,j-1];
      swapping = true;
if not swapping;
  break;
end
```

2.2 Insertion Sort

Insertion sort is elementary sorting algorithm and its time complexity in wore case is O(n2). Insertion Sort is a good choice when data is nearly related to each other due to adaptive algorithm. But it is not good when data size is small. Insertion Sort is recursive by nature when the problem size is small [14].

```
Insertion Sort Algorithm
for i = 2 to n;
   for (k = i; j > 1 and a[j] < a[j-1]; j--);
      swap a[j,j-1];
end
```

2.3 Merge Sort

Merge sort bases on divide and conquer algorithm. It divides the problem into two parts and sorts them, after sorting them it merges them. Merge Sort is best choice for different kinds of environments for example when sorting with link list, stability and fast result is required. But it is expensive when the problem is in array [15].

2.4 Heap Sort

Heap Sort algorithm bases on binary heap. In some sense it is similar to selection sort where we find the maximum value in start and rearrange it at the end; we repeat the same process for remaining numbers in the problem [16].

2.5 Quick Sort

Quick Sort also works in Divide and Conquer method just as Merge sort. It takes a number as pivot and make the partition according to the selected pivot. Selecting a pivot is grand work in this method. Different kinds of ways are to select the pivot.

- Pick first number as pivot.
- Pick last number as pivot.

- Pick random number as pivot.
- Pick middle number as pivot [17].

3 Problem Statement

We took an overview on different kinds of sorting algorithms and saw each algorithm is using a specific data structure such as arrays or link list. So every sorting algorithm has own its pros and cons, some need more memory but they are sharp in sorting the numbers, some algorithm needs less memory but it takes more time to sort the numbers. Different sorting algorithms take different time to sort the numbers. Every algorithm behaves different according to the nature of the problem. We are going to discuss specific method which behave constant in every nature of sorting data and takes less time against all sorting algorithms. In Main Index sorting algorithm we will use both kinds of data structure such as link list and arrays according to their efficient point of view.

4 The Proposed Solution

Now we are going to discuss a new sorting algorithm, we have following unsorted array.

32, 18, 45, 56, 22, 64, 8, 88, 67, 105, 77, 440, 11, 90, 16

We have these unsorted serials of numbers in array. To sort them in order first of all we should compute the maximum length in unsorted numbers. We are not going to find the maximum number in unsorted numbers. In our unsorted list 32 is the first number, in this we need two kind of information number one the value of first digit which is 3 in case of 32. Second we need the length of 32 which is 2. The length 2 numeric numbers should store in LIndex2. Later we will see what mean of LIndexN. Before going in detail of MainIndex first we should take an over view of the sketch of the MainIndex. Following is the basic detail figure of the MainIndex. As shown in the Fig. 1.

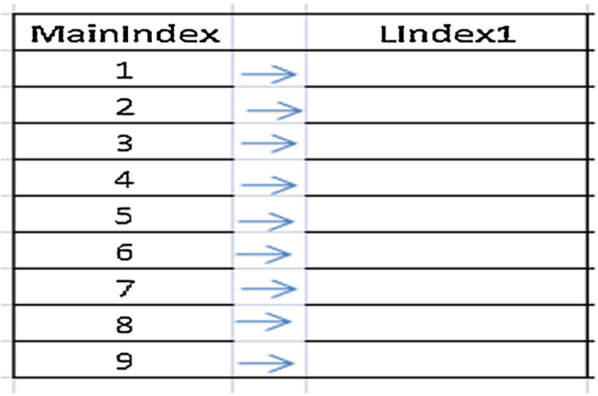


Fig. 1. Sketch of the MainIndex

4.1 MainIndex

MainIndex has a array from 1 to 9, this section will use appropriate position according to the value of the number, let's take 32 has example here again, 32 has first digit 3 so it should enter in array of 3 in MainIndex. The arrow is a pointer which is pointing to LIndex1. LIndex1 mean length of index 1 which is core structure with the MainIndex. All those numeric numbers which have its length one those should store in this index. Other LIndexN should create when a numeric number has different length against already created LIndexN. But LIndex1 should create automatically when MainIndex starts.

4.2 LIndexN

L mean length of the numeric number which is going to store in the sorting sequence, N is index number. Every LIndexN has its own sorting cells which should program in link list; we need compression of existing numbers with the new coming number, after compression it should store in its appropriate place. Internal structure of sorting cells is as following. It is given in the Fig. 2.

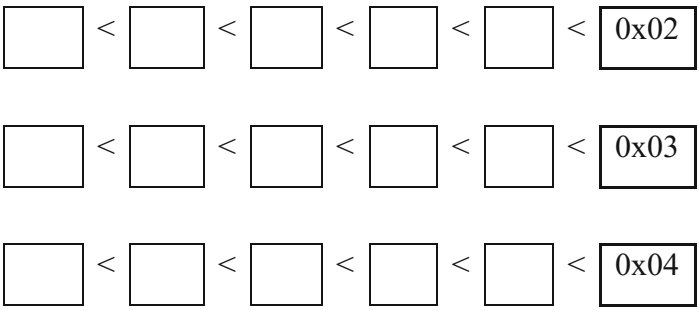


Fig. 2. Internal structure of the sorting cell

Above figure show that left sorting cells have the least number in its row, and the right side sorting cell has most number. 0x02, 0x03 and 0x04 are the addresses of next sorting cells. These addresses help us in scanning the sorted number to the next sorting cells. Now we are ready to sort the unsorted data according to our above discussion. Let's explain our method which is needed to sort the numbers.

Pick the first number and extract the following information.

- Value of the number by checking most left number.
- Length of the number.

The first number in our unsorted collection is 32, now we enter the number into MainIndex sorting algorithm with its value and length (Fig. 3).

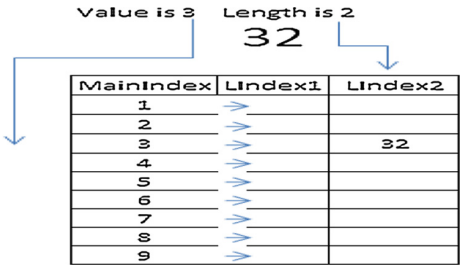


Fig. 3. Entering values in algorithm

Next 18 is the number in our unsorted array. Figure 4 shows the number 18 in the unsorted array and Fig. 5 shows the number 45.

Next number is 45.

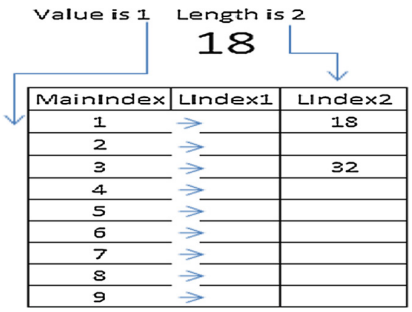


Fig. 4. Unsorted array with number 18

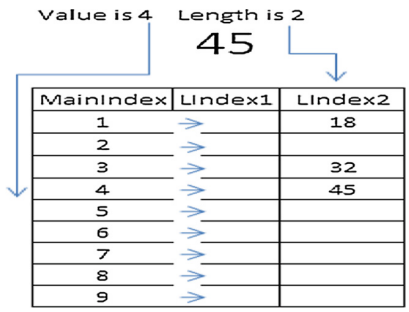


Fig. 5. Unsorted array with number 45

Next numbers are which have length two 56, 22, 64 we should fill these number according to above method. After filling the numbers Main Index should looks as depicted in the Fig. 6.

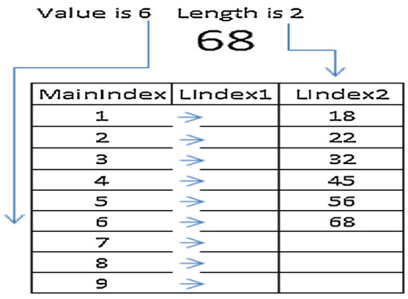


Fig. 6. Unsorted array with numbers of length two

Now the next digit is 8 which has length one, so it should store in LIndex1. It's given in the Fig. 7.

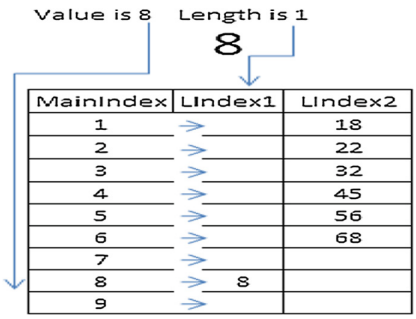


Fig. 7. Unsorted array with numbers of length one

Next numbers are 88, 67 these numbers should enter with length 2, after adding these two numbers. It's shown in the Fig. 8.

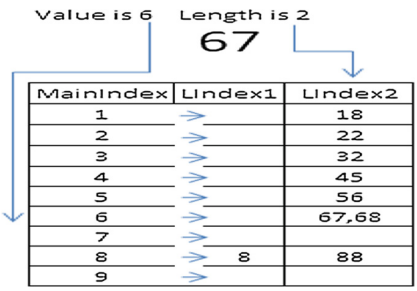


Fig. 8. Unsorted array with number 67

67 is less than already exists 68, so 67 should add before in sorting cell than 68. Now next number is 105 which have length three, which are not in our MainIndex, when the LIndexN is not available we should create it according to its length. If the new created LIndexN has less number already created LIndexN then we have to put the newly created before the LIndexN. We are working LIndexN with link list so it's easy to add a new list after and before of the LIndexN. Figures 9 and 10 shows the number with length of three.

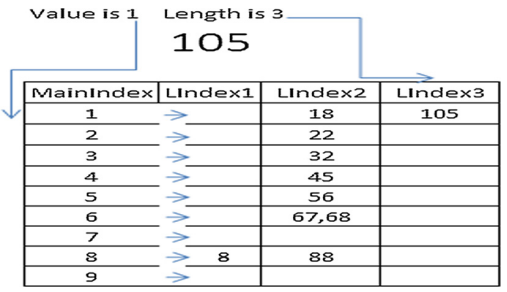


Fig. 9. Unsorted array with numbers of length three

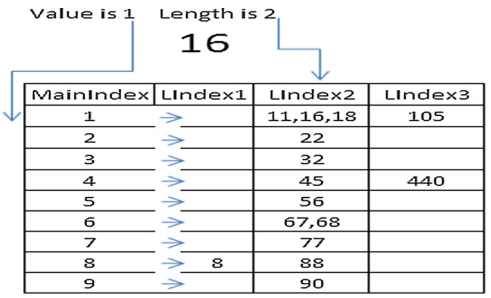


Fig. 10. Unsorted array with numbers of length three

Next numbers are 77, 40, 11, 90, 16 fill them.

Now we can see all numbers are entered into MainIndex according their appropriate place, now time to return these numbers from LIndex1 to LIndex3, keep in mind create LIndexN as they must be in sorted order. Now start scanning from LIndex1. If there is no number in the sorting cell its pointer should point to next sorting cell. The forwarding pointer must be in every sorting cell either it has a value or not. In LIndex 8 has first number next LIndex2 has 11, 16, 18 and LIndex2 next sorting cell has 22, next sorting cell has 32 and so on …

After completing above discus scanning we should see a sequence of the numbers. 8, 11, 16, 18, 22, 32, 45, 56, 67, 68, 77, 88, 90, 105, 440

So now a day's memory is not issue, the issue is how to improve more speed and performance of the system. We can see the other sorting algorithm Big O complexity in Fig. 11.

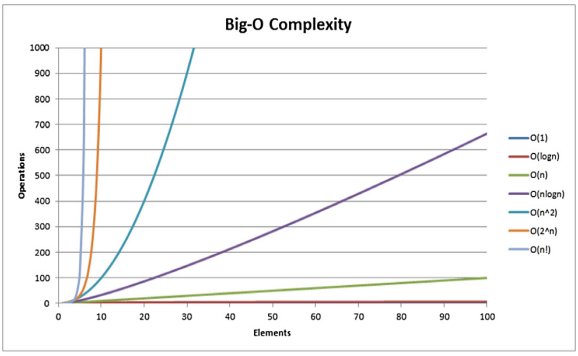


Fig. 11. Big O complexity of algorithms

5 Results and Discussion

MainIndex sorting algorithm efficiently works on unsorted orders numbers, we took an example of positive numbers, so if their negative numbers then we have to work in a different way, if we having the numbers are negative then we should check the number with its value to MainIndex number. If the value of the number is less than MainIndex target cell number. Then it should add in left side of the MainIndex, if it is positive number is should add in right side just we done in our example. MainIndex sorting algorithm work better in all kinds of sorting algorithm, just we discussed in the beginning more memory taking algorithms works better against such kinds of algorithms which takes less memory. MainIndex sorting algorithm required extra memory to map the LIndex1, LIndex2 and so on. The extra feature required memory to store the values in sorted order. The memory section is divided into different length Index to store the different numbers. So now a day's memory is not issue, the issue is how to improve more speed and performance of the system. we can see the other sorting algorithm Big O complexity. Lets see the pseudo code of MainIndex sorting algorithm.

```
For each number as integer in numberslist{
Firstvalueofnumber = leftnumber(number. (1)
Numberlength = number.length
```

// in above pseudo code for loop is used to input all the number which are in unsorted array. So the loop should continue up-to the total numbers of unsorted. So the loop should take O (n) time to process the numbes. Firstvalueofnumber is integer constant which is used to store the leftnumber of the received number from the unsorted array just if the number is 34, the firstvalueofnumber is 3 here. Firstvalueofnumber is a constant time operation. Next numberlength is also a constant time operation O (1) time. So the total time is required to peform this operation is O (n) time is required. //

```
If (numberlength ==1 and firstvalueofnumber ==1){
Void numberOneValueOne(number, firstvalueofnumber, numberlength)
}

// this section should execute on the time if given condition is true. This operation
also take O (1) time complete the execution and call the function.//

If (numberlength ==2 and firstvalueofnumber ==1){
Void numberTwoValueOne(number, firstvalueofnumber, numberlength)
}

// this section should execute on the time if given condition is true. This operation
also take O (1) time complete the execution and call the function.//

If (numberlength ==3 and firstvalueofnumber ==1){
Void numberThreeValueOne(number, firstvalueofnumber, numberlength)
}

// this section should execute on the time if given condition is true. This operation
also take O (1) time complete the execution and call the function.//

If (numberlength ==n and firstvalueofnumber ==1){
Void numberNValueOne(number, firstvalueofnumber, numberlength)
}

// this section should execute on the time if given condition is true. This operation
also take O (1) time complete the execution and call the function.//

If (numberlength ==N and firstvalueofnumber ==2){
Void numberNValueTwo(number, firstvalueofnumber, numberlength)
}

// this section should execute on the time if given condition is true. This operation
also take O (1) time complete the execution and call the function.//

If (numberlength ==N and firstvalueofnumber ==3){
Void numberNValueThree(number, firstvalueofnumber, numberlength)
}

// this section should execute on the time if given condition is true. This operation
also take O (1) time complete the execution and call the function.//

If (numberlength ==N and firstvalueofnumber ==N){
Void numberNValueN(number, firstvalueofnumber, numberlength)
}
// this section should execute on the time if given condition is true. This operation
also take O (1) time complete the execution and call the function.//

}
Void numberNValueN(number, firstvalueofnumber, numberlength){
While(number > (MainIndex.firstvalueofnumber, LIn-
dex.numberlength).linkedlist){
Repeatloop while number > if there are numbers are in specific location.
}
Add.newlinkedList(number)
}
```

// a while loop is used to check where the looped value should insert using linked list. It all depends on preexisting values in the linked list. If there is not any value then it should take O (1) time for the loop, if there are some values already there then should take m times to find the place to add the value. The current value may add in the middle of the linked list or might be possible to at start or at the end of linked list. So the calculated time for this operation is log (m) time. It is shown in the Table 2.

Table 2. MainIndex comparison with other algorithms

Sorting algorithm	Worst case	Average case	Best case
Bubble sort	O(n2)	O(n2)	O(n)
Insertion sort	O(n2)	O(n2)	O(n)
Merge sort	O(n log n)	O(n log n)	O(n log n)
Heap sort	O(n log n)	O(n log n)	O(n log n)
Quick sort	O(n2)	O(n log n)	O(n log n)
MainIndex	O(n log m)	O(n log m)	O(n log m)

6 Conclusions

MainIndex sorting algorithm work better in all kinds of sorting algorithm, just we discussed in the beginning more memory taking algorithms works better against such kinds of algorithms which takes less memory. MainIndex sorting algorithm takes extra memory but it works in a sharp and smart way. So now a day's memory is not issue, the issue is how to improve more speed and performance of the system. we can see the other sorting algorithm Big O complexity. MainIndex sorting algorithm efficiently works on unsorted orders numbers, we took an example of positive numbers, so if their negative numbers then we have to work in a different way, if we having the numbers are negative then we should check the number with its value to MainIndex number. If the value of the number is less than MainIndex target cell number. Then it should add in left side of the MainIndex, if it is positive number is should add in right side just we done in our example.

References

1. Mishra, A.D.: Selection of best sorting algorithm for a particular problem, Thapar University Patiala (2009)
2. Grover, R.S.: Programming with Java: A Multimedia Approach. Jones & Bartlett Publishers, Burlington (2011)
3. Drozdek, A.: Data Structures and Algorithms in C++. Cengage Learning, Boston (2012)
4. Puntambekar, A.A.: Data Structures and Algorithms. Technical Publications, Pune (2009)
5. Rolim, J.: Parallel and distributed processing: In: Proceedings of 15 IPDPS 2000 Workshops Cancun, Mexico, 1–5 May 2000. Springer (2003)
6. Cormen, T.H.: Algorithms Unlocked. Mit Press, Cambridge (2013)

7. Dale, N., Walker, H.M.: Abstract Data Types: Specifications, Implementations, and Applications. Jones & Bartlett Learning, Burlington (1996)
8. Karloff, H.: Proceedings of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms, vol. 95. SIAM (1998)
9. Sametinger, J.: Software Engineering with Reusable Components. Springer Science & Business Media, Heidelberg (1997)
10. Applegate, D., Brodal, G., Panario, D., Sedgewick, R.: Proceedings of the Ninth Workshop on Algorithm Engineering and Experiments and the Fourth Workshop on Analytic Algorithms and Combinatorics (2007)
11. Marciniak, J.J.: Encyclopedia of Software Engineering, vol. 2 OZ. Wiley-Interscience, Hoboken (1994)
12. Farrell, J.: Object-Oriented Programming Using C++. Cengage Learning, Boston (2008)
13. Beu, T.A.: Introduction to Numerical Programming: A Practical Guide for Scientists and Engineers Using Python and C/C++. CRC Press, Boca Raton (2014)
14. Mehlhorn, K.: Data Structures and Algorithms 1: Sorting and Searching, vol. 1. Springer Science & Business Media, Berlin (2013)
15. Deckard, A., Anafi, R.C., Hogenesch, J.B., Haase, S.B., Harer, J.: Design and analysis of large-scale biological rhythm studies: a comparison of algorithms for detecting periodic signals in biological data. Bioinformatics **29**, 3174–3180 (2013)
16. Cerri, S.A., Clancey, W.J., Papadourakis, G., Panourgia, K.K.: Intelligent tutoring systems. In: Proceedings of 11th International Conference, ITS 2012, Chania, Crete, Greece, 14–18 June 2012, vol. 7315. Springer (2012)
17. Cuesta-Infante, A., Colmenar, J.M., Bankovic, Z., Risco-Martín, J.L., Zapater, M., Hidalgo, J.I., et al.: Comparative study of meta-heuristic 3D floorplanning algorithms. Neurocomputing **150**, 67–81 (2015)

Deep Learning Tools for Human Microbiome Big Data

Oana Geman[1(✉)], Iuliana Chiuchisan[2], Mihai Covasa[1], Cris Doloc[3], Mariana-Rodica Milici[4], and Laurentiu-Dan Milici[4]

[1] Department of Health and Human Development, "Stefan cel Mare" University, 13 Universitatii Street, 720229 Suceava, Romania
oana.geman@usm.ro, mcovasa@gmail.com
[2] Computers, Electronics and Automation Department, "Stefan cel Mare" University, 13 Universitatii Street, 720229 Suceava, Romania
iuliana.chiuchisan@usm.ro
[3] Bio-Informatics Intelligence SRL, Chicago, USA
c.doloc@algomex.com
[4] Electrotechnics Department, "Stefan cel Mare" University, 13 Universitatii Street, 720229 Suceava, Romania
mami@eed.usv.ro, dam@usm.ro

Abstract. Deep Learning is a branch of Machine Learning, which focuses on a set of algorithms that model high-level abstractions in data by using a deep representation of multiple processing layers. The goal of Machine Learning is to map input patterns to output values. This paper will suggest a potential application of Deep Learning Algorithms for the analysis of large amounts of data produced by the research of the Human Microbiome. Humans have coevolved with microbes in the environment, and each body habitat has a unique set of microorganisms (microbiota). The most abundant and well-studied microbiota are found in the gut, where the bacterial density reaches 1011–1012 cells/g in the distal human colon. The number of bacteria in the human gut has been estimated to exceed the number of somatic cells in the body by an order of magnitude and that the biomass of the gut microbiota may reach up to 1.5 kg. This paper presents different methods that have been implemented and tested on a Human Microbiome Dataset. Besides the findings concerning accuracy and runtime, the results suggest that the Deep Learning algorithms could be successfully used to analyze large amounts of Microbiota data.

Keywords: Machine learning · Deep learning · Data Mining · Human Microbiome · Big Data

1 Introduction

To extract knowledge and meta-knowledge from Big Data in bioinformatics and biomedicine, Machine Learning has been successfully used. Machine Learning algorithms use training data to uncover underlying patterns, build models, and make pre-dictions based on the best fit model [1]. Indeed, some well-known algorithms (i.e., Support Vector Machines, Random Forests, Hidden Markov models, Bayesian net-works,

V.E. Balas et al. (eds.), *Soft Computing Applications*, Advances in Intelligent Systems and Computing 633, DOI 10.1007/978-3-319-62521-8_21

Gaussian networks) have been applied in genomics, proteomics systems, biology, and numerous other domains [2–8]. Deep Learning represents a category of automatic learning algorithms which are characterized by the fact that in its first stage it is learning how to process the input data and in the subsequent stages is processing the function.

Until very recently, for object recognition in images, researchers had to find all sorts of specific features of objects to come out of the images and then to apply an algorithm for classification such as SVM (Support Vector Machine) or Random Forest in order to determine which contains each image. The Deep Learning algorithm does not require the extraction of features about categories in advance because they learn how to make this in automatic mode [1].

The term "deep" comes from the fact that instead of having a single layer which receives the input data and produces the output, a series of levels are processing the data received from previous levels, picking the features of all higher level. The last level is the one that generates the result, after all the data has been processed and compressed.

The perceptron is a very simple computational model: the entry is multiplied by a weight and then all the results are added up. If the result is greater than a predefined, then the result is 1, otherwise is 0. The weights must be adapted for each data set. But this model has a major problem: it only operates for separable linear data. This is a limitation, which has led to the abandonment for a time of the neural networks [2]. In 1975, Paul Werbos has discovered the "backpropagation" algorithm, which enabled the use of multiple layers in the neural networks. Hornik has demonstrated theoretically that a neural network with a hidden layer and with a sufficiently large number of neurons is able to approximate any computational function, with an arbitrary precision. The algorithm operates in the following way: random initialization of all the weights of the neurons and you start to see what values would obtain for your data input; calculate the difference between the value of which must be obtained and for this you will do a backpropagation on the penultimate layer, by dividing in proportion the error to each neuron, depending on the weight of the neuron. This backpropagation shall be repeated until reaching the first level and then calculate the new values obtained. This cycle of correction is repeated until the error has been obtained on the experimental data is sufficiently low [3–5].

Deep-learning networks are distinguished from the more commonplace single-hidden-layer neural networks by their depth; that is, the number of node layers through which data passes in a multistep process of pattern recognition. Traditional machine learning relies on shallow nets, composed of one input and one output layer, and at most one hidden layer in between. More than three layers (including input and output) qualify as "deep" learning. So deep is a strictly defined, technical term that means more than one hid-den layer. In deep-learning networks, each layer of nodes trains on a distinct set of features based on the previous layer's output. The further you advance into the neural net, the more complex the features your nodes can recognize, since they aggregate and recombine features from the previous layer [7].

2 Human Microbiome Big Data

Intestinal microflora is a complex ecosystem and for a human body is an active cohabitant at the metabolic and the immune system levels, actively involved in the "metabolic inflammation" and in the energetic metabolism of the human. Microbiota or intestinal microbiome has a whole quantitatively genome superior the human genome and acts on the hormone level sending "messages" beyond the hematoencephalic barrier, influencing the essential features of our existence [9]. In the human body are 10 times more microbial cells than the human body cells, the number of microbial genes exceeding 150 times the human genes. All those entities that live in our body and on the surface it is called the "microbiome", that in terms of the modern science it is about an ecosystem. Knowledge of the "intestinal microbiome" and the interrelationships with the human body will bring major changes in the attitude of the therapeutic efficacy of many diseases, in nutrition, in pharmacology and in general in the biology of the human body.

Microbiome includes the intestinal bacteria, viruses, micelles, parasites. Some entities may be useful, others do not. Studies show that an obese person, with insulin resistance, with metabolic disorders, has a partly destroyed intestinal flora. A healthy person, with a normal body mass index, and without any particular health problems, has a "good" intestinal flora. From the point of view of the bacteria population, each individual has a unique intestinal flora. Several parts of the body are populated by almost the same bacteria, regardless of the individual, i.e. the nose, the nostrils, and the back.

The scientists from the National Institute of Health in the United States of America have carried out a research project called Human Microbiome, in which they analyzed the genomes of 178 germs that dwell in or on a healthy human body. The project has not been yet completed and its final goal is to come up with the so-called "diagnostic maps". Using the data provided by the Project Human Microbiome, the doctors will be able to associate the presence or the absence of microorganisms from the human body with the status of health or disease [10–18]. Most of the bacteria which reside in and on the human body are harmless.

Recent studies in the field of the intestinal bacteria have brought to light the food that influences the flora of the digestive tract. Moreover, it appears that the bacteria from the intestines is divided into three categories according to the predominantly "ecosystem", such the blood groups. Thus, for the persons which can be found in their diet a large quantities the meat, butter and fat dairy products have an increased level of Bacteroides bacteria which have an essential role in the processing of complex molecules of food and in their conversion into particles of simple and easy-to-assimilated by the body. Individuals who eat products especially rich in sugars (sweets, fruit, and cereals) have an intestinal flora in which prevail Prevotella bacteria, considered an "opportunist" microbe. And those who prefer the food with polyunsaturated fats (such as the fish and the olive oil) have the intestine settled especially of Ruminococcus bacteria, whose role is not fully elucidated. As a matter of fact, researchers at the University of Pennsylvania, those who have carried out the study, have not yet discovered all the health implications that this division has for the bacterial ecosystem. What is known is that some of the

prevailing micro-organisms are predisposing to the accumulation of kilograms, other influences the immune system and another category may cause imbalances of nutrients in the body [19–21].

3 Enterotypes and Personalized Medicine

Numerous studies show that the discovery of the enterotypes will lead to the transformation of the medicine as it happened in the case of blood-grouping reagents. Custom diets on the basis of the enterotype and the prescription of medicine suitable for the patient based on the enterotype are the first changes prefigured by specialists [22–28].

Also, the studies in this field show that the enterotypes will play an important role in the discovery of alternative treatments to antibiotics, which have become ineffective in recent years, as the bacteria have become more resistant. Thus, instead of trying to destroy all bacteria from the gut, doctors will be able to stimulate the bacteria, in order to restore the balance of the existing bacterial before the development of the medical condition. With time, enterotypes will allow the development of personalized medicine in which each of the patients will benefit from a treatment designed according to the needs of its own, identified on the basis of microbiome (Fig. 1).

	Human microbiota (10 times more microbial than human cells: 10^{14} vs 10^{13})	
Human microbial habitats	Most represented Phyla and their relative abundance (%)	Number of species
Oral cavity	*Firmicutes* (36.7), *Bacteroidetes* (17.3), *Proteobacteria* (17.1), *Actinobacteria* (11.9), *Fusobacteria* (5.2)	>500
Skin	*Actinobacteria* (52), *Firmicutes* (24.4), *Proteobacteria* (16.5), *Bacteroidetes* (6.3)	~300
Airways	*Actinobacteria* (55), *Firmicutes* (15), *Proteobacteria* (8), *Bacteroidetes* (3)	>500
Gut	*Firmicutes* (38.8), *Bacteroidetes* (27.8), *Actinobacteria* (8.2), *Proteobacteria* (2.1)	>1000
Urogenital tract[a]	*Firmicutes* (83), *Bacteroidetes* (3), *Actinobacteria* (3)	~150

[a] Mainly female.

Fig. 1. Human microbiota composition [9]

Moreover, in the papers [11, 12], professor Covasa and other researchers focused on the role of gut microbiota in the modulation of physiological, molecular and neuronal signals involved in regulating the energy balance in both, healthy and obesity conditions, using in vivo models and dietary manipulations. Bacteriotherapy is already a treatment which has begun to be implemented with encouraging results in the case of patients for which the antibiotics and other treatments do not give the results.

Therefore, by analysis of genes, clinicians can be identified the unique aspects to each patient which will enable the designing of custom treatments. Also, researchers expect that in the future the children will not be vaccinated only against viruses, but will be subject to detailed analysis of microbiome, to be identified bacteria key missing with a view to reintroducing them.

4 Classification Using Deep Learning Algorithms

With the rapid growth in the number and size of the databases, as well as of the applications of databases in the field of Human Microbiome, it is necessary to examine the automatic extraction of knowledge from large datasets (Fig. 2). On HMP Database there are available: 2471 Complete Genomes; 5543 Draft Bacteria & Archaea Genomes; 2399 Complete Virus Genomes; 26 Complete Fungi Genomes; 309 HMP Eukaryote Reference Genomes; Total 10,741 genomes, ~30 GB of sequences. With the advent of newer technologies, the HMP aim is to provide a catalog of microbes living in the human body and to establish their functions [9].

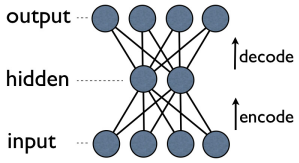


Fig. 2. Multi-layer ANNs using deep sparse auto-encode [2]

Some studies have demonstrated a clear association between the metabolic diseases and changes in the composition of the intestinal microbiota, evidenced by a low weighted of Firmicutes and a high proportion of Bacteroidetes and Proteobacteria in patients with diabetes of type 2, comparing the profile of the intestinal microbiota in persons with non-diabetes [11, 12].

For exemplification we used the data base on the Human Microbiome Project with a number of 124 obese persons who have diabetes of type 2 [9] and we used the WEKA environment, a collection of algorithms for Data Mining Learning. The WEKA algorithms were used directly on the data sets using the tools for the pre-processing the data, classification, regression, rules of association and viewing of data. In a preliminary version of our work [29] the best results were obtained with a SVM classifier (Support Vector Machine) and with the Artificial Neuronal Networks (ANN - Artificial Neural Networks) for all three classes (the class of patients with diabetes of type 2 on which are present Firmicutes micro-organisms (A), for the class of patients with diabetes of type 2 on which are present Bacteroidetes micro-organisms (B) and for the class of patients with diabetes of type 2 on which are present Proteobacteria micro-organisms (C)). For example, for class A were obtained very good results using SMO and ANN, and for class B and C were obtained similar results in favor of the use of Naïve Bayes and ANN [29].

In this paper we used the following Deep Learning algorithms: stack sparse coding algorithm; deep belief network (DBN); deep sparse auto-encoders. The new way to train multi-layer ANNs consist in the following procedure: EACH of the (non-output) layers is trained to be an auto-encoder. Basically, it is forced to learn good features that describe what comes from the previous layer. An auto-encoder is trained, with an absolutely standard weight-adjustment algorithm to reproduce the input (Fig. 2).

By making this happen with (many) fewer units than the inputs, this forces the "hidden layer" units to become good feature detectors. The intermediate layers are each trained to be auto-encoders (or similar). The final layer trained to predict class based on outputs from previous layers. To design a simple MATLAB neural network for classification ("Pattern Recognition") we used *nprtool* tool that opens a graphical interface that allows specification of a network element characterized by:

- A level of hidden units (the number of hidden units can be chosen by the
- user);
- Logistics activation (*logsig*) for both hidden units and for the output
- values ranged between (0.1);
- Backpropagation training algorithm based on minimization method of conjugate gradient.

The Artificial Neural Networks have the ability to learn, but the concrete way by which the process is accomplished is dictated by the algorithm used for training. A network is considered trained when application of an input vector leads to a desired output, or very close to it. Training consists of sequential application of various input vectors and adjusting the weights of the network in relation to a predetermined procedure. The results show that the best performance was obtained using a Multilayer Perceptron network (MLP). The MLP is a feedforward neural network comprising one or more hidden layers. Like any neural network, a network with backpropagation is characterized by the connections between neurons (forming the network architecture), activation of functions used by neurons and learning algorithm that specifies the procedure used to adjust the weights. Usually, a backpropagation neural network is a multilayer network comprising three or four layers fully connected [29–32].

Each neuron computes its output similar to Perceptron. Then input value is sent to the activation function. Unlike Perceptron, in a backpropagation Neural Networks the neurons have sigmoid type activation functions. Derivative function is very easy to calculate and ensure the [0, 1] output range. Each layer of a MLP neural network performs a specific function. The input layer accepts input signals and computational rarely contains neurons that do not process input patterns. Output layer supports output signals (stimuli coming from the hidden layer) and lays it out on the network. Detects hidden layer neurons traits and their weight is hidden patterns of input traits. These characteristics are then used to determine the output layer to the output pattern (Fig. 3).

The backpropagation algorithm is a supervised learning algorithm named generalized delta algorithm. This algorithm is based on minimizing the difference between the desired output and actual output by descending gradient method. The gradient tells us how the function varies in different directions. The idea of the algorithm is to find the minimum error function in relation to relative weights of connections. The error is given by the difference between the desired output and the actual output of the network.

Performance Metrics

Model Name	Training			Cross Validation			Testing		
	RMSE	r	Correct	RMSE	r	Correct	RMSE	r	Correct
MLPR-1-O-M (Regression MLP)	0.193136	0.504298	100.00%	0.40441	1.276899	100.00%	0.411551	0.828285	94.44%
MLPC-1-O-M (Classification MLP)	0.292957	0.659528	100.00%	0.423566	0.766415	100.00%	0.360862	0.77623	100.00%
LinR-0-B-R (Linear Regression)	0.334444	0.771683	100.00%	0.36981	0.774197	100.00%	0.378221	0.679267	94.44%
LinR-0-B-L (Linear Regression)	0.320609	0.723112	100.00%	0.382431	0.899123	100.00%	0.373497	0.713072	94.44%
LogR-0-B-R (Logistic Regression)	0.323498	0.802886	100.00%	0.411873	0.883485	100.00%	0.429729	0.777518	100.00%
LogR-0-B-L (Logistic Regression)	0.320243	0.756819	100.00%	0.399313	0.862804	100.00%	0.406476	0.735547	100.00%
MLPR-1-B-L (Regression MLP)	0.321919	0.701325	100.00%	0.448925	1.193378	100.00%	0.338979	0.690299	94.44%
MLPC-1-B-L (Classification MLP)	0.29544	0.929712	100.00%	0.504155	1.517705	100.00%	0.524958	1.351233	100.00%
PNN-0-N-N (Probabilistic Neural Network)	4.27E-06	2.41E-05	100.00%	0.578235	1	100.00%	0.522295	1	100.00%
RBF-1-B-L (Radial Basis Function)	0.391416	0.896709	100.00%	0.370999	0.691104	100.00%	0.468485	0.883554	100.00%
GFFR-1-B-L (Reg Gen Feedforward)	0.230875	0.608818	100.00%	0.513599	1.403539	100.00%	0.492597	1.227754	94.44%
GFFC-1-B-L (Class Gen Feedforward)	2.631665	11.12665	29.85%	2.33251	5.953767	42.11%	3.168973	10.87653	16.67%
MLPRPC-1-B-L (Reg MLP with PCA)	0.539245	1.436432	100.00%	0.505726	1.220976	100.00%	0.450536	0.946967	100.00%
MLPCPC-1-B-L (Class MLP with PCA)	0.604741	1.112989	100.00%	0.631973	1.093975	100.00%	0.440293	0.863168	100.00%
SVM-0-N-N (Classification SVM)	0.278716	0.667789	95.52%	0.324384	0.573036	94.74%	0.326516	0.558213	100.00%
TDNN-1-B-L (Time-Delay Network)	0.341169	0.862838	100.00%	0.721197	1.834166	100.00%	0.531575	0.985912	100.00%
TLRN-1-B-L (Time-Lag Recurrent Network)	1.009541	1.902801	100.00%	0.707728	1.619331	100.00%	0.906855	1.618652	100.00%
RN-1-B-L (Recurrent Network)	0.378266	0.882711	100.00%	0.465783	1.180841	100.00%	0.390453	0.773721	100.00%
MLPR-2-B-L (Regression MLP)	0.310913	0.896222	100.00%	0.476074	1.445478	100.00%	0.536759	1.04037	100.00%
MLPC-2-B-L (Classification MLP)	0.276627	0.854803	100.00%	0.497293	1.247617	100.00%	0.475513	1.028863	100.00%
MLPR-1-B-R (Regression MLP)	0.287202	0.768005	100.00%	0.406402	1.018289	100.00%	0.364087	0.870826	100.00%
MLPC-1-B-R (Classification MLP)	0.258878	0.717773	100.00%	0.414439	1.198915	100.00%	0.32167	0.687884	100.00%
MLPR-2-O-M (Regression MLP)	0.320195	0.69593	100.00%	0.445851	1.09194	100.00%	0.325504	0.728061	100.00%
MLPC-2-O-M (Classification MLP)	0.665406	1.262284	100.00%	0.586722	1.249546	100.00%	0.505312	1.243577	100.00%
MLPR-2-B-R (Regression MLP)	0.232608	0.747501	100.00%	0.410747	0.847326	100.00%	0.329586	0.68689	94.44%
MLPC-2-B-R (Classification MLP)	0.266694	0.745959	100.00%	0.344822	0.973972	100.00%	0.322039	0.714744	100.00%
MLPRPC-1-O-M (Reg MLP with PCA)	0.507237	1.213724	100.00%	0.477877	1.059424	100.00%	0.578924	1.194079	100.00%
MLPCPC-1-O-M (Class MLP with PCA)	0.57946	1.458307	100.00%	0.506006	1.161451	100.00%	0.445815	1.024323	100.00%
MLPRPC-1-B-R (Reg MLP with PCA)	0.49895	1.472947	100.00%	0.481242	0.966154	100.00%	0.555227	1.320027	100.00%
MLPCPC-1-B-R (Class MLP with PCA)	0.532171	1.479628	100.00%	0.440814	0.816954	100.00%	0.453173	0.812813	100.00%
GFFR-1-O-M (Reg Gen Feedforward)	0.270117	0.819621	98.51%	0.477297	0.929855	100.00%	0.458854	1.079618	94.44%
GFFC-1-O-M (Class Gen Feedforward)	0.3695	1.101388	100.00%	0.391954	0.822629	100.00%	0.319803	0.552367	100.00%
GFFR-1-B-R (Reg Gen Feedforward)	0.26564	0.766449	100.00%	0.486714	1.097254	100.00%	0.483942	1.302634	94.44%
GFFC-1-B-R (Class Gen Feedforward)	0.319684	0.808798	100.00%	0.37662	0.763517	100.00%	0.401291	0.794602	100.00%
RBF-1-O-M (Radial Basis Function)	0.551233	1.150324	100.00%	0.542021	0.974906	100.00%	0.393073	0.868909	100.00%
RBF-1-B-R (Radial Basis Function)	0.382085	0.841063	100.00%	0.37343	0.862982	100.00%	0.420303	0.819838	100.00%
TDNN-1-O-M (Time-Delay Network)	0.193776	0.849372	100.00%	0.541533	0.976413	100.00%	0.478553	0.881798	100.00%
TDNN-1-B-R (Time-Delay Network)	0.343269	1.009517	100.00%	0.489615	1.053666	100.00%	0.542653	0.973057	100.00%
RN-1-O-M (Recurrent Network)	0.566416	1.428649	100.00%	0.602502	1.37504	100.00%	0.653149	1.362624	100.00%
RN-1-B-R (Recurrent Network)	1.002518	2.107082	80.60%	1.266936	2.103606	73.68%	0.96296	1.911124	61.11%
TLRN-1-O-M (Time-Lag Recurrent Network)	0.403019	1.071385	100.00%	0.539168	1.007768	100.00%	0.40355	0.952376	100.00%
TLRN-1-B-R (Time-Lag Recurrent Network)	0.416302	0.921486	100.00%	0.589876	1.371814	100.00%	0.400352	0.779646	100.00%

Fig. 3. Performance Metrics - a Multilayer Perceptron network (MLP) best classification results (100% for training data vs. 100% for validation data vs. 100% for testing data)

The quality of a classifier in terms of correct identification of a class is measured using information from confusion matrix that contains the following [29–32]:

- The number of data correctly classified as belonging to the class interests: **True positive cases** (TP);
- The number of data correctly classified as not belonging to the class of interest: **True negative cases** (TN);
- The number of data misclassified as belonging to the class of interest: **False positive cases** (FP);
- The number of data misclassified as not belonging to the class of interest: **False negative cases** (FN);

Based on these values we calculated the following measures:

$$Sensitivity = TP/(TP + FN)$$

$$Specificity = TN/(TN + FP)$$

$$Precision = TP/(TP + FP)$$

$$Recall = TP/(TP + FN)$$

$$F = 2 * precision * recall/(precision + recall)$$

The most common error function is the mean square error. The RMSE is the mean square error and is used to characterize the scattering of the data in relation to the average. In our case, in all three stages of ANN testing, we obtained values for RMSE below 0.5, with 100% identification of classes as illustrated in Fig. 4.

Performance Metrics

	Training	Cross Val.	Testing
# of Rows	9344	2002	2002
RMSE	0.47151	0.281644	0.321151
Correlation (r)	1.538889	0.826787	0.84907
# Correct	9344	2002	2002
# Incorrect	0	0	0
% Correct	100.00%	100.00%	100.00%

Fig. 4. Performance results

Deep Learning seems to be a solution to Big Data problem because of the following characteristics: relates new knowledge to previous knowledge; the content is organized into a coherent way; focuses on problem-solving; synthesis, application, transfer; links concepts/principles to everyday experience; diagnosis using Deep Learning is typically more objective and accurate according to some studies; Deep learning does not need to provide features ahead of time; it learns features at different level by itself; the same deep learning architecture can be trained to accomplish different tasks. In today modern medicine, digital information security is an interdisciplinary issue, having to be constantly optimized, developed and innovated [33–35].

5 Conclusions

Research on microbiome is still in its incipient phase and a better understanding may allow for a more efficient fight against diseases. Therefore, the bacteria will play an essential role in ensuring human health. Our results and simulations made using the tools WEKA Machine Learning and MATLAB can help us to select suitable Data Mining, Knowledge Discovery Data and Deep Learning algorithms according to each data bases. We can find connections between different pathologies (Obesity and Diabetes) and Human Microbiome by using Deep Learning applications. Human Microbiome Project remains a global challenge priority, since it involves large communities of researchers and requires the latest approaches in biotechnology, bioengineering and bioinformatics.

Novel research that links microbiota to obesity, diabetes and other metabolic diseases is done in the project "Analysis of novel risk factors influencing control of food intake and regulation of bodyweight", headed by Prof. Mihai Covasa. Recent studies provide a significant contribution in Human Microbiome Big Data because it could lead to the discovery of alternative ways to treat especially those brain dysfunctions that does not always respond to usual prescribed medication. Diagnosis by Deep Learning is typically more objective and accurate according to our studies and Deep Learning can be a solution for Big Data Analysis, Clustering and Classification.

Acknowledgement. This work was supported by the Romanian National Program (PN-II-ID-PCE-2012-4-0608 no. 48/02.09.2013), "Analysis of novel risk factors influencing control of food intake and regulation of body weight" [36].

References

1. Schmidhuber, J.: Deep learning in neural networks: an overview. Neural Nets **61**, 85–117 (2015)
2. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553), 436–444 (2015)
3. Bhattacherjee, A., Sourav, R., Sneha, P., Pazel, R., Noreen, K., Nilanjan, D.: Classification approach for breast cancer detection using back propagation neural network: a study. In: Biomedical Image Analysis and Mining Techniques for Improved Health Outcomes, p. 12 (2016). doi:10.4018/978-1-4666-8811-7
4. Hore, S., et al.: Real time dactylology based selective image encryption using speeded up robust features extraction technique and artificial neural network. In: Image Feature Detectors Foundations, Innovations, and Applications. Studies in Computational Intelligence series (2015)
5. Samanta, S., et al.: Haralick features based automated glaucoma classification using back propagation neural network. In: Proceedings of the 3rd International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA), vol. 327 (2014)
6. Geman, O.: Nonlinear dynamics, artificial neural networks and neuro-fuzzy classifier for automatic assessing of tremor severity. In: E-Health and Bioengineering Conference (EHB), pp. 1–4 (2013)
7. Geman, O., Costin, H.: Automatic assessing of tremor severity using nonlinear dynamics, artificial neural networks and neuro-fuzzy classifier. Adv. Electr. Comput. Eng. **14**(1), 133–138 (2014)
8. Geman, O., Sanei, S., Chiuchisan, I., Graur, A., Prochazka, A., Vysata, O.: Towards an inclusive Parkinson's screening system. In: Proceedings of the International Conference on System Theory, Control and Computing, pp. 476–480 (2014)
9. Human Microbiome Project (2016). https://commonfund.nih.gov/hmp/overview
10. Samuel, B.S., Shaito, A., Motoike, T., Rey, F.E., Backhed, F., Manchester, J.K., Hammer, R.E., Williams, S.C., Crowley, J., Yanagisawa, M., Gordon, J.I.: Effects of the gut microbiota on host adiposity are modulated by the short-chain fatty-acid binding G protein-coupled receptor, Gpr41. Proc. Nat. Acad. Sci. **105**, 16767–16772 (2008)
11. Swartz, T.D., Duca, F.A., De Wouters, T., Sakar, Y., Covasa, M.: Upregulation of intestinal type 1 taste receptor 3 and sodium glucose luminal transporter-1 expression and increased sucrose intake in mice lacking gut microbiota. Br. J. Nutr. **107**, 621–630 (2012)

12. Duca, F.A., Covasa, M.: Current and emerging concepts on the role of peripheral signals in the control of food intake and development of obesity. Br. J. Nutr., 1–16 (2012)
13. Fettweis, J.M., Brooks, J.P., Serrano, M.G., Sheth, N.U., Vladimir, J.: Microbiology (Reading, England) (2014)
14. Eren, A., Murat, A., Borisy, G.G., Huse, S.M., Mark Welch, J.: Oligotyping analysis of the human oral microbiome. In: Proceedings of the National Academy of Sciences of the United States of America (2014)
15. Ma, L., Jungwoo, K., Hatzenpichler, R., Karymov, M.A., Hubert, N., Hanan, I.M., Chang, E.B., Ismagilov, R.F.: Gene-targeted microfluidic cultivation validated by isolation of a gut bacterium listed in Human Microbiome Project's Most Wanted taxa. In: Proceedings of the National Academy of Sciences of the United States of America, pp. 9768–9773 (2014)
16. Rooks, M.G., Veiga, P., Wardwell-Scott, L.H., Tickle, T., Segata, N., Michaud, M., Gallini, C.A., Beal, C., Van Hylckama-Vlieg, J.E.T., Ballal, S.A., Morgan, X.C., Glickman, J.N., Huttenhow, D.G.: Gut microbiome composition and function in experimental colitis during active disease and treatment-induced remission. ISME J. 1403–1417 (2014)
17. Spinler, J.F., Sontakke, A., Hollister, E.B., Venable, S.F., Phaik, L., Balderas, M.A., Saulnier, D.M.A., Mistretta, T.A., Devaraj, S., Walter, J., Versalovic, J., Highlander, S.K.: From prediction to function using evolutionary genomics: human-specific ecotypes of Lactobacillus reuteri have diverse probiotic functions. Genome Biol. Evol. 1772–1789 (2014)
18. Ma, L., Datta, S.S., Karymov, M.A., Pan, Q., Begolo, S., Ismagilov, R.F.: Individually addressable arrays of replica microbial cultures enabled by splitting SlipChips. Integr. Biol.: Quant. Biosci. Nano Macro (2014)
19. Huttenhower, C., Kostic, A.D., Xavier, R.: Inflammatory bowel disease as a model for translating the microbiome. Immunity, 843–854 (2014)
20. Franzosa, E.A., Xuan Zhang, X.C.: The oral and gut microbiomes are perturbed in rheumatoid arthritis and partly normalized after treatment. Nat. Med. (2015)
21. D'Argenio, V., Salvatore, F.: The role of the gut microbiome in the healthy adult status. Clin. Chim. Acta (2015)
22. Mandal, S., et al.: Analysis of composition of microbiomes: a novel method for studying microbial composition. Microb. Ecol. Health Dis. **26**, 1–7 (2015)
23. He, C., Shan, Y., Song, W.: Targeting gut microbiota as a possible therapy. Nutr. Res. **35**, 361–367 (2015)
24. Le Barz, M., Anhe, F.E., Varin, T., et al.: Probiotics as complementary treatment for metabolic disorders. Diab. Metab. J. **39**, 291–303 (2015)
25. Wu, Y., Ding, Y., Tanaka, Y.: Risk factors contributing to type 2 diabetes and recent advances in the treatment and prevention. Int. J. Med. Sci. **11**, 1185–1200 (2014)
26. Tremaroli, V., Backhed, F.: Functional interactions between the gut microbiota and host metabolism. Nature **489**, 242–249 (2012)
27. Everard, A., Cani, P.: Diabetes, obesity and gut microbiota. Best Pract. Res. Clin. Gastroenerol. **27**, 73–83 (2013)
28. Apostolescu, C., Moroti, R., Molagic, V., et al.: Gut microbiota and its complex role, the experience of the National Institute for Infectious Diseases "Prof. Dr. Matei Balş" in fecal bacteriotherapy for Clostridium difficile infection. BMC Infect. Dis. **13**(Suppl. 1), O19 (2013)
29. Geman, O., Chiuchisan, I., Covasa, M.: Data mining and knowledge discovery tools for human microbiome big data. In: Proceedings of IEEE - International Conference on Computers Communications and Control, Oradea, Romania, 10–14 May (2016)

30. Geman, O.: Data mining tools used in deep brain stimulation – analysis results. In: Artificial Intelligence Applications and Innovations. IFIP Advances in Information and Communication Technology Series, vol. 364, pp. 259–264. Springer, Heidelberg (2011)
31. Geman, O., Turcu, C.: Partitioning methods used in DBS treatments analysis results. In: Proceedings of International Joint Conference on Neural Networks, San Jose, California, USA, pp. 1788–1793 (2011)
32. Lungu, M.M., Bosancu, A., Geman, O.: Mini-review: human microbiome project – recent trends and future challenges. In: Proceedings of IEEE International Conference on E-Health and Bioengineering (EHB 2015), pp. 1–4 (2015)
33. Bucerzan, D., Raţiu, C.: Image processing with android steganography. In: Soft Computing Application, Proceedings of 6th International Workshop on Soft Computing Applications (SOFA 2014). Advances in Intelligent Systems and Computing, vol. 1, pp. 27–36. Springer International Publishing, Cham (2014). doi:10.1007/978-3-319-27179-8. eBook ISBN 978-3-319-18296-4, ISBN 978-3-319-27178-1
34. Bucerzan, D., Raţiu, C.: Contributions to steganographic techniques on mobile devices. In: Innovative Security Solutions for Information Technology and Communications. Lecture Notes in Computer Science, vol. 9522, pp. 242–252. Springer International Publishing, Cham (2016). doi:10.1007/978-3-319-27179. eBook ISBN 978-3-319-27179-8, ISBN 978-3-319-27178-1
35. Raţiu, C., Bucerzan, D., Manolescu, M.J.: SmartSteg: a new android based steganography application. Int. J. Comput. Commun. Control **8**(5), 681–688 (2013). ISSN 1841-9836
36. Project: Analysis of novel risk factors influencing control of food intake and regulation of body weight. (PN-II-ID-PCE-2012-4-0608 no. 48/02.09.2013) (2016). www.eed.usv.ro/idei_48

Evaluating the User Experience of a Web Application for Managing Electronic Health Records

Daniel-Alexandru Jurcău(✉) and Vasile Stoicu-Tivadar

Faculty of Automatics and Computer Science,
Politehnica University Timișoara, 2, Piața Victoriei Street,
300006 Timișoara, România
djurcau@gmail.com, vasile.stoicu-tivadar@aut.upt.ro

Abstract. This paper researches into ways of performing a user experience analysis of a web application that deals with the input of data related to electronic health records. The analysis methods presented are divided into two major categories. The first details mathematical methods that provide metrics characterizing the aesthetics of the application and also provide hints as to how long it would take an experienced user to perform various operations. The second method involves an actual usability test, where medics get to prototype the application under a specific usage scenario while having their activity closely monitored to provide information into how they interact with the web application. The paper provides details on this test, from how relevant data was collected to how it was processed and what was learned from it.

Keywords: Electronic health records · User experience · Heuristics · Activity analysis

1 Introduction

Providing a good experience is an essential characteristic for any digital product that aims to improve productivity and make the users' job easier. The theory of interaction design distinguished between [16]:

- *Usability goals* – concerned with meeting specific usability criteria
- *User experience goals* – largely concerned with explicating the quality of the use experience.

Even simple considerations like improving the screen clarity and readability by making screens less crowded has been proven to increase productivity by about 20% [4]. This is even more important in the case of fields like medical informatics, in which "complex tables, lists, charts, and diagrams are inevitable" [10].

Previous studies show that the "lack of good user interfaces has been long recognized as a major impediment to the acceptance and routine use of clinical informatics applications" [8, 18]. One of the main causes of complaints in the case of early user interfaces proved to be the absence of "guidance as to a desired workflow" [18].

V.E. Balas et al. (eds.), *Soft Computing Applications*, Advances in Intelligent Systems and Computing 633, DOI 10.1007/978-3-319-62521-8_22

This paper investigates ways of evaluating and improving a web user-interface prototype that deals with electronic health records. As can be seen in Fig. 1, the page structure shows navigation links at the top, allowing the user to select between multiple input forms (for chronic illness) and then goes on to present the actual inputs in a two column manner.

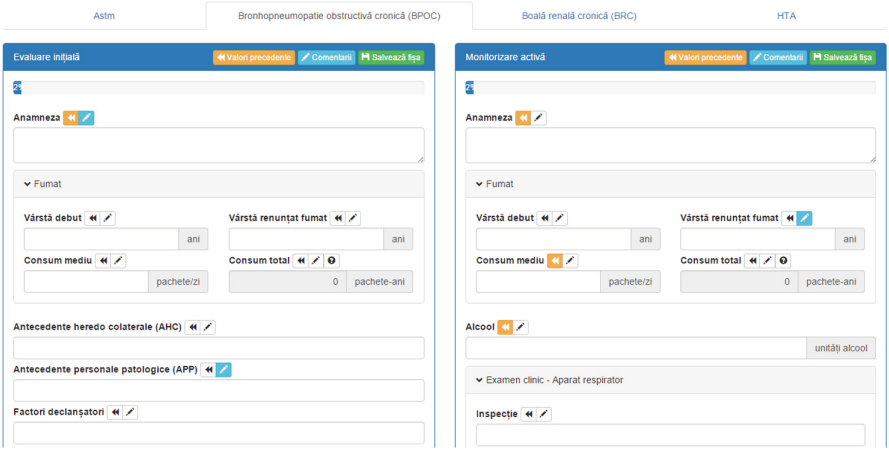


Fig. 1. Two-column layout showing the selected input form. The elements for the initial evaluation are displayed on the left and those of the active one are on the right

Modern usability practices, such as the user centered design approach, request that the usability of a product be tested early in the development process [17]. There are multiple types of evaluations possible, as described in [9]:

- *Expert evaluation* – analysts systematically step through a user interface
  - Heuristic evaluation: compares the interface against a set of usability guidelines and heuristics
  - Cognitive walk-through: steps through the user interface for a task, noticing problems or responses
- *Usability testing* – representative end-users are observed while interacting with the system to carry out representative tasks
- *Clinical simulation* – a type of usability testing conducted in a simulation laboratory or in real setting to ensure a higher fidelity of the evaluation
- *Post implementation surveillance* – high fidelity evaluation based on the actual usage of a live system.

From a technical point of view, the "data left by many interactive applications in the servers log file is minimal and not sufficient for extracting detailed information about the actual usage of the application" [1]. Such logs only record error events or information about when a user has accessed a specific page. Information about various forms that have been filled, if available at all, does not show a specific order in which the inputs have been filled out or whether the user has encountered any difficulties in doing so.

Recording every action the users makes (what keyboard keys are clicked, where is the mouse moved, etc.) enables the tracking of all user activity. However, because this data is very concrete, more abstract information needs to be inferred from it, while also taking events, such as the user pausing to answer the phone, into consideration [1].

The following sections present ways of performing automated evaluations as well as a usability test on the developed prototype.

2 Performing Automated Evaluations

Performing automated evaluations involves applying various mathematical methods that calculate metrics based on the detailed structure of the web page.

2.1 Evaluating Aesthetics

As people prefer attractive interfaces, the use of automatic tools for the evaluation of interface aesthetics has become a way to obtain good a design on a tighter budget [12]. Existing studies have introduced mathematical formulas that enable an objective treatment of the aesthetic issues of graphical user interfaces [11, 13, 14].

These aesthetic measures are confined "to examining only the dimension and position of rectangular regions in order to control content effects and to facilitate interpretation of the data analyses" [13]. Their result is a numeric value between 0 (worst) and 1 (best) [13].

A first step in calculating the measures presented above was to render a screenshot of the web application showing all the inputs of one input form. This resulted in a 1643 × 3265 bitmap which was then processed to extract the coordinates of all its inner rectangles.

Because the amount of text displayed with an empty form is quite small and usually in the form of labels for each input, the analysis was conducted once without taking input labels into account and once with them in mind as small bounding rectangles for each individual word. The results are presented in Table 1.

Table 1. Analysis of aesthetics measures

Measure	With input labels	Without input labels
Balance	0.7704	0.7706
Equilibrium	0.9790	0.9842
Unity	0.4797	0.4776
Proportion	0.7344	0.5601
Simplicity	0.0051	0.0192
Density	0.9353	0.9745
Economy	0.0227	0.0833

Not surprisingly, the values describing *economy* and *simplicity* are low. They decline rapidly as the number of objects on the screen grows, a situation with is inevitable due to the complexity of medical inputs.

The *proportion* score is above average; however it shows room for improvement as the value is negatively influenced by the high width/height ratio of most input forms.

Unity is another measure which can be improved, in this case by increasing the margin of the layout in regards to the whole page and, at the same time, reducing the space between individual elements, compacting the screen.

The chosen layout shows a good *balance*, avoiding the placement of heavy objects on only one side. Further improvements can be made by reducing the size of the header shown at the top of the page, or by adding a footer to balance it better.

The two-column layout chosen to present the initial recordings at the same time as the active ones leads to a very good *equilibrium*. A similar high score is also obtained for *density*.

With a few exceptions, taking the input labels into account or not when computing the scores has little effect on the outcome.

2.2 Goals, Objects, Methods and Selection Rules

A renowned method for performing quantitative analyses of user interface design is the classic model of goals, objects, methods and selection rules (GOMS) presented by [2]. This method allows for the prediction of how long an experienced worker will need to perform an operation when using a specific interface design [15].

By means of laboratory experiments, the authors came up with timings for different gestures. They found these typical timings to be enough for comparative analyses (when a keyboard and graphical input device are involved in solving the task) rather the having to measure the exact time it takes each individual [15]. As can be seen in Table 2, each timing is designated by a one-letter mnemonic.

Table 2. Timings used by the GOMS model [2]

Mnemonic	Duration	Description
K	0.2 s	The time it takes to tap a key on the keyboard
P	1.1 s	The time it takes to point to a position on the display
H	0.4 s	The time it takes the user's hand to switch from the keyboard to the graphical input device or vise versa
M	1.35 s	The time it takes the user to prepare mentally for the next step
R		The time it takes the computer to respond to input

Table 3 presents a GOMS analysis performed on the following inputs:

1.	Anamneză	Expectorație în cantități reduse
2.	Hemoglobina	11 g/dl
3.	Leucocite	10 000/mm^3
4.	Trombocite	265 000/mm^3
5.	Vaccinare antigripală	Neefectuată

Table 3. Quantitative analysis using the GOMS model

Input	GOMS representation	Duration	
		Total	Actual input
#1	HPKHKKKKKKKKKKKKKKK KKKKKKKKKKKKKKK	8.5 s	6.4 s
#2	HPKHKK	2.5 s	0.4 s
#3	HPKHKKKKK	3.1 s	1.0 s
#4	HPKHKKKKKK	3.3 s	1.2 s
#5	HPKHKKKKKKKKKK	4.3 s	2.2 s

The results of the analysis indicate that a good amount of the time is spent entering the actual information in the input fields. When navigating from one input field to another using the mouse, there is a constant delay of 2.1 s until the new input is selected. This delay is kept under control because all the inputs are displayed on the same page and are not hidden under various windows or subpages.

The more information the user needs to input into one specific field, the less significant the initial delay becomes. In the case of small inputs, using the mouse to move from one input to another can cause some delays, as it takes more than entering the actual input. The solution to this is to benefit from the fact that numeric inputs are often placed close to one another and can be reached via keyboard navigation. Thus, a single key press (0.2 s) is sufficient to move to the next input, significantly reducing the delay.

3 Developing a Framework for Automated User Activity Tracking

Recording and logging the interactions between the user and the application is the first step towards analyzing the user experience provided by an interface. For this purpose, we developed a web framework that transparently tracks the user's actions while using the application and provides information such as:

- How was the web page rendered on the client machine?
- Where and how often did the user move the mouse?
- What did the user type using the keyboard and in which field?

3.1 Existing Tracking Solutions

The behavior of traditional web pages was to request a new page from the server every time a link was clicked. This led to the development of industry standard analytics frameworks like Google Analytics (http://www.google.com/analytics/) which records by whom and when a web page is accessed and also provides the administrator with the interpretation of the data.

The major drawback of such tracking solutions is that they are intended for public facing web-sites and not intranet applications:

- The recording is handled by an external server that must be reachable from the client machine. This can be a problem especially if, from a security point of view, one does not want external services to come in possession of information on users' activities.
- The analytics provided (what pages does a user navigate to, how often does one return to the website) are targeted at marketing departments, not user-experience engineers.

With the introduction of single-page applications, the task of recording relevant information becomes harder because, from the browser's perspective, the user is still on the same page even if, after clicking a link, the content of that page has changed almost entirely.

As such, scripts such as the one provided by Google Analytics have been updated to allow the developer to write code that manually notifies the tracking service that a new page view is being shown to the user: "To track dynamically loaded content as distinct pageviews you can send a pageview hit to Google Analytics and specify the Document Path by setting the page field" [5]. The need to explicitly provide this kind of information increases the development time proportionally to the number of distinct views.

Other companies offer commercial online analytics that are more centered on detecting user experience issues: ClickTale (http://www.clicktale.com/), ExtraWatch™ (http://www.extrawatch.com/), Mouseflow (https://mouseflow.com/), Seevolution (https://www.seevolution.com/). Among the services offered are: heat map visualizations of mouse movement, session replays and analysis on the completion of forms.

3.2 Creating a Custom Framework

The main reason for choosing to create a custom solution for recording and analyzing the data instead of opting for an already available solution is that recordings can be stored on the same server used to host the web application being tested. This in turn protects the privacy of the data which no longer travels to external servers, has the same up/downtime as the web application and allows for lower latency and lower bandwidth usage.

Another important reason is that the type of data being collected is not limited to presets and the way in which the tracking information is being analyzed can be improved without having to wait on 3rd parties to update their code.

Creating a framework for tracking user activities on a web page involves two major steps: recording the events and sending the records to a server for storage.

3.2.1 Recording Events

In order for the solution to work with rich client single-page applications, all recordings are done on the client side. This is quite trivial to implement as current web browsers already offer JavaScript API's for listening for a variety of events [1].

Moving the mouse can create a lot of events, especially if the users move it from one part of the screen to another in a very short amount of time. As such, the current cursor coordinates are stored in a temporary variable and only recorded on the server once every 50 ms, if their value has changed.

The simple value of cursor coordinates does not reveal much information. In order for it to be useful, it needs to be correlated to whatever the cursor was hovering upon. Because the way a web page is rendered relies heavily on the client screen resolution and also the type of web browser used, we decided to also record a "screenshot" showing how the page is actually rendered on the client. This is accomplished by using an open-source library called *html2canvas* (https://html2canvas.hertzen.com/) to render the current page on a canvas, the content of which can then be exported to a bitmap.

3.2.2 Sending the Event Records to the Server

An easy approach would be to send HTTP requests on each event with a small payload detailing the action taken by the user (e.g. click x = 815; y = 231 target = id: SPAN16 [1]).

However, because the HTTP protocol is stateless [7], this approach ends up creating a lot of requests in the process, thus requiring the transfer of much more data (TCP handshakes, HTTP headers) than the actual intended payload. A simple check using a network sniffer shows that a 94 byte payload actually requires 388 bytes to be transferred due to the HTTP headers.

A better solution was found by using the newer *web sockets* standard which "enables two-way communication between a client running untrusted code in a controlled environment to a remote host that has opted-in to communications from that code" [6]. This way a connection is kept open between the client and the server for the duration of the page being displayed in the browser, much like how a normal TCP connection works. Except for the initial handshake, this method adds very little extra overhead when transferring the data. As such, the user's experience while using the web application is not affected by the activity tracking that takes place at the same time.

The records containing the client's actions are serialized using the JSON format, easy for both the client-side JavaScript code to create, and for the server-side code to interpret. Each event is an entry in an array, with properties describing at least the type of event and the time stamp (milliseconds elapsed since the recording was started).

As can be seen in Fig. 2, property names are kept short to preserve bandwidth, a mouse move ("m") event contains the new coordinates, a key input ("k") event contains the id of the DOM element where the key was pressed, along with the key's character code (100 – d); and a screen content changed ("s") event contains the new size of the page and the DOM elements that comprise it (each with id and bounding rectangle). Assigning id's to each distinct input element has proven to simplify identifying them.

In order to further reduce the amount of bandwidth required for transferring the recordings, the JSON documents are compressed using the deflate algorithm (https://tools.ietf.org/html/rfc1951). As compression is more effective the greater the amount of data that can be scanned for potential duplicates, the JSON entries are stored in a buffer where they accumulate for 5 s before the buffer is compressed and sent over the wire.

The compression ratio obtained is about 23% for the average entries and 56% for the initial entries which contain the screen elements and the screenshot inserted as a base64 string in the JSON document.


```
{
    "t" : "m", "ts" : 1673959, "x" : 92, "y" : 1603
}, {
    "t" : "k", "ts" : 1674762,
    "c" : 100, "id" : "inputMURMURVEZICULARi",
}, {
    "t" : "s", "ts" : 1675815,
    "w" : 1349, "h" : 3137,
    "e" : [{
            "id" : "labelinputANAMNEZAi",
            "x" : 36, "y" : 343, "w" : 71, "h" : 20
        }, {
            "id" : "buttonPrevinputANAMNEZAi",
            "x" : 111, "y" : 348, "w" : 12, "h" : 18
        }, {
            "id" : "inputANAMNEZAi",
            "x" : 36, "y" : 373, "w" : 582, "h" : 20
        }
    ]
}
```


Fig. 2. Event data recorded as JavaScript Object Notation (JSON)

4 Analyzing the Users' Activity

The evaluation of the users' experience in using the web application was conducted with the help of a medical scenario detailing information about a fictitious patient consultation that needs to be inputted into the forms.

The scenario, a screenshot of which is shown in Fig. 3, contains information relevant to two chronic diseases, ranging between textual, numeric and even boolean inputs. The same scenario was presented to all candidates and it also included an introductory guide on how to use the application.

Twenty medical doctors and residents have accepted the invitation to pilot the prototype. Their interactions with the web application have been logged in detail and later analyzed using various methods.

Câmp	Valoare
Anamneza	Muncitor în mină de cărbune timp de 30 de ani Fumător, 1 pachet/zi Dispnee la eforturi medii Expectoraţie în cantităţi reduse Tuse cronică de aproximativ 5 ani
Examen clinic - Aparat respirator	
Inspecţie	Torace mărit de volum antero-posterior, cu creşterea spaţiilor intercostale
Palpare	Transmiterea vibraţiilor vocale
Percuţie	Hipersonoritate
Auscultaţie	Raluri crepitante
Spirometrie	70%
HLG completă	Hemoglobina = 11 g/dl Hematii = 4,7 x 10^6 / mm^3 Leucocite = 10.000 / mm^3 Trombocite = 265.000 / mm^3
Rgr pulmonară	Cardiomegalie (ICT>0.5), accentuarea hilurilor pulmonare, hipertransparenţa câmpurilor pulmonare
Vaccinare antigripală	Neefectuată

Fig. 3. Example information shown in the user experience evaluation scenario

4.1 Heat Maps

An important method in understanding how a user interacts with something shown on a computer screen is by tracking the users' eye movements, what they are focusing at.

Performing such an analysis is difficult as it implies the use of expensive monitoring equipment and a controlled environment. However, studies have found that plotting the users' mouse movements gives a good indication of what elements the users' eyes were focusing at [3].

A popular method for visualizing mouse movement offered by commercial solutions is in the form of heat maps. These are 2D plots, overlaid upon the screenshot of the web page, which use color to indicate the relative time the mouse pointer has hovered over specific areas.

When using the default, linear scale, one issue that became immediately apparent was that a lot of users left the mouse cursor in a stationary point for a longer period of time. This led to mouse movement barely visible in other areas of the page, as the cursor stationed for far less in those areas. A solution was to use a logarithmic scale which allows one to easily distinguish areas in which the cursor was stationed a lot, was just passing through or wasn't present at all.

Comparing the heat maps generated for multiple user sessions shows that users hovered a lot over click-able areas like the navigation links at the top of the page, or the actual input forms. Figure 4 underlines the link between the eye and mouse movement as the user moves the cursor along a line of text while reading it.

As can be seen in Fig. 5, areas where the mouse is stationed for a longer time are also accompanied by mouse clicks. The users activate the inputs with their mouse which then remains stationary while the keyboard is used to enter data.

Fig. 4. Heat map showing the cursor following a text while the user reads

A more detailed analysis shows that 78% of inputs are activated using mouse-clicks with only 22% using the keyboard by means of the TAB key. Figure 5 also shows that, by placing the measurement unit on the right side of the numeric inputs, the user was able to correctly deduce that the text of numeric inputs is aligned to the right.

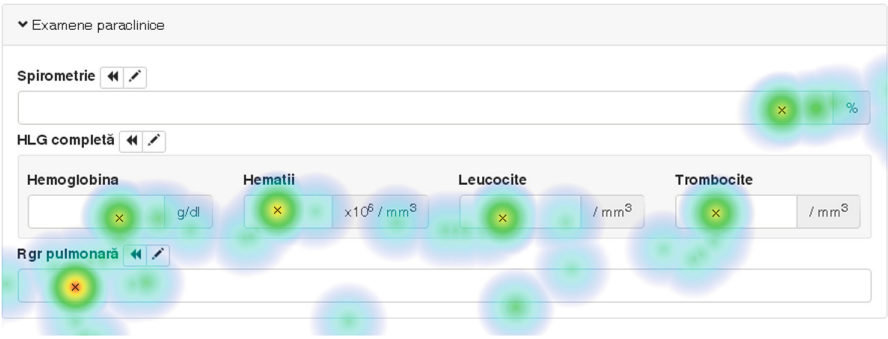


Fig. 5. Heat map also highlighting mouse clicks

4.2 Input Analysis

The order in which the user fills the inputs, along with the ease to find the exact input one searches is important for boosting productivity by reducing the time spent navigating around the page.

Figure 6 uses arrows to indicate the temporal relation between inputs. In order to indicate the chronological order, the arrows are colored using different hues, in ascending order. The majority (90%) of inputs are entered in a left-to-right, top-to-bottom order, as is natural for the Romanian (and/or English) language. Figure 7 details this percentage.

When it comes to how fast the users were able to input the data, Fig. 8 contains a histogram showing how many occurrences there are of various durations between inputs, taking the entries for all users into account. Based on it, one can safely consider that entries taking longer than 2500 ms are caused by other factors and can be excluded from calculating the average writing speed.

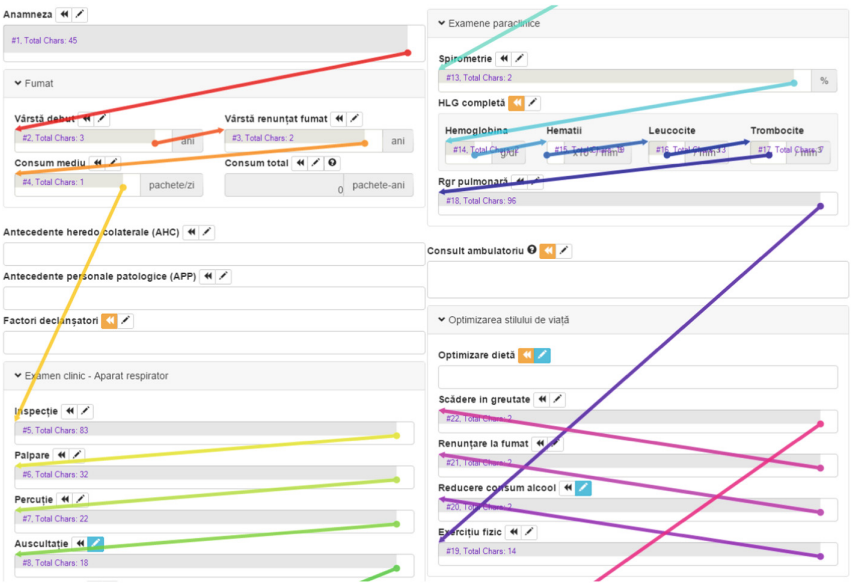


Fig. 6. Order in which inputs are filled: Left-to-Right, Top-to-Bottom (left) and Mixed order (right)

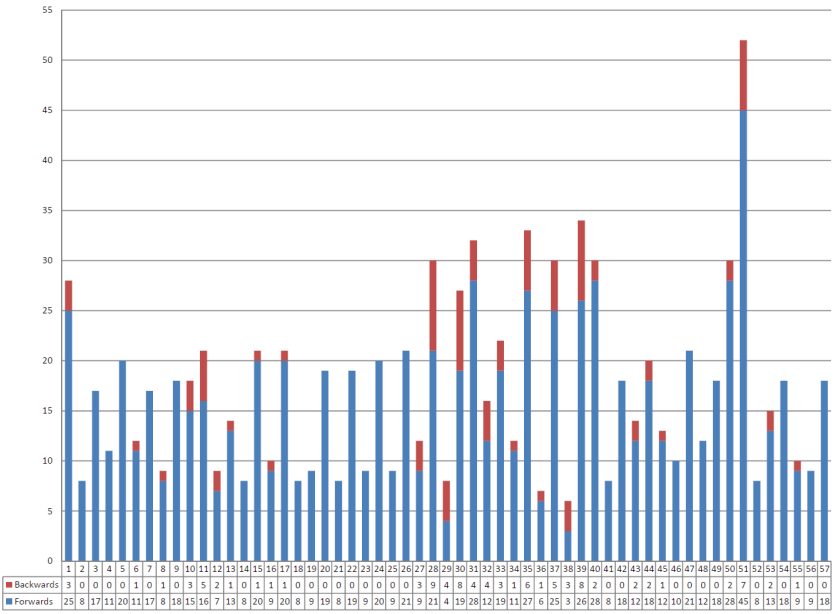


Fig. 7. Direction of navigation between adjacent inputs

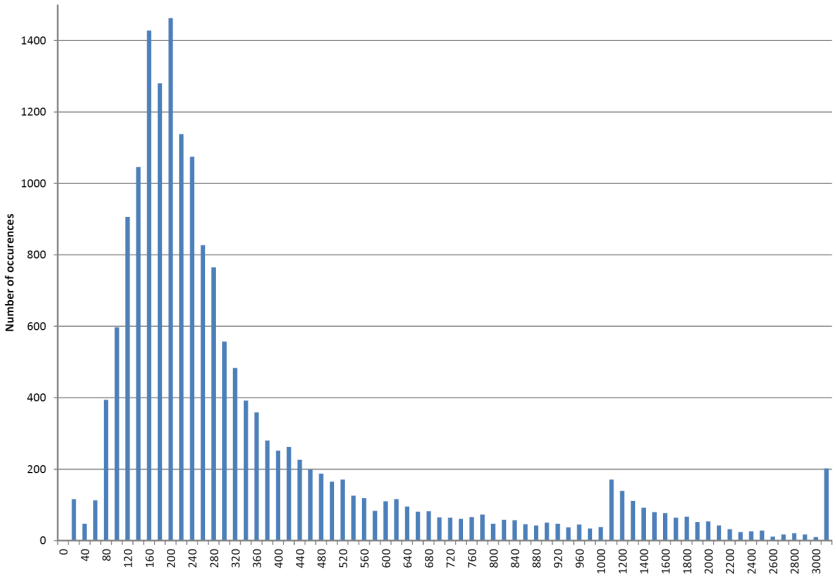


Fig. 8. Histogram of the duration between user key presses

The results of analyzing the average writing speed across all users reveals an average of 175.2 characters/minute, situated between a minimum of 135.0 and a maximum of 245.3 characters/minute.

5 Conclusions

In conclusion, this research on using multiple ways to evaluate the user experience of a web application dealing with electronic health records has proven successful.

Analyzing the web page aesthetics using simple mathematical formulas reveals valuable information about areas that can be improved at an early stage, before starting more elaborate and time consuming analysis. Including the GOMS method into the workflow helps avoid adding features that accidentally lower the users' productivity. Using a generic, efficient and transparent framework for collecting data while performing usability tests encourages such activities to be performed more often and also prevent sensitive data from leaking to 3rd parties.

The results of the user experience analysis highlight aspects and use cases that can benefit from improvements:

- Increasing the page margin from 20 pixels to 100 pixels allows the unity score to exceed 50%
- Compacting the page header while also duplicating the form selection tabs at the bottom of the page helps prevent unnecessary scrolling to the top of the page and also achieves symmetry, thus increasing the balance

- Compacting group headers reduces the space occupied by input groups and allows more inputs to be displayed on screen
- Allowing medics to individually customize the order of inputs gives them more control and helps adapt the application to their way of work.

References

1. Atterer, R., Wnuk, M., Schmidt, A.: Knowing the user's every move: user activity tracking for website usability evaluation and implicit interaction. In: Proceedings of the 15th International Conference on World Wide Web, WWW 2006, pp. 203–212. ACM, New York, NY, USA (2006). doi:10.1145/1135777.1135811. http://doi.acm.org/10.1145/1135777.1135811
2. Card, S., Moran, T.P., Newell, A.: The Psychology of Human-Computer Interaction. Lawrence Erlbaum Associates, Hillsdale (1983)
3. Chen, M.C., Anderson, J.R., Sohn, M.H.: What can a mouse cursor tell us more?: correlation of eye/mouse movements on web browsing. In: CHI 2001 Extended Abstracts on Human Factors in Computing Systems, CHI EA 2001, pp. 281–282. ACM, New York, NY, USA (2001). doi:10.1145/634067.634234. http://doi.acm.org/10.1145/634067.634234
4. Galitz, W.O.: The Essential Guide to User Interface Design. Wiley, USA (2002)
5. Google Developers: Analytics for Web (analytics.js) – Single Page Application Tracking (2015). https://developers.google.com/analytics/devguides/collection/analyticsjs/single-page-applications
6. Internet Engineering Task Force (IETF): Request for Comments: 6455 – The WebSocket Protocol (2011). https://tools.ietf.org/html/rfc6455
7. Internet Engineering Task Force (IETF): Request for Comments: 7230 – Hypertext Transfer Protocol (HTTP/1.1): Message Syntax and Routing (2014). https://tools.ietf.org/html/rfc7230
8. Johnson, C.M., Johnson, T.R., Zhang, J.: A user-centered framework for redesigning health care interfaces. J. Biomed. Inf. **38**, 75–87 (2005). Elsevier
9. Marcilly, R., Kushniruk, A.W., Beuscart-Zephir, M.C., Borycki, E.M.: Insights and limits of usability evaluation methods along the health information technology lifecycle. In: Proceedings of MIE 2015 Digital Healthcare Empowering Europeans, pp. 115–119. IOS Press, Amsterdam, Netherlands (2015)
10. Marcus, A., Wieser, K., Armitage, J., Frank, V., Guttman, E.: User-interface design for medical informatics: a case study of kaiser permanente. In: Proceedings of the 33rd Hawaii International Conference on System Sciences. IEEE (2000)
11. Mbenza Buanga, P.: Automated Evaluation of Graphical User Interface Metrics. Universit catholique de Louvain, Louvain-la-Neuve (2011)
12. Miniukovich, A., De Angeli, A.: Computation of interface aesthetics. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, CHI 2015, pp. 1163–1172. ACM, New York, NY, USA (2015). doi:10.1145/2702123.2702575. http://doi.acm.org/10.1145/2702123.2702575
13. Ngo, D.C.L., Byrne, J.G.: Another look at a model for evaluating interface aesthetics. In: International Journal of Applied Mathematics and Computer Science, pp. 515–535. University of Zielona Gora Press (2001). http://pldml.icm.edu.pl/pldml/element/bwmeta1.element.bwnjournal-article-amcv11i2p515bwm
14. Ngo, D.C.L., Teo, L.S., Byrne, J.G.: A mathematical theory of interface aesthetics. In: Visual Mathematics, vol. 8. Mathematical Institute SASA (2000). http://eudml.org/doc/256723

15. Raskin, J.: The Humane Interface. New Directions for Designing Interactive Systems. Addison-Wesley, USA (2000)
16. Rogers, Y., Sharp, H., Preece, J.: Interaction' design. Wiley, USA (2002)
17. Rueda, D., Hoto, R., Conejero, A.: Human Factors in Computing and Informatics. In: Proceedings of the First International Conference, SouthCHI 2013, Maribor, Slovenia, 1–3 July 2013, pp. 122–136. Springer Berlin Heidelberg, Germany (2013). doi:10.1007/978-3-642-39062-3_8. http://dx.doi.org/10.1007/978-3-642-39062-3_8
18. Zheng, K., Padman, R., Johnson, M.P.: User interface optimization for an electronic medical record system. In: Proceedings of the 12th World Congress on Health (Medical) Informatics, pp. 1058–1062. IOS Press (2007)

Multi Framing HDR for MRI Brain Images

Marius M. Balas(✉) and Adelin Sofrag

"Aurel Vlaicu" University of Arad, Arad, Romania
marius@drbalas.ro, adelins22@yahoo.com

Abstract. The paper aims to improve the quality of the RMN brain images by extending their dynamic range. A Fast Multiple Frame HDR is applied: a sequence of conventional low dynamic range frames with increasing exposures are merged in order to obtain a HDR file, containing more details, easing such way the segmentation stage.

Keywords: High Dynamic Range · Multi framing HDR · Magnetic Resonance Imaging · Neuroimaging

1 Introduction

In the last decade, computer assisted medical image analysis techniques were used to provide a better understanding of the available data.

Magnetic Resonance Imaging (MRI) has become a primary technique widely used for imaging the anatomy of the human soft tissues. Its applications extend across all parts of the human body and it is the most common method to study the human brain. MRI has several advantages over other imaging techniques, providing 3D images with high contrast for soft tissues.

MRI is based on the principles of nuclear magnetic resonance (NMR), a spectroscopic technique used by scientists to obtain physical information about molecules.

One of the most inaccessible and complicated organ, the human brain is the main beneficiary of this new imaging techniques. Excellent soft tissue contrast, the non-invasive and high-quality spatial resolution made it possible to study brain disorders such as Alzheimer's disease, multiple sclerosis, schizophrenia, congenital abnormalities, visual or hearing loss, brain tumors, etc. The brain contains three main classes of soft tissue: white matter, gray matter and cerebrospinal fluid. It is necessary to measure the quantities of white matter, gray matter and cerebrospinal fluid as well as their distribution in space and over time the changes for the diagnosis of various diseases of the brain.

2 High Dynamic Range Imagery

Our goal is to improve the quality of MRI images by endowing them with *High Dynamic Range* (HDR) features.

The *dynamic range* DR of an image is the ratio between the smallest and the largest possible values of the lighting. The DR of usual photographic subjects can be as high as

V.E. Balas et al. (eds.), *Soft Computing Applications*, Advances in Intelligent Systems and Computing 633, DOI 10.1007/978-3-319-62521-8_23

100 dB. The DR of the human sight may get close to 90 dB, meaning that we can see objects in dark (with reduced color differentiation) or in bright sunlight. It is to mention that we need few seconds to adjust our eyes to different light levels. We expect our visual display systems (image acquisition, processing and rendition) to approach DR = 100 dB, being able to capture both shadow details in dark and bright areas of sunny scenes, with controlled distribution of the resolution and of the DR all over the image surface. However, the highest DR provided by the best image sensors (the high end CCDs) can reach only 80 dB. CMOS sensors are even weaker from the DR point of view, but unlike CCDs, they can combine the image sensor function with image processing, within the same integrated circuit, due to the use of the cheap, reliable, fast and energy-intensive standard CMOS integrated circuits technology.

That is why the HDR techniques were developed in order to enhance the final images' DR, capturing from the photographic subject as much information as possible, from dark areas and bright areas in the same time. In the particular case of the neuroimaging, higher DR make possible more accurate demarcations between different tissues, and thereby better diagnoses and better surgical operations.

The conventional desktop *multi-frame HDR* needs multiple standard exposures of the same still scene: *low dynamic range* frames (LDR). LDRs with different exposures are obtained by *exposure bracketing*, with camera mounted on tripod. Each LDR needs its own shutter cycle.

The idea of using multiple exposures to extend the luminance range appeared in 1850. The idea belongs to Gustave Le Gray and it involves combining details sky and the sea, all in the same image (Fig. 1). Le Gray used a negative for sky and another one, overexposed, for the sea [3].

Fig. 1. The first multiple frame HDR image (Le Gray, 1850)

Starting from the desktop HDR the next step was to realize the image enhancement by built-in-camera HDR or HDR image sensors. Such way HDR is generated in the very moment of the exposure and the amount of information maximizes.

The speed is important in order to cope with moving subjects. In medical imagery fast captures give sharper images and reduce the patients' discomfort.

The fastest photographic HDR algorithm is the *Burst Readout Multiple Exposure* (BROME) [4], which takes full advantage of the processing capability of the CMOS image sensors. While conventional cameras read an image frame once during the frame period T_F, by one mechanical shutter cycle, BROME produces m readouts during one shutter cycle. One long and three short accumulation time signals are read during a frame period: LA = $3T_R$ (long accumulation), SA (short), VSA (very short) and ESA (extremely short) (Fig. 2.). Such way a HDR capture can be obtained within a single shutter cycle, which is substantially increasing speed. On the other hand one observes that reset impulses are applied to pixels after each reading out, at the end of the time slots 3, 4, 5 and 6.

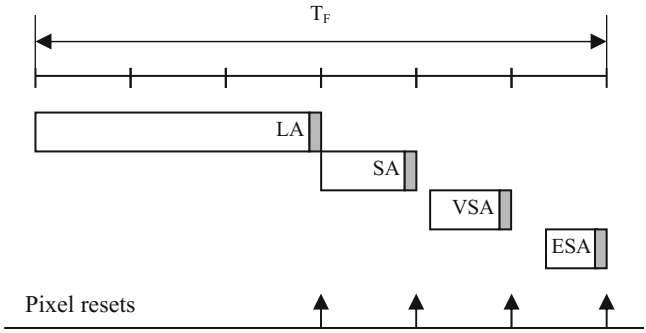


Fig. 2. The BROME readout sequence for a pixel

The *Cumulative Burst Readout Multiple Exposure* (CBROME) is even faster, taking advantage of the intrinsic CMOS technology assets: insulated gate and onchip image processing. The duration of a well exposed LDR frame (LA) is split into several partial exposures, which by their succession, are capturing the continuous loading of the pixels gates with electrical charges, corresponding to the irradiance of the photographic subject. After each partial exposure the current information of the pixels is read and memorized, but only one reset impulse is applied to the pixel, at the end of LA [5] (Fig. 3).

There were developed and tested numerous merging methods for the m LDRs, starting from Debevec initial approach [6]. The weighted sum is already producing very good applicative results and allows some simple developments, such as the fuzzy-interpolative custom HDR characteristics [7, 8] that may adjust the result according to the photographer intentions: best details in dark arias, best details in bright arias, great contrast etc.

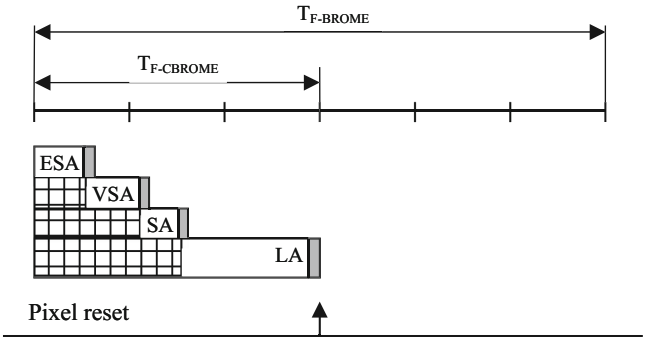


Fig. 3. The CBROME readout sequence for a pixel

3 Multiple Frame HDR for MRI Brain Images

The following processing is realized in Matlab, using the Image Processing Toolbox, starting from three frames (Figs. 4, 5 and 6):

- Main = imread('mri.jpg') – the normally exposed frame
- Sub = imread('mriSub.jpg') – the under exposed frame
- Supra = imread('mriSupra.jpg') – the over exposed frame

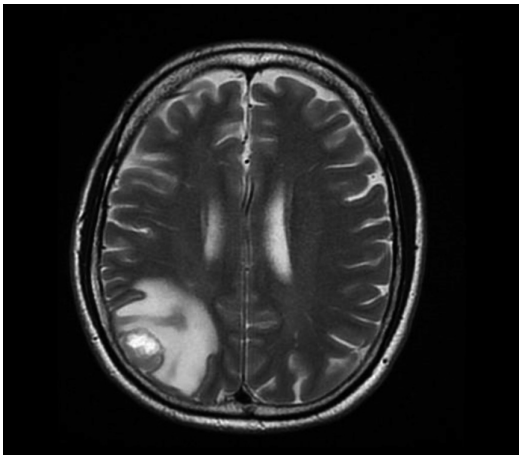


Fig. 4. The Main frame

The Matlab Image Processing Toolbox offers several functions for merging images. Figure 7 shows an HDR image obtained with the function *Fimlincomb* (linear combinations of the input images). Although very simple, the linear combinations are very convenient in applications, the improvement of the images is easily visible and no distortions of the dynamic characteristics are generated.

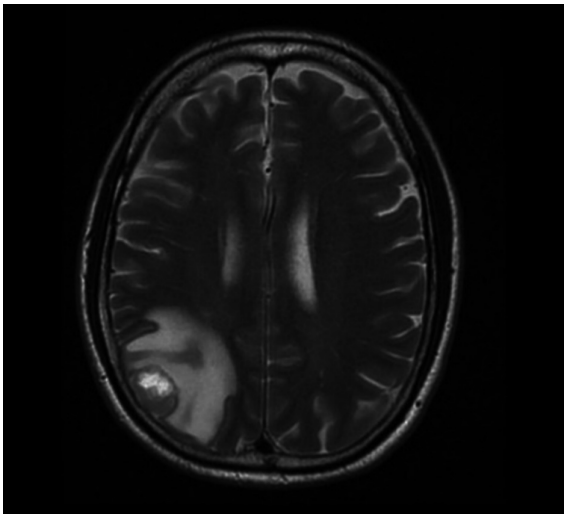


Fig. 5. The Sub frame

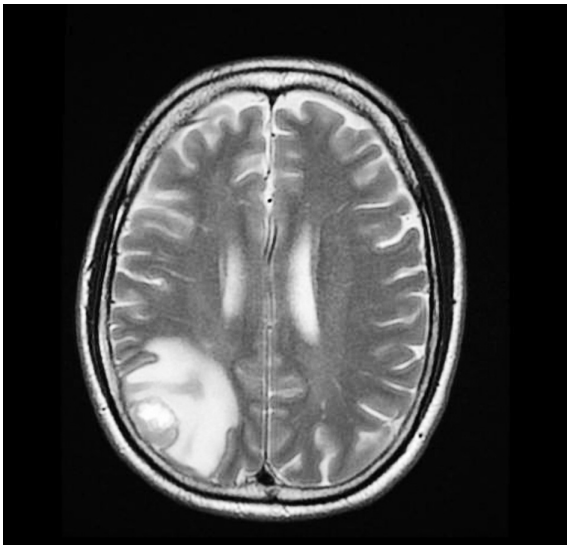


Fig. 6. The Supra frame

With two exposures on brain MRI image, we used Matlab software to combine images to obtain a new MRI image. The output image provides a larger amount of details than a single MRI image, and offers a better perspective for segmentation.

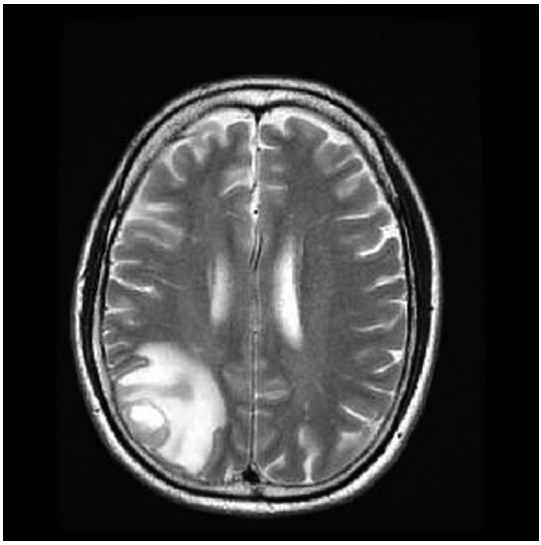


Fig. 7. The HDR frame (*fimlincomb*)

4 The Segmentation

The segmentation refers to the process of extracting meaningful objects in a region and it involves division or separation of image regions with similar attributes.

Medical image segmentation is one of the most important task in diagnosis, surgery planning, postsurgical evaluation and other medical actions. MRI brain images are essential in the early detection of tumors and necrotic tissue.

The previous HDR image was successfully used for a very simple, therefore fast and effective segmentation procedure, illustrated in Figs. 8, 9 and 10.

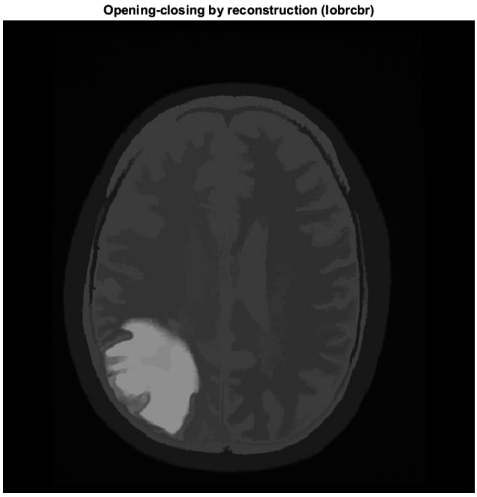


Fig. 8. Opening-closing by reconstruction

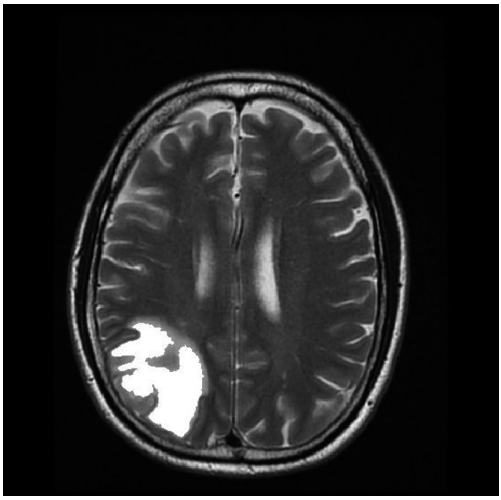


Fig. 9. Regional maxima of opening-closing by reconstruction (fgm)

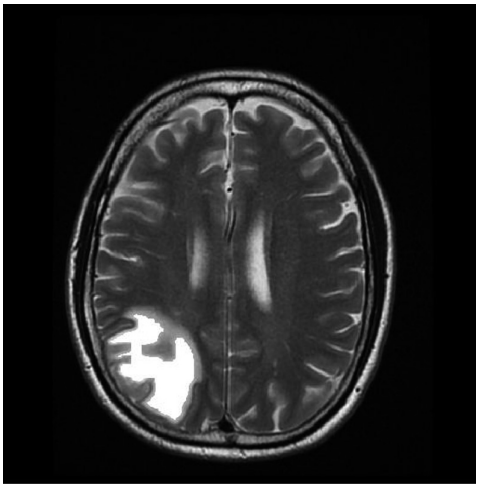


Fig. 10. Regional maxima overlapping the initial image

5 Discussion

The paper is presenting a method to obtain more details on MRI brain images by applying a particular fast multi-frame HDR technique, merging two or more frames with increasing exposures, obtained during a single exposure of the subject.

The objective of this work is to improve the quality of the brain MRI images and of the segmentation procedures applied for their detailed analysis.

Acknowledgement. This work was co-funded by European Union through European Regional Development Funds Structural Operational Program "Increasing of Economic Competitiveness" Priority axis 2, operation 2.1.2. Contract Number 621/2014.

References

1. Krishna Priya, R., Thangaraj, C., Kesavadas, C., Kannan, S.: Fuzzy entropy based MR brain image segmentation using modified particle swarm optimization. Int. J. Imaging Syst. Technol. **23**(4), 281–288 (2013)
2. Rajesh, C.P., Bhalchandra, A.S.: Brain tumour extraction from MRI images using MATLAB. Int. J. Electron. Commun. Soft Comput. Sci. Eng. **2**(1), 1–4 (2012)
3. Paul Getty Museum, J.: Gustave Le Gray, Photographer, 9 July–29 September (2002). http://wwwgetty.edu/art/exhibitions/le_gray/. Accessed 12 July 2016
4. Sasaki, M., Mase, M., Kawahito, S., Tadokoro, Y.: A wide-dynamic-range CMOS image sensor based on multiple short exposure-time readout with multiple-resolution column-parallel ADC. IEEE Sens. J. **7**(1), 151–158 (2007)
5. Balas, M.M.: Adaptive and fast HDR image sensors. Resources of the CMOS multiple exposure. In: Proceedings of 8th International Conference on Computing Technology and Information Management, ICCM 2012, Seoul, Korea, 24–26 April, pp. 643–647 (2012)
6. Debevec, P.E., Malik, J.: Recovering high dynamic range radiance maps from photographs. In: Proceedings of SIGGRAPH 1997, Los Angeles, August, pp. 369–378 (1997)
7. Balas, M.M., Ciugudean, M.A., Lobont, M.: Custom high dynamic range characteristics developed by fuzzy-interpolative methodology. In: Proceedings of 4th International Workshop on Soft Computing Applications, SOFA 2010, Arad, 15–17 July, pp. 185–190 (2010)
8. Balas, V.E., Balas, M.M.: Fuzzy-interpolative linguistic controllers for custom HDR photo cameras. In: Proceedings of the World Conference on Soft Computing, WConSC 2011, San Francisco, 23–26 May (2011)
9. Sharma, A., Sharma, P., Rashmi, H.K.: Edge detection of medical images using morpholgical algorithms. Int. J. Sci. Emerg. Technol. Latest Trends (2012). http://www.mathworks.com/help/fuzzy. Accessed 15 May 2016
10. Chan, T.F., Vese, L.A.: Active contours without edges. IEEE Trans. Image Process. **10**(2), 266–277 (2001)
11. Caselles, V., Kimmel, R., Sapiro, G.: Geodesic active contours. Int. J. Comput. Vis. **22**(1), 61–79 (1997)
12. Whitaker, R.T.: A level-set approach to 3D reconstruction from range data. Int. J. Comput. Vis. **29**(3), 203–231 (1998)
13. Zuiderveld K.: Contrast limited adaptive histograph equalization. In: Graphic Gems IV, pp. 474–485. Academic Press Professional, San Diego (1994)
14. Gonzalez, R.C., Woods, R.E., Eddins, S.L.: Digital Image Processing Using MATLAB. Gatesmark Publishing, Knoxville (2009)
15. Haralick, R.M., Shapiro, L.G.: Computer and Robot Vision, vol. I, pp. 158–205. Addison-Wesley, Boston (1992)
16. van den Boomgard, R., van Balen, R.: Methods for fast morphological image transforms using bitmapped images. Comput. Vis. Graph. Image Process.: Graph. Models Image Process. **54**(3), 254–258 (1992)
17. Vincent, L.: Morphological grayscale reconstruction in image analysis: applications and efficient algorithms. IEEE Trans. Image Process. **2**(2), 176–201 (1993)

Computer Aided Posture Screening Using the Microsoft Kinect on Professional Computer Users for Malicious Postures

Norbert Gal-Nadasan[1(✉)], Vasile Stoicu-Tivadar[1], Dan V. Poenaru[2], Diana Popa-Andrei[3], and Emanuela Gal-Nadasan[3]

[1] Department of Automation and Applied Informatics, Politehnica University of Timisoara, Timisoara, Romania
{norbert.gal, vasile.stoicu-tivadar}@aut.upt.ro
[2] Clinic II of Orthopaedic and Traumatology, University of Medicine and Pharmacy "Victor Babes", Timisoara, Romania
danvpoenaru@gmail.com
[3] Department of Medical Rehabilitation, University of Medicine and Pharmacy "Victor Babes", Timisoara, Romania
andreidiana8l@gmail.com, emma.nadasan@gmail.com

Abstract. This paper suggest the usage of the Microsoft Kinect in the screening of the professional computer users for malicious postures. Since the Microsoft Kinect is a marker less body scanning sensor based on an infra-red structured light architecture the harmful exposure of the screened patients is reduced to zero. During the screening process several key joints of the human body are tracked like the hips and shoulders. The Cartesian 3 dimensional coordinates of the track joints are used to digitally recreate the patient's posture. The postural data is statistically processed using IBM Watson's Analytics to analyze if the screened patients are susceptible to develop malicious postures or if they already have some kind of malicious posture.

Keywords: Sitting posture · Microsoft Kinect · IBM Watson

1 Introduction

When the patient presents at a cosultation, the medical doctor evaluates the patient form the clinical point of view. This evaluation includes a discussion about the chronological age, the time spent in the front of the computer screen, and the medical family history about spine diseases, with congenital hip dysplasia, with neurological illness, or other major diseases.

In the case of young adults it is nessesary the screening of the the spine in colleges, to diagnose early any vertebral axial deviation. In the begging it can be only a postural deviation but after a few years of evolution it transforms to a structural scoliosis and vertebral body modifications.

In the moment when we stay a long time at computer we develop a bad posture [1, 2], our back develops a C shape in the top of the spinal column, the shoulders are elevated and the head has a pushed-forward position, the curve of the lumbar spine is

V.E. Balas et al. (eds.), *Soft Computing Applications*, Advances in Intelligent Systems and Computing 633, DOI 10.1007/978-3-319-62521-8_24

bigger and we have a protruding stomach [3]. In this position we put pressure over the hip joints and the lower back presented in Fig. 1. The correct seating position to maintain the correct body posture is presented in Fig. 2.

Fig. 1. Incorrect seating position

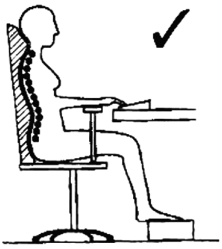


Fig. 2. Correct seating position

For an accurate examination the patient must be in a standing position. When we have a leg inequality the patient must be examined from a seated position, because like that we correct the leg inequality.

In the first moment we make a posterior body examination [4] we follow: a marked hump in the middle of the back, asymmetry of the shoulders and the legs, asymmetry of iliac crests, asymmetry of the buttocks, marked hump at the convexity of spinal curvature, the existence of spina bifida, a birth defect or other injuries of the spine.

At an anterior examination of the body we follow: asymmetry of the shoulders, asymmetry of the sternum, asymmetry and inequality of the breasts and asymmetry of anterior-superior iliac spines.

Palpation, when the patient is palpated we may found painful places and contracted muscles.

The most important symptom that determines the adult patient to came to the doctor is the pain.

This pain may be an acute one or a chronic one. Many patiens take NSAID (Nonsteroidal anti-inflammatory drug) at home and they came to the medical doctor only when the NSAID don't fix the pain.

The pain can be [5]:

- localized median or lateral,
- irradiating ascending and descending, latteraly, to the shoulder and thoracic region ot to the base of the thorax and abdomen,
- like a burning sensation or like an pinching,
- daily or during the night,
- agravated by the lying down of the patient, by movement, by seated position or by orthostatic position,
- Reduced in some favorite positions.

2 Posture Screening Method

As more and more young adults chooses to become an engineer in the IT domain like choosing software engineering, software programming or just work as a computer operator they have the risk of developing some kind of malicious body postures. The most common malicious posture malformations are the kyphotic and scoliotic body postures. In these postures the natural curvatures of the spinal cord are deformed. In kyphosis the cervical curvature is overly deformed to the front and in the scoliotic posture the spinal cord is deformed sideways as shown in Fig. 3.

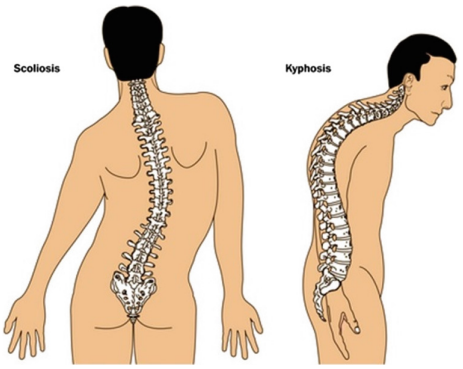


Fig. 3. Scoliotic and Kyphotic posture.

With regular screening these malicious body postures can be detected in early stages and corrected using medical rehabilitation exercises.

Each patient has to fill out a questionnaire to measure the subjective impact of the malicious body posture over the patient. It is advised to start these screening as early as possible.

2.1 The Screening Method

The proposed screening method uses a non-invasive, non-irradiant and marker less human body tracking method based on the Microsoft Kinect 3D sensor. This sensor has

already proven it's usability in the medical rehabilitation domain with several medical applications [6].

The sensor uses a structured light system based on an IR grid projected from the IR (infra-red) laser diode. Using an IR camera the system detects the grid and creates a depth map of the surrounding space. The system is capable to separate the human body from the rest of the objects. The detected body is represented as a "matchstick" skeleton like in Fig. 4.

Fig. 4. Kinect skeleton data

The markerless tracking system can track 20 joints of the human body and for each joint assigns a 3 dimensional value which represents the joint position in Cartesian space. The X coordinate represents the horizontal space, the Y coordinate represents the vertical space and the Z coordinate represent the depth space.

To get a relevant view about the patient's posture several correlations between the tracked joints must be analyzed. The most important correlations are given by the angles between the interested joints. These angles are calculated using the 2 vector method (Fig. 5).

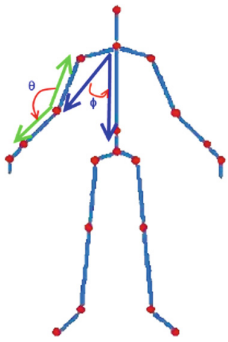


Fig. 5. The 2 vector method

The method implies 3 joints, the first joint called the middle joint represents the point where the angle is measured and two other adjacent joints between which the angle is measured. The formulas are presented below:

$$A \bullet B = (Ax \bullet Bx + Ay \bullet By) \tag{1}$$

$$|A| = \sqrt{Ax \bullet Ax + Ay \bullet Ay} \tag{2}$$

$$theta = \cos^{-1}\left(\frac{A \bullet B}{|A||B|}\right) \tag{3}$$

where A and B are the adjacent joints of the joint of interest and Ax, Ay, Bx and By are the Cartesian coordinates of the A and B points. $|B|$ is calculated using Eq. (2). *Theta* is the searched angle.

Using these angles at key points and there depth data an image of the posture can be created.

2.2 Tracked Body Points and Measured Values

The proposed screening method is a non-irradiant markerless tracking method which reduces the harmful radiation emitted from x-ray machines to zero.

The presence of scoliosis and kyphosis which affects most of the people that work in the IT domain [7] can be observed by tracking the following joints and properties of the joints:

- Height of the left and right shoulder and the difference between the two heights.
- The angle created by the two shoulders at the neck.
- The rotation of the shoulders: if one of the shoulder is significantly deeper than the other one.
- Height of the left and right hip and the difference between the two heights.
- The angle created by the two hips at the center of the hip.
- The rotation of the hips: if one of the hip is significantly deeper than the other one.

The measurements are saved into a *.CSV file. To get correct measurements the Microsoft Kinect must be set up in the following way:

- The patient must at 2.7 m from the sensor.
- The sensor must be at 65 cm from the ground, perpendicular to the patient's hip. If the patient is taller than 2 m the height of the sensor must be adjusted.
- The patient must not wear high heels otherwise affects the overall measurements.
- The patient must not wear loose clothing, it is recommended to wear a sports top and shorts.

The saved data from the CSV file is imported in an EXCEL compatible file and then analyzed using IBM's Watson analytics software [8].

2.3 Subjective Pain Questionnaire

To measure the lower back pain intensity of the screened patients 2 type of questionnaires were used. These questioners not only measure the subjective pain intensity but the impact of these malicious postures on the patients daily routine.

The first questionnaire that need to be answered was developed to measure the subjective correlation between the patients working environment the appearance of the low back pain associated with malicious body postures. This questioner has 19 questions. Three of the questions measure the time spent by the patient in front of the computer screen and the active daily physical activity.

The questions are the following:

- How many hours do you spend in front of a computer screen?
- How long have you been working/studying in the IT domain?
- How many hours do you spend for physical activity on a daily bases?
- Your pain is generated by your physical activity?
- Your physical activity can increase the lower back pain?
- Your physical activity can be harmful for your back?
- I can't fulfill my physical activity because is aggravating my low back pain.
- My low back pain was induced by my work or a work related accident.
- My work aggravates my low back pain.
- It is too hard for me to fulfill my work related tasks.
- With my low back pain I should not fulfill my job related task.
- Until treated I can't fulfill my job related tasks.
- In the following three month's I can't return to my job.
- I will never be able to return to my current job.

Another questioner is the based on the Roland – Morris questioner [9] that measures the impact of the low back pain on the patients daily routine. Several of the questions are presented below:

- I stay at home most of the time because of back pain.
- I change position frequently to try to alleviate my back pain.
- Back pain prevents me carry my usual housework.
- Because of lower back pain are forced to use the handrail when climbing stairs.
- Back pain often leads me to sleep, to rest.
- Because of lower back pain should I support something to lift me.
- Because back pain trying to find people to help me in ordinary activities.
- Back pain forces me to get dressed more slowly than usual.
- Because of lower back pain are forced to stand shorter periods of time.
- I find it hard to get back into bed from one side to another, due to lower back pain.
- I can't put on my shoes with socks (stockings) because of lower back pain.
- Because of lower back pain may not go through than short distances.
- Because of back pain I need another person to help me dress.
- Back pain causes me to avoid domestic work requiring great physical exertion.
- Because of back pain I became irritable and I behave badly with others.

From these two questionnaires an objective image can be obtained about how and when the malicious postures and the associated low back pain affects the patients on a daily bases. Using these data preventive actions can be formulated and implemented to prevent the progress and correct the malicious postures.

3 Results

To test the proposed system a lot 69 computer engineering students between the ages of 19 and 22 year old was measured. From this lot 28 were male students and 41 female students.

The raw data from the evaluations were analyzed using IBM's Watson analytics software. The data was introduced in the form of excel tables and evaluated as 92% high quality data according to the IBM Watson analytics data analysis system.

3.1 Posture Evaluation

The posture was evaluated statistically by calculating the relationship of the height of the shoulders and the height of the hips to the age and the patient's height.

The relationship of the shoulders height and hips height by age is presented in Figs. 6 and 7.

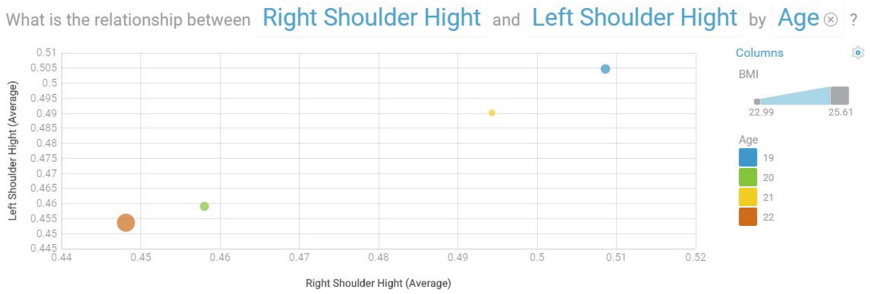


Fig. 6. Shoulders height in correlation with the age

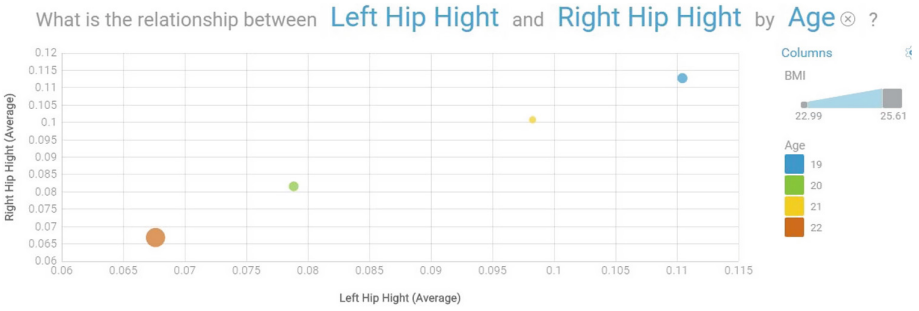


Fig. 7. Hips height in correlation with the age

According to these result the younger adults, in our case those at age of 19 suffer less from malicious postures and those at the age of 22 have a more pronounced malicious posture.

If we look at the correlation between the shoulders height and the height of the person the data is more scattered but in the screened lot those who were between 1.60 m and 1.75 m tall have a susceptibility of malicious posture. The results are presented on the image below (Fig. 8).

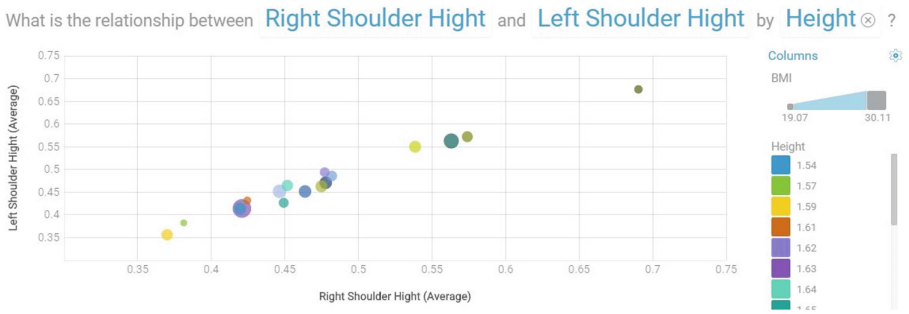


Fig. 8. Correlation between the shoulders height and the height of the person

The single person from the lot who is 1.92 m tall shows himself predisposition to a malicious position (Fig. 9).

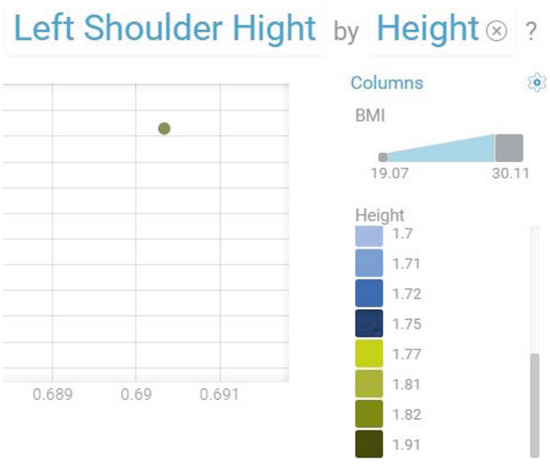


Fig. 9. Correlation between the shoulders height and the height of the 1.92 m tall person

3.2 Questionar Results

The target group was at the age of 19 and 22 year. Due to their young age the long time 4–7 h a day spent in front of the computer screen the impact of the malicious posture associated with low back pain is relatively low. As the majority of the screened lot goes to school by foot the time spent walking can considered as a slight physical activity.

At the end of the day there is no significant pain, only a small discomfort at the lumbar part of the spinal cord. Even this slight discomfort indicates the presents of a fatigue at the lumbar part of the torso which must not be ignored. The fatigue can be eliminated with regular exercises.

4 Conclusions

From the results it can be observed that the measured lot developed light symptoms of malicious postures due to long sitting hours in front of a poorly adjusted chair and computer monitor. The most affected from the malicious posture are the tall patients because of the abnormal sitting position. They tend to lean forward by bending the torso to get a proper viewing angle of the computer screen. Another wrong sitting position is when the patients sit in a semi seated position on the chair.

These siting positions favor the appearance of malicious postures associated with low back pain. The best prevention method against the malicious postures and the associated low back is a 10 min break form work and 30 min of daily physical exercises.

References

1. Ankrum, D.R., Nemeth, K.J.: Head and neck posture at computer workstations - what's neutral? In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 44, no. 30, pp. 565–568, July 2000
2. Claus, A.P., Hides, J.A., Moseley, G.L., Hodges, P.W.: Thoracic and lumbar posture behaviour in sitting tasks and standing: progressing the biomechanics from observations to measurements. Appl. Ergon. **53**, 161–168 (2016)
3. Straker, L.M., Smith, A.J., Bear, N., O'Sullivan, P.B., de Klerk, N.H.: Neck/shoulder pain, habitual spinal posture and computer use in adolescents: the importance of gender. Ergonomics **54**(6), 539–546 (2011)
4. Ramirez, N., Charles, E., Johnston, I.I., Browne, R.H.: The prevalence of back pain in children who have idiopathic scoliosis. J. Bone Joint Surg. Am. **79**(3), 364–368 (1997)
5. Silva, J.A.P., Woolf, D.A.: Rheumatology in Practice (2010). ISBN 978-1-84882-580-2
6. Clark, R.A., Pua, Y.-H., Fortin, K., Ritchie, C., Webster, K.E., Denehy, L., Bryant, A.L.: Validity of the microsoft kinect for assessment of postural control. Gait & Posture **1**, 372–377 (2012)
7. Brink, Y., Louw, Q., Grimmer, K., Jordaan, E.: The relationship between sitting posture and seated-related upper quadrant musculoskeletal pain in computing South African adolescents: a prospective study. Manual Ther. **20**(6), 820–826 (2015)

8. Winters-Miner, L.A., Bolding, P.S., Hilbe, J.M., Goldstein, M.R., Hill, T., Nisbet, R., Walton, N., Miner, G.D.: IBM watson for clinical decision support. In: Practical Predictive Analytics and Decisioning Systems for Medicine, pp. 1038–1040. Academic Press (2015)
9. Roland, M.O., Morris, R.W.: A study of the natural history of back pain. Part 1: Development of a Reliable and Sensitive Measure of Disability in Low Back Pain. Spine, vol. 8, pp. 141–144 (1983)

IT Complex Solution Supporting Continuity of Care

Mihaela Crișan-Vida(✉), Liliana Bărbuț, Alexandra Bărbuț, and Lăcrămioara Stoicu-Tivadar

Automation and Computer Science Faculty, University Politehnica Timișoara, Bd. Vasile Parvan no. 2, 300223 Timisoara, Romania
mihaela.vida@upt.ro, barbutliliana@gmail.com, alexandrambarbut@gmail.com, lacramioara.stoicu-tivadar@aut.upt.ro

Abstract. The paper presents a lab tests information system with two modules: a desktop one (for laboratory staff) and a web application accessed by the patient on different devices (Smartphone or Tablet). Continuity of care is ensured through cloud computing (Windows Azure) and standardized communication (HL7 Clinical Document Architecture). The benefits of the system are that the laboratory results are always available for the medical staff or the patients; in case of emergency the medical staff has access to patient lab results history and may improve the patient treatment based on the evolution of the lab results.

Keywords: Laboratory information system · HL7 CDA · Cloud storage · Interoperability

1 Introduction

A data management information system supporting specific laboratory activities is referred as a laboratory information system (LIS). The paper presents an application focused on interoperability and continuity of care in relation with activities performed in a clinical laboratory. The main characteristics in such a system are: workflow management, support for processing, storing and searching data, and providing a flexible architecture and interface to exchange data [1].

LIS is an evolving concept that frequently adds features and functionalities. Given the continuing demand for change in laboratory work and technological changes, the LIS functions evolve continually. Despites these changes, LIS has a basic set of main functions. These functions are [2]:

- Identifying the patient based on ID, and identification of the laboratory sample associated with the day of the test
- Distribution, scheduling and tracking of the test sample in progress
- Processing and quality control of the test sample and testing equipment

V.E. Balas et al. (eds.), *Soft Computing Applications*, Advances in Intelligent Systems and Computing 633, DOI 10.1007/978-3-319-62521-8_25

– Storing the resulting data related to the sample analysis
– Inspection, validation and interpretation of results for reporting and/or the need of further analysis.

The advantages of using an information system are: instant access to the patient results anytime-anyplace, easy update of the system with new functionalities, continually update of the prices and test analysis, access to patient history by the medical staff and by the patient.

[3, 4] describe functionalities and patient data management for two laboratory information systems.

To provide continuity of care it is important that the laboratory results are always available and the system which accesses the data storage has the right privileges and the data is available in a standardized format. To fulfill these requirements, the solution uses cloud computing and standardization. Our system uses Microsoft Azure cloud and as standard, the HL7 Clinical Document Architecture.

It is important to provide an advanced communication between different medical units and the supporting information system in a complex and highly dynamic environment that has to fulfill several requirements: to be open, scalable, flexible, portable, distributed, standard conformance, service oriented semantic interoperability and appropriate security and privacy services [5].

The standard supporting our system is HL7 Clinical Document Architecture, a document markup standard that specifies the structure and semantics of "clinical documents" for data exchange, a complete information object which can include text, images, sounds, and other multimedia data and could be any of the following: discharge summary, referral, clinical summary, history/physical examination, diagnostic report, prescription, or public health report [6, 7].

The application uses Microsoft Azure cloud, a cloud computing platform and infrastructure developed by Microsoft. It uses the blob storage containing the XML in HL7 CDA format, a collection of binary data stored as a single entity in a database management system. This storage may be accessed by other medical units, having authorization and can read HL7 CDA documents.

The paper presents two modules for LIS, the communication with other medical units using standardized communication and the cloud storage solution.

2 System Architecture

Figure 1 presents the system architecture. The system has two components: the desktop application for the medical staff and the web application for the patient. The database is on the Windows Azure cloud. This is the storage from where other medical units may access the laboratory results stored in the cloud container.

The desktop application is developed with Visual Studio.net 2015, using C# language.

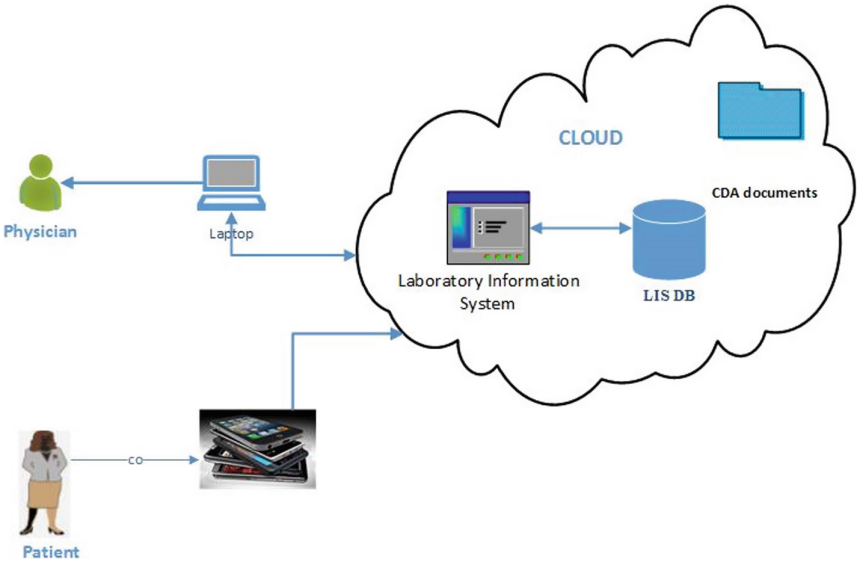


Fig. 1. System architecture

The web application is developed with Visual Studio.net 2015, using ASP.NET pages and C# language, and Bootstrap for the patient module. The patient accesses the web application on different devices and Bootstrap is suitable because is a framework for developing responsive, mobile-first web sites [8].

In the Windows Azure containers exists a folder to store the XMLs in HL7 CDA format accessed by different medical units.

3 Modeling the Laboratory Information System

Figure 2 presents the real workflow of current LIS using Business Process Modeling and Notation (BPMN) language [9]. This language is a graphical notation that describes the logic of steps in a business process. The workflow was modeled using Bizagy BPMN Modeler [10].

In Process 1 the main actor is represented by the laboratory staff. Input of new patient data implies authentication of test lab staff and validation. If he/she has the correct role than may input patient data (demographic, laboratory results, invoices) into a Patient Record, also if a record is previously created.

In Process 2 the main actor is the patient. If he/she wants to visualize laboratory results or the history chart for a specific analyze (e.g. glycaemia), the first step is to authenticate and after the system verifies if the actor has access to visualize his/her personal data, the patient can visualize the laboratory results.

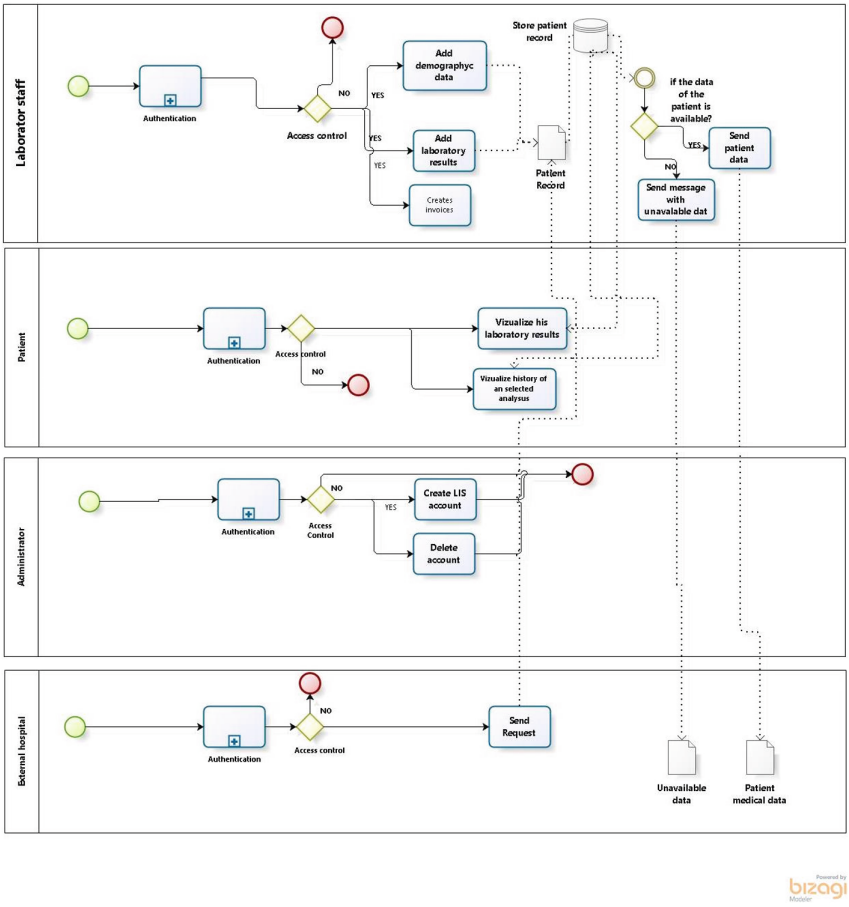


Fig. 2. Real workflow in LIS

In Process 3 the main actor is the administrator. If he wants to make changes into the system the first step is authentication and after the system verifies if the actor has access, the changes into the system are operated. If the actor has the correct role than he/she has the possibility to create new accounts or delete accounts.

In Process 4 the main actor is an external hospital which requests data about a patient. Firstly, the actor must authenticate in the system and if the credentials are correct then the system can send a XML in HL7 CDA format.

4 Laboratory Information System

Figure 3 presents a print screen from the LIS for desktop where an analyst adds a new patient.

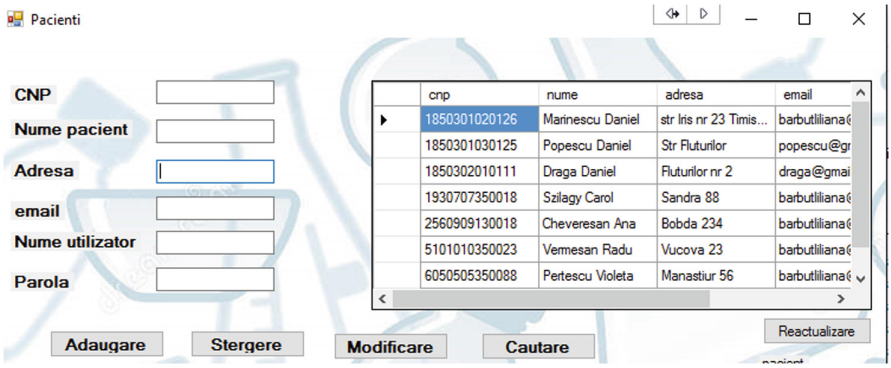


Fig. 3. Interface where an analyst can add a patient

Figure 4 presents a print screen from the LIS for the patient module where a patient can visualize his/her results on mobile.

Fig. 4. Patient system interface to view analysis results

5 Using HL7 Clinical Document Architecture for a Standardized Communication

Communication between LIS and other medical units accessing the laboratory results uses the HL7 CDA standard. Figure 5 presents how a medical unit can access the lab results of a patient. The XML in HL7 CDA format is stored in the Windows Azure blob and may be accessed by other information systems with rights to receive this information.

Figure 5 shows how the information system for the general practitioner and cardiology department accesses the XML in HL7 CDA format.

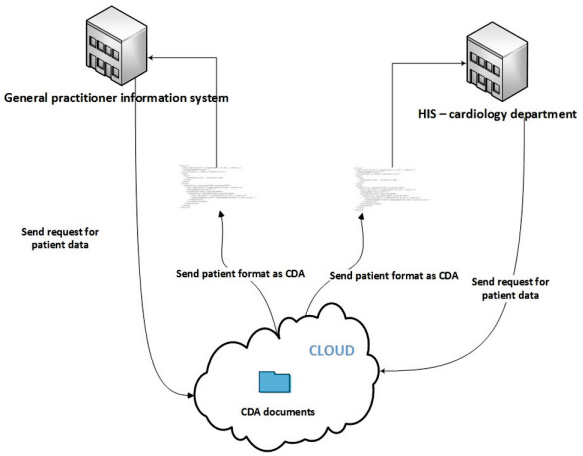


Fig. 5. Medical units accessing XML in HL7 CDA format

An XML in HL7 CDA format is composed by a header and a body, the major components of a HL7 CDA is presented after [11] in Fig. 6.

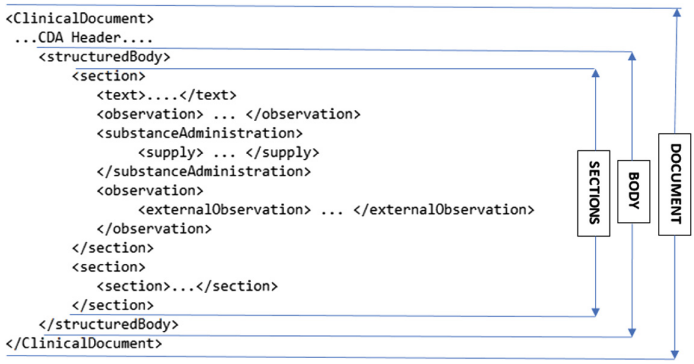


Fig. 6. Major components of HL7 CDA after [11]

Figure 7 presents the body part in a laboratory result, the beta-HCG. It is important that in the header of the XML in the HL7 CDA format to have the patient identification, the institution which created the document and the person responsible for the results.

```
<!--******************Laborator***************-->
<component>
    <section>
        <code code="11502-2" codeSystem="2.16.840.1.113883.6.1" codeSystemName="LOINC" displayName="Laborator" />
        <title>Test de laborator</title>
        <entry>
            <observation classcode="OBS" moodCode="EVN">
                <code code="19180-9" codeSystem="2.16.840.1.113883.6.1" codeSystemName="LOINC" displayName="beta-HCG" />
                <effectiveTime value="20160505" />
                <value xsi:type="PQ" value="115000" unit="mUI/ml" />
            </observation>
        </entry>
    </section>
</component>
```

Fig. 7. Section of a HL7 CDA body

6 Cloud Storage of Laboratory Results

The results from a laboratory tests are saved in a database and also the information system creates the XML in HL7 CDA format and saves it in a container on the Microsoft Azure cloud.

The system uses Windows Azure cloud which is a cloud computing platform and infrastructure. This cloud is a Microsoft product, where a software developer can build and manage applications and services through a global network of Microsoft datacenters.

Figure 8 presents the cloud storage where the XML in HL7 CDA are stored.

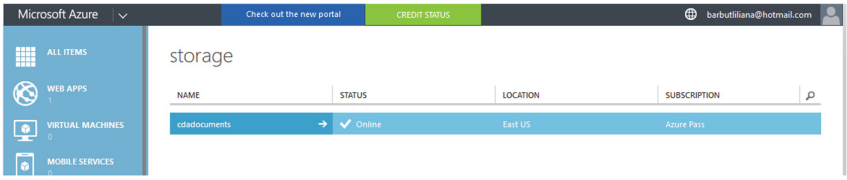


Fig. 8. Windows Azure storage

Figure 9 presents the container where the XML in HL7 CDA format will be available for other information systems from other medical units can access.

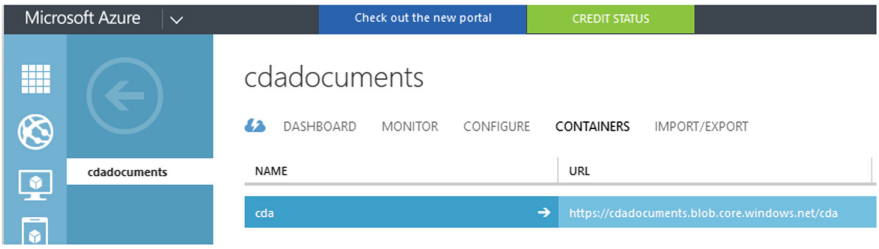


Fig. 9. Windows Azure container

7 Conclusions

A LIS in the presented framework helps the lab test staff to store and access the patient's lab test results, to create the invoices for the patient, and also to support the patient receiving the results quicker and easier, having a history of the results, accessing different results on different times or visualizing the chart with a specific analysis.

Cloud computing offers many opportunities to improve healthcare services and valuable data will be available in critical moments, resulting in a better care for the patient.

Using a standardized communication, currently HL7 CDA, will improve the communication between LIS and other medical units ensuring continuity of care. The lab test results will be always available and the other medical units will understand the data received.

The system is also cost-effective because reduces the cost with unnecessary analysis test. Sometimes is not need to repeat a test so frequently and the physician can access the LIS and check if the patient has that specific test already done.

References

1. Laboratory Information Management: So what is a LIMS, Sapio Sciences (2010). Accessed 15 Apr 2016
2. Skobelev, D.O., Zaytseva, T.M., Kozlov, A.D., Perepelitsa, V.L., Makarova, A.S.: Laboratory information management systems in the work of the analytic laboratory. Meas. Tech. **53**(10), 1182–1189 (2011)
3. e-Lab. http://www.ebiosys.com/Products.aspx. Accessed 10 Apr 2016
4. ParcPlus&MedClinica. http://www.cormedic.ro/software-laborator/. Accessed 10 Apr 2016
5. Blobel, B., Engel, K., Pharow, P.: HL7 version 3 compared to advanced architecture standards. Methods Inf. Med. **45**, 343–353 (2006)
6. HL7 Clinical Document Architecture, Release 2.0, HL7 version 3 Interoperability Standards, Normative Edition 2009, Disk 1 – Standards Publication
7. Vida, M., Gomoi, V., Stoicu-Tivadar, L., Stoicu-Tivadar, V.: Generating medical computer-based protocols using standardized data transmission. In: 4th International Workshop on Soft Computing Applications, Arad, pp. 155–158 (2010)
8. Microsoft Azure. https://azure.microsoft.com/en-us/. Accessed 20 Apr 2016
9. Business Process Modeling and Notation. http://www.omg.org/spec/BPMN/. Accessed 03 May 2016
10. Bizagy BPMN Modeler. http://www.bizagi.com/en/products/bpm-suite/modeler. Accessed 10 Apr 2016
11. Dolin, R.H., Alschuler, L., Boyer, S., Beebe, C., Behlen, F., Biron, P., Shabo, A.: HL7 clinical document architecture, release 2. J. Am. Med. Inform. Assoc. **13**(1), 30–39 (2006)

The Importance of Quantification of Data in Studies on the Health Effects of Exposure to Electromagnetic Fields Generated by Mobile Base Stations

S.M.J. Mortazavi[1], Valentina Emilia Balas[2(✉)], A. Zamani[3], A. Zamani[4], S.A.R. Mortazavi[5], M. Haghani[1], O. Jaberi[6], and A. Soleimani[7]

[1] Ionizing and Non-ionizing Radiation Protection Research Center (INIRPRC), Shiraz University of Medical Sciences, Shiraz, Iran
mmortazavi@sums.ac.ir
[2] Department of Automatics and Applied Informatics, Aurel Vlaicu University, B-dul Revolutiei 77, 310130 Arad, Romania
balas@drbalas.ro
[3] Department of Statistics, Shiraz University, Shiraz, Iran
a_zamani2@yahoo.com
[4] Medical Physics and Medical Engineering Department, Shiraz University of Medical Sciences, Shiraz, Iran
atefeh.zamani@hotmail.com
[5] Student Research Committee, Shiraz University of Medical Sciences, Shiraz, Iran
a.mortazavi.72@gmail.com
[6] Occupational Health Department, School of Health, Shiraz University of Medical Sciences, Shiraz, Iran
mefomp@sums.ac.ir
[7] Epidemiology Department, School of Medicine, Shiraz University of Medical Sciences, Shiraz, Iran
aslami_epid@yahoo.com

Abstract. This study aimed at developing simple methods for quantification of the data in studies on the health effects of exposure to electromagnetic fields of mobile base stations to prevent different interpretations of the findings. We have previously conducted a cross-sectional study on people living in 10 different districts of Shiraz city. Based on how frequent the symptoms were (always/usually/occasionally/never), a score was determined for each self-reported symptom and the total score for each individual was calculated. In the next stage, instead of reporting the effect of RF-EMF exposure on the occurrence of each symptom, the impact of exposures on the human health were assessed by evaluation of the total scores. Moreover, principal component analysis (PCA) was used for assessing the factors with greatest correlation. We found that exposure to RF-EMF emitted by mobile base stations significantly affected the residents' health. Furthermore, the distance between the mobile phone base stations and the homes had a significant effect on the residents' health. The frequency of self-reported symptoms of myalgia, palpitation, early fatigue, nervousness and low back pain decreased with increasing the distance from mobile phone base stations.

V.E. Balas et al. (eds.), *Soft Computing Applications*, Advances in Intelligent Systems and Computing 633, DOI 10.1007/978-3-319-62521-8_26

Keywords: Data quantification · Mobile base stations · Electromagnetic fields (EMFs) · Radiofrequency (RF) · Health effects

1 Introduction

Today, humans produce, transmit and utilize electricity in a exponentially increasing manner that is believed to be an essential component of the modern life. This electricity-linked modern life has raised rapidly growing concerns about the adverse health effects of exposure to different levels of electromagnetic fields. The modern life now is associated with rapid development of different wireless communications systems such as satellites, cell phones, Internet, and Wi-Fi. In this light, continuous exposure to radiofrequency radiation can be considered as an inevitable part of civilization and radiofrequency radiation from mobile phone base stations and satellite antennas is among the main sources to which the *world's population is exposed.*

Mobile phones and their base stations use electromagnetic radiation for transmitting and receiving radiofrequency (RF) signals. Currently, the largest radiofrequency electromagnetic fields (RF-EMF) exposures are attributed to mobile phone use [1–4]. The exponential growth in the use of mobile phones has led to significant expansion of mobile phone base stations. This growth has raised global concerns about the safety of mobile phones and their base stations and prompted scientists to investigate different aspects of the health issues of exposure to these common sources of RF-EMF [5–13]. Over the past decade, the health effects of exposure to RF-EMF from mobile phone base stations have been investigated in humans [14–19] and animals [20–24]. However, due to existence of numerous confounding factors, the findings are still controversial. While some scientists have reported that the incidence of most of the self-reported symptoms in the people living in the vicinity of mobile base stations was related to RF-EMF exposure levels [25], other researchers who reviewed the recent literature have concluded that there is no sufficient data to draw strong conclusions regarding these common long-term low-level exposures [26].

Romania is among the European countries in which several studies have been performed on the safety issues of mobile phone stations [27, 28]. To date, all measured electric field values were far below the reference levels suggested by the *International Commission on Non-Ionizing Radiation Protection* (*ICNIRP*) [28]. Although Romanian scientists haveperformed a number of measures to protect the general public against exposure to RF-EMFs, currently there is no specific national standard in Romania and the guidelines of ICNIRP [29] have been adopted as standard for limiting human exposures [30].

We have previously criticized the validity of the studies reporting a significant relationship between the incidence of most of the symptoms and exposure levels [31]. Our lab at the non-ionizing department of the Ionizing and Non-ionizing Radiation Protection Research Center (INIRPRC) has performed experiments on the health effects of exposure of animal models and humans to different sources of electromagnetic fields such as cellular phones [32–39], mobile base stations [40], mobile phone jammers [41, 42], laptop computers [43], radars [33], dentistry cavitrons [44] MRI [38, 45] and Helmholtz coils [46]. Our research team has also shown that exposure to

some common sources of electromagnetic fields (through increased release of mercury from dental amalgam restorations) might be linked to higher incidence of diseases such as speech problems [47]. Based on our findings, we also hypothesized that these exposures through increased release of mercury from amalgam fillings) can be associated with higher incidences of ADHD or *Autism spectrum disorder* (ASD) [48, 49].

A cross-sectional study on people living in 10 different districts of Shiraz city was previously performed by our team [50]. This study aimed at developing simple methods for quantification of the data in studies on the health effects of exposure to electromagnetic fields of mobile base stations to prevent different interpretations of the findings.

2 Materials and Methods

A total of 755 randomly selected people (488 males and 267 females) were enrolled. Each participant was interviewed by an expert team and *self-reported symptoms were recorded* using a standardized questionnaire. Participants were asked about presence/absence of a variety of symptoms. Residential addresses were recorded and distance from each antenna was calculated. Based on how frequent the symptoms were (always/usually/occasionally/never), a score was determined for each self-reported symptom and the total score for each individual was calculated. In the next stage, instead of reporting the effect of RF-EMF exposure on the occurrence of each symptom, the impact of exposures on the human health were assessed by evaluation of the total scores. Therefore, the effect of factors such as the distance (d) from base stations on the frequency of subjective symptoms (symp) such as headache, vertigo, sleep problem, myalgia, palpitation, fatigue, tinnitus, concentration problem, attention problem, nervousness, back pain and toothache was calculated as follows:

$$\text{Total Score}_{\text{d}} = \left(\text{Score}_{\text{symp1}} + \text{Score}_{\text{symp2}} + \ldots + \text{Score}_{\text{symp n}}\right)/\text{n}$$

In the next stage, principal component analysis (PCA), a statistical procedure which uses an orthogonal transformation to convert a set of findings (possibly correlated variables) into principal components (a set of values of linearly uncorrelated variables), was used for assessing the factors with greatest correlation. This study was conducted in accordance with the guidelines of the Helsinki declaration [51] and the medical ethics guidelines of Shiraz University of Medical Sciences.

3 Results

Demographic characteristics of the participants are indicated in Table 1. The mean (±SD) age of the participants was 31.92 ± 11.96 years (ranged 11–79 years). The education level of the participants was relatively moderate, with more than 70% holding a high school diploma or a college degree compared to 26.2% of the population who did not graduate from high school. When instead of reporting the effect of RF-EMF exposure on the occurrence of each symptom, the impact of exposures on the

Table 1. Demographic characteristic of the participants

Demographic characteristic	Categories	Frequency (%)
Sex	Male	488 (64.6%)
	Female	267 (35.4%)
Age	Mean ± SD (years)	31.92 ± 11.56
	(min–max)	(5–79)
	5–20	87 (11.5%)
	21–30	348 (46.1%)
	31–40	158 (20.9%)
	40+	162 (21.5%)
Educational level	High school diploma	198 (26.3%)
	Diplomato Bachelor	521 (69.1%)
	Graduate	33 (4.4%)
	Missing	2 (0.02%)

human health were assessed by evaluation of the total scores, we found that exposure to RF-EMF emitted by mobile base stations significantly affected the residents' health.

Principal component analysis (PCA) of the self-reported subjective symptoms is shown in Table 2. As presented in this table, myalgia, palpitation, fatigue, nervousness and back pain are categorized as class-I (the symptoms which showed the strongest correlation with each other). Among other symptoms, concentration problem, and attention problem are categorized as class-II (the symptoms which showed a 2nd level of correlation with each other) and tinnitus and toothache are in class-III (the symptoms which showed a 3rd level of correlation with each other). Moreover, the distance

Table 2. Principal component analysis (PCA) of the self-reported subjective symptoms.

Rotated component matrix[a]			
	Component		
	1	2	3
Myalgia	0.772	−0.062	0.041
Palpitation	0.531	0.089	0.293
Fatigue	0.745	0.255	−0.019
Tinnitus	0.279	0.089	0.616
Concentration problem	0.179	0.901	0.106
Attention problem	0.181	0.894	0.107
Nervousness	0.573	0.309	0.038
Back pain	0.573	0.200	0.209
Toothache	−0.037	0.080	0.841
Extraction method: principal component analysis Rotation method: varimax with Kaiser normalization[a]			

a. Rotation converged in 5 iterations

between the homes and base stations had a significant effect on the residents' health. The frequency of self reported symptoms of myalgia, palpitation, early fatigue, nervousness and low back pain decreased with increasing the distance from mobile phone base stations (P = 0.029).

4 Discussion

It is not yet clear whether exposure to RF-EMFs emitted by mobile phone base stations can adversely affect human health. Currently, there are great discrepancies in the evaluation of the health effects of exposure to RF-EMFs emitted by mobile phone base stations [52]. The findings of this study showed that exposure to RF-EMF emitted by mobile base stations significantly affected the residents' health. It was also revealed that the distance between the mobile phone base stations and homes had a significant effect on the residents' health. It is worth mentioning that in our previous report which used conventional analysis (analysis of the effect of exposure on each self-reported symptom), there was no association between the distance from base stations and the frequency of nearly all the subjective symptoms (we had only found a significant association between the distance from base stations and the frequency of tooth ache). In this light, we believe that conventional analysis can be considered as the reason for controversial findings regarding the role of the distance from base stations. The following reports show some of these controversial results. Statistically significant health effects in people living within 300 m of mobile phone base stations have been reported by Santini et al. in 2002. These researchers suggested that a minimum safe distance of 300 m should be considered between mobile phone base stations and inhabited areas [19].

On the other hand, in Poland, Bortkiewicz et al. have reported that among 57% of their study participants who had headache, 36.4% lived 100–150 m away from the base stations. Moreover, 24.4% of the study participants, mostly living distances above 150 m, had reported memory problems [15]. Blettner et al. in Germany also showed that participants living in the vicinity of a mobile phone base station at distances less than 500 m, as well as those who were concerned about the health effects of RF-EMF exposures, reported slightly more health complaints than other participants [17]. Kundi et al. in Austria also reported that despite it was impossible to determine a threshold below which no health effect occurs, the power densities of mobile phone base station must be above 0.5–1 mW/m^2 to observe adverse health effects [16].

From the other point of view, a recent study which was performed by aggregating the data from two previous studies, aimed at answering whether exposure to RF-EMFs produced by mobile phone base stations causes adverse health effects in people with electromagnetic hypersensitivity (EHS) and control individuals. In this study, no causal relationship between short-term exposure to EMFs and subjective well-being in EHS and control individuals was found [53].

As discussed by Coggon, a major weakness of the studies conducted so far, is that the data about both exposure and subjective symptoms are collected by questioning participants. Therefore, Coggon believes that as a major limitation in these studies, risk estimates may have been exaggerated through biased recall [54]. A recent cross-sectional

study that is conducted on 440 individuals in Brazil, studied the psychiatric complaints of the residents of areas around mobile phone base stations. This study showed an association between psychiatric symptoms and residential proximity to the base station. In this study it was revealed that factors such as gender, schooling, and smoking status could not affect the significant association between exposure to RF-EMF from base stations and the psychiatric symptoms [55]. In 2015, Klaps et al. carried out a meta-analysis on the findings of 17 different studies. They reported that double-blind studies could not found any adverse effects on human well-being. However, field or unblinded studies were capable of showing detrimental effects. These researchers concluded that some effects are due to nocebo effect [52].

In light of the findings obtained in this study, it can be claimed that the previous reports which denied any risk to the health of people living in the vicinity of mobile phone base stations, are no longer valid. For example an independent expert group reported that as the levels of exposure to mobile phone base stations are only small fractions of currently accepted guidelines, the balance of evidence shows no general risk to the health of people living near to base stations [56]. Moreover, the British Advisory Group on Non-Ionizing Radiation in 2003 reported that exposure levels coming from living in the vicinity of mobile phone base stations are very low, hence based on current evidence, they generally cannot make any threats to people living in the vicinity of base stations [57]. In England, the National Radiological Protection Board (NRPB) in 2004 reported that their measurements showed that the people exposure to radiofrequency radiation emitted by macrocell base stations are only a small fraction of current guidelines. It is worth noting that NRPB has studied the radiation emission pattern of 60 base stations and reported that exposure levels in the vicinity of picocells were only a few per cent of the guidelines for general public [58].

In this study, as indicated in Table 2, the first axis of the PCA identified a series of intercorrelated symptoms (myalgia, palpitation, fatigue, nervousness and back pain). These are the symptoms which showed the strongest correlation with each other. The second axis identified two intercorrelated symptoms of concentration problem, and attention problem. And finally, the third axis identified two other intercorrelated symptoms of tinnitus and toothache. These findings can suggest that the mechanism of the RF-EMF effects on body tissues for occurrence of these symptoms have similar patterns.

At a broader scale, our findings are in line with the substantial evidence which show exposure to RF-EMF affects the incidence of self reported symptoms such as myalgia [39], fatigue [59–65], and nervousness [66]. However, in contrast with some previous studies, we had no complaints of headache, earache, warmth sensation, cognitive disturbances, vertigo/dizziness, sleep disturbance-insomnia, tension-anxiety, joint and bone pain, lacrimation of the eyes, hearing loss and tinnitus [15, 59–61, 63, 65, 67–70].

5 Conclusion

It is not yet well understood if exposure to RF-EMFs emitted by mobile phone base stations adversely affects human health. Currently, there are discrepancies in the assessment of the health effects of exposure to RF-EMFs from a base station. Although the RF-EMF levels of mobile base stations do not exceed the international

limits, if people are exposed to these low-intensity RF-EMFs for a very long time, serious health problems can occur [71]. Altogether, these findings indicate that developing simple methods for quantification of the data can help scientist better evaluate the controversial issue of the health effects of exposure to RF-EMF of mobile base stations.

Acknowledgement. This study was supported by the Ionizing and Non-ionizing Radiation Protection Research Center (INIRPRC), Shiraz University of Medical Sciences (SUMS), Shiraz, Iran and Aurel Vlaicu University, Arad, Romania.

References

1. Bolte, J.F., Eikelboom, T.: Personal radiofrequency electromagnetic field measurements in The Netherlands: exposure level and variability for everyday activities, times of day and types of area. Environ. Int. **48**, 133–142 (2012)
2. Joseph, W., Frei, P., Roosli, M., Thuroczy, G., Gajsek, P., Trcek, T., et al.: Comparison of personal radio frequency electromagnetic field exposure in different urban areas across Europe. Environ. Res. **110**(7), 658–663 (2010)
3. Frei, P., Mohler, E., Neubauer, G., Theis, G., Burgi, A., Frohlich, J., et al.: Temporal and spatial variability of personal exposure to radio frequency electromagnetic fields. Environ. Res. **109**(6), 779–785 (2009)
4. Joseph, W., Vermeeren, G., Verloock, L., Heredia, M.M., Martens, L.: Characterization of personal RF electromagnetic field exposure and actual absorption for the general public. Health Phys. **95**(3), 317–330 (2008)
5. Zhu, Y., Jin, W., Liu, H., Peng, D., Ding, Z., Tang, Z., et al.: Effects of electromagnetic fields from mobile phones on depression and anxiety after titanium mesh cranioplasty among patients with traumatic brain injury. Brain Inj. **30**(1), 66–73 (2016). doi:10.3109/02699052.2015.1089594
6. Yoon, S., Choi, J.W., Lee, E., Ahn, H., Kim, H.S., Choi, H.D., et al.: Mobile phone use and risk of glioma: a case-control study in Korea for 2002–2007. Environ. Health Toxicol. **30**, e2015015 (2015). doi:10.5620/eht.e2015015. eCollection 2015
7. Morgan, L.L., Miller, A.B., Sasco, A., Davis, D.L.: Mobile phone radiation causes brain tumors and should be classified as a probable human carcinogen (2A) (review). Int. J. Oncol. **46**(5), 1865–1871 (2015)
8. Malek, F., Rani, K.A., Rahim, H.A., Omar, M.H.: Effect of short-term mobile phone base station exposure on cognitive performance, body temperature, heart rate and blood pressure of Malaysians. Sci. Rep. **5**, 13206 (2015)
9. Ikinci, A., Mercantepe, T., Unal, D., Erol, H.S., Sahin, A., Aslan, A., et al.: Morphological and antioxidant impairments in the spinal cord of male offspring rats following exposure to a continuous 900 MHz electromagnetic field during early and mid-adolescence. J. Chem. Neuroanat. **75**(Pt. B), 99–104 (2016). doi:10.1016/j.jchemneu.2015.11.006
10. Huss, A., van Eijsden, M., Guxens, M., Beekhuizen, J., van Strien, R., Kromhout, H., et al.: Environmental radiofrequency electromagnetic fields exposure at home, mobile and cordless phone use, and sleep problems in 7-year-old children. PLoS ONE **10**(10), e0139869 (2015)
11. Yogesh, S., Abha, S., Priyanka, S.: Mobile usage and sleep patterns among medical students. Indian J. Physiol. Pharmacol. **58**(1), 100–103 (2014)
12. Sadetzki, S., Langer, C.E., Bruchim, R., Kundi, M., Merletti, F., Vermeulen, R., et al.: The MOBI-kids study protocol: challenges in assessing childhood and adolescent exposure to electromagnetic fields from wireless telecommunication technologies and possible association with brain tumor risk. Front. Public Health **2**, 124 (2014)

13. Carlberg, M., Hardell, L.: Decreased survival of glioma patients with astrocytoma grade IV (glioblastoma multiforme) associated with long-term use of mobile and cordless phones. Int. J. Environ. Res. Public Health **11**(10), 10790–10805 (2014)
14. Gandhi, G., Kaur, G., Nisar, U.: A cross-sectional case control study on genetic damage in individuals residing in the vicinity of a mobile phone base station. Electromagn. Biol. Med. **34**(4), 344–354 (2015)
15. Bortkiewicz, A., Gadzicka, E., Szyjkowska, A., Politanski, P., Mamrot, P., Szymczak, W., et al.: Subjective complaints of people living near mobile phone base stations in Poland. Int. J. Occup. Med. Environ. Health **25**(1), 31–40 (2012)
16. Kundi, M., Hutter, H.P.: Mobile phone base stations-effects on wellbeing and health. Pathophysiology **16**(2–3), 123–135 (2009)
17. Blettner, M., Schlehofer, B., Breckenkamp, J., Kowall, B., Schmiedel, S., Reis, U., et al.: Mobile phone base stations and adverse health effects: phase 1 of a population-based, cross-sectional study in Germany. Occup. Environ. Med. **66**(2), 118–123 (2009)
18. Bortkiewicz, A., Zmyslony, M., Szyjkowska, A., Gadzicka, E.: Subjective symptoms reported by people living in the vicinity of cellular phone base stations: review. Med. Pr. **55** (4), 345–351 (2004)
19. Santini, R., Santini, P., Danze, J.M., Le Ruz, P., Seigne, M.: Investigation on the health of people living near mobile telephone relay stations: I/Incidence according to distance and sex. Pathologie-biologie **50**(6), 369–373 (2002)
20. Hassig, M., Jud, F., Spiess, B.: Increased occurrence of nuclear cataract in the calf after erection of a mobile phone base station. Schweiz. Arch. Tierheilkd. **154**(2), 82–86 (2012)
21. Ohtani, S., Ushiyama, A., Maeda, M., Ogasawara, Y., Wang, J., Kunugita, N., et al.: The effects of radio-frequency electromagnetic fields on T cell function during development. J. Radiat. Res. **56**(3), 467–474 (2015)
22. Klose, M., Grote, K., Spathmann, O., Streckert, J., Clemens, M., Hansen, V.W., et al.: Effects of early-onset radiofrequency electromagnetic field exposure (GSM 900 MHz) on behavior and memory in rats. Radiat. Res. **182**(4), 435–447 (2014)
23. Lerchl, A., Klose, M., Grote, K., Wilhelm, A.F., Spathmann, O., Fiedler, T., et al.: Tumor promotion by exposure to radiofrequency electromagnetic fields below exposure limits for humans. Biochem. Biophys. Res. Commun. **459**(4), 585–590 (2015)
24. Jeong, Y.J., Kang, G.Y., Kwon, J.H., Choi, H.D., Pack, J.K., Kim, N., et al.: 1950 MHz electromagnetic fields ameliorate abeta pathology in Alzheimer's disease mice. Curr. Alzheimer Res. **12**(5), 481–492 (2015)
25. Gomez-Perretta, C., Navarro, E.A., Segura, J., Portoles, M.: Subjective symptoms related to GSM radiation from mobile phone base stations: a cross-sectional study. BMJ Open **3**(12), e003836 (2013)
26. Roosli, M., Frei, P., Mohler, E., Hug, K.: Systematic review on the health effects of exposure to radiofrequency electromagnetic fields from mobile phone base stations. Bull. World Health Organ. **88**(12), 887–896 (2010)
27. Miclaus, S., Bechet, P., Goiceanu, C., Olariu, O.V., Demeter, S.: Practical issues concerning cellular base station emissions and population exposure in Romania. In: The 3rd European Medical and Biological Engineering Conference EMBEC 2005 and IFMBE, IFMBE Proceedings CD-ROM ISBN 1727-1983 Prague, Czech Republic, November 2005. ISSN 0363-6135
28. Miclaus, S., Bechet, P.: Estimated and measured values of the radiofrequency radiation power density around cellular base stations. Rom. J. Phys. **52**(3/4), 429 (2007)
29. Protection ICoN-IR: ICNIRP statement on the guidelines for limiting exposure to time-varying electric, magnetic, and electromagnetic fields (up to 300 GHz). Health Phys. **97**(3), 257–258 (2009)

30. GSM Association: Base station planning permission in Europe. GSM Europe Report (2004)
31. Mortazavi, S.M.: Subjective symptoms related to GSM radiation from mobile phone base stations: a cross-sectional study. J. Biomed. Phys. Eng. **4**(1), 39–40 (2014)
32. Mortazavi, S.M.J., Motamedifar, M., Namdari, G., Taheri, M., Mortazavi, A.R., Shokrpour, N.: Non-linear adaptive phenomena which decrease the risk of infection after pre-exposure to radiofrequency radiation. Dose-Response **12**(2), 233–245 (2014)
33. Mortazavi, S.M.J., Taeb, S., Dehghan, N.: Alterations of visual reaction time and short term memory in military radar personnel. Iran. J. Publ. Health **42**(4), 428–435 (2013)
34. Mortazavi, S.M.J., Rouintan, M.S., Taeb, S., Dehghan, N., Ghaffarpanah, A.A., Sadeghi, Z., et al.: Human short-term exposure to electromagnetic fields emitted by mobile phones decreases computer-assisted visual reaction time. Acta Neurol. Belg. **112**(2), 171–175 (2012)
35. Mortazavi, S.M.J., Mosleh-Shirazi, M.A., Tavassoli, A.R., Taheri, M., Mehdizadeh, A.R., Namazi, S.A.S., et al.: Increased radioresistance to lethal doses of gamma rays in mice and rats after exposure to microwave radiation emitted by a GSM mobile phone simulator. Dose-Response: Publ. Int. Hormesis Soc. **11**(2), 281–292 (2013)
36. Mortazavi, S., Mosleh-Shirazi, M., Tavassoli, A., Taheri, M., Bagheri, Z., Ghalandari, R., et al.: A comparative study on the increased radioresistance to lethal doses of gamma rays after exposure to microwave radiation and oral intake of flaxseed oil. Iran. J. Radiat. Res. **9** (1), 9–14 (2011)
37. Mortazavi, S.M.J., Habib, A., Ganj-Karimi, A.H., Samimi-Doost, R., Pour-Abedi, A., Babaie, A.: Alterations in TSH and thyroid hormones following mobile phone use. OMJ **24**, 274–278 (2009)
38. Mortazavi, S.M.J., Daiee, E., Yazdi, A., Khiabani, K., Kavousi, A., Vazirinejad, R., et al.: Mercury release from dental amalgam restorations after magnetic resonance imaging and following mobile phone use. Pak. J. Biol. Sci. **11**(8), 1142–1146 (2008)
39. Mortazavi, S.M.J., Ahmadi, J., Shariati, M.: Prevalence of subjective poor health symptoms associated with exposure to electromagnetic fields among university students. Bioelectromagnetics **28**(4), 326–330 (2007)
40. Mortazavi, S.M.J.: Safety Issue of Mobile Phone Base Stations. J. Biomed. Phys. Eng. **3**(1), 1–2 (2013)
41. Parsanezhad, M.E., Mortazavi, S.M.J., Doohandeh, T.: Exposure to radiofrequency radiation emitted from mobile phone jammers adversely affects the quality of human sperm. IJRR **15**(1), 63–70 (2017)
42. Mortazavi, S.M.J., Parsanezhad, M.E., Kazempour, M., Ghahramani, P., Mortazavi, A.R., Davari, M.: Male reproductive health under threat: Short term exposure to radiofrequency radiations emitted by common mobile jammers. J. Hum. Reprod. Sci. **6**(2), 124–128 (2013)
43. Mortazavi, S.M.J., Tavasoli, A.R., Ranjbari, F., Moamaei, P.: Effects of laptop computers' electromagnetic field on sperm quality. J. Reprod. Infertil. **11**(4), 251–258 (2011)
44. Mortazavi, S.M., Vazife-Doost, S., Yaghooti, M., Mehdizadeh, S., Rajaie-Far, A.: Occupational exposure of dentists to electromagnetic fields produced by magnetostrictive cavitrons alters the serum cortisol level. J. Nat. Sci. Biol. Med. **3**(1), 60–64 (2012)
45. Mortazavi, S.M.J., Neghab, M., Anooshe, S.M.H., Bahaeddini, N., Mortazavi, G., Neghab, P.: High-field MRI and mercury release from dental amalgam fillings. THEIJOEM **5**(2), 101–105 (2014)
46. Haghnegahdar, A., Khosrovpanah, H., Andisheh-Tadbir, A., Mortazavi, G., Saeedi Moghadam, M., Mortazavi, S., et al.: Design and fabrication of helmholtz coils to study the effects of pulsed electromagnetic fields on the healing process in periodontitis: preliminary animal results. J. Biomed. Phys. Eng. **4**(3), 83–90 (2014)

47. Zarei, S., Mortazavi, S.M., Mehdizadeh, A.R., Jalalipour, M., Borzou, S., Taeb, S., et al.: A challenging issue in the etiology of speech problems: the effect of maternal exposure to electromagnetic fields on speech problems in the offspring. J. Biomed. Phys. Eng. **5**(3), 151–154 (2015)
48. Mortazavi, G., Mortazavi, S.M.: Increased mercury release from dental amalgam restorations after exposure to electromagnetic fields as a potential hazard for hypersensitive people and pregnant women. Rev. Environ. Health **30**(4), 287–292 (2015)
49. Mortazavi, G., Haghani, M., Rastegarian, N., Zarei, S., Mortazavi, S.: Increased release of mercury from dental amalgam fillings due to maternal exposure to electromagnetic fields as a possible mechanism for the high rates of autism in the offspring: introducing a hypothesis. J. Biomed. Phys. Eng. **6**(1), 41–46 (2016)
50. Mortazavi, S.M.J., Tavakkoli-Golpayegani, A., Jaberi, O., Haghani, M., Soleimani, A., Mortazavi, S.A.R., et al. (eds.): Management of the Risks Associated with Long Term Exposure to Radiofrequency Radiations of Mobile Base Stations. International Conference on Biological, Chemical and Environmental Sciences (BCES-2014) 2014, Patong Beach, Phuket (Thailand), 21–22 January 2014. IICBE
51. World medical association declaration of helsinki: Recommendations guiding physicians in biomedical research involving human subjects. JAMA **277**(11), 925–926 (1997)
52. Klaps, A., Ponocny, I., Winker, R., Kundi, M., Auersperg, F., Barth, A.: Mobile phone base stations and well-being - a meta-analysis. Sci. Total Environ. **544**, 24–30 (2015)
53. Eltiti, S., Wallace, D., Russo, R., Fox, E.: Aggregated data from two double-blind base station provocation studies comparing individuals with idiopathic environmental intolerance with attribution to electromagnetic fields and controls. Bioelectromagnetics **36**(2), 96–107 (2015)
54. Coggon, D.: Health risks from mobile phone base stations. Occup. Environ. Med. **63**(5), 298–299 (2006)
55. Silva, D.F., Barros, W.R., Almeida, M.D., Rego, M.A.: Exposure to non-ionizing electromagnetic radiation from mobile telephony and the association with psychiatric symptoms. Cadernos de saude publica **31**(10), 2110–2126 (2015)
56. Stewart, W.: Independent Expert Group on Mobile Phones: Report on Mobile Phones and Health (2000)
57. Swerdlow, A.: Health effects from radiofrequency electromagnetic fields. Doc. NRPB. **14**(2), 1–181 (2003)
58. NRPB: Mobile Phones and Health 2004: Report by the Board of NRPB. Doc. NRPB **15**, 1–114 (2004)
59. Pall, M.L.: Microwave frequency electromagnetic fields (EMFs) produce widespread neuropsychiatric effects including depression. J. Chem. Neuroanat. (2015)
60. Lamech, F.: Self-reporting of symptom development from exposure to radiofrequency fields of wireless smart meters in victoria, australia: a case series. Altern. Ther. Health Med. **20**(6), 28–39 (2014)
61. Kucer, N., Pamukcu, T.: Self-reported symptoms associated with exposure to electromagnetic fields: a questionnaire study. Electromagn. Biol. Med. **33**(1), 15–17 (2014)
62. Hagstrom, M., Auranen, J., Ekman, R.: Electromagnetic hypersensitive Finns: Symptoms, perceived sources and treatments, a questionnaire study. Pathophysiology **20**(2), 117–122 (2013)
63. Kato, Y., Johansson, O.: Reported functional impairments of electrohypersensitive Japanese: a questionnaire survey. Pathophysiology **19**(2), 95–100 (2012)
64. Fragopoulou, A.F., Samara, A., Antonelou, M.H., Xanthopoulou, A., Papadopoulou, A., Vougas, K., et al.: Brain proteome response following whole body exposure of mice to mobile phone or wireless DECT base radiation. Electromagn. Biol. Med. **31**(4), 250–274 (2012)

65. Baliatsas, C., Van Kamp, I., Bolte, J., Schipper, M., Yzermans, J., Lebret, E.: Non-specific physical symptoms and electromagnetic field exposure in the general population: can we get more specific? A Syst. Rev. Environ. Int. **41**, 15–28 (2012)
66. Roosli, M., Moser, M., Baldinini, Y., Meier, M., Braun-Fahrlander, C.: Symptoms of ill health ascribed to electromagnetic field exposure–a questionnaire survey. Int. J. Hyg. Environ. Health **207**(2), 141–150 (2004)
67. Korpinen, L.H., Paakkonen, R.J.: Self-report of physical symptoms associated with using mobile phones and other electrical devices. Bioelectromagnetics **30**(6), 431–437 (2009)
68. Wu, L., Feng, J.W., Peng, X.W., Mei, Y., Shi, Y.Q., Jiang, G.F., et al.: A cross-sectional study about the effects of residential exposure to power transmission line on human nervous system symptoms and hematology. Zhonghua lao dong wei sheng zhi ye bing za zhi = Zhonghua laodong weisheng zhiyebing zazhi = Chin. J. Ind. Hyg. Occup. Dis. **30**(5), 361–363 (2012)
69. McCarty, D.E., Carrubba, S., Chesson, A.L., Frilot, C., Gonzalez-Toledo, E., Marino, A.A.: Electromagnetic hypersensitivity: evidence for a novel neurological syndrome. Int. J. Neurosci. **121**(12), 670–676 (2011)
70. Heinrich, S., Thomas, S., Heumann, C., von Kries, R., Radon, K.: Association between exposure to radiofrequency electromagnetic fields assessed by dosimetry and acute symptoms in children and adolescents: a population based cross-sectional study. Environ. Health: Glob. Access Sci. Source **9**, 75 (2010)
71. Sorgucu, U., Develi, I.: Measurement and analysis of electromagnetic pollution generated by GSM-900 mobile phone networks in Erciyes University. Turkey. Electromagn. Biol. Med. **31**(4), 404–415 (2012)

Graphical Method for Evaluating and Predicting the Human Performance in Real Time

Mariana-Rodica Milici[1(✉)], Oana Geman[2], Iuliana Chiuchisan[3], and Laurentiu-Dan Milici[1]

[1] Electrotechnics Department, "Stefan cel Mare" University, 13 Universitatii Street, 720229 Suceava, Romania
mami@eed.usv.ro, dam@usm.ro

[2] Department of Health and Human Development, "Stefan cel Mare" University, 13 Universitatii Street, 720229 Suceava, Romania
oana.geman@usm.ro

[3] Computers, Electronics and Automation Department, "Stefan cel Mare" University, 13 Universitatii Street, 720229 Suceava, Romania
iuliana.chiuchisan@usm.ro

Abstract. Theoretical-scientifical significance of the paper consists in developing and reasoning a complex method for predicting sportive performance, during a training stage, that complements the theoretical concept of structure and content of education. Through the theoretical and experimental results presented in this work, one can see that the mathematical model and algorithm developed and applied in training the athletes has led to significant results regarding the predicting methodology in sports and improving the psychomotor and psychological parameters. This methodology can be effectively applied by extension to other levels of human practice because human performance can manifest differently. The theoretical and methodological concept may be included in the theoretical and methodical training of specialists in sports but also in other areas of human performance (education, artistic and scientific creativity etc.). Considering that training is a complex process and has the priority aim to prepare performance, we consider that in its structure and content must be implemented a whole range of advanced methodologies including computers. The mathematical methods developed by computing technique must include sufficient information on subjects' developing level and how to provide their improvement. The methodology of predicting based on mathematical modeling methods must highlight the most important directions of developing the individual both in the initial stages and in the performance stages of training. The method aims to reduce the multidimensional representation given by an extensive package of tests at a two-dimensional graphical representation, easily analyzed and very suggestive, by minimizing the extrapolation errors of numerical values involved.

Keywords: Human performance · Graphical method · Approximation functions · Reduction the total sizes

V.E. Balas et al. (eds.), *Soft Computing Applications*, Advances in Intelligent Systems and Computing 633, DOI 10.1007/978-3-319-62521-8_27

1 The Necessity of Using an Intuitive Graphical Methods for Evaluating the Human Performance

Sport, like any other segment corresponding to human activity, is an area which has as dominant the competition and the motivation for obtaining sports results determined the increasing in the number of hours of intense effort. Education also has become more complex due to the aid granted by the specialists and researchers in the field. There is now a broad data base about athletes, base that is reflected in the methodology of sports training [1–6].

Most part of scientific knowledge, whether it is taken from practical activity or from research activity, targets the understanding and optimizing the effects of sportive exercises on the body. The research in various sciences comes to enrich the theory and methodology of education, which was transformed into an independent science. Education is a complex pedagogical and biological process, conducted systematically, continuous and gradual, in order to adapt the human body to physical and mental efforts of various intensities, for obtaining results with a certain value. It is based on the competitive practicing forms of physical exercise thus is defined as a complex process whose goal is reproducing the sports performance in a systematic manner and goal-oriented. In this context, *the complex training* is an action process performed in order to obtain precise effects in relation to all the features that determine the performance of an individual or a team. During the training, the subject reacts to various stimuli, of which some are predictable, others less predictable. From the direct process of training, the physiological, biochemical, psychological and social information is registered as requirements, rules, principles, etc. The education principles are the foundation of this complex process.

A person is born with different predispositions which implies the achievement of a specific educational process. Talent is, in its most part, of genetic nature. The aptitudes of strength, speed, resistance and skill inherited plays an important role in achieving high levels of performance. In training does not happen that a single quality to dominate the effort and the specific movement. From this point of view, in training, not only one quality is educated/trained, but the combination of two or more. The researches conducted till now emphasize that the human mind allows a faster analysis in case of a graphical representation than of a series of numerical values. This is why it appeals to an intuitive graphical method, a method that can be applied in other fields of activities. For analysis, at the end of the training period, it can use the same graphical representation, which is based on two indicators that put each individual in a plane through a point.

The terms of model and modeling have penetrated deep into the theory and practice of education. The functions (cognitive, formative-educational, instrumental and normative) fulfilled by the models used in solving the problems of theory and practice, can have different shares. The efficient management of the educational process is related to the use of different models. The models used in education fall into two basic groups. First group includes the models that characterize the structure of competitive activity, namely those aimed at different sides of training, the morphological models that reflects the morphological and functional features of body, thereby assuring the achievement of

level required by sport performance. In the second group are included the models that reflect the continuity and dynamics of establishing performance and of designing a plan for short, medium, long and very long term and also the models of different training exercises with the providing of their complexity [1].

A. *The general models* reflect the features of the object or process, obtained from the study carried on a large group of individuals of a particular gender, age and that are practicing a certain type of activity.
B. *The group models* are built based on a study of a group of individuals (or companies), being distinguished by a specific index in the areas of each type of activity.
C. *The individual models* are developed for each individual in part and they are based on the lengthy research data, on its separate training and its reactions to various tasks, etc.

The human performances are assessed by periodic tests. Based on these tests it can draw conclusions as to how the individual has responded to a specific training program, to the parameters that can be developed, to the degree of accumulated fatigue, etc. In order to extrapolate these data for monitoring the future progress and for predicting future performance is necessary to find a law of progess of values regularly monitored up to a certain point. This prediction can be made for one or more tests, at the middle of training period in order to infer the progress or in the second half or at the end of this period in order to prepare a new training program.

From specialized literature it results that the issue of predicting human performance is focussed on using various prediction methods (method of extrapolation, of modeling and of expertise). The prediction carried out by applying the method of extrapolation allows the hierarchization of the results based on the study of some regularities corresponding previous period. The accuracy of achieving the prediction can be corrected if the prediction period is shorter and if the intermediary data are numerous [8]. The extrapolation is widely used as a method of modeling and obtaining expertise data, that's why, will find, in the future, an increasingly wide applicability of it in prediction. The prospect to apply the prediction will be more accurate if it will be used more quickly, more effectively and if it will use the possibilities of information technology, with which the sports results obtained during an educational macrocycle or more macrocycles will be processed and analyzed.

Knowing that the evolution of human performance is achieved by leaps (discontinuous functions, with thresholds of variable level, occurring at different times for the same sportsman and with variations hardly approximated from one individual to another) [1–6] the time periods for which the approximation is done must be longer than the time between two performance jumps of the tested athlete, but also must not include more than three levels of progress, because the prediction for a too long time can not be done. Because the functions of type stage of development can not be approximated, is trying the approximation using continuous functions that coincide with the evolving functions for both ends of the approximation range or for maximum two other points, approximately equal chosen in this range (Fig. 1) [4].

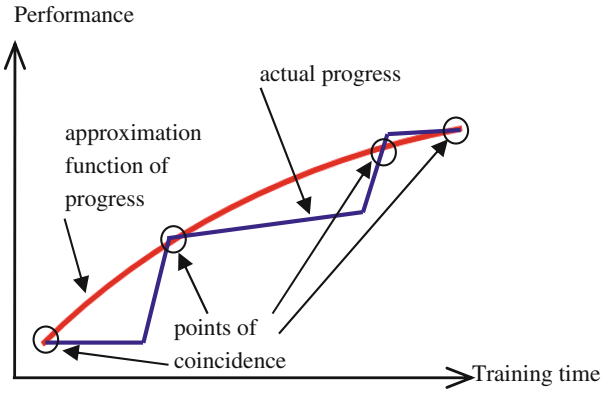


Fig. 1. Approximation function of progress

Starting from the idea that the psychomotor training should be based on the integration of the motor and mental functions, under the effect of the nervous system maturation, in general, and of the individuals' age, we carried out a study on the functional, motor, psycho-motor and psychological development of a lot of swimmers of small age. According to the theory and methodology of training, sports training aims to optimize and maximize the ability to obtain sport performance. In this respect it is necessary to develop the particularities of the growing body, through sports training as long-term activity, which involves many techniques and specific means of intervention.

In training the performance athletes, is necessary a periodical evaluation of their performance, with the purpose to optimize the training program depending on their evolution. If we refer to swimming, it's about monitoring the athletes performance in the psychomotor tests that allow a complex analisys on preparedness of each sportsman and a training personalization depending on the specifics of each sportsman [9].

The following describes an original method to analize the sportsmen lot using a grafichal representation that allows an intuitive evaluation of the potential sportsmen able of performance, evaluation that is made on the half of the 8 month training period (the fourth month) or at the end of this period (Fig. 2). The sportsmen placed as close to the top-right corner are those that allow certainly obtaining superior time intervals. The sportsmen of which the punctiform representation is placed toward the left-bottom corner didn't have progress during the 4 past training months, but they had from the beginning a performance superior to the average of the group. The sportsmen placed under the horizontal axis, at the right, didn't have progress after training and didn't start with data that to assure them a convenient position into the tested sportsmen group. Depending on this representation, it can draw conclusions regarding the sportsmen who deserve to continue the training or regarding to those sportsmen which either don't have aptitudes for this sport or it requires a changing in the training program in order to record the desired performance [7].

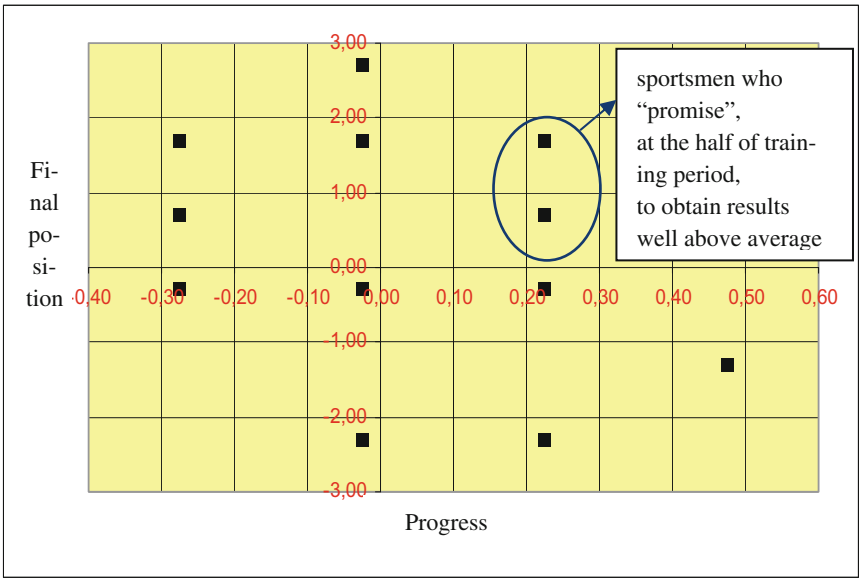


Fig. 2. The proposed graphical representation

For analysis at the end of training period, it can use the same graphical representation. The representation is based on two indicators that place each sportsman, through a point, within a plane. The researches carried out up to the present emphasize that the human mind allows a faster analisys in case of a graphical representation than in case of a series of numerical values. This is why it appealed to this intuitive representation, method applied successfully also in other analisys cases.

2 Presentation of Graphical Analysis Method

The two status indicators of a sportsman, used in the graphical representation proposed, are [7]:

- the progress during the training period (the vertical axis) - computed as difference between the slope of the linear function of progress, between the last two tests of sportsman, and the slope of the linear function of progress of the average of whole group of sportsmen. In this case the slope was computed for the second half of the training period.
- the final position in comparison with the average (the horizontal axis) – computed as difference between the final value, obtained by sportsman during the test, and the average of the group, computed at the end of the training period.

The obtained results were weighted by a coefficient allowing the evaluation and easy reading of the representations made. The two indicators divide the representation plan into four quadrants as following (Fig. 2) [7]:

- the left-bottom quadrant, where will be find the sportsmen which are situated, at the end of the training period, over the average of the group, but which have obtained a progress under the average during the last half of the training period;
- the right-top quadrant, in that will be represented the sportsmen which obtained, at the end of the training period, performance over the average of the group and also had a progress over the average of the group during the second half of the training period;
- the right-top quadrant will contain the sportsmen which, although registered a progress over the average during the last period, still obtain results under the average of the training colleagues;
- the left-bottom quadrant that includes the sportsmen which, after the training period, failed to be above the average of the group, neither regarding the results obtained, nor regarding the training progress.

Based on the proposed representation it can make an analysis of the sportsmen activity at the end of training period in order to classify the members of the group as following (Fig. 2) [7, 10]:

- the sportsmen placed within the right-top quadrant are those who had progress and results over the average and therefore who can accede to performance in sports. The more close to the right-top corner of the representation surface is the point that represents the performance, the more notable is the performance of this sportsman.
- the sportsmen placed within the left-top quadrant are those who perform above average but have not achieved notable progress during the training. For these sportsmen is recommended to change the training program to see if they have the potential to increase their performance or have already reached a level that can not be exceeded.
- the right-bottom quadrant represents the sportsmen who have achieved major progress in training, but their results are still below the group's average due to the low performance with that they began the training. For these sportsmen is recommended to extend the training period or to carry out additional training, but not exceeding the fatique level tolerated by the human body.
- the persons who, after the training program, are within the left-bottom quadrant don't have the skills to achieve performance. They can continue to do sports but, on the material grounds, they should be excluded from the group trained for performance.

Based on the data sets taken during the training period, for 12 girls aged between 14 and 16 years, there were made graphical representations for 14 psychomotor tests. Table 1 shows, for example, the results from one of these tests namely the establishing of the capacity to control the minimum speed. It must to noted that the values are computed based on the result of exploration. To obtain more accurate results an effective testing is recommended to be done at the end of the training period [12, 14]. Figure 3 shows, for this test, the proposed graphic representation, at the end of the training period, for the considered lot.

Table 1. Establishing the capacity to control the minimum speed

Establishing the capacity to control the minimum speed [m/s]					Prediction
Period		Sept 05	Dec 05	Jan 06	May 06
Name and surname	Age	0	3	4	8
A.A.	14	0.17	0.1	0.08	0.04
A.R.	15	0.15	0.14	0.11	0.09
C.N.	14	0.19	0.15	0.11	0.07
B.R.	14	0.1	0.13	0.09	0.11
S.D.	16	0.13	0.12	0.09	0.07
V.E.	15	0.16	0.09	0.08	0.04
R.S.	15	0.11	0.12	0.09	0.09
G.B.	16	0.12	0.1	0.12	0.10
S.B.	16	0.16	0.11	0.1	0.06
C.V.	15	0.15	0.13	0.06	0.04
S.A.	14	0.14	0.11	0.09	0.06
T.A.	14	0.11	0.09	0.08	0.06
Average	14.83	0.14	0.12	0.09	0.0693

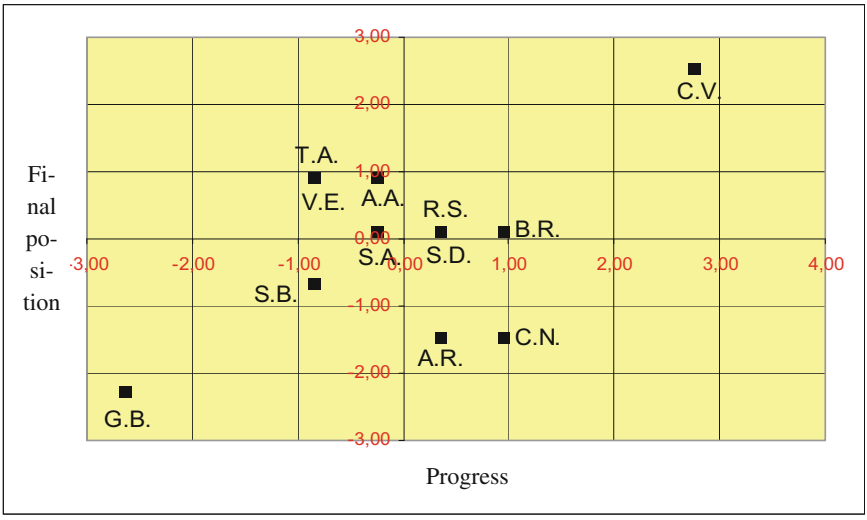


Fig. 3. Establishing the capacity to lead the minimum speed

Thus, it can observe a sportsman's behaviour, in terms of performance achieved, and also from the perspective of the progress recorded for all the tests proposed. An unitary representation of the sportsman's performance, including all tests, should be recommended, but it is impossible to be carried out because it would be necessary to

place a point in an N-dimensional space (N – number of tests). But, for a limited number of tests, it can try to represent in the same plan, by different colors, the sportsmen's results, to have an overview on the progress for different tests or different times [5].

From the representations carried out, it can observe that a sportsman can not achieve significant results for all tests. However, there are individuals who are in areas of maximum performance in a significant number of tests, which leads to the conclusion that they are closer to become "a complete sportsman". In these cases it is preferable for a sportsman to achieve positive results in a large number of tests, even if these results are not the best of the group, than to obtain superior results into group but for a limited number of tests [3].

Concerning the psychomotor tests, one may notice a nonuniform distribution of the athletes in the plane of representation, where from it can be deduced that every athlete has its own evolution, different, during the training; notice that, during the paddling step length test, no any athlete was placed in the top-right quadrant, which means that these skills can not still develop in this part of the training period or that the athletes tested couldn't focus enough during the tests because of fatigue or other elements that have distracted their attention. The proposed representations can quickly highlight the performance of an athlete from several points of view, by highlighting the performance for that it must work overtime, thus establishing an individual training program, specific to the athlete, in the prospect of obtaining significant sports results (Fig. 4) [5].

3 Assessing the Method Errors

Practically, the approximation accuracy is checked subsequently, also by experimental determinations, for that will be counted also the inevitable errors of testing methods. In these situations, as presented, the approximation function can be chosen, being several methods and techniques for determining the transfer static function.

The choice of method depends on the user experience and the computers available to it. In the rare cases when the derivative functions in nodes are known as well, can be used also other methods, and the case of functions of several variables could be dealt eventually if multiple sizes, or the same size as a source of errors, are taken into consideration. Considering, therefore, that the mathematical function expressing the graphical representation is an approximation function, arises naturally the question of accuracy with which this approximation is done, hence of approximation error [3].

The existence and uniqueness of a best approximation functions must be studied separately for each case, and the actual calculation of the best approximation function is a special problem as it requires, in its turn, carrying out operations. Even in case of calculations made by computer, the results are obtained by truncation and rounding, so by approximation the numerical values.

Due to those mentioned above, it proposes addressing the problem of determination the best approximation from technical standpoint, of an experimental nature, which consists of estimating, from metrological standpoint, the error occurring through the function approximation. The absolute error in measurement technology is defined as

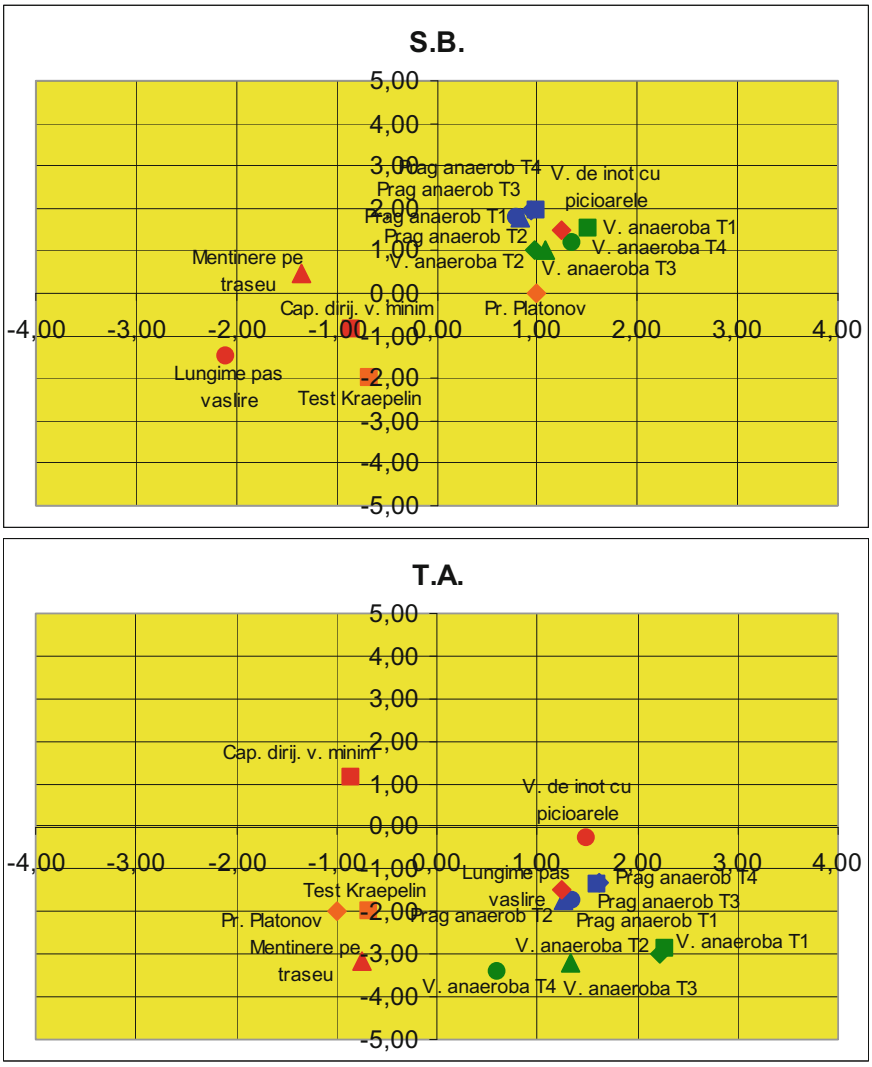


Fig. 4. Example of representation the intermediate results for two athletes with different performance

the difference between the value measured by the measuring device X_m and the real value of the measured quantity X_r:

$$\varepsilon_a = |X_m - X_r| \tag{1}$$

Because the information provided by the absolute error is not conclusive in terms of the measurement accuracy, the relative error is determined according to the equation:

$$\varepsilon_r = \frac{\varepsilon_{a\max}}{X_{\max}} \cdot 100 \quad (2)$$

with X_{max} – the maximum size measured [7].

It is proposed, for determination the optimal approximation method, to use the relative error in sense of the foregoing definition, considering in computing the actual value of the measured quantity as being the value of the assesed quantity, this being measured for a certain number of points n using specific accurate methods, and the measured value, in sense of the previous definition, being computed for the respectiv input points, using the approximation function determined through the methods described above.

This methodology is virtually identical to that used in establishing the accuracy class of the measurement devices, according to the metrological rules, but the dispersion of the numerical values is significantly affected by the natural specificity of the biological regnum [2].

4 Conclusions

In case of the tested athletes, from functional point of view, the coordination functions of the neuroendocrine system are maturing, significant fact in balancing the motive acts and actions and in the superior adjustment of them. Amid developing the skills in sensing the significant elements for an effective motric behavior, the responses become complex and analytical. The main forms of motric strain can be assessed through the motor skills, which can be divided into conditional and coordinative qualities.

From the proposed representation it can quickly highlight the performance of an individual, from as many points of view, highlighting the elements that it must train them overtime, thus establishing a specific individual training program, in the prospect of obtaining significant athletic performance [13]. Regarding the continuation of training the lot, the method recommends [5, 7]:

- to continue the program, as it was started, for the sportsmen placed in the right-top quadrant;
- to change the training program for the sportsmen placed in the left-top quadrant, sportsmen who didn't obtain a satisfactory progress, although they have good results;
- to supplement the training program for the sportsmen represented in the right-bottom quadrant, sportsmen who, although have a superior training progress, still obtain results under the average;
- waiving of the training for obtaining performance for those sportsmen placed in left-bottom quadrant;
- all these being valid at significant majority of the tests done.

From the processed data follows that there aren't sportsmen who don't have results at any test, therefore is recommended a complex analysis of the information obtained, in order to assess the possibility that a sportsman who has no chance of progress in the sport chosen to be recruited into the training programs of another sport where the positive results already achieved matter substantially.

This analisys method can be applied repeatedly in different periods of training the sporstmen, in order to emphasize the reaching by them of certain performance thresholds that can not be overtaken, to highlight the need to change the training way with another one that ensures a faster progress, to reduce the group so that it can focus only on the individuals capable of performance, to highlight the areas that require a higher volume of training in prospect of obtaining complete sportsmen. Another advantage of using this method consists in carrying out some classifications of sportsmen from the perspective of a particular test [11]. If we refer only to the psychological and psychomotor tests, we can deduce an important feature of the sportmen's age namely the existence of some moments when their concentrating diminishes. Regarding these tests, it observes an uniform distribution of sportsmen in the plan of representation from where it can deduce that, during the training period, the sportsmen have a very different progress from one individual to another.

References

1. Milici, D.: Computerized system for testing and formation the speed of backward pusf of sportmen. In: 13th International Symposium on Measurements for Research and Industry Applications Organisating by International Measurement Confederation Athens, Greece, pp. 673–677 (2004)
2. Milici, L.D., Milici, M.R.: The main ethical strategies in human performance monitoring field. In: 3rd International Conference EPHES2015, The Challenges and the Crises of a Technological World (2015)
3. Milici, M.R., Milici, L.D., Cretu, M., Pentiuc, R.: Using the continuous extrapolation functions of measurement data on prediction of the sportman performances. In: 16th IMEKO TC4 Symposium Exploring New Frontiers of Instrumentation and Methods for Electrical and Electronic Measurements, Florence, Italy, pp. 888–893 (2008)
4. Milici, L.D., Milici, M.R., Cernomazu, D., Popa, C.: Modeling of physical and psyhological human performance evolution. In: Proceedings of the 6th International Conference on Electrical & Power Engineering, EPE 2010, Iasi, Romania, vol. 1, pp. 15–20 (2010)
5. Milici, L.D., Rata, E., Milici, M.R.: Study of new graphical method for sportman evaluation. Int. J. Comput. Commun. **1**(4), 99–107 (2007). University Press
6. Pentiuc, S.G., Milici, L.D., Pentiuc, R.D., Milici, M.R.: Unsupervised learning algorithms for decision making support in physical education. In: Proceedings of the First European Conference on the Use of Modern Information and Communication Technologies ECUMICT 2004, Gent, Belgium, pp. 21–28 (2004)
7. Rata, E., Risneac, B., Milici, L.D.: Prognoza pregatirii psihomotrice in antrenamentul sportivilor prin aplicarea modelarii matematice. Publishing House of the State University of Physical Education and Sport, Chisinau (2007)
8. Astrov, I., Tatarly, S., Tatarly, S.: Simulation of two-rate neural network control for stochastic model. Advances in Electrical and Computer Engineering, no. 1, pp. 75–84, (2006). ISSN 1582-7445
9. Ilioi, C.: Analiza numerica, Tehnici de aproximare. "Al. I. Cuza" University of Iasi (1983)
10. Iorga, V., et al.: Programare numerica. Teora Publishing House (1996)
11. Khovanski, A.: Applications des fractions continues et de leurs generalisations aux problemes de l'analyse approchee, Gostekhizdat (1956)

12. Kunzi, H., et al.: Numerical Methods of Mathematical Optimization. Academic Press, Cambridge (1968)
13. Larson, R.J., Marx, M.L.: An Introduction to Mathematical Statistics and its Application. Prentice Hall, Upper Saddle River (1986)
14. Popovici, P., Civa, O.: Rezolvarea numerica a ecuatiilor neliniare. Signata Publishing House (1992)

Tremor Measurement System for Neurological Disorders Screening

Iuliana Chiuchisan[1], Iulian Chiuchisan[1], Oana Geman[2(✉)], Rodica-Mariana Milici[3], and Laurentiu-Dan Milici[3]

[1] Computers, Electronics and Automation Department, Stefan cel Mare University, 13 Universitatii Street, 720229 Suceava, Romania
iuliana.chiuchisan@usm.ro, iulian@eed.usv.ro
[2] Department of Health and Human Development, Stefan cel Mare University, 13 Universitatii Street, 720229 Suceava, Romania
oana.geman@usm.ro
[3] Electrotechnics Department, Stefan cel Mare University, 13 Universitatii Street, 720229 Suceava, Romania
mami@eed.usv.ro, dam@usm.ro

Abstract. In this paper is described an acquisition method for tremor signal using accelerometer in patients with Parkinson's disease. The system acquires the acceleration information from Wii™ accelerometer sensor using the Bluetooth connection. The tremor signal data are automatically uploaded on a server for further analysis, using FTP protocol. This tremor signal is processed using FFT and Wavelet filters in order to assists and helps the specialists in differential diagnosis between Parkinson's disease and other neurological diseases.

Keywords: Neurological disorders · Healthcare system · Tremor · Accelerometer · Signal processing

1 Introduction

The analysis of the demographic data for the European countries shows that in the last decades the population of over 65 years of age is in a continuing increase [1]. The number of the elderly population will increase in all European countries and within the groups of old age, the population with ages for over 80 years will record an ascending increase much more accentuated. The neurological diseases, frequently encountered in elderly patients, increase the functional disability, decreasing the quality of life and increasing the cost in the healthcare of long duration or of the health services in general [1]. Parkinson's disease (PD) is a disorder of the central nervous system [2], involving a degeneration of certain nerve cells in deep parts of the brain, and in particular a loss of nerve cells in a part of the brainstem [3]. These cells make the neurochemical messenger dopamine, which is partly responsible for starting a circuit of messages that coordinate normal movement [3]. PD is a chronic and slowly progressive illness, but the rate of progression can vary from patient to patient [4]. PD symptoms consist in tremor, slowness of movement (bradykinesia), rigidity, and posture instability. The

V.E. Balas et al. (eds.), *Soft Computing Applications*, Advances in Intelligent Systems and Computing 633, DOI 10.1007/978-3-319-62521-8_28

resting tremor is the main visible symptom and is one of the cardinal features of Parkinson's disease, although proximally 10% of the patients do not present tremor.

The tremor is a common disturbance of movement, and can vary in frequency and amplitude (e.g. the tremor frequency can vary from low values, 4–5 Hz, to high values, 8–10 Hz) [5–7]. The physiological, psychological or motor factors, the consumption of alcohol, drugs, or other chemical substances, can influence the tremor. The expression of the motor symptoms may vary according to the circumstances under which they occur: the frequency at which the tremor occurs and the body part that is involved [7]. Also, the tremor may occur in cases such as at rest, during postural holding or during voluntary movements [2].

A classification scheme that categorizes the tremor includes three tremor syndromes associated with Parkinson's disease, described in [8]. Different techniques were used to measure the amplitude and frequency of tremor [9–18]. In order to detect and quantify tremor, different tools have been used, such as electromyography (EMG) [19], accelerometer and gyroscope [6, 20]. Some types of tremor that indicates an early stage of a neurological disease are summarized in Table 1.

Table 1. Types of tremor

Movement	Frequency	Part of body	Neurological disorder
Rest tremor	4–6 Hz	Arms, legs	Parkinson's disease
Postural tremor	7–12 Hz	Hands	Essential tremor
Intention tremor	2–5 Hz	Arms, legs	Cerebral lesions

2 Tremor Measurement Method

The accelerometry is a method used to quantify human movement patterns using the accelerometer sensors which measure static or dynamic acceleration forces such as the force of gravity that actuates on a body part or the movement caused by the tremor [6]. The benefits of using accelerometers compared to more traditional gait analysis instruments include low cost; accelerometers are small and wearable; the direct measurement of 3D accelerations eliminates errors associated with differentiating displacement and velocity data. Figure 1 presents an example of tremor measurement using accelerometer sensors.

In order to measure the tremor, a system containing a Wii™ remote controller (Wiimote) and new software for measurement and analysis, were described in this paper The Wiimote used in the measurements contains a three-axis accelerometer from Analog Devices Inc. (Fig. 2), an IR image sensor able to track up to five objects simultaneously, and Bluetooth™ connectivity [21].

The accelerometer (model ADXL330) has a range of ±3 G which is sufficient for tremor recording. It can measure the static acceleration of gravity as well as dynamic acceleration resulting from motion, shock, or vibration [21].

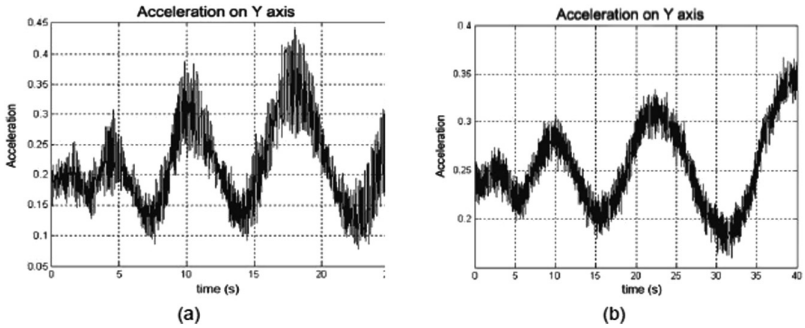


Fig. 1. Tremor measurement: (a) Patient with tremor; (b) Normal patient [6]

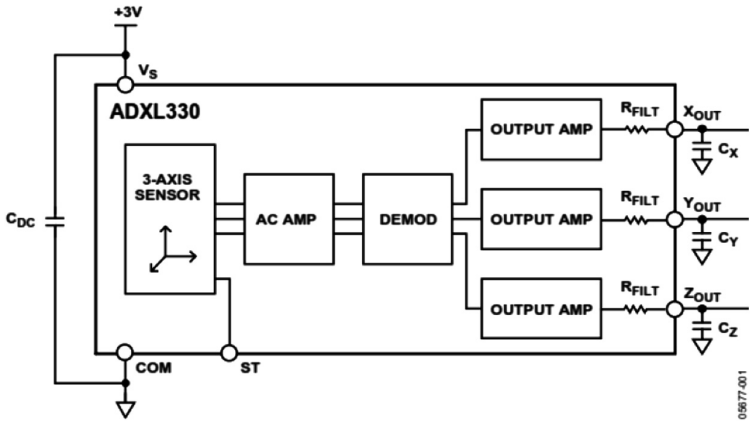


Fig. 2. Function block of three-axis accelerometer

3 Healthcare System for Parkinson's Disease

The NeuroParkinScreen (Fig. 3) is a complex system, described initially in [22]. The system can be used to manage health information and to support the specialists in diagnosis, rehabilitation and at-home monitoring of patients with Parkinson's disease. The system facilitates also the interaction at long distance between doctors and the patients. The system described in [22] involves:

– A PC/laptop/tablet with video camera; the patient can connect to NeuroParkinScreen application using Wiimote and will transfer the tremor recorded data to the network server. The PD screening system analyzes and processes the Parkinson's tremor data and send it to the medical specialists;
– The medical specialists who can access and analyze the patient's health records, and therefore can manage the patient's medical history, evolution, reaction to medication, etc. and can send notification to the patient.

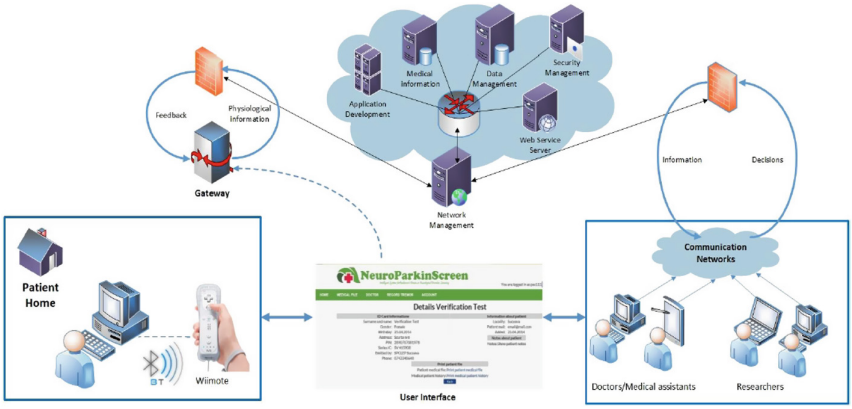


Fig. 3. General architecture of NeuroParkinScreen system [22]

4 Tremor Signal Acquisition

In this paper we have approached the tremor acquisition and the signal processing part of the NeuroParkinScreen application described in [22]. The system acquires the acceleration data from Wii™ accelerometer sensor using Bluetooth™ connection. The information is automatically uploaded on a server for further analysis, using the FTP protocol.

In order to measure the tremor using NeuroParkinScreen application we have established a set of steps that the patient has to follow:

1. The patient has to download from the specified website (Fig. 4) the following files: *wiipair* and *parkinwiidatacollection*; the web-based application can be accessed using a specific user and password;
2. The patient will run the *wiipair* file. After installing the application, the patient will press and release the red SYNC button from inside of battery cover on the back of the Wiimote in order to enable the Bluetooth connection between Wiimote and PC/laptop/tablet;
3. The patient will run the *parkinwiidatacollection* file;
4. Using the Wiimote the patient will start the tremor measurement stage. In order to make the tremor acquisition, it's necessary for the patient to keep the hand with Wiimote in a horizontal position facing to PC for about 5 or 10 s. To begin the tremor acquisition session the patient will press the A button from Wiimote and will keep the A button pressed for 60 s and then will release it.
5. After the patient releases the A button, the tremor acquisition session is finished (Fig. 5). The tremor signal data are stored into an external file and is automatically saved on the server.

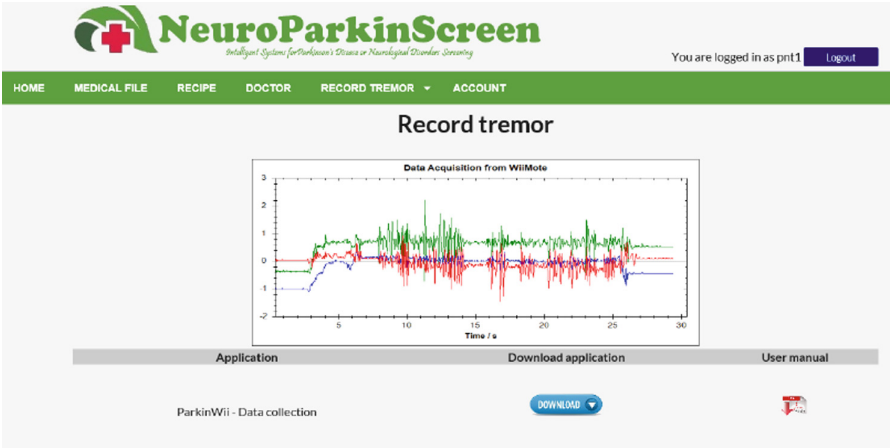


Fig. 4. Tremor measurement GUI

Fig. 5. Tremor acquisition interface

5 Signal Processing and Results

The tremor signal information, acquired using built-in accelerometer of the Wii™ remote controller, are processed using Fast Fourier transform (FFT) and Wavelet 1-D technique in order to determine the frequency and amplitude of the tremor signal: the normal (N) tremor is between 5–12 Hz and the PD tremor is between 4–6 Hz. In Fig. 6 we present the time series for the Normal tremor and the PD tremor processed in MATLAB, for three components: x (lateral) = 0, y (anteroposterior) = 0 and z (vertical) = 1 (equilibrium state).

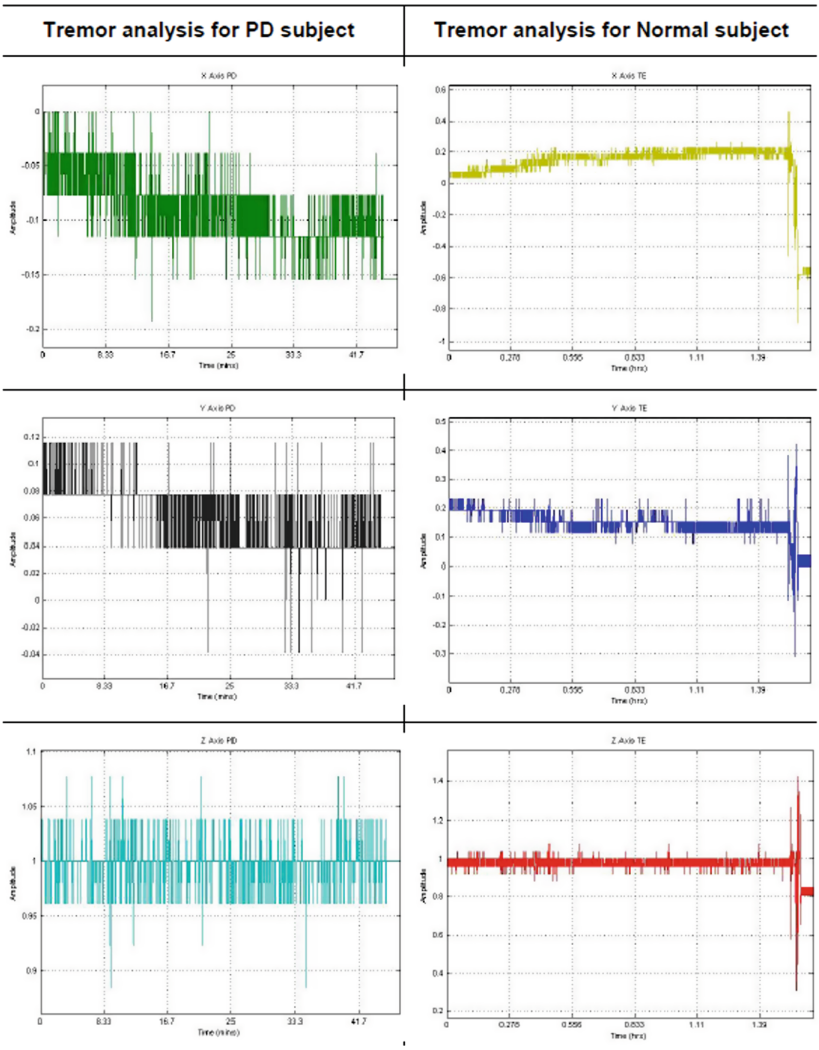


Fig. 6. Time series for the PD tremor and the normal tremor

Using MATLAB we analyzed Y axis recordings from the PD patients and the nor-mal patients and we applied two types of filters: FFT filter (Figs. 7 and 8) and wavelet 1-D filter (Figs. 9 and 10).

In the upcoming research, the tremor signal acquired using accelerometer sensors will be segmented, filtered, and processed in order to determine other features associated with movement.

In future research, a different group of patients will be used, represented by patients with obesity problems [23], knowing that the obesity increases the risk of various diseases. In today modern medicine, digital information security is an interdisciplinary issue, having to be constantly optimized, developed and innovated [24–26].

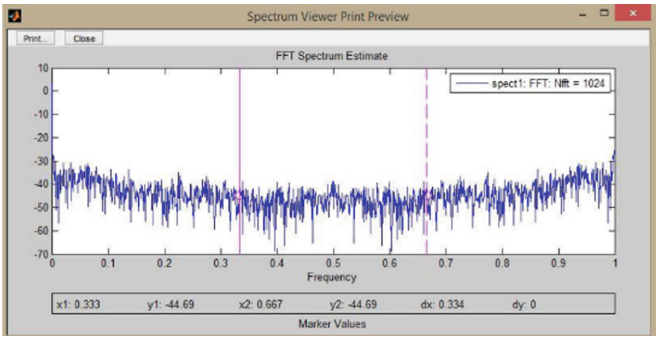


Fig. 7. FFT filter (PD patient)

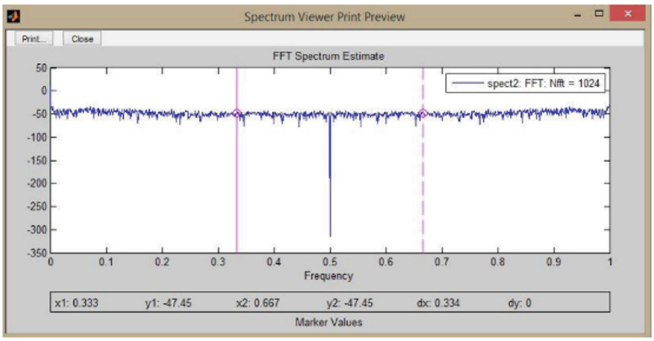


Fig. 8. FFT filter (normal patient)

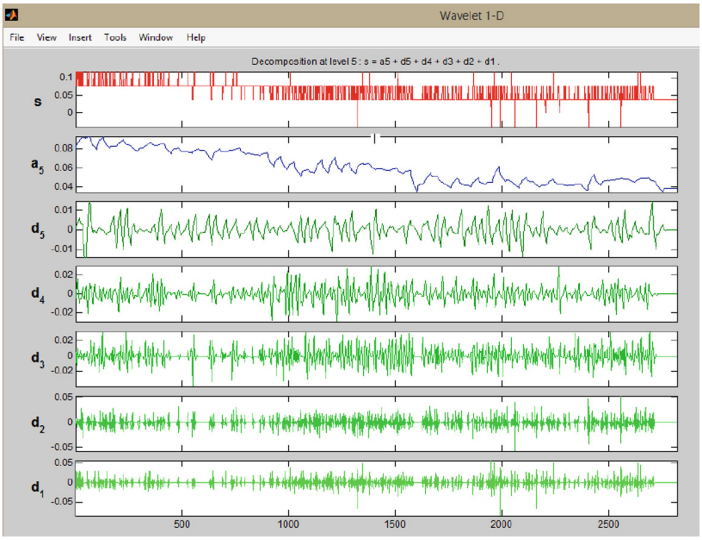


Fig. 9. Wavelet 1-D filter (PD patient)

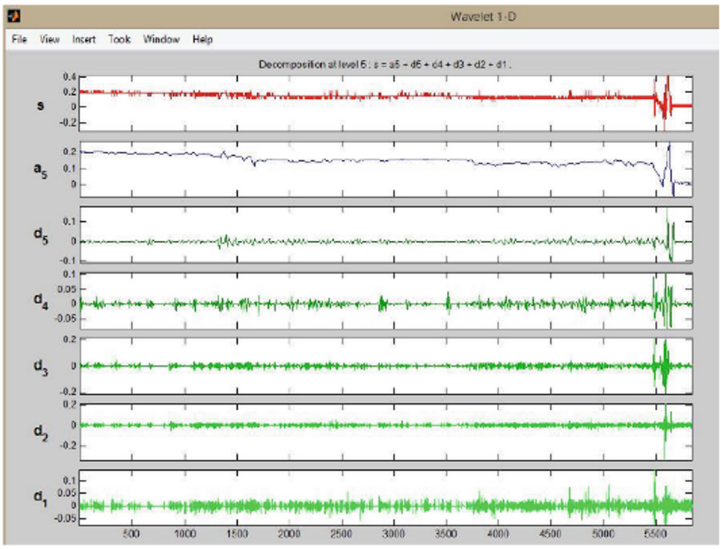


Fig. 10. Wavelet 1-D filter (normal patient)

6 Conclusions

The system presented in this paper presents the following features: gather data from sensors; support user interface; network connectivity for access to infrastructural services; low power, robustness, durability, accuracy and reliability. Using this system the medical specialists can keep the patients under observation from distance for a certain period of time in order to evaluate the severity of symptoms.

Also the system can assists and helps the specialists in differential diagnosis between Parkinson's disease and other neurological diseases. In future research, a statistical analysis part will be designed and added to the described system in order to compare data information from Parkinson's patients with data information from normal (control) patients and also to generate an early prediction of the neurological disease.

Acknowledgement. This work was supported by the Romanian National Program (PN-II-ID-PCE-2012-4-0608 no. 48/02.09.2013), "Analysis of novel risk factors influencing control of food intake and regulation of body weight" [23].

References

1. European Commission: Directorate-General for Research and Innovation, Population ageing in Europe. Technical report (2014)
2. Olesen, J., Gustavsson, A., Svensson, M., Wittchen, H.-U., Jonsson, B.: The economic cost of brain disorders in Europe. Eur. J. Neurol. **19**, 155–162 (2012)
3. Golbe, L., Mark, M., Sage, J.: Parkinson's Disease Handbook. American Parkinson Disease Association, Inc., New York (2010)

4. National Parkinson Foundation: Understanding Parkinson's. Reports (2016)
5. Fishman, P.S.: Paradoxical aspects of parkinsonian tremor. Mov. Disord. **23**(2), 168–173 (2008)
6. Mamorita, N., Iizuka, T., Takeuchi, A., Shirataka, M., Ikeda, N.: Development of a system for measurement and analysis of tremor using a three-axis accelerometer. Methods Inf. Med. **48**(6), 589–594 (2009)
7. Lauk, M., Timmer, J., Lucking, C.H., Honerkamp, J., Deuschl, G.: A software for recording and analysis of human tremor. Comput. Methods Programs Biomed. **60**, 65–77 (1999)
8. Deuschl, G., Bain, P., Brin, M.: Consensus statement of the movement disorder society on tremor. Mov. Disord. **13**(Suppl. 3), 2–23 (1998)
9. Geman, O.: Screening and Rehabilitation System for Patients with Parkinson's Disease. Advances in Biomedicine and Health Science Series. WSEAS Press (2013)
10. Geman, O., Costin H.N.: Tremor and gait screening and rehabilitation system for patients with neurodegenerative disorders. Autom. Control Comput. Sci. **LIX (LXIII)**(2) (2013)
11. Farkas, Z., Csillik, A., Szirmai, I., Kamondi, A.: Asymmetry of tremor intensity and frequency in Parkinson's disease and essential tremor. Parkinsonism Relat. Disord. **12**(1), 49–55 (2006)
12. Piboolnurak, P., Rothey, N., Ahmed, A., Ford, B., Yu, Q., Xu, D., Pullman, S.L.: Psychogenic tremor disorders identified using tree-based statistical algorithms and quantitative tremor analysis. Mov. Disord. **20**(12), 1543–1549 (2005)
13. Akin, M.: Comparison of wavelet transform and FFT methods in the analysis of EEG signals. J. Med. Syst. **26**(3), 241–247 (2002)
14. Issartel, J., Marin, L., Gaillot, P., Bardainne, T., Cadopi, M.: A practical guide to time-frequency analysis in the study of human motor behavior: the contribution of wavelet transform. J. Mot. Behav. **38**(2), 139–159 (2006)
15. Spyers-Ashby, J.M., Bain, P.G., Roberts, S.J.: A comparison of fast Fourier transform (FFT) and autoregressive (AR) spectral estimation techniques for the analysis of tremor data. J. Neurosci. Methods **83**(1), 35–43 (1998)
16. Geman, O., Zamfir, C.: Using wavelet for early detection of pathological tremor. In: European Association for Signal Processing (EUSIPCO 2012), pp. 1723–1727 (2012). ISSN 2076-1465
17. Geman, O., Costin, H.N.: Automatic assessing of tremor severity using nonlinear dynamics, artificial neural networks and neuro-fuzzy classifier. Adv. Electr. Comput. Eng. **14**(1), 133–138 (2014). ISSN 1582-7445
18. Geman, O., Costin, H.N.: Parkinson's disease prediction based on multistate markov models. Int. J. Comput. Commun. Control **8**(4), 525–537 (2013). ISSN 1841-9836
19. Caviness, J.N., Liss, J.M., Adler, C., Evidente, V.: Analysis of high-frequency electroencephalographic-electromyographic coherence elicited by speech and oral non-speech tasks in Parkinson's disease. J. Speech Lang. Hear. Res. **49**(2), 424–438 (2006)
20. Machowska-Majchrzak, A., Pierzchata, K., Pietraszek, S.: Analysis of selected parameters of tremor recorded by a biaxial accelerometer in patients with parkinsonian tremor, essential tremor and cerebellar tremor. Neurol. Neurochir. Pol. **41**(3), 241–250 (2007)
21. Analog Devices, Small, Low Power, 3-Axis MEMS Accelerometer (2007)
22. Chiuchisan, I., Geman, O., Chiuchisan, I., Iuresi, A.C., Graur, A.: NeuroParkinScreen – a health care system for neurological disorders screening and rehabilitation. In: International Conference on Electrical and Power Engineering (EPE 2014), 16–18 October 2014, pp. 536–540 (2014)
23. Project "Analysis of novel risk factors influencing control of food intake and regulation of body weight" (PN-II-ID-PCE-2012–4-0608 no. 48/02.09.2013) (2016). www.eed.usv.ro/idei_48

24. Bucerzan, D., Raţiu, C.: Image processing with android steganography. In: Soft Computing Application, Proceedings of 6th International Workshop on Soft Computing Applications (SOFA 2014). Advances in Intelligent Systems and Computing, vol. 1, pp. 27–36 (2014). Springer International Publishing. eBook ISBN 978-3-319-18296-4, ISBN 978-3-319-27178-1. doi:10.1007/978-3-319-27179-8
25. Bucerzan, D., Raţiu, C.: Contributions to steganographic techniques on mobile devices. In: Innovative Security Solutions for Information Technology and Communications. LNCS, vol. 9522, pp. 242–252 (2016). Springer International Publishing. eBook ISBN 978-3-319-27179-8, ISBN 978-3-319-27178-1. doi:10.1007/978-3-319-27179
26. Raţiu, C., Bucerzan, D., Manolescu, M.J.: SmartSteg: a new android based steganography application. Int. J. Comput. Commun. Control **8**(5), 681–688 (2013). ISSN 1841-9836

Novel Method for Neurodegenerative Disorders Screening Patients Using Hurst Coefficients on EEG Delta Rhythm

Roxana Toderean (Aldea)[1], Oana Geman[1(✉)], Iuliana Chiuchisan[2], Valentina Emilia Balas[3], and Valeriu Beiu[3]

[1] Department of Health and Human Development, Stefan cel Mare University, 13 Universitatii Street, 720229 Suceava, Romania
roxana.toderean@gmail.com, oana.geman@usm.ro
[2] Computers, Electronics and Automation Department, Stefan cel Mare University, 13 Universitatii Street, 720229 Suceava, Romania
iuliana.chiuchisan@usm.ro
[3] "Aurel Vlaicu" University, Arad, Romania
balas@drbalas.ro, valeriu.beiu@uav.ro

Abstract. Parkinson's disease (PD) is the most common degenerative movement disorder and a progressive nervous system disorder. It affects body movement, memory, speaking and daily mental and physical activities, being directly connected to the dopaminergic loss. The EEG signal can be modulated under the effect of neurotransmitters, e.g., using dopamine (L-dopa), the alpha and beta frequencies (the characteristic of normal EEG activity at rest) increase, and reduce delta-theta activities. Joint EEG and EMG signals from 22 patients (16 men and 6 women of 62 years on average) have been acquired on several mental tasks. These patients were diagnosed with PD according to the UK PD Society Brain Bank diagnostic criteria, while the EEG recording was performed according to the International 10–20 System. Nowadays, there is no screening test for early detection of PD. Symptoms become increasingly visible as the disease progresses. The Hurst coefficient is one of the indicators that could be used for the characterization of EEG signals, as these signals may be seen as processes with extended memory. The results obtained for the patients with PD, PD and Dementia, PD and diabetes present lower Hurst coefficient values for the delta rhythm. In the case of healthy subjects a decrease of Hurst coefficient values for this rhythm was not observed. In conclusion, the Hurst coefficient applied to EEG signals can be a good marker for the early diagnosis of PD, and maybe other neurodegenerative diseases.

Keywords: Hurst coefficient · EEG signals · Parkinson's disease · Multiresolution wavelet analysis · Aggregated variance method

1 Introduction

Neuropsychological dysfunctions affect more than 160 million Europeans (i.e., 38% of the population of Europe) every year, according to a report presented in 2014 by the European Brain Council [1] and the European College of Neuropsychopharmacology and World Health Organization [2].

V.E. Balas et al. (eds.), *Soft Computing Applications*, Advances in Intelligent Systems and Computing 633, DOI 10.1007/978-3-319-62521-8_29

On top of this, 27% of the adult population (including people aged between 18 and 65 years) suffer from a mental disorder such as psychosis, anxiety, or eating behaviour disorder. It was estimated that approximately 83 million people suffer from neurological disorders, which underestimates the actual number as taking into account only those under the age of 65 years [2].

Parkinson's disease (PD) is the most common degenerative movement disorder and a progressive nervous system disorder that affects: body movement, memory, speaking and daily mental and physical activities [1]. PD is caused by progressive decay of neurons that produce dopamine (a neurotransmitter responsible for the body's ability to control movements). In PD, it is considered that the symptoms start manifesting themselves when the dopamine levels in the brain decreases by more than 50% of the normal values [3, 4]. The first/early symptoms of the disease can be either non-motor (e.g., fatigue, pain, excessive salivation, speech problems, constipation, sleep disorders, urinary incontinence), or symptoms of neuropsychological order common to many patients (e.g., depression, anxiety, confusion, apathy), and even cognitive symptoms (e.g., memory impairment, attention, difficulty in making decisions) [2].

In this paper we focus on the delta rhythm as we expect that brain rhythms may help detect PD early on. We have used wavelet multiresolution analysis to decompose the EEG signal into sub-bands representing EEG rhythms: delta (0–4 Hz), theta (4–8 Hz), alpha (8–12 Hz), beta (13–30 Hz) and gamma (30–60 Hz). For an adult man in sleep status, at the alert threshold, the EEG recorded in the bipolar derivation shows typically two types of rhythms: alpha and beta. If the subject is in the sensorial (eye closed) and mental rest state, we are witnessing the entry of the alpha rhythm that has a frequency of maximum 12 Hz and amplitude of 50 μV (10–100 μV). Typically, the amplitude increases and decreases regularly and the waves are grouped into trunnions (bursts). Most of the alpha rhythm originates from the occipital area [5].

The sensors influenced by external activities, and in particular light (opening of the eyes), act to inhibit the alpha rhythm and allow entering the beta rhythm. The same phenomenon is happening during sustained intellectual effort or strong emotions. Beta rhythm is characterized by a frequency between 13 and 30 Hz and amplitude of 5–30 μV. Unlike the alpha rhythm, the beta rhythm is highly irregular and signifies a desynchronization of the neurons activity. In 15% of normal subjects the theta rhythm originates from the front region. It is characterized by maximum amplitude of 20 μV and a frequency of 8 Hz, under the form of isolated waves, not exceeding 25% of the total length of the trends [5].

During deep sleep, the degree of synchronization increases further and the predominant frequency is given by the delta rhythm. The delta rhythm is considered pathological if appearing in the sleep state. It can be observed in cases like lesions and brain tumours, hypoglycemia, hypocalcemia, hypoxemia palsy, barbiturate coma, etc. In fact, any serious shortcomings of the brain food processing, metabolic endocrine pathology, toxicity, etc., are indicated by the appearance of these slow waves, non-specific type delta rhythm [5].

In this paper we have started by using multiresolution decomposition of the EEG signals into sub-bands, and then we have calculated the values of the Hurst coefficient (for different rhythms) using the aggregated variance method.

2 Medical Database

Joint EEG and EMG recordings from 22 patients (16 men and 6 women of 62 years on average) have been acquired by applying the following mental tasks:

1. Cortical activity: the subject is instructed to relax as much as possible and to perform some simple lifting and moving of a light object (**Task 1**);
2. Eye blinking: the subject has to blink regularly to calibrate the EEG (**Task 2**);
3. Lifting and placing an object with open eyes: the subject has to lift and place a light object (a plastic bottle) on a box located in front of him/her at a distance of 15–20 cm (with his/her eyes open) (**Task 3**);
4. Lifting and placing an object with eyes closed: Task 3 is repeated with the eyes closed. This is to include the effect of visual feedback within the process (**Task 4**).

The acquisition of signals was achieved using g.GAMMAsys [6], developed by g.TEC. The product, called g.GAMMAsys, is a complete research/development system. It is using MATLAB, and includes all the hardware and software components necessary for: the acquisition of signals, online/offline analyses, classification of signals, and also predict the neuro-feedback. The G.MOBIlab is available with up to 32 EEG channels, is portable and allows for wireless connectivity. The main advantages are:

– Complete system for researching EEG and ECG;
– Reading the letters and the control of the cursor on a monitor screen;
– Integration of trials in real-time and offline analysis;
– Runs using hardware signal acquisition (g.MOBIlab);
– Access to the source code makes it easy to develop new applications.

The pre-processing of brain signals is performed before feature extraction to enhance the signal-to-noise ratio. Selecting the characteristics and reducing the size of the data was carried out without losing relevant information. By preprocessing we aimed to reduce the number of features and/or channels to a compact and free of noise form. There are many ways for feature extraction, the most widely used being: band power, cross-correlation between bands of power, frequency representation, time-frequency representation, spectral analysis, parametric modeling, and non-linear dynamics [7]. In this paper we have decided to use the wavelet multiresolution analysis for feature reduction.

3 Multiresolution Wavelet Analysis

In case of continue wavelet transform, the signal is analysed using scaling and translation of a wavelet function. In the case of discrete wavelet transform, the representation of the time-scale of a discrete signal is performed by a method based on digital filtering. Discrete wavelet transform analyses the signal at different resolutions (multiresolution analysis), through the successive decomposition of the signal in frequency bands.

The algorithm of Mallat [7] calculates the DWT by complementary low pass filtering (LPF) and high pass filtering (HPF) and successive sub-sampling. The impulse

response of the LPF is *h(n)*, and the one of the HPF filter is *g(n)*. If the signal is *x(n)*, after each level of decomposition we obtain [8, 9]:

$$y_L(n) = x(n) * g(n) = \sum_k x(k)g(2n-k), \tag{1}$$

$$y_H(n) = x(n) * h(n) = \sum_k x(k)h(2n-k), \tag{2}$$

where $y_H(n)$ şi $y_L(n)$ represents the impulse response of the HPF and LPF filters at every level, after sub-sampling by 2. The impulse response $y_H(n)$ is represented by the detail coefficient of the first order, while the impulse response of the filter $y_L(n)$ is the approximation coefficient of the first order. These are further decomposed and filtered, generating the detail and approximation coefficient of the second order. At each level HPF generates detail, while LPF produces approximation coefficients. Equations (1) and (2) correspond to the convolution of the signal with the impulse response of the filters. At each decomposition level the filters generate output signals that cover only half of the band of the input signals. This process doubles the frequency resolution, and the uncertainty in frequency is reduced by half.

4 Hurst Coefficient

The Hurst coefficient is used in applied mathematics in fields like chaos theory as well as in the case of processes with memory and spectral analysis. The EEG signals are signals that have nonlinearities and, like all biological signals, have areas of similarity. In order to be able to identify these areas, one should not rely only on classical approaches such as Fourier analysis. This is because of the non-stationary feature of the signals. However, there is a generally accepted opinion as regards deterministic or random character of these signals. This is the reason why other methods for revealing the characteristics of interest for a particular research area have started to be investigated [10].

The Hurst coefficient is one of the indicators that could be used for characterizing EEG signals due to the fact that these signals may be regarded as processes with extended memory. It was introduced in 1951 and is in equal measure an indicator of the similarity and also of the extended memory of a process, the two aspects that emphasize autocorrelations [11]. The complexity of similarity (of structures) is quantified using a fractal dimension, which is, in this case, a decimal number. The size of the fractal dimension, D, can be calculated using the relation [10]:

$$D = \frac{5-\gamma}{2} \tag{3}$$

where γ represents the self-similarity parameter.

The Hurst coefficient and the size of the fractal dimension are linked as:

$$D = 2 - H \tag{4}$$

The size of the fractal dimension shows the roughness of a surface [11]. A lower Hurst coefficient corresponds to a larger size of the fractal dimension and a rougher surface, while a Hurst coefficient of higher value corresponds to a smaller size of the fractal dimension and a smoother surface area. In our case the surface is represented by the distribution of the signal and a small Hurst coefficient would indicate a disorder of the cerebral system [11]. Form Eqs. (3) and (4) it is possible to deduce a simple relation between the Hurst coefficient and the self-similarity, γ:

$$H = \frac{\gamma - 1}{2} \tag{5}$$

According to the inter-relations among these three parameters, it follows that the size of the fractal dimension of the EEG signals may be obtained indirectly from the Hurst coefficient. The Hurst coefficient varies between 0 and 1. A Hurst coefficient of 0.5 indicates that we have no correlation whatsoever between the time series. Hurst coefficients with values in the interval [0, 0.5) indicate a low correlation between the time series. If they are in between 0.5 and 1 an increase of the correlation is to be expected. The Hurst coefficient cannot be determined exactly; hence it has to be estimated rather than calculated [10].

There are several methods which are used to estimate the Hurst coefficient: differential scattering; dispersion; rescaled range; Higuchi; absolute moment; aggregated variance, and others. In this paper we will detail only the aggregated variance method which is the one we have decided to use.

5 Estimating the Hurst Coefficient Using the Aggregated Variance Method

The aggregated variance method is based on the property of self-similarity between samples of a given process. X is considered to be a time series of length N, which is divided into sub-series of length m. For each sub-series, the aggregated series are formed using [12]:

$$X_k^{(m)} = \frac{1}{m} \sum_{i=(k+1)m-1}^{km} X_{i,} \qquad k = 1, 2, \ldots, M \tag{6}$$

Because of the self-similarity, the $X^{(m)}$ has the same distribution as $m^{H-1}X$ for large values of m. In particular:

$$Var\left(X_k^{(m)}\right) = m^{2H-2} Var(X_k) \tag{7}$$

The variance of $X_k^{(m)}$ is equal for each k (from the stationary hypothesis), and the estimator is:

$$Var\left(X_k^{(m)}\right) = \frac{1}{M}\sum_{i=0}^{M-1}\left(X_i^{(m)} - \overline{X(m)}\right)^2 \tag{8}$$

where $\overline{X(m)}$ represents the average of $X^{(m)}$:

$$\overline{X^{(m)}} = \frac{1}{M}\sum_{i=0}^{M-1} X_i^{(m)} \tag{9}$$

An estimation of H is obtained using the graphic representation of the size of the $Var\left(X_k^{(m)}\right)$ with respect to m, on double logarithmic scale. When the variance estimation is equal to the actual value, all points (estimators) are located on a straight line with slope $2H - 2$. In practice, the slope is estimated by fitting a straight line through the points (estimators). These H estimators are positioned between the estimators slope [13].

6 Results

The method proposed in this paper combines the estimation of the Hurst coefficient with the multiresolution wavelet analysis. The estimation of the Hurst coefficient is done using the aggregated variance method for the EEG sub-component signals with frequencies appropriate to the EEG rhythms frequencies. The method goes as follows:

1. Import the registrations into MATLAB;
2. Signal filtering with a Butterworth bandpass filter between 0 and 60 Hz. We used this particular band because the cerebral rhythms have the frequencies within this range (delta 0–4 Hz, theta 4–8 Hz, alpha 8–12 Hz, beta 13–30 Hz, and gamma 30–60 Hz);
3. Multiresolution wavelet decomposition is performed for signals on each channel using successively the wavelets: *Daubechies2*, *Coiflet4* and *Symlet6*. To gain access to every rhythm, signals are subdivided up to the fourth order. After the first level of decomposition the EEG signal is subdivided into the detail coefficient with high frequency D1 (30–60 Hz) and the approximation coefficient with low-frequency A1 (0–30 Hz). At the next level of decomposition the approximation coefficient A1 is subdivided in the detail coefficient D2 (15–30 Hz) and the approximation coefficient A2 (0–15 Hz). By following this process we obtain the coefficients: D3 (7.5–15 Hz), A3 (0–7.5 Hz), D4 (3.75–7.5 Hz) and A4 (0–3.75 Hz).
4. Estimate the Hurst coefficient using the aggregate variance method for the sub-components: the approximation coefficient of the fourth order - A4 (corresponding to the delta rhythm), the detail coefficient of the fourth order - D4

(corresponding to the theta rhythm), the detail coefficient of the third order - D3 (corresponding to the alpha rhythm), the detail coefficient of the second order - D2 (corresponding to the beta rhythm), the detail coefficient with high frequency of the first order - D1 (corresponding to the gamma rhythm).

We report in this paper the results for 5 patients. The multiresolution wavelet decomposition has been carried out using *Daubechies2 Wavelet*.

Case 1 - Normal patient, 35 years old

In the case of this normal subject no significant differences between the Hurst coefficients for the five different rhythms were observed (Fig. 1).

Canal\Ritm	Delta	Theta	Alpha	Beta	Gamma
1	0.41344	0.282152	0.572185	0.226345	0.469695
2	0.733837	0.585582	0.574346	0.361392	0.534047
3	0.581247	0.352636	0.435798	0.343489	0.605689
4	0.582633	0.0143	0.426255	0.435643	0.56005
5	0.157247	0.107506	0.755766	0.393572	0.468823
6	0.713434	0.358174	0.565215	0.317037	0.453256
7	0.535087	0.665839	0.607862	0.284297	0.621691
8	0.47745	0.207965	0.486062	0.28647	0.482568
9	0.597001	0.024688	0.392881	0.439256	0.565167
10	0.325933	0.041337	0.699869	0.389076	0.486011
11	0.628133	0.339444	0.549035	0.387977	0.433668
12	0.650093	0.805822	0.489494	0.322344	0.614294

Fig. 1. Hurst coefficient for a normal patient

Case 2 – Female patient, 80 years old

This patient was diagnosed with PD since 2008 (right hand tremor). Low values of the Hurst coefficients for the delta rhythm as compared to the other rhythms were observed (Fig. 2). For alpha and beta rhythms the Hurst values are close to 0.4.

Case 3 – Male patient, 75 years old

This patient was diagnosed with PD (since 2000) and dementia. He is taking medication while presenting intense tremor on both hands and legs. The patient diagnosed with PD and dementia also exhibits low values of the Hurst coefficient for delta rhythm (Fig. 3) and values of about 0.4 for the alpha and beta rhythms.

Canal\Ritm	Delta	Theta	Alpha	Beta	Gamma
1	0.0345353	0.6493754	0.4529862	0.4163538	0.750628
2	0.0347058	0.6373908	0.4588761	0.4101119	0.7552795
3	0.0348821	0.6299943	0.4564389	0.4097972	0.7538377
4	0.0348671	0.6398564	0.4479044	0.4018675	0.7556911
5	0.0337932	0.639655	0.4516111	0.416684	0.7524031
6	0.0332902	0.6500335	0.4654084	0.4288559	0.7405627
7	0.0702669	0.6826998	0.4619992	0.4123469	0.7062632
8	0.0218751	0.5306822	0.2071775	0.2731241	0.0224777
9	0.0345576	0.6389065	0.4545839	0.408236	0.7556786
10	0.0282654	0.6002317	0.4479862	0.4074394	0.7475454
11	0.0337663	0.6404656	0.4468172	0.3836081	0.7115622
12	0.0341592	0.6553298	0.4484005	0.4227536	0.6971204

Fig. 2. Hurst coefficient for a PD patient

Canal\Ritm	Delta	Theta	Alpha	Beta	Gamma
1	0.035173	0.636394	0.455721	0.408772	0.758554
2	0.033914	0.636304	0.455643	0.407921	0.757503
3	0.037354	0.636756	0.457515	0.408191	0.756727
4	0.037575	0.637918	0.45811	0.409024	0.758602
5	0.035089	0.636469	0.455738	0.408565	0.758732
6	0.035876	0.637512	0.456861	0.408624	0.758011
7	0.035055	0.636058	0.455993	0.408707	0.75862
8	0.034864	0.635929	0.455865	0.408656	0.758688
9	0.036405	0.637568	0.456421	0.407296	0.758952
10	0.036105	0.637175	0.455873	0.408532	0.758501
11	0.035526	0.637572	0.457358	0.408262	0.758245
12	0.034929	0.63577	0.456146	0.408604	0.758696

Fig. 3. Hurst coefficient for a PD patient with dementia

Case 4 – Female patient, 80 years old

This patient was diagnosed with PD (since2008) and diabetes, and is taking medication. She has difficulty to move, shortness of movements, and presents the rest tremor. Similar results have been obtained in this case. After analysing the results recorded the Hurst coefficient for delta rhythm was low as compared to the theta, alpha, beta, and gamma rhythms (Fig. 4).

Case 5 – Female patient, 79 years old

In this case, the patient has not been diagnosed with any neurological disorder, i.e., there is no particular visible tremor (stroke), and no medication is being administered.

Canal\Ritm	Delta	Theta	Alpha	Beta	Gamma
1	0.034527	0.636451	0.451609	0.408431	0.75765
2	0.033318	0.634832	0.443923	0.402649	0.741965
3	0.034417	0.636844	0.452452	0.406722	0.756732
4	0.032154	0.629469	0.449439	0.402988	0.739543
5	0.033758	0.635709	0.454412	0.406852	0.757004
6	0.034522	0.635599	0.453877	0.407715	0.758022
7	0.034482	0.634835	0.44924	0.409035	0.757131
8	0.034382	0.636086	0.453428	0.407989	0.75732
9	0.034296	0.635966	0.451661	0.409377	0.75711
10	0.033862	0.634293	0.450273	0.403751	0.755215
11	0.034457	0.635782	0.453878	0.408253	0.758133
12	0.033849	0.634524	0.449838	0.416245	0.743109

Fig. 4. Hurst coefficient for a PD patient with diabetes

Canal\Ritm	Delta	Theta	Alpha	Beta	Gamma
1	0.033741	0.629329	0.452378	0.406019	0.743491
2	0.02525	0.617794	0.449289	0.383941	0.475423
3	0.033476	0.635108	0.453796	0.411484	0.738768
4	0.035887	0.614309	0.439057	0.371912	0.601492
5	0.033797	0.639819	0.454791	0.416885	0.73304
6	0.035561	0.636796	0.454259	0.410808	0.75252
7	0.034532	0.639792	0.4553	0.413344	0.749604
8	0.035414	0.63491	0.452203	0.407306	0.738758
9	0.035269	0.635791	0.453742	0.418174	0.748701
10	0.03399	0.636091	0.453714	0.411605	0.749271
11	0.037353	0.642568	0.449422	0.413821	0.662166
12	0.034985	0.637815	0.455518	0.41081	0.753955

Fig. 5. Hurst coefficient for a non-diagnosed patient

Analysing the Hurst coefficient values we have observed similarities to the corresponding delta, alpha and beta ones of PD patients (Fig. 5).

7 Conclusions

The Hurst coefficient is a method that could be used to extract features from EEG signals acquired from patients with neurodegenerative disorders. This could reveal characteristics of sensorimotor rhythms. In order to estimate the Hurst coefficients we have relied on the aggregate variance method. Multiresolution wavelet analysis was applied to the subcomponents of the signals within the frequency band corresponding to EEG signals. Taking into account these preliminary results for patients with PD,

PD and dementia, PD and diabetes, as well as normal patients, it can be speculated that for non-diagnosed patients one could rely on the Hurst coefficients. In particular, non-diagnosed patients with low Hurst coefficient for the delta rhythm, have a high risk of neurological problems. Finally, a normal patient exhibits a way higher Hurst coefficient for the delta rhythm, which should be considered as a normal value. Considering the fact that nowadays there is no screening test for the early detection of PD, we have started to investigate the Hurst coefficient as a possible marker for neurological disorders screening.

References

1. World Health Organization (2016). http://www.who.int/research/en/
2. National Parkinson Foundation (2016). www.parkinson.org
3. Geman, O., Sanei, S., Chiuchisan, I., Graur, A., Prochazka, A., Vysata, O.: Towards an inclusive Parkinson's screening system. In: Proceedings of International Conference on System Theory, Control and Computing, pp. 476–480 (2014)
4. Iuliana, C., Geman, O., Chiuchisan, I., Iuresi, A.C., Graur, A.: NeuroParkinScreen – a health care system for neurological disorders screening and rehabilitation. In: International Conference on Electrical and Power Engineering (EPE 2014), 16–18 October 2014, pp. 536–540 (2014)
5. Campeanu, E., Abrudan, M.: Neurologie Clinica. Ed. Dacia, Cluj-Napoca, Romania (1980)
6. g.GAMMAsys (2016). http://www.gtec.at/Products/Electrodes-and-Sensors/g.GAMMAsys-Specs-Features
7. Mallat, S.: A Wavelet Tour of Signal Processing (1999)
8. Lazar, A.M.: Prelucrarea Discreta a Semnalelor Medicale Unidimesionale. Ed. Gh. Asachi, Iasi, Romania (2001)
9. Aldea, R.: Wavelet-based EEG subbands decomposition to highlight sensorimotor rhythms. Buletinul Institului Politehnic Iaşi, vol. LIX (LXIII), Fasc. 3, pp. 49–58 (2013)
10. Aldea, R.: Multiresolution wavelet analysis and Hurst estimation used for highlighting sensorimotor rhythms. Buletinul Institului Politehnic Iaşi, vol. LIX (LXIII), pp. 63–71 (2013)
11. Hurst, H.: Long-term storage capacity of reservoirs. Trans. Am. Soc. Civil Eng. **116**, 770–799 (1951)
12. Aldea, R., Tarniceriu D.: Estimating Hurst coefficient in motor imagery-based brain computer interface. In: International Conference on Speech Technology and Human-Computer Dialogue (SpeD 2013), Cluj-Napoca, Romania, 16–19 October 2013, pp. 215–219 (2013)
13. Akay, M. (ed.): Nonlinear Biomedical Signal Processing. Dynamic Analysis and Modeling, vol. II, pp. 1–25. Wiley-IEEE Press (2000)

Environment Description for Blind People

J.S. Park[1], D. López De Luise[2(✉)], D.J. Hemanth[3], and J. Pérez[1]

[1] CI2S Lab of Buenos Aires, Buenos Aires, Argentina
zeroalpha2000@gmail.com, perezjavier72@gmail.com
[2] CI2S Lab of Buenos Aires, UAI, AUSTRAL and UADER, Buenos Aires, Argentina
mdldl@ci2s.com.ar
[3] Department of ECE, Karunya University, Coimbatore, India
judespecialissue@gmail.com

Abstract. Visual processing is very efficient, letting people to use vision as the first approach to get information about environment. For blind people that information must be complemented with another very powerful data collection: sounds. In order to complement the white stick sounds, the prototype HOLOTECH gathers and segments video images and produces specific sounds to acknowledge about potential hazards. The underlaying model is based on a set of Neural Networks coordinated by an Expert System to make it possible to react to any new event in real time. This paper presents an outline of the model, the project and a test set to evaluate one of the Neural Networks specialized to detect and evaluate faces and other objects like cars. The main contribution of this work is automate the selection model for proper combination of information, discarding unnecessary data and defining the minimum precision requirements to fulfill the current goal.

Keywords: Computational intelligence · Blind people · Expert systems · Neural Networks

1 Introduction

It is hard to move around in a city or complex environment, especially for visually impaired people.

There are many different assistive technologies for impaired people. They usually try to detect and process any specific stimulus from the environment, in any situation. Among them it worth mention some proposals of assistive technology for students [1], orientation and mobility [2], recreation and leisure [3], perception enhancing and learning [4], etc. In the case of assistants for displacement, most of the approaches try to guide the visually impaired offering indications of how and where to move or how to translate from one place to another. All of them aim to help blind people to actively participate in society and to perform activities easily and safely [5]. SMART Vision [6] is a system for assisting the blind and visually impaired while navigating autonomously. It integrates a GIS with GPS, Wi-Fi, RFID tags and computer vision. In the other side, the current proposal requires a simple, light and cheaper technology. There is an electronic cane for navigating in indoor environment [7], but it requires a

V.E. Balas et al. (eds.), *Soft Computing Applications*, Advances in Intelligent Systems and Computing 633, DOI 10.1007/978-3-319-62521-8_30

customization of the cane to get an appropriate efficiency of the tool. Some proposals require the use of GPS or a previously sensed environment to provide a good displacement guide while tracking the user's movements [8–11]. These types of solutions are not reliable for very close obstacles in in-door environments or out-door short distance mobile threats. Most of these solutions has voice based warnings, but not a coding system in sounds. They usually also give instructions or perform alerts that take too much time during an emergency event. They may even perform repetitive signals making the user be confused and overwhelmed by the alerts [12, 13]. The current proposal provides a sound based language instead a voice commands or alerts, making it simpler to understand and reducing the time response and amount of information required to successfully avoid obstacles.

To help detecting surrounding obstacles there are several proposals, but most of them try to fully describe the environment becoming computational intensive and requiring hardware that is unpractical for a daily and all-day basis usage [14, 15].

This paper presents a navigation system that models part of its environment by a human fashion approach that takes relevant features and works them as main cues. Information comes from a cell phone camera, ultrasound sensors and the mobile inset microphone and the software runs in an Arduino nano GRAVITECH(c). Its implementation is a prototype named HOLOTECH [21] that communicates with the user by a very reduced sound language. Through this language, the prototype is able to alert the user with a simple set of sounds, about risk levels, proximity and type of objects nearby. The model combines Haar like Feature-base Cascade Classifier, sound features extraction, ultrasound information and Expert System (ES) implemented to decide if the nearby object is a future risk for the user, before selecting this kind of model we tested using a fuzzy multi criteria for robotic system to determine different approach [28]. It is out of the scope of this paper the evaluation of the language of sounds produced. The goal of this paper is to implement a lightweight, accessible to build and easy to adapt system to assist blind people in moving in in-door and out-door places, and reducing risks of injuries. It does not intend to replace the traditional white cane but to complement it.

2 The HOLOTECH Model and Prototype

Mobile devices are useful many indoor/outdoor environments. Besides they constitute a simple and accessible technology, which nowadays can be expanded with android applications. The prototype uses also a Nano Arduino GRAVITECH [16], two ultrasonic sensors HC-SR04 and a pedometer. Figure 1 is the global configuration of the hardware required to implement the model.

The goal is to collect information with redundancy, in order to improve inference precision about which objects are in the environment. The prototype is working on Android SDK 22, with Matlab for Cascade Object detection connected to an Expert System though a simple feature vector. The arduino programs are C++ [17].

The device is located on the chest of the user, to complement the position and activity of the eyes while walking.

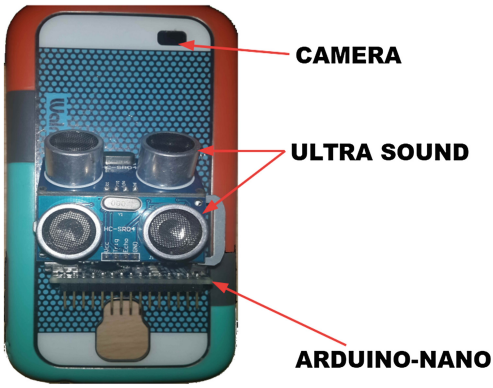


Fig. 1. HOLOTECH global disposition

Video sampling from the built-in camera feeds the Android module located in the cell phone, and forwarded to the Arduino's software. The core of the system is a set of Neural Networks (NN) connected to an Expert System (ES). The NN are trained filters using the features derived by Cascade Object Detection (COD) for specific obstacles recognition.

The workflow has many steps:

- ***First step***, input video. For testing purposes a recording is taken with the camera at a resolution of 320 × 240 pixels. Input ultrasonic from two sensors located at a 45 position respecting the floor with a 90° each other.
- ***Second step***, slice into images every 30 s.
- ***Third step***, after slicing, the images are processed to determine which objects are present in the sequence.
- ***Fourth step***, segment objects: Convert into a grey scale, segment, chop and extract the border [18, 19].
- ***Fifth step***, evaluate location and distance of obstacles: process with Haar-like feature objects detection. Evaluate a short subsequence of images with a fixed sliding window in time; it is possible to obtain the acceleration, orientation and displacements in the scene.

The system uses a coordinate with its origin in the bottom-center of the video camera of the device, to analyse and for a better understanding of the movements for detected objects. The x axis has its origin in the center of the image, and the Y axis at the bottom. This way images may be associated to sound alarms easily, since objects on the right side of the individual have positive coordinate while objects on the other side will always have a negative location. Alarm sounds have different frequency according to the side.

When an object is detected (using Cascade Object Detector, that is a NN), it is separated from its background, sized and located in the space. With data derived, the prototype creates a vector representing the image like the one in Table 1.

Table 1. Vector information

X/T1	Y/T1	X/T2	Y/T2	DX	DY	V/X	V/Y
100	120	110	120	20	10	10	5

In the table, X/T1 stands for X position in frame 1 at time T1, and Y/T1 stands for Y position in frame 2 at time T1. In a similar way, X/T2 and Y/T2 represent location of the same obstacle at time T2. DX and DY are the displacement for both coordinates during T2–T1 and V/X, V/Y are velocity in pixels/sec. This information added with the data sensed with both ultrasonic sensor is fed to the expert system in order to evaluate if any object would strike the person and the severity of this approximation in time.

- ***Sixth step***, encode position/distance and activate the proper sound in the alarm system (see [21]): the system implements an specific language varying frequency, tone and vibration. To do so, an ES [20] analyses the information extracted and determines the urgency of the communication. It implements a context-free grammar that is flexible enough to cover more new situations.
- ***Seventh step***, sound production restated by the model's grammar. It communicates any risk situation (as evaluated by the ES). Rules and patterns [22] are defined in a unique way, in order to avoid confusion or ambiguity.

3 Test Set and Results

In order to test the system's precision, a set of 15 short videos of 10 s. All of them are real situations walking in an outdoor environment.

3.1 Image Acquisition and Feature Extraction

The 15 real-time videos are obtained using a video camera (2X). The duration for each video is 10 s. Among the 15 videos, some of the videos consist of human faces and others of moving cars. The human faces and the cars are identified as objects which are the region of interest. Initially, each video is sliced into images which results in 200–300 frames for each video data. The dimension of each image is 720 $\times$ 480 pixels and the average size is 30 KB. The format of raw image is .jpg which is converted to .avi format for further processing [23]. Among these huge number of frames, 10 frames are chosen for further processing. The selection of these frames is done randomly and they are different for each of the input video. It may be noted that frame selection is less significant since the objective of this work is only to detect the objects, for validation purpose some others techniques like features detection method were used [27].

Further, feature extraction is done in each of these frames. The features are nothing but the motion vector of the region of interest in each frame. The motion vector is the position of the first pixel of the object in each frame. In order to estimate the motion vector, it is necessary to zero down on the object of interest (faces/cars). The region of interest is selected using a semi-automated method which portrays the object within a

rectangular box. The software used for this selection is MATLAB. Now, the co-ordinates of the upper left most pixel within the rectangular box is considered to be the motion vector of that frame [24].

These motion vector values are noted down. The same process is repeated for all the frames and the motion vector values are observed. There are sufficient changes between the motion vectors of each frame of the same input video data. These motion vector values are tabulated in an array form. These values are further used to train the classifier.

3.2 Classifier Training and Precision Evaluation

The classifier used for the training is haar-cascade classifier; this classifier was selected for the fast response and high success rate to recognize the object that was trained for.

Because the classifier doesn't support multiple classifications of different objects at the same time, we used many classifier to classifier all possible object and later build a expert system to integrate all specific classifier. For this paper only the description of the training done to recognize a car is descripted.

Most of the procedure for the training is similar between the object. In each case a collection of positive images, were the object is present and a collection of negative images, were the object is not present was used for the training.

To Start with and to test the efficiency of the classifier an image bank was used to train and the data recorded in real-time. The image bank UIUC Image Database [25] have a collection images of cars and a random set of image that later is used as the negative set. The classifier configuration wasn't changed, and was made to use the his default values were, the stages were the initial value is 14, splits 1, minimum hit rate 0.995, max false alarm 0.5, weight trimming 0.95, max tree splits 0 and minimum positive sample per cluster is 500.

In the training each stage the information displayed are:

N = current feature for this cascade (seq- > total),
%%SMP = percentage of samples used, if trimmings enabled (v_wt)
F = '+' if is Flipped, if symmetry is specified (v_flipped), '−' otherwise
ST.THR = stage threshold,
HR = Hit Rate based on Stage threshold (v_hitrate/numpos),
FA = False alarm based on Stage threshold (v_falsealarm/numneg)
EXP.ERR = Strong classification error of adaboost algorithm, based on threshold = 0 (v_experr)

This is a description of information provided from classifier.

An example result of the stage one of the training is that the more feature use the HR progressively get worse (Table 2).

POSITIVE SAMPLE: 500 502 0.996016
NEGATIVE SAMPLE: 500 0.45045
BACKGROUND PROCESSING TIME: 0.00
Precalculation time: 0.00
Stage training time: 152.00

Table 2. Training statistical information for cascade detector

N	%SMP	F	ST.THR	HR	FA	EXP.ERR
1	100%	−	−0.703392	1.000000	1.000000	0.168000
2	100%	+	−1.186522	1.000000	1.000000	0.168000
3	100%	−	−1.753016	1.000000	1.000000	0.132000
4	84%	+	−1.960341	1.000000	1.000000	0.130000
5	86%	−	−1.567593	0.998000	0.734000	0.121000
6	74%	+	−1.470981	0.996000	0.516000	0.108000
7	69%	−	−1.175175	0.996000	0.382000	0.075000

Number of used features: 7
Parent node: 0
Chosen number of splits: 0
Total number of splits: 0
Result of stage 1 training.

As can be seen the speed is high, which is important since it is intended for real time events (a human walking in the street). The HR and FA are pretty balanced, giving an excellent EXP.ERR rate.

Once the training is done an XML containing the classification logic is generated. Using this XML and some testing images, the real-time image collected, the next step is to test performance of the model for real-time cases.

Using 10 sets of images to get an overall test, the classifier with highest Hit rates was used to calculate the kappa statistics value (k) of the model.

The formula for such statistics is [26]:

$K = P(a) - P(e)/(1 - P(e))$

where P(a) is the probability of an event called "a", and P(e) is the statistical probability of its occurence.

The output of the performance evaluation is as follow:

As can be seen from the Table 3, hits are balanced with missed cases but in just one case there is a false positive, that indicates the approach avoid most of false alarms.

With the data obtained to estimate a partial statistical of the model.

Table 3. Performance testing for the detector

File name	Hits	Missed	False
Tests/img1.jpg	0	1	0
Tests/img2.jpg	1	0	0
Tests/img3.jpg	1	0	0
Tests/img4.jpg	1	0	0
Tests/img5.jpg	0	1	1
Tests/img6.jpg	1	0	0
Tests/img7.jpg	0	1	0
Total	5	4	1

4 Conclusions and Future Work

With the result obtained with the classifier we can estimate a helpful decision of the obstacle to assimilate with the expert system. And including a Sound language is useful to encode the information that describes environment's risk, as it is faster and unambiguous. It is important to recognize and analyze an object's behavior to take quick action in real time.

For future work the system needs to apply the sound language to a more extended catalogue of objects. Training the classifier with more data and change the Data bank image with only real-time image situation. It is also pending to consider other versions of android platform to make it more flexible. Besides, it can be evaluated the possibility of customizing certain details (for instance type of sounds and specific set of obstacles of interest). Also an extensive statistical analysis and usability tests are pending.

References

1. Lahav, O., Schloerb, D.W.: Virtual environments for people who are visually impaired integrated into an orientation and mobility program. Am. J. Vis. Impair. (2015)
2. Kelly, S.M., Ajuwon, P.M., Wolffe, K.E.: The recreation and leisure pursuits of employed adults with visual impairments in Nigeria: part 1. Am. J. Vis. Impair. (2014)
3. Szubielska, M., Marek, B.: The role of visual experience in changing the size of objects in imagery processing. Am. J. Vis. Impair. (2015)
4. Erin, J.N.: Practice perspectives. Different paths to success: an individualized approach to effective teaching. Am. J. Vis. Impair. (2015)
5. Hersh, M.A.: The design and evaluation of assistive technology products and devices. Part 1: design. Int. Encycl. Rehabil. (2010)
6. Faria, J., Lopes, S., Fernandes, H., Martins, P., Barroso, J.: Electronic white cane for blind people navigation assistance. In: 2010 World Automation Congress (WAC), pp. 1–7 (2010)
7. Ali, A.M., Nordin, M.J.: SIFT based monocular SLAM with multi-clouds features for indoor navigation. In: TENCON 2010 - 2010 IEEE Region 10 Conference, pp. 2326–2331 (2010)
8. Lahav, O.: Improving orientation and mobility skills through virtual environments for people who are blind: past research and future potential. In: Proceedings of the 9th International Conference on Disability, Virtual Reality & Associated Technologies Laval, France, 10–12 September 2012
9. Tihamér, S., Brassa, I.: Assistive Technologies for Visually Impaired People (2011)
10. Evangeline, J.: Guide systems for the blind pedestrian positioning and artificial vision. IJISET – Int. J. Innov. Sci. Eng. Technol. **1**(3), 42–44 (2014)
11. Duarte, K., Cecilio, J., Sá Silva, J., Furtado, P.: Information and assisted navigation system for blind people. In: Proceedings of the 8th International Conference on Sensing Technology, Liverpool, UK, 2–4 September 2014
12. Arditi, A., YingLi, T.: User interface preference in the design of a camera-base navigation and wayfinding aid. J. Vis. Impair. Blind. **107**, 118 (2013)
13. Salazar, E., Ceres, R.: El sensor ultrasónico como potenciador de procesos comunicativos en personas con limitación visual. In: Congreso Iberoamericano de Comunicación Alternativa y Aumentativa, Lisboa (1993)

14. Gueuning, F.E., Varlan, M., Eugene, C.E., Dupuis, P.: Accurate distance measurement by an autonomous ultrasonic system combining time-of-flight and phase-shift methods. IEEE Trans. Instrum. Meas. **46** (1997)
15. Velázquez, R.: Wearable assistive devices for the blind. In: Lay-Ekuakille, A., Mukhopadhyay, S.C. (eds.) Wearable and Autonomous Biomedical Devices and Systems for Smart Environment: Issues and Characterization. LNEE, vol. 75, pp. 331–349. Springer (2010)
16. Arduino Nano. http://arduino.cc/en/Main/ArduinoBoardNano (2015)
17. Igoe, T.: Making Things Talk. 2nd edn. (2012)
18. Krasula, L., Klima, M., Rogard, E., Jeanblanc, E.: MATLAB-based Applications for Image Processing and Image Quality Assessment Part II. Experimental Results (2012)
19. Bala, A.: An improved watershed image segmentation technique using MATLAB. Int. J. Sci. Eng. Res. **3**(6), 1–4 (2012)
20. http://www.mathworks.com/help/vision/ug/train-a-cascade-object-detector.html (2015)
21. Park, J.S., De Luise, D.L., Pérez, J.: HOLOTECH prototype. Sound language for environment's understanding. Int. J. Learn. Technol. Interscience Publishers (2015)
22. De Luise, D.L.: Ingeniería en inteligencia computacional (Computational Intelligence Engineering). In: Rovarini, P. (ed.) (UTN-FRT), pp. 104 (2012). Zadeh, L.A.: Interpolative reasoning as a common basis for inference in fuzzy logic, neural network theory and the calculus of fuzzy If/Then rules. Opening talk. In: Proceedings of 2nd International Conference on Fuzzy Logic and Neural Networks, Iizuka, pp. XIII–XIV (1992)
23. Chen, Q., Kotani, K., Lee, F., Ohmi, T.: A fast search algorithm for large video database using HOG based features
24. Mahdi, H.S.: Image Understanding Using Object Identification and Spatial Relationship
25. https://cogcomp.cs.illinois.edu/Data/Car/
26. Carletta, J.: Assessing agreement on classification tasks: the kappa statistic. Comput. Linguist. **22**(2), 249–254 (1996)
27. Abada, L., Aouat, S.: Facial shape-from-shading using features detection method, pp. 3–19. doi:10.1504/IJAIP.2016.074774
28. Bairagi, B., Dey, B., Sarkar, B., Sanyal, S.: Selection of robotic systems in fuzzy multi criteria decision-making environment, pp. 32–42. doi:10.1504/IJCSYSE.2015.067798

Specialized Software System for Heart Rate Variability Analysis: An Implementation of Nonlinear Graphical Methods

E. Gospodinova[1], M. Gospodinov[1], Nilanjan Dey[2(✉)], I. Domuschiev[3], Amira S. Ashour[4], Sanda V. Balas[5], and Teodora Olariu[6]

[1] Institute for System's Engineering and Robotics, Bulgarian Academy of Sciences, Filial Veliko Tarnovo, Sofia, Bulgaria
{jenigospodinova, mitgo}@abv.bg
[2] Department of Information Technology, Techno India College of Technology, Kolkata, India
neelanjandey@gmail.com
[3] Department of Internal Diseases, Multiprofile Hospital for Active Treatment, Plovdiv, Bulgaria
voper@dir.bg
[4] Department of Electronics and Electrical Communications Engineering, Faculty of Engineering, Tanta University, Tanta, Egypt
amirasashour@yahoo.com
[5] BTM Resources SRL, Arad, Romania
dzanda@yahoo.com
[6] Vasile Goldis Western University of Arad, Arad, Romania
olariu_teodora@yahoo.com

Abstract. Heart Rate Variability (HRV) designates the progressive variation in the intervals between successive heartbeats in the sinus rhythm. The HRV analysis is a non-invasive/effective tool to demonstrate the influence of the autonomic nervous system over the heart rhythm regulation. Recently, researches are interested with developing advanced software systems for HRV analysis. For cardiac disease patients' investigation, nonlinear modeling and analysis can keep track in real time and foresee possible changes in circadian heart rate. The current work presents a novel created software system for HRV analysis based on 24-h Holter ECG signals of group of healthy and unhealthy subjects. The nonlinear analysis of heart intervals was performed with the implementation of original high performance algorithms and software to quantify the heart rate irregularity. The proposed software system achieved the short time ability for parametric estimation of patients' cardiac status. It is based on long-term (24-h) Holter ECG signals with implementation of mathematical nonlinear methods. The experimental results established that the designed software system for analysis of 24-h Holter recordings is appropriate for diagnostic, forecast and prevention of the pathological cardiac statuses. The developed and implemented graphical representation and visualization approaches for the results can be stored in specialized data base.

V.E. Balas et al. (eds.), *Soft Computing Applications*, Advances in Intelligent Systems and Computing 633, DOI 10.1007/978-3-319-62521-8_31

Keywords: Heart Rate Variability (HRV) · ECG (Electrocardiography) signal · Holter signal · Nonlinear graphical methods

1 Introduction

Various cardiovascular diseases are characterized by different graphical images after processing the heartbeat of software programs. Ischemic heart disease, stroke, chronic obstructive lung have remained the top major causes of death. Researchers depicted that cardiovascular diseases can be mathematically analyzed and predicted by cardiac screening for early detection of possible complications to prevent disease. The HRV is known as a diagnostic parameter that defined by the ECG through measuring time intervals between the heartbeats. In 1996, the European Society of Cardiology and North American Society of Pacing and Electrophysiology [1] provided recommendations on clinical usage of the HRV method for the evaluation of cardiology disease risk including the myocardial infarction (heart attacks), sudden cardiac death and essential hypertension. The HRV signals can be analyzed and calculated in real time by non-invasive manner, while all other biomarkers, used in clinical practice, are discrete and imply blood sample analysis [2].

Methods of HRV analysis are divided in two groups, namely linear and non-linear methods. Linear methods can be used in time- or frequency- domain for HRV analysis [3–5]. Nevertheless, significant characteristics of the signal dynamics are missed during the use of conventional (linear) methods. Therefore, recently increased interest is directed toward the non-linear analysis, where the the HRV measurements are non-linear and non-stationary. In addition, a considerable part of information is coded in the dynamics of their fluctuation in different time periods [5].

Consequently, the current work is conducted on selected control groups of patients with cardiovascular diseases, each with 24-h Holter recordings. Each record contains information of around 100,000 heartbeats. Fluctuations of the physiological signals possess hidden information in the form of self-similarity, scale structure and fractality [5]. Due to the large volume of research information, it is important to correctly determine the trend of the disease in each patient. Such tests are performed periodically to compare the graphic characteristics of images from clinical studies undertaken as a result of treatment and to give an idea for the patient's condition and treatment quality. Nonlinear analysis of cardiology data is a relatively new scientific approach that provides new approach to assess dynamics of heart activity. The main objectives of this paper are described, as follows:

- Development of software for HRV analysis of 24-h Holter ECG records of healthy and unhealthy subjects. The software offered analysis in time-domain, frequency-domain, nonlinear and wavelet analysis. The nonlinear analysis results are illustrated by applying the Detrended Fluctuation Analysis (DFA), Rescaled adjusted range Statistics plot (R/S) and Poincaré plot.
- Demonstration of the results' graphical representation using different approaches on clinical studies of HRV and assessment of health status of patients with cardiovascular disease with multiple and periodic 24-h Holter studies of treatment.

2 Methodology

In the present work, two groups of signals are analysed, namely the RR time series of 16 normal subjects and 16 congestive heart failure (CHF) patients. These signals are consisting of around 100,000 data points corresponding to 24-h Holter ECG recordings. The proposed software system deployed several methods as follows, where these methods are realized by the Matlab software developed in research project for nonlinear analysis of ECG signals.

2.1 Detrended Fluctuation Analysis (DFA)

The DFA is a technique for detecting correlations in time series [6]. These functions are able to estimate several scaling exponents from the RR time series being analyzed. The scaling exponents characterize short or long-term fluctuations. The relationship on a double-log graph between the investigated signal's fluctuations F(n) and the time scales n can be approximately evaluated by a linear model that provides the scaling exponent α given by: $F(n) \approx n^{\alpha}$. The parameter α depends on the correlation properties of the signal, while changing 'n' affects the fluctuations change of the signal. Linear behavior of the dependence F (n) indicates the presence of signal scalable behavior. The slope of the straight line determines the value of α. For uncorrelated signals, the value of α is within the range [0, 0.5]. Whereas, $\alpha > 0.5$ indicates the presence of correlation, and $\alpha = 1$ when the signal is 1/f (noise), while $\alpha = 1.5$ usually Brownian motion.

Typically, in the case of RR time series, DFA shows two ranges of scale invariance that are quantified by two separate scaling exponents α_1 and α_2 reflecting the short-term and long-term correlation; respectively [6]. The short-term fluctuation are characterized by the slope α_1 obtained from the (log n, log F(n)) graph within range $4 \leq n \geq 11$ and the slope α_2 obtained from the range $12 \leq n \geq 64$.

2.2 Rescaled Adjusted Range Statistics Plot (R/S)

The rescaled range is a statistical measure of the time series variability. The Hurst exponent is one closely associated method with the R/S [7]. This exponent is a measure that has been widely used to evaluate the self-similarity and correlation properties of fractional Brownian noise, the time series produced by a fractional Gaussian process. The self-similarity means that the statistical properties (all moments) of a stochastic process do not change for all aggregation levels. The main properties of self-similar processes include: slowly decaying variance, long-range dependence and Hurst effect. The sample variance is decreased more slowly than the reciprocal of the sample size. This process is called a stationary process with long-range dependence if its autocorrelation function is non-summable. The autocorrelations decay speed is more hyperbolic than exponential. The Hurst effect expresses the degree of self-similarity. Based on the Hurst exponent value, the following classifications of time series can be realized:

- $H = 0.5$ indicates a random series;
- $0 < H < 0.5$ indicates that the signal data is anti-correlated;
- $0.5 < H < 1$ indicates that the signal data is long-range correlated.

The R/S method for the time series X(n) is asymptotically given by a power law: $R(n)/S(n) \propto n^H$, where R(n) is the difference between the minimum and maximum accumulated values range; S(n) is the standard deviation estimated from the observed data X(n), and H is the Hurst exponent. In order to estimate the Hurst exponent, R(n)/S(n) versus n in log-log axes is plotted. The slope of the regression line approximates the Hurst exponent [7].

2.3 Poincaré Plot

The Poincaré plot analysis is a graphical nonlinear method to assess the dynamic of HRV [8]. This method provides summary information as well as detailed beat-to-beat information on the behaviour of the heart. It is a graphical representation of temporal correlations within the *RR* intervals derived from ECG signal. The Poincaré plot is known as a return maps or scatter plots, where each *RR* interval from the time series $RR = \{RR1, RR2, \ldots, RRn, RRn+1\}$ is plotted against next *RR* interval. In the current work, the used Poincaré plot parameters are $SD1$, $SD2$ and the $SD1/SD2$ ratio, where $SD1$ and $SD2$ are the standard deviation of Poincaré plot projection on the line perpendicular to the line of identify and the on the line of identify (y = x); respectively. Typically, $SD1$ is correlated with high frequency power, while $SD2$ is correlated with both low and high frequency powers. The ratio $SD1/SD2$ is associated with the randomness of the HRV signal. It has been suggested that the ratio $SD1/SD2$, which is a measure of the randomness in HRV time series, has the strongest association with mortality in adults.

Consequently, these methods are used to implement a new software tool for the HRV based classification of healthy subjects and CHF patients, where several researchers were interested with the HRV analysis and cardiac diseases [9–11].

3 Results and Discussion

The described mathematical methods are implemented in the form of specialized software for the calculation and assessment of HRV parameters. The analyzed data from the medical study of patients were combined into two groups of signals and records, namely for 16 CHF (Congestive Heart Failure) patients and 16 normal subjects. The main software menu for each patient is illustrated in Fig. 1.

Figure 1 shows the distribution of QRS complexes, RR intervals and results of time-domain analysis (the values of investigated parameters, RR and HR histograms) in the menu. This page allows the selection of the required analysis type including the time-domain, frequency-domain, nonlinear and wavelet analysis. This work is interested with the nonlinear analysis results.

Figures 2(a) and (b) illustrate values of scaling exponents and the line slope F(n) on double logarithmic plot obtained by using the DFA method for the investigation of signals.

Figure 2 depicts a significant difference between patients with CHF and healthy controls in short- and long- time scales. Healthy subjects typically show fractal behavior of heartbeat dynamics, while patients with CHF show an alteration in fractal

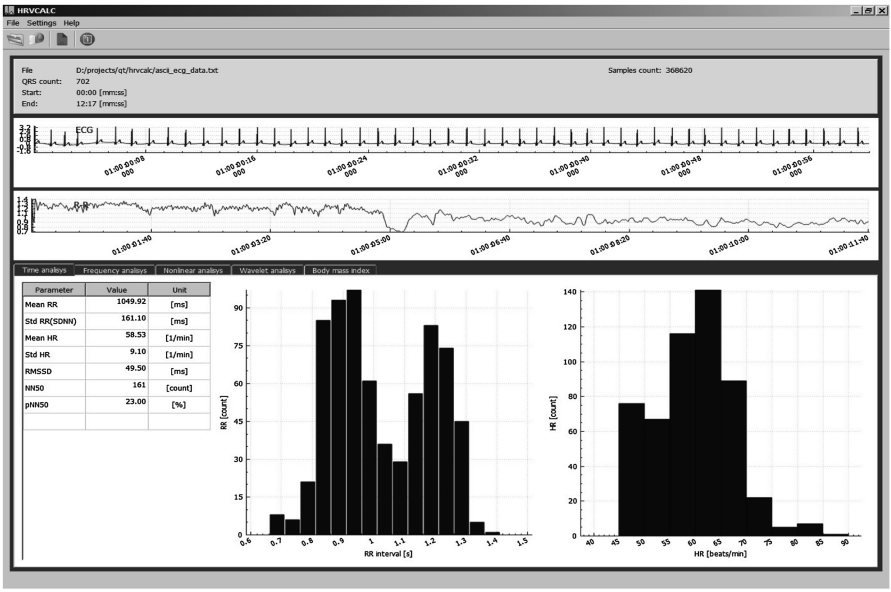
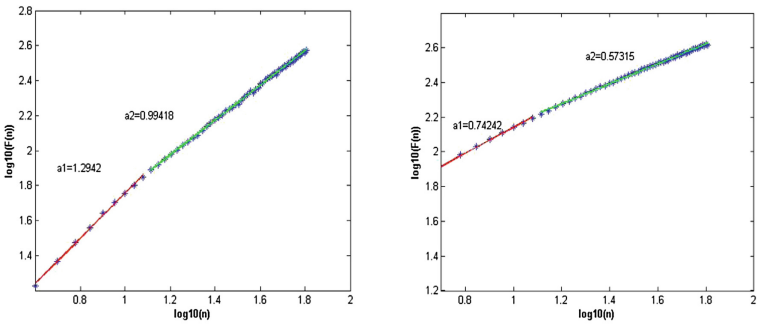


Fig. 1. Main menu of the proposed HRV analysis software



a) b)

Fig. 2. (a) scaling exponents, and (b) the line slope on double logarithmic plot

correlation properties. Moreover, Fig. 3(a) and (b) demonstrate the obtained results of the R/S method that applied to the studied signals to determine the value of the Hurst exponent.

Figure 3 establishes that the RR time series are correlated, i.e. they are fractal time series. For normal subject, the value of Hurst exponent is high due to the variation being chaotic, and for CHF patient this value decreases due to the low RR variation. The results of the Poincaré plot analysis of RR time series for healthy subject and CHF patient are presented in Figs. 4(a) and (b).

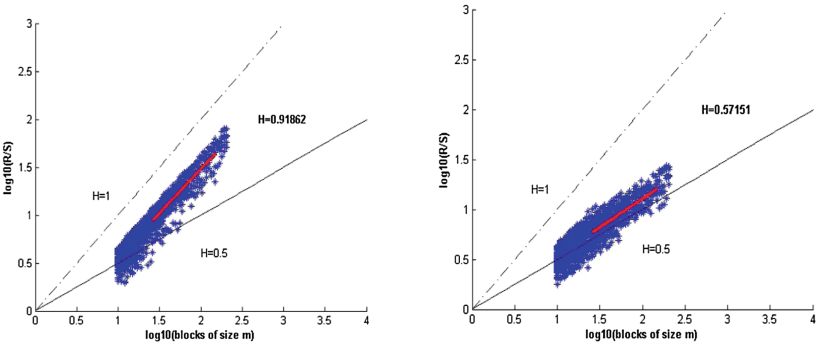


Fig. 3. The Hurst exponent value for (a) the normal subjects, and (b) the CHF patients.

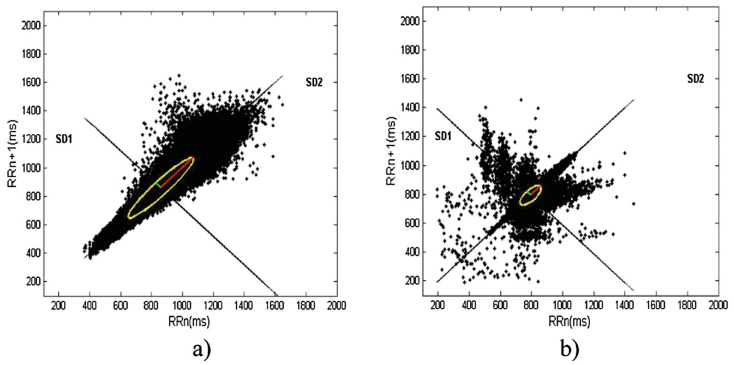


Fig. 4. The Poincaré plot analysis of RR time series for (a) healthy subject and (b) CHF patient

Figure 4 illustrates that the Poincaré plot for healthy subject is a cloud of points in the shape of an ellipse ('comet' shaped plot). However, points for the CHF patient are a cloud of points in shape of a circle ('complex' shaped plot). The geometry of these plots can be used to distinguish between healthy and unhealthy subjects. The obtained Poincaré plot parameters are directly related to the physiology of the heart. The parameter $SD1$ represents the semi-minor axis length of the ellipse that reflects short term variability of heart rate. The parameter $SD2$ represents the semi-major axis length of long term variability measurements. Since, the values of the $SD1$ and $SD2$ depend on the RR intervals, the ratio of $SD1/SD2$ is used to make comparison among Poincaré plots from different subjects. The values of $SD1$ and $SD2$ are higher in normal subjects compared to the congestive heart failure subjects. Finally, Table 1 reports the investigated parameters values (mean $\pm$ standard deviation).

Table 1 illustrates the significant difference between the different parameter's values for the corresponding method for the healthy subjects and the CHF patients. It is noticeable that values of all parameters have higher values in the healthy subjects compared to the CHF patients (which have lower values) except in the vase of the

Table 1. Parameters for healthy subjects and CHF patients

Parameter	Healthy subjects	CHF patients
Alpha 1 (DFA)	1.2942 ± 0.205	0.7424 ± 0.311
Alpha 2 (DFA)	0.9942 ± 0.319	0.5731 ± 0.376
Alpha (DFA)	1.0578 ± 0.198	0.6192 ± 0.231
Hurst (R/S)	0.9186 ± 0.018	0.5715 ± 0.219
SD1 [ms] (Poincare plot)	46.393 ± 22.02	35.123 ± 18.23
SD2 [ms] (Poincare plot)	295.295 ± 35.9	87.5673 ± 29.12
SD1/SD2 (Poincare plot)	0.15711 ± 0.32	0.40121 ± 0.41

$SD1/SD2$ ratio. Consequently, the proposed software is significant to distinguish between the two classes of the healthy and CHF patients.

The preceding results established the effectiveness of the proposed approach, thus it is recommend designing a system that integrates the current work with other techniques as in [12–24].

4 Conclusions

The current work implemented software tool for HRV analysis. It represented the results of nonlinear mathematical methods based on the fractal theory for selected group of patients. The experimental results established that the used nonlinear graphical methods for HRV analysis are an effective tool to visualize HRV fluctuations. The cardiac data results and the selected methods to manage the analysis providing both detailed information for the patients physiological status proved the proposed software significance. Additionally, this software developed a step towards a framework for prognosis and prevention of pathology status in case of cardiovascular disease. The use of graphic-oriented methods for analysis and visualization of cardiac diseases facilitates decision-making by health professionals, especially in case of large amounts of information such as the analysis of diurnal heart rhythms.

References

1. Heart rate variability. Standards of measurement, physiological interpretation, and clinical use. Task force of the European society of cardiology and the North American Society of pacing and electrophysiology. Eur. Heart J. **17**, 354–381 (1996)
2. Dimitrova, M., Lahtchev, L., Lozanova, S., Roumenin, C.: Cloud computing approach to novel medical interface design. In: Handbook of Medical and Healthcare Technologies, pp. 245–265. Springer, New York (2013)
3. Ernst, G.: Heart Rate Variability. Springer, London (2014)
4. Acharya, U.-R., Suri, J.-S., Spaan, J.-E., Krishnan, S.-M.: Advances in Cardiac Signal Processing. Springer, Heidelberg (2007)
5. Gospodinova, E., Gospodinov, M., Dey, N., Domuschiev, I., Ashour, A., Sifaki-Pistolla, D.: Analysis of heart rate variability by applying nonlinear methods with different approaches for graphical representation of results. Int. J. Adv. Comput. Sci. App. **6**(8), 39–45 (2015)

6. Peng, C.-K., Havlin, S., Stanley, H.-E., Goldberger, A.-L.: Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series. Chaos: an interdisciplinary. J. Nonlinear Sci. **5**(1), 82–87 (1995)
7. Gospodinova, E.: Graphical methods for nonlinear analysis of ECG signals. Int. J. Adv. Res. Comput. Sci. Softw. Eng. **4**(12), 40–44 (2014)
8. Smith, R.L., Wathen, E.R., Abaci, P.C., Bergen, N.H.V., Law, I.H., Dick II, M.D., Connor, C., Dove, E.L.: Analyzing heart rate variability in infant using non-linear Poincarè techniques. Comput. Cardiol. **36**, 673–876 (2009)
9. Dey, N., Das, A., Chaudhuri, S.S.: Wavelet based normal and abnormal heart sound identification using spectrogram analysis. Int. J. Comput. Sci. Eng. Technol. (IJCSET) **3**(6), 186–192 (2012). ISSN 2229-3345
10. Dey, N., Mishra, G., Nandi, B., Pal, M., Das, A., Chaudhuri, S.-S.: Wavelet based watermarked normal and abnormal heart sound identification using spectrogram analysis. In: 2012 IEEE International Conference on Computational Intelligence & Computing Research (ICCIC), pp. 1–7 (2012)
11. Araki, T., Ikeda, N., Dey, N., Acharjee, S., Molinari, F., Saba, L., Suri, J.-S.: Shape-based approach for coronary calcium lesion volume measurement on intravascular ultrasound imaging and its association with carotid intima-media thickness. J. Ultrasound Med. **34**(3), 469–482 (2015)
12. Aarthishree, S., Jayashree, M., Fathima, J.R.: A quick approach to detect epilepsy and seizure in brain. Int. J. Adv. Intell. Paradig. **8**(4), 412–424 (2016)
13. Arokiaraj, S.P., Robert, L.: RDNAS: a simple DNA sequence squeezer using enhanced run length encoding. Int. J. Adv. Intell. Paradig. **8**(4), 443–450 (2016)
14. Khan, S., Rahman, S.M., Tanim, M.F., Ahmed, F.: Factors influencing K means algorithm. Int. J. Comput. Syst. Eng. **1**(4), 217–228 (2013)
15. Aguiar, Y.P., Maria de Fátima, Q.V., Galy, E., Santoni, C.: Accounting for individual and situation characteristics to understand the user behaviour when interacting with systems during critical situations. Int. J. Comput. Syst. Eng. (IJACI) **6**(2), 29–55 (2014)
16. Bersch, S., Azzi, D., Khusainov, R., Achumba, I.E.: Artificial immune systems for anomaly detection in ambient assisted living applications. Int. J. Ambient Comput. Intell. (IJACI) **5**(3), 1–15 (2013)
17. Kumar, S.U., Inbarani, H.H., Azar, A.T., Hassanien, A.E.: Identification of heart valve disease using bijective soft sets theory. Int. J. Rough Sets Data Anal. (IJRSDA) **1**(2), 1–14 (2014)
18. Kanungo, D.P., Nayak, J., Naik, B., Behera, H.S.: Hybrid clustering using elitist teaching learning-based optimization: an improved hybrid approach of TLBO. Int. J. Rough Sets Data Anal. (IJRSDA) **3**(1), 1–19 (2016)
19. Hudlicka, E.: Guidelines for designing computational models of emotions. Int. J. Synth. Emot. (IJSE) **2**(1), 26–79 (2011)
20. Shivakumar, G., Vijaya, P.A.: Analysis of human emotions using galvanic skin response and finger tip temperature. Int. J. Synth. Emot. (IJSE) **2**(1), 15–25 (2011)
21. Hore, S., Chakroborty, S., Ashour, A.S., Dey, N., Ashour, A.S., Sifaki-Pistolla, D., Bhattacharya, T., Chaudhuri, S.R.: Finding contours of hippocampus brain cell using microscopic image analysis. J. Adv. Micros. Res. **10**(2), 93–103 (2015)
22. Dey, N., Das, A., Chaudhuri, S.S.: Wavelet based normal and abnormal heart sound identification using spectrogram analysis. arXiv preprint arXiv:1209.1224 (2012)
23. Roy, P., Goswami, S., Chakraborty, S., Azar, A.T., Dey, N.: Image segmentation using rough set theory: a review. Int. J. Rough Sets Data Anal. (IJRSDA) **1**(2), 62–74 (2014)
24. Dey, A., Bhattacha, D.K., Tibarewala, D.N., Dey, N., Ashour, A.S., Le, D.N., Gospodinova, E., Gospodinov, M.: Chinese-chi and Kundalini yoga meditations effects on the autonomic nervous system: comparative study. Int. J. Interact. Multimedia Artif. Intell. **3**(7), 87–95 (2016)

Classifier Ensemble Selection Based on mRMR Algorithm and Diversity Measures: An Application of Medical Data Classification

Soraya Cheriguene[1,2], Nabiha Azizi[1,2], Nilanjan Dey[3(✉)], Amira S. Ashour[4], Corina A. Mnerie[5], Teodora Olariu[6], and Fuqian Shi[7]

[1] Labged Laboratory, Annaba, Algeria
{cheriguene,azizi}@labged.net
[2] Computer Science Department, BadjiMokhtar University, PO BOX 12, 23000 Annaba, Algeria
[3] Department of Information Technology, Techno India College of Technology, Kolkata, India
neelanjandey@gmail.com
[4] Department of Electronics and Electrical Communications Engineering, Faculty of Engineering, Tanta University, Tanta, Egypt
amirasashour@yahoo.com
[5] Faculty of Engineering, Aurel Vlaicu University of Arad, Arad, Romania
corina.mnerie@gmail.com
[6] Vasile Goldis Western University of Arad, Arad, Romania
olariu_teodora@yahoo.com
[7] College of Information & Engineering, Wenzhou Medical University, Wenzhou, People's Republic Of China
sfq@wmu.edu.cn

Abstract. Classifier selection is a significant problem in machine learning to reduce the computational time and the number of ensemble members. Over the past decade, multiple classifier systems (MCS) have been actively exploited to enhance the classification accuracy. Finding a pertinent objective function for measuring the competence of base classifier is a critical issue to select the appropriate subset from a pool of classifiers. Along with the accuracy, diversity measures are designed as objective functions for ensemble selection. This current work proposed a new selection method based on accuracy and diversity in order to achieve better medical data classification performance. The classifiers correlation was calculated using Minimum Redundancy Maximum Relevance (mRMR) method based on relevance and diversity measures. Experiments were carried out on five data sets from UCI Machine Learning Repository and LudmilaKuncheva Collection. The experimental results proved the superiority of the proposed classifiers selection method.

Keywords: Medical data classification · Classifiers selection · Diversity measures · Relevance · Minimum Redundancy Maximum Relevance Method

V.E. Balas et al. (eds.), *Soft Computing Applications*, Advances in Intelligent Systems and Computing 633, DOI 10.1007/978-3-319-62521-8_32

1 Introduction

Automatic disease diagnosis is an emerging and evolutionary research domain that receives growing attention from both research community and medicine industry [1, 2]. Early detection and accurate diagnosis are the two significant key factors that affect the patients' health. In medical classification modalities, accuracy is very important [3], which led to the development of several medical data applications and intelligent classifiers [4–6].

The MCSs are very efficient technique as the classification performance of an ensemble often outperforms their base models [7, 8]. The three main issues in classifier ensemble are the diversity of individual classifiers, classifier selection, and the combination rule for the outputs of these classifiers [9]. Many researchers have demonstrated that ensembles can be better than its members if the base models make uncorrelated errors [10]. Thus, assuring the diversity in the generation phase of the ensemble is an imperative issue [11]. Generally, the diversity in classifier ensemble is generated by using different randomly selected training sets, such as bagging and boosting [12, 13] or using different feature sets for each base classifier, i.e. Random Subspaces (RSS) [14]. The RSS creates various component classifiers using different random subsets of features to train them. In each pass, such a selection is made and a subspace is fixed. Then, all samples are projected to this subspace, and a classifier is trained by using the projected training samples [15].

In machine learning, the classifier selection is a complex problem that aims to reducing the computational time as well as the number of ensemble members by choosing an appropriate subset within a pool of classifiers that maximizes the performance [16]. The classifier selection can be divided into two general categories, namely (i) static, where the ensemble members are defined during the training phase once and used for the classification of unseen patterns, and (ii) dynamic that are defined during the classification based on training performances and also various parameters of the actual sample to be classified [17]. The key issue in classifier ensemble selection is to find a pertinent objective function for measuring the competence of base classifier. Diversity measures are designed as objective functions for ensemble selection [18, 19]. The ideal situation is when individual classifiers are the most accurate where the probability of correct classification for objects recognition is the greatest, but are possibly different from each other at the same time. The individual classifiers must be both diverse and accurate [20, 21].

After selecting the best classifiers, it is significant to choose an effective aggregation rule that is related to how combine the individual classifiers' results. Typically, voting is a simple widely combination method whose effectiveness has been proven empirically [12, 22]. Given a new pattern x, each classifier votes for one specific class, and the final output class label is the one that receives the highest number of votes. Recently, various approaches have been developed for gene selection which achieves promising results. A Minimum Redundancy Maximum Relevance method (MRMR) that aims to maximize the relevancy of a gene subset, while minimizing the redundancy among the genes to find the optimal subset of multiple genes was developed [23, 24]. Peker *et al*. [25] proposed a novel technique for the Parkinson's disease (PD) diagnosis.

For effective attributes identification, the mRMR attribute selection procedure was applied. Afterward, the resulting attributes were considered as input to the complex-valued artificial neural network (CVANN).

Therefore, this work focused on employing multiple classifier systems (MCS) to improve the accuracy of medical image classification. Originally, an initial pool of classifiers was trained using the random subspace method to assure initial diversity among base classifiers. Afterward, classifier selection that uses modified version of the mRMR method was applied. The proposed Maximum Relevance Maximum diversity (MRMD) approach combining both mutual information and diversity measures to select the best N classifiers from the original classifiers set. The classifiers can be selected to be the most accurate keeping at the same time high level of diversity between them. The decisions provided by the selected classifiers are merged using majority voting rule to produce the final classification result.

The remaining sections organization is as follows. Section 2 provided an outline of the mRMR method. Section 3 described the classifier selection method in detail and presented the base classifier competence measurements. The proposed system presented in Sect. 4, followed by the results and discussion in Sect. 5. Finally, Sect. 6 addressed the conclusion of the proposed work.

2 The mRMR Method

The mutual information is a quantity that measures the mutual dependence between two random variables X and Y. In this case, information is thought of as the uncertainty reduction associated of a random variable due to knowledge of the other random variable. Formally, the mutual information of two discrete random variables X and Y can be defined as:

$$I(X;Y) = \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\frac{p(x;y)}{p(x)p(y)} \tag{1}$$

where, $p(x,y)$ is the joint probability distribution function of X and Y, while $p(x)$ and $p(y)$ are the marginal probability distribution functions of X and Y; respectively. In the continuous case, the summation is replaced by a double integral.

From the mutual information concept, the mRMR method proposed by Peng *et al.* [23] aimedto select the candidate genes with both the maximum relevance for the target concerned and the minimum redundancy among the genes themselves. Given g_i. which represents the gene i in S, where g_j represents the gene j, and the class label c, then the redundancy '*Red*' and the pertinence '*Per*' are defined as:

$$Per(i) = \frac{1}{|S|}\sum_{g_i\in S} I(g_i, \mathrm{c}) \tag{2}$$

$$Red(i) = \frac{1}{|S|}\sum_{g_i,g_j\in S} I\left(g_i, g_j\right) \tag{3}$$

Equations (2) and (3) are combined into the score function [24] to obtain the gene i with maximal relevance for the target c and minimal redundancy relative to the others genes in S. The score function is given by:

$$Score(i) = Per(i)/Red(i) \tag{4}$$

In practice, the important factors to achieve better classification performance in ensemble learning are the accuracy and diversity of base learners. However, balancing these two factors lacks theoretical guidance as many multi-objective optimization problems [18, 19]. The current work presented a classifier selection method based on both mutual information and diversity using mRMR method to control the balance between the accuracy and diversity of base learners.

3 Methodology

The proposed approach relies on the idea of forming an ensemble of classifiers by selecting them from a wider set according to a performance measure. It integrates the use of individual classifier accuracy and diversity measure as selection criteria. According to this measure, a subset of classifiers is selected that correlate strongly with the classification labels, which is normally called maximum relevance selection. On the other hand, the selected members must be divers from each other to enhance the classification accuracy. Maximum diversity is therefore utilized to address this problem and can also remove the redundant classifiers. Thus, the proposed method is named maximum relevance-maximum diversity (MRMD).

The main idea of the proposed algorithm MRMD takes a pool of classifiers $C = \{c_1, c_2, \ldots, c_N\}$ with size N as an input. At the beginning, each classifier's pertinence is evaluated on validation set by calculating the relevance using mutual information as per Eq. 3. Also, the diversity or the redundancy of each classifier is calculated as per Eq. 5 that given by:

$$Div(i) = \frac{1}{|C|} \sum\nolimits_{c_i, c_j \in S} div\left(c_i, c_j\right) \tag{5}$$

where,$div\left(c_i, c_j\right)$ is the diversity between the c_i, c_j classifiers in C.

The diversity measures include the Disagreement measure (DS), and the Weighted count of errors and correct results (WCEC). The DS represents the number of times that one of the classifiers was incorrect and the other correct [18]. For any two classifiers, the DS can be given by:

$$Dis(i,j) = \frac{N^{10} + N^{01}}{N^{00} + N^{01} + N^{10} + N^{11}} \tag{6}$$

where, N_{00} is the number of patterns that both classifiers wrongly classified; N_{11} stands for the number of patterns that both classifiers correctly classified; N_{10} is the number of

patterns classified correctly by classifier D_i but not by D_j; and N01 is the total of patterns classified correctly by classifier D_j but not by D_i.

Moreover, in the WCEC), the measured information is not only the incorrect, but also the correct results are considered with more emphasis placed on the situation where classifiers agree on either the correct or incorrect result. Simply a count for the occurrences of the different combinations can be given with suitable weight on the "both correct" which is a favorable situation, and "both same incorrect", which is an unfavorable situation [18].

$$\mathrm{WCEC}_{a,b} = N^{11} + \frac{1}{2}\left(N^{01} + N^{10}\right) - \mathrm{N}^{00}_{\mathrm{different}} - 5\mathrm{N}^{00}_{\mathrm{same}} \tag{7}$$

4 Proposed Approach

Researches were interested with developing several classification systems in the medical domain [26–29]. The current work proposed a novel selection method based on accuracy and diversity to realize superior medical data classification performance. Five different medical datasets from the UCI machine learning data repository [24] and Ludmila Kuncheva Collection [30] of real medical data are selected to evaluate the performance of the proposed method. Details about these datasets can be found in Table 1 that illustrated the selected datasets in the class of binary classification problems. In order to minimize the influence of variability in the training set, 5-fold cross validation is applied on the five datasets. Each dataset is partitioned into five subsets with similar sizes and distributions.

The base classifiers are generated by using MLP (Multi-Layer Perceptron) and functional trees (FT), which are taken from the Waikato Environment for Knowledge Analysis (Weka) version 3.4. The parameters used for each algorithm in this study are set at the default settings.

In the present study, the proposed approach performance is compared to three multiple classifier systems, namely (i) the mRMR that selects the best classifiers with the high score using mRMR method, (ii) the DIV (DS/WCEC) that defines the competence of the classifier subsets according to the diversity measures and select the most diver ensemble of classifiers, and (iii) the PER that based on selecting the best pertinent classifiers. All the learning and combination methods used in this study were conducted

Table 1. Datasets used for binary classification

Dataset		#Attribute	#Class	#Simple	#Positive	#Negative
Wisconsin Breast Cancer	WBC	9	2	699	485	241
Parkinson's disease	PK	22	2	195	48	147
Echocardiogram	ECG	12	2	74	24	50
Heart disease	HD	13	2	270	120	170
Respiratory	REP	17	2	85	45	40

using Java Language using WEKA tools and an initial pool of twenty classifier are trained using the Random Subspace method with the aim of assuring initial diversity among base classifiers. The selected classifiers size is fixed at 4 in all methods.

5 Results and Discussion

The classification accuracies of MRMD-DM, mRMR, MRMD-WCEC, DIV-DM, DIV-WCEC and PER using FT or MLP as base classifiers are illustrated in Table 2. The best result for each database is bolded.

Table 2. Accuracy using FT and MLP ensembles

Data set		MRMD-DM	mRMR	MRMD-WCEC	DIV-DM	DIV-WCEC	PER
WBC	FT	**95.85**	92.14	93.71	91.71	92.29	92.14
	MLP	**94.85**	91.14	93.57	91.43	91.17	90.14
PK	FT	**87.17**	84.61	86.67	85.64	83.59	86.18
	MLP	**82.56**	81.82	80.51	82.05	77.44	81.03
ECG	FT	92.00	90.67	**94.66**	90.67	92.00	90.67
	MLP	**96.00**	92.00	94.66	89.33	86.66	90.99
HD	FT	**81.11**	76.29	74.44	79.26	77.41	76.30
	MLP	81.11	**82.22**	73.33	80.37	79.63	**82.22**
REP	FT	**89.41**	87.06	88.24	88.24	87.06	87.06
	MLP	82.35	81.18	84.71	**85.53**	80.06	81.18

Table 2 depicts that proposed method based on FT provides better performance compared to the others method in the five datasets. The achieved accuracies are 94.66% with the ECG dataset using WCEC measure, 95.85% with the WBC dataset, 87.17% with the PK dataset, 81.11% with the HD dataset, and 89.41% with the REP using disagreement measure. In addition, the proposed method based on MLP has the highest classification accuracy with the four datasets, where 94.85%, 82.56%, 96%, and 84.71% accuracies are obtained with the WBC, PK, ECG using disagreement measure and RES using WCEC measure; respectively. In the HD dataset, the accuracy has significantly enhanced using the mRMR method.

Since, the sensitivity is defined as the number of true positive classifications divided by all positive classifications. Specificity represents the true negative rate and it is calculated by the division of true negative classifications by true negative and false positive classifications. Thus, it is important to observe the specificity and sensitivity of the medical diagnostic system as demonstrated in Tables 3 and 4.

Tables 3 and 4 depicts that the MRMD offered better results compared to the others ensemble learning. Based on FT ensembles, the proposed approach has the highest rate of sensitivity over the HD and REP datasets using WCEC, and also over ECG, WBC datasets using Disagreement measures. For the PK datasets, FT by DIV-WCEC outperforms the other systems. On the other hand, the MRMD using MLP classifier has the best sensitivity over four datasets: WBC, PK, REP and ECG. The MRMD system

achieves the highest specificity for PK, ECG, and REP datasets using FT ensembles and also for WBC, and ECG datasets using MLP ensembles. Consequently, it is recommended to apply the proposed classifier in several applications as included in [31–46].

Table 3. Sensitivity using FT and MLP ensembles

Data set		MRMD-DM	mRMR	MRMD-WCEC	DIV-DM	DIV-WCEC	PER
WBC	FT	**96.57**	87.82	95.56	86.22	88.14	87.82
	MLP	**92.22**	86.86	**92.22**	87.82	86.86	86.86
PK	FT	76.19	78.57	73.81	71.43	**83.33**	83.33
	MLP	**66.67**	57.14	64.29	52.38	54.76	57.14
ECG	FT	92.00	88.13	**93.99**	92.22	89.83	88.14
	MLP	**93.99**	89.63	**93.99**	88.14	83.05	88.83
HD	FT	**75.00**	64.17	60.00	70.00	64.17	64.17
	MLP	**70.83**	70.00	56.00	69.17	67.50	**70.83**
REP	FT	87.17	89.13	89.13	**91.30**	84.78	89.13
	MLP	82.61	82.61	**86.96**	84.78	82.61	82.61

Table 4. Specificity using FT and MLP ensembles

Data set		MRMD-DM	mRMR	MRMD-WCEC	DIV-DM	DIV-WCEC	PER
WBC	FT	95.35	95.62	95.56	**96.13**	95.62	95.62
	MLP	**96.00**	94.59	94.66	94.33	93.62	94.50
PK	FT	**90.20**	86.27	**90.20**	89.54	83.66	88.24
	MLP	86.93	87.58	84.97	**90.20**	83.66	87.58
ECG	FT	92.00	**100**	**100**	81.25	**100**	**100**
	MLP	**100**	**100**	96	93.75	**100**	**100**
HD	FT	86.00	86.00	86.00	86.67	**88.00**	86.00
	MLP	90.00	**91.33**	86.67	89.33	89.33	**91.33**
REP	FT	**91.30**	84.62	87.18	84.62	85.74	84.62
	MLP	79.49	79.49	82.05	82.05	**92.31**	79.49

6 Conclusion

Classifier selection is an imperative problem in machine learning. In order to select appropriate subset from a pool of classifiers, it is necessary to find a pertinent objective function for measuring the competence of base classifier. The experimental results established that the proposed novel scheme of classifier selection based on mutual information and diversity achieved better medical data classification performance. Classifiers correlation was ensured using mRMR method and diversity measures. Experiments were carried out on five data sets from UCI Machine Learning Repository and Ludmila Kuncheva Collection. The experimental results are encouraging and validate the effectiveness of the proposed classifiers selection method.

References

1. Lahmiri, S., Boukadoum, M.: Alzheimer's disease detection in brain magnetic resonance images using multiscale fractal analysis. ISRN Radiol. **2013**, 627303 (2013)
2. Zhang, S., Cohen, I., Goldszmidt, M., Symons, J., Fox, A.: Ensembles of models for automated diagnosis of system performance problems. In: 2005 International Conference on Dependable Systems and Networks (DSN 2005), pp. 644–653). IEEE (2005)
3. Onan, A.: A fuzzy-rough nearest neighbor classifier combined with consistency-based subset evaluation and instance selection for automated diagnosis of breast cancer. Expert Syst. Appl. **42**, 6844–6852 (2015)
4. Abdel-Aal, R.-E.: Improved classification of medical data using abductive network committees trained on different feature subsets. Comput. Methods Programs Biomed. **80**, 141–153 (2005)
5. Sekar, B.-D., Dong, M.-C., Shi, J., Hu, X.-Y.: Fused hierarchical neural networks for cardiovascular disease diagnosis. IEEE Sens. J. **12**, 644–650 (2012)
6. Cheriguene, S., Azizi, N., Zemmal, N., Dey, N., Djellali, H., Farah, N.: Optimized tumor breast cancer classification using combining random subspace and static classifiers selection paradigms. In: Applications of Intelligent Optimization in Biology and Medicine (2016)
7. Kuncheva, L.-I.: Combining Pattern Classifiers: Methods and Algorithms. Wiley, Hoboken (2004)
8. Kittler, J., Roli, F.: Multiple Classifier Systems. Springer, Heidelberg (2010)
9. Cruz, R.-M., Sabourin, R., Cavalcanti, G.-D., Ren, T.-I.: META-DES: a dynamic ensemble selection framework using meta-learning. Pattern Recognit. **48**(5), 1925–1935 (2015)
10. Zhang, L., Zhou, W.-D., Li, F.-Z.: Kernel sparse representation-based classifier ensemble for face recognition. Multimed. Tools Appl. **74**, 123–137 (2013)
11. Kuncheva, L-I., That elusive diversity in classifier ensembles. In: Lecture Notes in Computer Science, pp. 1126–1138 (2003)
12. Breiman, L.: Bagging predictors. Mach. Learn. **24**, 123–140 (1996)
13. Freund, Y., Schapire, R.: Experiments with a new boosting algorithm. In: Proceedings 13th International Conference on Machine Learning, pp. 148–156 (1996)
14. Ho, T.: The random subspace method for constructing decision forests. IEEE Trans. Pattern Anal. Mach. Intell. **20**, 832–844 (1998)
15. Wang, G., Zhang, Z., Sun, J., Yang, S., Larson, C.-A.: POS-RS: a random subspace method for sentiment classification based on part-of-speech analysis. Inf. Process. Manag. **51**, 458–479 (2015)
16. Şen, M.-U., Erdogan, H.: Linear classifier combination and selection using group sparse regularization and hinge loss. Pattern Recognit. Lett. **34**, 265–274 (2013)
17. Britto, A.-S., Sabourin, R., Oliveira, L.-S.: Dynamic selection of classifiers - a comprehensive review. Pattern Recognit. **47**, 3665–3680 (2014)
18. Aksela, M., Laaksonen, J.: Using diversity of errors for selecting members of a committee classifier. Pattern Recognit. **39**, 608–623 (2006)
19. Visentini, I., Snidaro, L., Foresti, G.L.: Diversity-aware classifier ensemble selection via f-score. Inf. Fusion. **28**, 24–43 (2016)
20. Chiu, C.-Y., Verma, B.: Effect of varying hidden neurons and data size on clusters, layers, diversity and accuracy in neural ensemble classifier. 2013 IEEE 16th International Conference on Computational Science and Engineering, pp. 455–459 (2013)
21. Bi, Y.: The impact of diversity on the accuracy of evidential classifier ensembles. Int. J. Approx. Reason. **53**, 584–607 (2012)

22. Lam, L., Suen, C.-Y.: A theoretical analysis of the application of majority voting to pattern recognition. IEEE Trans. Syst. Man Cybern. **27**, 418–420 (1994)
23. Peng, H., Long, F., Ding, C.: Feature selection based on mutual information. IEEE Trans. Pattern Anal. Mach. Intell. **27**, 1226–1238 (2005)
24. Learned-Miller, E.-G.: Entropy and mutual information. Technical report, vol. 4, pp. 1–4. University of Massachusetts Amherst (2013)
25. Peker, M., Şen, B., Delen, D.: Computer-aided diagnosis of parkinson's disease using complex-valued neural networks and mRMR feature selection algorithm. J. Healthc. Eng. **6** (3), 281–302 (2015)
26. Ghosh, A., Sarkar, A., Ashour, A.-S., Balas-Timar, D., Dey, N., Balas, V.E.: Grid color moment features in glaucoma classification. Int. J. Adv. Comput. Sci. Appl. **6**(9), 1–4 (2015)
27. Nath, S., Kar, J., Chakraborty, S., Mishra, G., Dey, N.: A survey of image classification methods and techniques. In: International Conference on Control, Instrumentation, Communication and Computational Technologies (2014)
28. Nawel, Z., Azizi, N., Sellami, M., Dey, N.: Automated classification of mammographic abnormalities using transductive semi supervised learning algorithm. In: Mediterranean Conference on Information & Communication Technologies 015, Saïdia, Morocco, pp. 7–9, May 2015
29. Kotyk, T., Ashour, A-S., Chakraborty, S., Dey, N., Balas, V-E.: Apoptosis analysis in classification paradigm: a neural network based approach. In: Healthy World Conference 2015 - A Healthy World for a Happy Life, Kakinada (AP) India, pp. 17–22 (2015)
30. Kuncheva, L.: Ludmilakuncheva collection (2004). http://pages.bangor.ac.uk/∼mas00a/activities/real_data.html
31. Beagum, S., Dey, N., Ashour, A.S., Sifaki-Pistolla, D., Balas, V.E.: Nonparametric de-noising filter optimization using structure-based microscopic image classification. Microsc. Res. Tech. **80**, 419–429 (2016)
32. Anusha, M., Sathiaseelan, J.G.R.: An empirical study on multi-objective genetic algorithms using clustering techniques. Int. J. Adv. Intell. Paradig. **8**(3), 343–354 (2016)
33. Anter, A.M., El Souod, M.A., Azar, A.T., Hassanien, A.E.: A hybrid approach to diagnosis of hepatic tumors in computed tomography images. Int. J. Rough Sets Data Anal. (IJRSDA) **1**(2), 31–48 (2014)
34. Beldjehem, M.: A unified granular fuzzy-neuro min-max relational framework for medical diagnosis. Int. J. Adv. Intell. Paradig. **3**(2), 122–144 (2011)
35. Kapoor, N., Ohri, J.: GA and PSO optimised SVM controller for manipulator. Int. J. Comput. Syst. Eng. **2**(3), 121–130 (2016)
36. Singh, V.P., Srivastava, S., Srivastava, R.: An efficient image retrieval based on fusion of fast features and query image classification. Int. J. Rough Sets Data Anal. (IJRSDA) **4**(1), 19–37 (2017)
37. Ahmed, S.S., Dey, N., Ashour, A.S., Sifaki-Pistolla, D., Bălas-Timar, D., Balas, V.E., Tavares, J.M.R.: Effect of fuzzy partitioning in Crohn's disease classification a neuro-fuzzy-based approach. Med. Biol. Eng. Comput. **55**, 1–15 (2016)
38. Muralidharan, V., Sugumaran, V.: Fault diagnosis of centrifugal pump using wavelet features–fuzzy-based approach. Int. J. Comput. Syst. Eng. **1**(3), 175–183 (2013)
39. Sambyal, N., Abrol, P.: Feature based text extraction system using connected component method. Int. J. Synth. Emot. (IJSE) **7**(1), 41–57 (2016)
40. Fouad, K.M., Hassan, B.M., Hassan, M.F.: User authentication based on dynamic keystroke recognition. Int. J. Ambient Comput. Intell. (IJACI) **7**(2), 1–32 (2016)
41. Kishor, D.R., Venkateswarlu, N.B.: A novel hybridization of expectation-maximization and k-means algorithms for better clustering performance. Int. J. Ambient Comput. Intell. (IJACI) **7**(2), 47–74 (2016)

42. Trabelsi, I., Bouhlel, M.S.: Comparison of Several Acoustic Modeling Techniques for Speech Emotion Recognition. Int. J. Synth. Emot. (IJSE) **7**(1), 58–68 (2016)
43. Virmani, J., Dey, N.; Kumar, V.: PCA-PNN and PCA-SVM based CAD systems for breast density classification. In : Applications of Intelligent Optimization in Biology and Medicine, pp. 159–180 (2016)
44. Kausar, N., Palaniappan, S., Samir, B.B., Abdullah, A., Dey, N.: Systematic analysis of applied data mining based optimization algorithms in clinical attribute extraction and classification for diagnosis of cardiac patients. In: Applications of Intelligent Optimization in Biology and Medicine, pp. 217–231 (2016)
45. AlShahrani, A.M., Al-Abadi, M.A., Al-Malki, A.S., Ashour, A.S., Dey, N.: Automated system for crops recognition and classification. In: Applied Video Processing in Surveillance and Monitoring Systems, pp. 54–69 (2016)
46. Saba, L., Dey, N., Ashour, A.S., Samanta, S., Nath, S.S., Chakraborty, S., Sanches, J., Kumar, D., Marinho, R., Suri, J.S.: Automated stratification of liver disease in ultrasound: An online accurate feature classification paradigm. Comput. Methods Programs Biomed. **130**, 118–134 (2016)

Personal Health Record Management System Using Hadoop Framework: An Application for Smarter Health Care

Bidyut Biman Sarkar[1(✉)], Swagata Paul[1], Barna Cornel[2], Noemi Rohatinovici[2], and Nabendu Chaki[3]

[1] Techno India College of Technology, Rajarhat, Newtown, Kolkata, India
bidyutbiman@gmail.com, swagatapaul@hotmail.com
[2] Aurel Vlaicu University of Arad, B-Dul Revolutiei 77, 310130 Arad, Romania
barna.cornel@gmail.com, noemi_c91@yahoo.com
[3] University of Calcutta, Kolkata, India
nchaki@cucse.org

Abstract. Health care informatics now can map a rural health center to 'n' number of city hospitals is no more a research agenda. The remote diagnostics, alert services, and finding ambulances can be talented over mobile devices. However, to support such services massive data sets are to be processed in a short span of time. It is possible only when un-structured data sets can be processed in parallel with some degree of fault tolerance. The Hadoop distributed framework and Map/Reduce engine can process such hundreds of terabytes of data at a low cost. An electronic health record (EHR) is a real-time, patient-centered archive. A personal health record (PHR) is maintained by patient securely. In this work, a web enabled distributed EHR and PHR management framework using Hadoop HBase, is proposed. It will assist patient's data exploration for advanced data analytics on demand.

Keywords: Big data · Hadoop · HBase · PHR · MapReduce · Data analytics

1 Introduction

There are e service disciplines in medical sciences in practice [1]. Some of them are: medical care, surgical, consultation, diagnostic X-Ray, restorative, MRI Scan, Speech Therapy, etc. In recent years, it is a challenge to extend cost effective remote health care treatment and monitoring services like Google and Microsoft [2, 3]. Disease pattern like diabetes, etc. can be detected for early treatment as prospective benefits. It can be used to predict adverse outcomes in case of epidemics, can prevent health care frauds, manage crises in emergencies, and can be used for disease surveillance for the benefit of the inhabitants. Sizeable data volumes can be transformed into actionable information more quickly and efficiently [4]. The electronic health data sets are multifaceted and hefty. It is difficult to manage with conventional methods, software's, hardwires, and regular data management tools [5]. A new paradigm shift is introduced to process huge volume of data in a short span time and is named as 'Big data' [6]. The Big data

V.E. Balas et al. (eds.), *Soft Computing Applications*, Advances in Intelligent Systems and Computing 633, DOI 10.1007/978-3-319-62521-8_33

characteristics are defined as variety, velocity and, veracity [7, 8]. The raw data sets needs to be pooled, processed, transformed, and integrated with web services from the view point of big data analytics [9]. There are quite a few architectures, platforms, and tools are already available and some are in incubation to exploit and maximize the potential of big data analytics in healthcare. In this work an attempt is made to build a holistic reliable health service framework of EHR and PHR management in a distributed domain using HBASE [10]. In Sect. 2 some literature review on the PHR creation, maintenance and Big-Data processing is done. In Sect. 3 each of the state of the art used to build the PHR is explained. In Sect. 4 cluster architectural framework for PHR management is described along with its configuration and process flow methods. In Sect. 5 data analytics with some result sets are presented in the form of example. In Sect. 6 we concluded the computation process with a view to support the society for a better civilization.

2 Literature Review

The Electronic Record Development and Implementation Project (ERDIP) of Manchester Business School convert the dream of health record mechanization to a reality in early 2000 [11]. The concept of Health space was introduced in 2007. It is an internet-accessible PHR. The main idea is self updation and access of personal data by oneself for self-management [12]. The independent sources in 2009 reported on internet usage behaviour of US college students. The study revels nationwide 26.1% depressive symptoms among college students those who use the Internet. This study is suggesting the need for compulsory creation of PHR and business analytics are necessary for predicting such situations and prevent debacles [13]. This research is advancing the process of PHR creation for future healthcare applications for monitoring aged people in the smart home environment [14]. Due to the existence of high volume of human health informatics data, this paper introduces Big data tools to analyse Health Informatics at multiple levels by using different human-scales; like biology, clinical-scale, epidemic-scale [15]. Large size distributed data handling at a faster pace is a challenge at this instance. Web services and social media turn out together a striking amount of daily data, reaching the scale of petabytes [16]. The phrase Big Data is used to refer to a group of large datasets that may not be routed through traditional database management tools to process. The challenge involved in dealing with Big Data is storage, and subsequent analysis in no time, which is empowering its heterogeneity, multi-dimensionality, scale, complexity, volume and speed of data generated in the realm of health-care [17]. Some of the Big-Data processing challenges are dependency, integrity, cluster load balancing, and task scheduling [18]. Hadoop is a tool for Big-Data analytics, understanding Hadoop architecture, its design, implementation, scalability and fault tolerance capabilities and transformation of spatial SQL queries into its Map-reduce equivalent is discussed. Hadoop performance, operations into Map-Reduce and their short comings are discussed in [19–21].

3 Hadoop Component Used

Hadoop is a platform providing cloud computing services consisting of the components presented in Table 1. Partitioning of data and computations in parallel and across thousands of hosts can be done close to their data location. A Hadoop cluster scales up computation capacity, storage capacity and I/O bandwidth by simply adding commodity servers. For example, Hadoop clusters at Yahoo span 40,000 servers, and store 40 petabytes of application data, with the largest cluster being 4000 servers [22, 23].

Table 1. Hadoop components

Components	Descriptions
HDFS	Hadoop Distributed File System (HDFS) enables the cluster storage. It divides the data into smaller parts and distributes it across the various nodes
HBase	A non-relational, NoSQL distributed database that runs on top of HDFS. It is a column-oriented database management system
Zookeeper	ZooKeeper is a centralized service for maintaining configuration information, naming and providing distributed synchronization
MapReduce	MapReduce provides the way of distribution of sub-tasks and the gathering of outputs
Pig	A compiler called Pig Latin saves MapReduce programs writing and used for manipulating data stored in HDFS, which includes data extractions, transformations and loading

4 Cluster Architecture Framework

In our work we have developed a cloud for distributed processing that can process a huge amount of data. We used Java based open source big data processing tools that can retrieve and analyze data efficiently. The system is shown in Fig. 1.

The components of this cluster architecture are: (I). *Client and Client Application*: A Client is an end user with any one the following roles– patient, hospital operator. A Client Application communicates with Application Server. (II). *Application Server*: Runs the services for client application. Temporarily stores the new records in ZoneDb Server. After a certain amount of data volume or after a certain amount of time the new data is send to a specific NHZ server based on the patient location. If a patient is relocated from location x to y, all the existing data is copied from NHZx to NHZy server. Application layer API built for Client are: AddNewCitizen(), SetCitizenLocation(), AddNewCitizenData(), AddBulkCitizenData(), AddFactSheets(), QueryCitizenData(), AnalyzeData(). (III). *ZoneDb Server*: Keep info of all patients' location information. Temporarily stores all the new records before it goes to HBase. (IV). *NHZ Servers*: All NHZ servers communicated with application server. All works as master node for HDFS and HBase. They store NameNode data for HDFS, Hmaster data for HBase. They run Zookeeper to coordinate distributed synchronization of data stored in RD servers. There are n number of NHZ servers provide services to application server. (V). *RD servers*: Each NHZ server is connected to RD server set. Each set contains m

number of region server called Rgn. One NHZ server stores data in m Rgn Server. HBase Memory store (MStore) and HBase file (HFile) works as distributed database using HDFS to store data permanently. There are multiple HFile and MStore in each Rgn server.

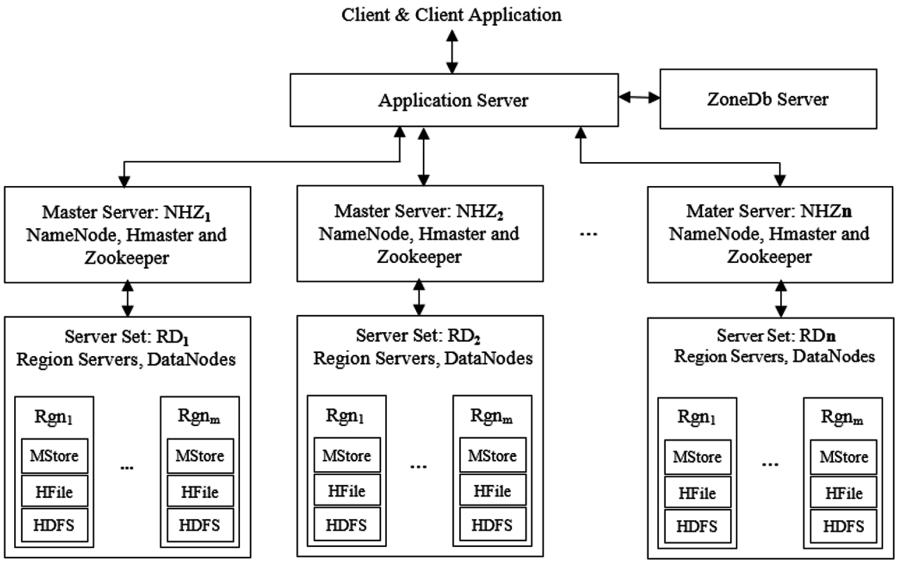


Fig. 1. Cluster architecture

4.1 Process and Data Flow Methods

In this work we developed Application Program Interface (API) for data access from lower layer to client [Fig. 1]. The APIs runs on the application server and provides services to upper layer i.e. client application. The APIs connects to the lower layer Master Server NHZi servers to read and write data. The master server connects to its own server set RDi. NHZi is dedicated for one RDi server set, where each RDi server set has m number of Rgnj server. One NHZi server and its RDi set represents a working unit for a geographical area of hospice. Each API accepts XML input, process it and returns XML output. The used API functions are namely: (I). *AddNewCitizen()*: Generate new global id called CitizenID for the new citizen. Send citizens personal data to corresponding NZH server to be stored in HBase. Once the data is stored the NZH sends an acknowledgement to application server, which is recorded in ZoneDb Table. (II). *SetCitizenLocation()*: Save the record (CitizenID, Location) in ZoneDb Table. (III). *AddNewCitizenData()*: Send citizens healthcare data to corresponding NZH server and its RD server set to be stored in HBase. Once the data is stored the NZH sends an acknowledgement to application server, which is recorded in ZoneDb Table. (IV). *AddBulkCitizenData()*: Send a large set of citizen's healthcare record to NHZ servers. (V). *AddFactSheets()*: Add globally known and established facts on various diseases.

This data is currently soured in ZoneDb server. (VI). *QueryCitizenData()*: Get the location info from ZoneDb. If record is not there, it broadcast to all NHZ servers to get the same. Only one NHZ servers reply to that broadcast. Application server then send the query to the specific NZH server. The NHZ server and its RD Server set work together using HBase and HDFS to get the result quickly and return the result to application server. (VII). *AnalyzeData()*: Based on the XML input parameter in the AnalyzeData query, one or multiple NZH servers works in parallel along with their RD server set to get the result summery. We use map-reduce techniques, AggregationClient methods to reduce it to result set.

4.2 System Configuration

Table 2 shows the overall cluster configuration of Fig. 1 with 90 TB data space, 360 GB primary memory, 180 CPU cores connected with 1 Gbps Ethernet link. Table 3 presents the configuration of Hadoop Filesystem which is used by HBase to store the data. In this work we used 32 MB, 64 MB and 128 MB HDFS block size on different set of servers to study the behavior and performance of the system. The HBase is the non-relational NoSQL distributed database we used in this work. The configuration is presented in the Table 4. This database stores the data in Hadoop File system. HBase uses multiple group of memory store (MStore) and HFile to make the system more stable and faster. In our configuration we have used 50 such groups called regions. The benefits of using HBase and HDFS is given in Table 5.

Table 2. The cluster

Items	Details
Total number of nodes	100, including NHZ servers and RD servers
Node H/W configuration	4 GB DDR3 RAM, 1024 GB HDD, Intel Core i3
Node OS configuration	Ubuntu 14 64bit, Ext3 Filesystem
LAN	Gigabyte Ethernet, 48port switch
Total NHZ server	10, each represents master server for a geographical area/hospital
Total Rgn servers	90, 9 servers under one NHZ Server, each set of 10 node represents data nodes for a geographical area/hospital
Total data space on disk (Rgn Servers)	90 × 1 TB = 90 TB
Total primary memory (Rgn Servers)	90 × 4 GB = 360 GB
Total physical core (Rgn Servers)	90 × 2 = 180 Cores, Intel core i3 is dual-core processor
Total space on master servers	Disk 10 TB, primary memory 40 GB
Node connection	The set of RD servers and the corresponding NHZ server represents a single distributed system in a geographical zone or in a hospital

Table 3. The HDFS

Items	Details
Hadoop	Version 2.7.1 on Ubuntu 14 46bit OS
Optimized for	Streaming access of large files
HDFS read write and append	Supported
The NameNode of HDFS, YARN	Runs on NHZ nodes
The DataNode of HDFS, NodeManager	Runs on Rgn nodes
DataNode Block Size	32 MB, 64 MB and 128 MB
Java virtual machine	64Bit on Ubuntu 14 64Bit
Hadoop	Version 2.7.1 on Ubuntu 14 64bit OS
Optimized for	Streaming access of large files

Table 4. The HBase

Items	Details
HBase on Hadoop	Version 1.1.5 on Hadoop 2.7.1
Optimized for	Low latency access to small amounts of data from within a large data set. We can access single rows quickly from a billion row table
Data model	Flexible data model to work with and data is indexed by the row key. Fast scans across tables
Scalable	Scale in terms of writes as well as total volume of data
MStore of HBase	Runs on Rgn nodes
HFile of HBase	Runs on Rgn nodes
No of regions on each Rgn servers	50, each regions contains MStore & HFile

Table 5. List of HDFS and HBase benefits

Items	Details
Low cost	The open-source framework is free and uses commodity hardware to store large quantities of data
Computing power	Rapidly it can process very large volumes of data. More the computing nodes more the processing power
Scalability	System can easily grow by adding more nodes with minimum system administration
Portability	HDFS & HBase has its portability between various Hadoop distributions, which helps in minimizing vendor specific locking constraints
Storage flexibility	Unstructured data like text, images and videos can be stored directly as much data as you need without any prior formatting and decide later how to use it
Inherent data protection and self-healing capabilities	If a node doesn't respond, jobs are by design redirected to other replicated nodes to make sure the distributed computing does not fail. It implies hardware independence i.e. Data and application processing are protected against hardware failure

5 Example Analysis

We collected the datasets from the websites shown in Table 6. We used two main APIs AddBulkCitizenData() and AddFactSheets() to load bulk data from these data sources. Once the data is stored in RD server set as described in cluster architecture framework. We used AnalyzeData() function to test "Detection of Heart Failure" and "High Blood Pressure Statistics" with XML input parameter on a total data volume of 5 TB, 64 MB HDFS block size, 3 NHZ server and their RD server set. The result analysis is shown in Table 7.

Table 6. The data source

Dataset	Description
i2b2 informatics https://www.i2b2.org [24]	Clinical notes used for NLP challenges
TX health data www.dshs.state.tx.us [25]	Dataset for cardiovascular disease
VHA medical datasets http://www.va.gov [26]	Patient care data from VHA facilities
Inpatient sample www.hcup-us.ahrq.gov [27]	Hospitals discharge data
CA patient data http://www.oshpd.ca.gov [28]	Demographic diagnostics
MIMIC II database mimic.physionet.org [29]	ICU clinical measurements, lab results
Medicare data https://data.medicare.gov [30]	Hospital, nursing home health data

Table 7. Data diagnostics

Item	Result: detection of heart failure
Heart failure count	5.12 million
New cases in each year	0.52 million
Life risk after 40 years old	19.45%
5 years mortality rate	47.66%
HBase aggregation client	Methods: avg, max, sum, median, rowCount, std
Map-Reduce	Yes
Time taken to calculate (5 TB)	**5.56 s**
Item	**Result: HIGH BLOOD PRESSURE**
High blood pressure count	70 million
Condition under control count	52%
Primary or contributing cause of death	1,000 deaths each day in 2013
HBase aggregation client	Methods: avg, max, sum, median, rowCount.
Map-Reduce	Yes
Time taken to calculate (5 TB)	**8.4 s**

There are many more XML input parameters for different APIs. We tested all APIs with their proper parameters and got result in stipulated amount of time.

6 Conclusion

The data volume, velocity and variety are increasing every day. The standalone application with a single database server will be no more efficient and effective in near future. Even if the service provider grows vertically, it will not be cost effective. The solution we presented to store, analyze and retrieve a huge set of Electronic Healthcare Record can grow horizontally and therefore cost effective and more powerful for data analysis. The flexible architecture we presented can be expanded by adding more NHZ and RD Server sets. Therefore healthcare organization can setup NHZ and RD Server set and connect to our Application server to get all APIs for data analytics. The current framework has not been compared with other distributed processing frameworks. In our future work, we intend to further reduce the processing complexity to improve the query performance for effective data analytics.

References

1. Yang, X.: UI Design for PAIN-OUT Online System. Mälardalen University, Diss (2014)
2. Kitchens, B., Harle, C.A., Li, S.: Quality of health-related online search results. Decis. Support Syst. **57**(1), 444–453 (2014). Elsevier
3. Ozdemir, Z.D., Bandyopadhyay, S., Barron, J.M.: Adoption of electronic and personal health records: an economic analysis. In: Proceedings of the AMCIS (2009). http://aisel.aisnet.org/amcis2009/755
4. Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Byers, A.H.: Big Data: The Next Frontier for Innovation, Competition, and Productivity (2011)
5. Yadav, V., Verma, M., Kaushik, V.D.: Big data analytics for health systems. In: 2015 International Conference on Green Computing and Internet of Things (ICGCIoT). IEEE (2015)
6. Sheriff, C.I., Naqishbandi, T., Geetha, A.: Healthcare informatics and analytics framework. In: 2015 International Conference on Computer Communication and Informatics (ICCCI). IEEE (2015)
7. Connolly, S., Wooledge, S., Aster, T.: Harnessing The Value of Big Data Analytics. Teradata, Hortonworks, Santa Clara (2013)
8. Courtney, M.: Puzzling out big data [information technology analytics]. Eng. Technol. **7**(12), 56–60 (2013)
9. Raghupathi, W., Kesh, S.: Interoperable electronic health records design: towards a service-oriented architecture. e-Service J. **5**, 39–57 (2007)
10. Raghupathi, W., Raghupathi, V.: Big data analytics in healthcare: promise and potential. Health Inf. Sci. Syst. **2**(1), 3 (2014)
11. Patel, V.L., Rogers, R., Haux, R.: Patient Empowerment and the Electronic Health Record. In: Proceedings of the MEDINFO, IOS Press, pp. 663–665 (2001). ISBN 1 58603 1945
12. Greenhalgh, T.: Adoption, non adoption, and abandonment of a personal electronic health record: case study of Health Space. Br. Med. J. (2010). doi:10.1136/bmj.c5814, http://bit.ly/boP
13. Katikalapudi, R., Chellappan, S., Montgomery, F., Wunsch, D., Lutzen, K.: Associating internet usage with depressive behavior among college students. IEEE Technol. Soc. Mag. **31**(4), 73–80 (2012). doi:10.1109/MTS.2012.2225462

14. Roy, N., Pallapa, G., Das, S.K.: A middleware framework for ambiguous context mediation in smart healthcare Application. In: IEEE conference on Wireless and Mobile Computing, Networking and Communications, pp. 72 (2007). ISBN: 0-7695- 2889-9
15. Herland, M., Khoshgoftaar, T.M., Wald, R.: A review of data mining using big data in health informatics. J. Big Data **1**(2) (2014). doi:10.1186/2196-1115-1-2. SpringerOpen Journal
16. Jim, H.: Broad data: exploring the emerging web of data. Big Data **1**(1), 18–20 (2013)
17. White, T.: Hadoop The Definitive Guide, 3rd edn. O'Reilly Media Inc, Sebastopol (2012). Xie, J., Tian, Y., Yin, S., Zhang, J., Ruan, X., Qin, X.: Adaptive preshuffling in hadoop clusters. Procedia Comput. Sci. **18**, 2458 –2467 (2013). International Conference on Computational Science
18. DeWitt, D.J., Paulson, E., Robinson, E., Naughton, J., Royalty, J., Shankar, S., Krioukov, A.: Clustera: an integrated computation and data management system. Proc. VLDB Endow. **1**(1), 28–41 (2008)
19. Karlff, H., Suriet, S., et al.: A model of computation for MapReduce. In: Proceedings of the 20th SODA (2010)
20. Deanand, J., Ghemawat, S.: MapReduce: simplified data processing on large clusters. In: Proceedings of the 6th Conference on Symposium on Operating Systems Design & Implementation, SanFrancisco, vol. 6, p. 10 (2004)
21. Zhang, S., Han, J., Liu, Z., et.al.: Parallelizing spatial join with mapreduce on clusters. In: Proceedings of CLUSTER, pp. 1–8 (2009)
22. Khare, R., Cutting, D., Sitaker, K., Rifkin, A.: Nutch: a flexible and scalable open-source web search engine. Or. State Univ. **1**, 32 (2004)
23. Jurik, B.A., Blekinge, A.A., Ferneke-Nielsen, R.B., Møldrup-Dalum, P.: Bridging the gap between real world repositories and scalable preservation environments. Int. J. Digit. Libr. **16**(3–4), 267–282 (2015)
24. Murphy, S.N., Weber, G., Mendis, M., Gainer, V., Chueh, H.C., Churchill, S., Kohane, I.: Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J. Am. Med. Inform. Assoc. **17**(2), 124–130 (2010)
25. Code, T.A., Rules, L.: Department Of State Health Services Regulatory Licensing Unit Facility Licensing Group (2010)
26. Department of Veterans Affairs: National center for veterans analysis and statistics (2012)
27. Overview of the Nationwide Inpatient Sample (NIS): Healthcare Cost and Utilization Project, Agency for Health Care Policy and Research (2007)
28. Weech-Maldonado, R., Elliott, M.N., Pradhan, R., Schiller, C., Dreachslin, J., Hays, R.D.: Moving towards culturally competent health systems: organizational and market factors. Soc. Sci. Med. **75**(5), 815–822 (2012)
29. Saeed, M., Lieu, C., Raber, G., Mark, R.G.: MIMIC II: a massive temporal ICU patient database to support research in intelligent patient monitoring. In: Computers in Cardiology 2002, pp. 641–644. IEEE, September 2002
30. Brennan, N., Oelschlaeger, A., Cox, C., Tavenner, M.: Leveraging the big-data revolution: CMS is expanding capabilities to spur health system transformation. Health Aff. **33**(7), 1195–1202 (2014)

Gene-Disease-Food Relation Extraction from Biomedical Database

Wahiba Ben Abdessalem Karaa[1,2], Monia Mannai[2,3], Nilanjan Dey[4(✉)], Amira S. Ashour[5], and Iustin Olariu[6]

[1] College of Computer and IT, Taif University, Taif, Saudi Arabia
wahiba.bak@gmail.com
[2] RIADI-GDL Laboratory, ENSI, Tunis, Tunisia
[3] High Institute of Management, Tunis, Tunisia
moniamannai@gmail.com
[4] Department of Information Technology, Techno India College of Technology, Kolkata, India
neelanjandey@gmail.com
[5] Department of Electronics and Electrical Communications Engineering, Faculty of Engineering, Tanta University, Tanta, Egypt
amirasashour@yahoo.com
[6] Faculty of Dentistry, Vasile Goldis Western University of Arad, Arad, Romania
iustin_olariu@yahoo.com

Abstract. Through the past years, an incredible increase in the biomedical data amount presented on the web is enlarged due to the increased data volume in the medical and biological domains. Hence, the search for documents and information on the internet became increasingly complicated. In the current work, a new approach for information extraction using the Natural Language Processing (NLP) tools and ontology was proposed. It described a system to extract relations between the concepts from biomedical texts using morphological analysis and information extraction techniques. In the first step, the system segmented the input text into sentences. Each sentence is then segmented into words that were tagged with part-of-speech labels and concept classes (food, drug, and gene). A set of relation extraction rules (regular expression patterns) are applied on the annotated sentences. If a pattern matches, the concepts and relations are extracted. The system has been tested on a set of 700 MEDLINE abstracts. For performance evaluation, the precision, recall and F-score were calculated. The proposed approach created by information retrieval from MEDLINE to gather a set of abstracts related to a given domain. Then, these texts were annotated using an automaton and ontology via recognizing interesting concepts for morphological analysis. After the annotation step, some rules were summarizing in an automaton that help gene-disease-food relationships discovery. This work proposed an approach for identifying relations between medical concepts using NLP tools. An evaluation experiment reported good effectiveness results.

Keywords: Biomedical information · Semantic reasoning techniques · Information extraction · MEDLINE · Natural language processing (NLP) · Text mining

V.E. Balas et al. (eds.), *Soft Computing Applications*, Advances in Intelligent Systems and Computing 633, DOI 10.1007/978-3-319-62521-8_34

1 Introduction

Text is the main standard for information exchange among specialists. In the biomedical domain, knowledge is available to researchers in unstructured published scientific text. However, a lot of relevant information that is otherwise hidden can be inferred from these texts using a range of techniques. Therefore, text analysis is one of the text mining research field topics that can be used for the automation of the information extraction from texts.

Generally, text mining refers to the process of extracting and analysing unstructured text, in order to discover interesting information and knowledge [1]. A lot of researchers have applied text mining in the biomedical field to find new relationships between concepts in the medical domain [2, 3]. Scientists published their articles in various media such as MEDLINE; nevertheless is difficult to extract embedded knowledge in these articles using basic models [4]. Consequently, a variety of methods has been proposed to exploit the new relationships between concepts [5] in MEDLINE.

Generally, the pre-processing procedure is a common step for all text mining treatments due to the nature of the textual data that are unstructured and mostly available in natural language form. The text pre-processing goal is to optimize the performance of the next step. It is performed for each abstract and consists of selecting and cleaning as a first step of the approach.

In this paper, a new method is described to extract relationships between three associations, namely disease, gene and food from MEDLINE records. In order to clarify the importance of the connection between these concepts, the text mining techniques are combined with ontology in order to annotate the texts. A morphological analysis to determine each concept role is accomplished. It is considered as a formal annotation, where it focuses on how the term was constructed and its role in the sentence. The morphological analysis helps to detect and to understand the relation between the concepts. However, it is a complex task, because this kind of relations is always described and explained in different ways with natural language. Hence, morphological analysis greatly facilitates the mission by normalizing words that compose sentences. Each sentence consists of a subject, verb, and complement.

The rest of this paper is organized as follows. Section 2 discussed related literature and previous works. Then, the methodology is presented in Sect. 3. In Sect. 4, the proposed approach is presented, followed by the experimental results in Sect. 5. Finally, in Sect. 6 the conclusion and future work are presented.

2 Related Works

Numerous efforts are interested in extracting associations automatically in order to support the discovery of new relationships between concepts in MEDLINE. These approaches are often limited to explore relationships between two concepts, such as drugs-disease associations or disease-gene relations. In 1991, Sperzel *et al*. [6] conducted an experiment to investigate the feasibility of using the Unified Medical Language System (UMLS), which is a repository of biomedical vocabularies) resources

to link databases (DB) in clinical genetics and molecular biology. References from the Mendelian Inheritance in Man (MIM) were lexically mapped to the equivalent citations in MEDLINE.

In 2004, Gall and Brahmi [7] tested the search capabilities of the EndNote search engine for retrieving citations from MEDLINE in support of importation into EndNote version 7.0. As EndNote is a citation management software package. Ovid MEDLINE and PubMed were selected for comparison. In 2005, Rak and Kurgan [8] used the mining technology to classify articles from the largest medical repository, MEDLINE. This method was based on a novel associative classification technique that considers the recurrent items and most importantly multi-label characteristic of the MEDLINE data. In 2006, Al-Mubaid and Nguyen [9] adapted information-based semantic similarity measures from general English and applied them into the biomedical domain to measure the similarity between biomedical terms. The experimental results depicted that by using MEDLINE and MeSH ontology, the information-based similarity measures performs very well and produced high correlations with human ratings. Yoo and Xiaohua [10] improved the clustering quality for MEDLINE articles by investigating the biomedical ontology MeSH. For this investigation, the authors performed a comprehensive comparison study of various document clustering approaches, such as hierarchical clustering methods (single-link, complete-link, and complete link), bisecting K-means, K-means, and suffix tree clustering (STC) in terms of efficiency, effectiveness, and scalability. In 2006, Plikus *et al*. [11] developed citation's prioritisation algorithm based on journal impact factor, forward referencing volume, referencing dynamics, and author's contribution level. It could be applied either to the primary set of PubMed search results or to the subsets of these results identified through key terms from controlled biomedical vocabularies and ontologies. The National Cancer Institute (NCI) thesaurus and Mouse Genome Database (MGD) mammalian gene ontology had been implemented for key terms analysis. PubFocus provides a scalable platform for the integration of multiple available ontology databases. The PubFocus analytics can be adapted for input sources of biomedical citations other than PubMed.

In 2007, Névéol *et al.* [12] reported the latest results of an Indexing Initiative effort addressing the automatic attachment of subheadings to MeSH main headings recommended by the NLM's Medical Text Indexer. In 2008, Booth and Rourke [13] concluded whether the information retrieval might be improved by assembly each constituent of a structured abstract. In 2008, Kim *et al*. [14] presented a novel search engine called MedEvi that imposes positional restriction on occurrences matching multi-term queries. Based on the observation, that terms with explicitly stated semantic relations in text were not found too far from each other. MedEvi further identifies additional keywords of biological and statistical significance from local context of matching occurrences in order to help users reformulate their queries for better results. In 2009, Humphrey *et al*. [15] evaluated and compared the performance of these systems against a gold standard of humanly assigned categories for one hundred MEDLINE documents using six measures.

A variety of studies have been published using different approaches for exploring the relationships between biomedical concepts. Some studies developed a system based on co-occurrence grouping co-mentioned drugs, diseases, genes and genomic variations [16]. Although, other work [17] used the disease/drug co-occurrence in MEDLINE abstracts to discover the drug out and to construct a network explaining disease and drug relation. Other researches [18] described a system that extracts the relations between protein and sub-cellular localization using natural language processing techniques. In 2012, Yeganova *et al*. [19] described and compared two methods for automatically learning meaningful biomedical categories in MEDLINE. Rather than imposing external ontology on MEDLINE, the methods allowed categories to emerge from the text. Other studies proposed the extraction of gene relation using a new technology [20] rule-based text-mining algorithm with keyword matching. For example, Chun *et al*. [21] presented a system for extracting disease-gene relations from Medline using medical dictionaries and machine learning. The system consists of gathering a dictionary of disease along with the gene names from different public databases and extracting relation candidates by dictionary matching. In 2012, Raj and Prasanna [22] extracted relations between diseases and treatments using a machine learning approach based on the Naïve Bayes algorithm to enhance automatic identification of the disease in the medical field and to improve text classification using an integrated model. In 2013, Jimeno *et al.* [23] had built translation data sets in the biomedical domain that could easily be extended to other languages available in MEDLINE. These sets could successfully be applied to train statistical machine translation models. In 2013, Gu *et al*. [24] proposed a new semi supervised spectral clustering method, i.e., SSNCut, for clustering over the LC (link clustering) similarities with two types of constraints, namely must-link (ML) constraints on document pairs with high MS (or GC) similarities and cannot-link (CL) constraints on those with low similarities. The authors empirically demonstrated the performance of SSNCut on MEDLINE document clustering using 100 data sets of MEDLINE records. In 2013, Zhang *et al*. [25] illustrated the framework through the construction of disease-specific networks from Semantic MEDLINE, an NLP-generated association database, followed by the analysis of network properties, such as hub nodes and degree distribution.

Xu and Wang [5] developed a system based on pattern-learning approach to extract treatment-specific drug-disease pairs from biomedical abstract available on MEDLINE.

The most straight-forward approaches to detect medical concept relationships are the hybrids approaches. For example, Bchir and Ben Abdessalem Karaa [26] suggested an approach to extract disease-drug relations using machine learning combined with natural language processing. In the first step, natural language processing techniques for the abstracts' pre-processing and extracting a set of features from the pre-processed abstracts were employed. In the second step, a disease-drug relation using machine learning classifier was extracted. The Anno-pharma system introduced by Benzartiand Ben Abdessalem Karaa [27] used natural language processing techniques, ontology and dictionaries to detect the substances responsible for adverse reaction on the organs of the human body. A new methodology has been presented for the extraction of the hidden relationships from MEDLINE. In 2014, Kwon *et al*. [28] determined the value

and efficacy of searching biomedical databases beyond MEDLINE for systematic reviews. The authors suggested that expanding the range of databases to search beyond MEDLINE, Embase, and CINAHL. A systematic review was conducted on the efficiency of ward closure as an infection control practice. In consultation with the primary investigators who are subject experts, databases were selected. The search strategy was developed by librarians. YK and SEP searched Ovid MEDLINE including In-Process/other Non-Indexed Citations, OvidEmbase, CINAHL Plus, Cochrane Database of Systematic reviews (CDSR), LILACS, and IndMED for any study type discussing the implementation of ward closure in the case of an outbreak.

Table 1 reported a comparison of different studies related to the extraction relations from MEDLINE abstracts. These literatures have concentrated on specific types of relations. Therefore, a standard model for semantic relations is proposed.

Table 1. Tabulated comparison of extraction relations from MEDLINE abstracts

Author	Year	Used technique	Extracted relation
Chun *et al.* [21]	2006	Medical dictionaries and machine learning	Disease-gene
Li and Chen [17]	2009	Co-occurrence disease and drug in MEDLINE abstracts	Drugs-disease
Chun *et al.* [18]	2010	Natural language processing techniques	Protein and sub-cellular localization
Ben Abacha and Zweigenbaum [29]	2011	A hybrid approach	Cure and side effect relations
Raj and Prasanna [22]	2012	Machine learning approach based on the Naïve Bayes algorithm	Diseases and treatments
Xu and Wang [5]	2013	Pattern-learning approach	Treatment-specific drug-disease
Bchir and Ben Abdessalem Karaa [26]	2013	Machine learning combined with natural language processing	Disease-drug
Benzartiand Ben Abdessalem Karaa [27]	2013	Natural language processing techniques, ontology and dictionaries	The substances responsible for adverse reaction on the organs of the human body
Jung *et al.* [20]	2014	Rule-based text-mining algorithm with keyword matching	Gene-gene
Nhung *et al.* [30]	2015	Deep syntactic patterns	Diverse types of binary relations

Consequently, the current proposed a novel method for information extraction using the NLP tools and ontology. It depicted a system that extracted relations between the three concepts from biomedical texts via morphological analysis and information extraction techniques.

3 Methodologies

MEDLINE is a massive biomedical corpus that includes many topics, such as biology, biochemistry, medicine, nursing, dentistry, health care system, and other fields. Therefore, the main steps for the proposed method are included in the following procedure.

Proposed System Procedure
Start
Select and *Clean* the MEDLINE abstract
Perform morphological analysis
Annotation
Extract the information
End

A detailed description for these steps is as follows:

3.1 MEDLINE Abstract Selecting and Cleaning

The selecting step consists of extracting a set of pertinent abstracts from the MEDLINE database. At this stage, it must be guaranteed that all abstracts are randomly selected without any user intervention. For this purpose, JabRef tool is used for importing data from online scientific databases like MEDLINE. This allows building personal biomedical database of pertinent abstracts. Afterward, the output must be cleaned as it contains a lot of irrelevant information that affects the performance of the results. The cleaning step consists of eliminating useless information from abstracts, such as the stop words (an, a, and, so, any, etc.). Thus, only the important data will be kept.

3.2 Morphological Analysis

This analysis determines the syntactic structure of words in a text. The sentence analysis consists of two stages: (i) automata construction step, and (ii) recognition automata step. The sequences of the different phases of this step are as follows.

1. Cut the text into sentences and split each one into words.
2. Determine for each word already segmented the morphological characteristics.

Automata construction: the proposed work starts by creating the own automata. The automaton is directed by a word supplied as input that passes from one state to another, reading the input alphabet [31]. The automaton is described, where an automaton A parses the input text and produces an output. At first, the text is to be cut in sentences to facilitate the analysis. The standard rule applied is as follows: each time there exist "." or "?" or "!" or "…" followed by capital letter inserts marker "#". Each one sentence contains a subject, a verb and a complement. Thereafter, splits the sentences into words to establish the initial structure of the automaton (automaton reads a string in a sentence each time there is a space it considers the string as a word). For example, the sentence: "Iron deficiency anemia occurs." is represented by the following automaton in Fig. 1.

Fig. 1. Example of an initial automaton

Recognition automata: In this step, the automaton is to be tagged from a morph-syntactic point of view. Then, for each word already segmented, determines the morphological characteristics. This will allow the various relations identification and the recognition concepts in the next step. Since, the INTEX dictionaries [32] include two sets of dictionaries. Thus, these dictionaries are used; one dictionary describes simple words, and the second presents compound words. Each word is tagged with all its occurrences in dictionaries. The needed information is added in some cases to be more comprehensive. Figure 2 shows an example of a tagged sentence.

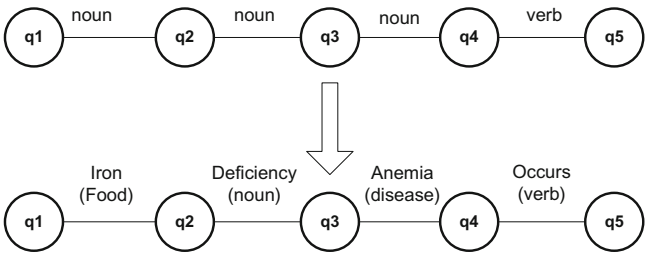


Fig. 2. Labelled automaton

3.3 Annotation

In the current work, ontologies are used to provide formal representations of knowledge about given fields that could be processed by machines. Ontology usually represents classes of entities, their properties, their relationships and roles, their decomposition into parts, and the events/processes in which entities could participate. The ontology supports the extraction of information based on the annotations concept. This will give the chance to learn more about a very specific field. In this step, the existing ontologies are reused using the relationships between genes, disease and foods. Such ontologies are: Gene Ontology (GO) (http://www.geneontology.org/) to identify all appearances of gene names in the text and annotated them to be recognized in the next step. Disease Ontology (http://disease-ontology.org/downloads/) (DO) is to find a link between the development of gene and the disease. Finally, Food Ontology:

(http://data.lirmm.fr/ontologies/food#Food) (FO) is used to describe food products that will be used later. These ontologies can be modified to the needs, if the case requires having more efficient results during the annotation step. Thus, GATE (https://gate.ac.uk/) is a framework and development environment for language engineering. In order to accomplish this task, to GATE the ontology, cover areas of the research already present (food, gene and disease) and the annotation performed is provided.

Once food, gene and disease names are identified with ontology, the next target is to analyse the output to find out the interaction between these concepts that are usually expressed by frequently seen verbs in the domain such as "cause", "generate", "occurs", etc., in addition the frequently seen adjectives, such as "precarious", "poor", "defect", etc., or the adverbs "accumulatively", "abusively", "highly", etc. For example, a relation in the following sentence can be identified: "*Precarious in glucose causes defects of genes substrate, hence we can have diabetes.*"

3.4 Information Extraction

In this step, an algorithm used to extract the relationships from the prepared corpus. This algorithm has two steps, namely define rules that are validated by an expert, aiding gene disease-food relationships discovery and extracts relationships from the abstract that correspond to the defined rules. The rules are written in the following patterns:

Rule1: <Gene> develop <Disease> prevented by <food>
Rule2: <Food> verb <Disease> adjective <Gene>
Rule3: <Food> Lower <Disease> Associate <Gene>
Rule4: <Food> cause <Disease> mutated <Gene>
Rule5: <Food> verb <gene> verb <disease>
etc.

where, verb, adverb, and adjective are the most frequent types of relationships between the food, disease and gene. The rules are gathered in an automaton to facilitate the information extraction step. The Fig. 3 shows an extract of the automata.

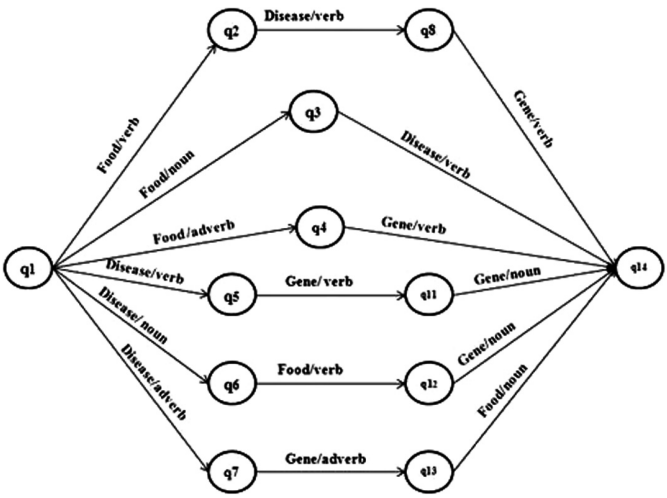


Fig. 3. Automaton summarize rules

4 The Proposed Approach

From the above mentioned methodology the proposed method algorithm can be summarized as:

Algorithm: Relations Extraction from Biomedical DB
Start
Input: *MEDLINE abstract*
Pre-processing for each abstract including:
- *Selecting*
- *Cleaning*

Morphological analysis
- *Automata construction*
- *Recognition automata*

Annotation using Ontology to identify the food, gene and disease names and to represent classes of entities, their properties, their relationships and roles
Information extraction
- *Define rules (gene-disease-food relationships discovery)*
- *Extract relationships*

Output: *Identified relations between medical concepts using NLP tool*
End

Figure 4 illustrates the overall architecture of the extraction of gene–disease-food relationships approach, where the essential steps description for the proposed model is introduces as follows: first a pre-processes step for each abstract in the corpus is done.

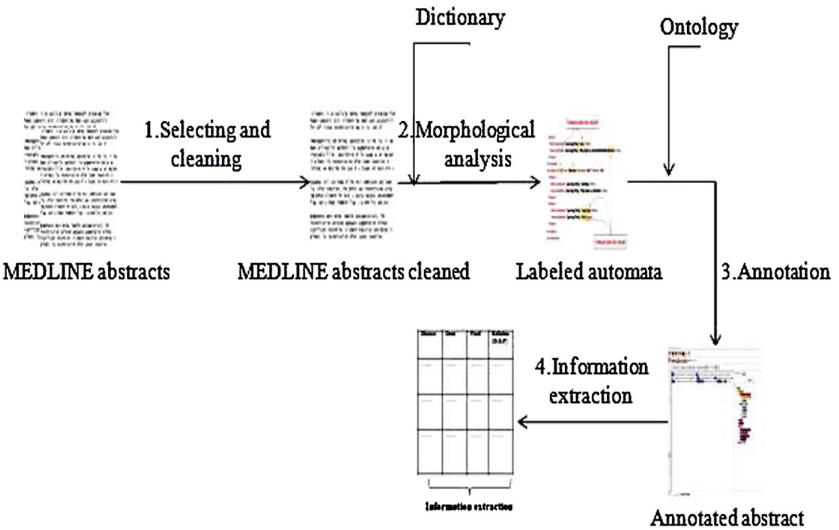


Fig. 4. Proposed system overview

Secondly, a morphological analysis to identify the role of each concept in the sentence is conducted. Thirdly, an annotation that maps each word has been made to ontology and expresses the relationships between concepts, finally the information extraction.

5 Experimental Results

Typically, biomedical analysis and mining techniques for improved health outcomes is an essential domain [33–35]. Thus, this work proposed a novel method to extract relations between gene, disease and food, which are medical concepts using NLP tools. As mentioned in the proposed approach, the first step is a pre-processing step. It is common to all text mining treatments due to the nature of textual data that are unstructured and mostly available in natural language form. In this case, pre-processing consists of two steps: selecting and cleaning. The selecting step consists of extracting 700 abstracts from the MEDLINE database. The output of the pre-processing step is a set of abstracts that are ready to be the input of the morphological analysis. The output of the previous step is a tagged corpus ready to be the input of the annotation processes. GATE is to be used in the annotation step to provide a reusable design for language engineering. It contains a set of prefabricated software integrated into blocks that language engineers can use, or expand and adapt to their own needs. The final step is the information extraction.

The result after the implemented tool execution is represented in a structured format Table 2 and Fig. 5. The relation between food, disease and gene is expressed in natural language that makes the task difficult as each one has its own way of expressing this relationship. Therefore, it must interpret the results to evaluate the approach. Validating results gives researchers the ability to verify the accuracy of the system. Among the results provided by the system, it is found that an excess of sodium causes hypertension. Also, an excess of glucose causes defects in gene substrate, thus leading to diabetes. In addition, performance measures such as precision, recall and F-measure of food are calculated for gene and diseases separately to ensure the accuracy and reliability of the system as illustrated in Fig. 5.

Table 2. Performance measurements of relation detection

	Precision	Recall	F-measure
Food	0.9	0.84	0.86
Disease	0.98	0.91	0.94
Gene	0.93	0.88	0.90
Relations (F, G, D)	0.77	0.69	0.72

The preceding results establish that the proposed approach allows the extraction of genes, diseases, and food names and the relations between them. An evaluation is conducted to interpret the performance of the proposed approach. Moreover, it is observed promising results. About: 0.86, 0.94, and 0.90 of F-measure related to

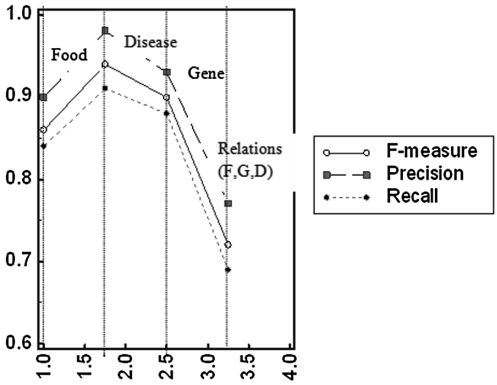


Fig. 5. Graphical measures of relation detection

respectively Food, Disease, and Gene concepts. However, the system has lower F-measure for detecting relations (0.72). The reason can be the ambiguity of natural language, since abstracts are written by humans and each one has its own way of expressing ideas.

As a future work, the extraction of information from the full-length papers using other technique to discover the hidden knowledge will be tried. Furthermore, the extraction of information from the full-length papers using other techniques to discover the hidden knowledge can be performed in the future work. In addition, researchers can apply the used approach in several datasets that included in various applications as in [36–44].

6 Conclusions

The development of an information extraction system helps users to explore various relationships between concepts in an easy and logical manner. Thus, discovering relationships between biomedical concepts plays a significant role in human health to develop new medical diagnostic techniques and more effective treatment for prevention. Since, the natural language processing tools such as text mining have considerable significance in biomedical field by extracting hidden knowledge from text using textual information. Therefore, this study focused on the importance of the connection between MEDLINE concepts by combining text mining techniques with ontology in order to annotate texts.

In this work, the concept of textual information extraction data is more precisely extracting information from MEDLINE. This work proposed a new approach for information extraction using natural language processing tools. It originated simply by information retrieval to gather a set of abstracts related to the area. Then, uses an automaton and the ontology were annotated to gather a set of abstracts from the text by recognizing interesting concepts and making a morphological analysis. After the annotation step, some rules were defined to help gene disease-food relationships discovery.

In the experimental study, a sample of 700 abstracts is used. The returned effectiveness measures were of a good value, thus, the proposed approach can be applied in several applications [36–44]. Nonetheless, this did not prevent the system from making some mistakes.

References

1. Feldman, R., Dagan, I.: Knowledge discovery in textual database (KDT). In: The First International Conference on Knowledge Discovery and Data Mining (KDD-95), Montreal, Canada (1995)
2. Sassi, D.-B., Karaa, W.-A.: Genetic algorithm for clustering Medline abstracts. In: The First International Conference on Knowledge Management Information and Knowledge Systems (KMIKS), Hammamet, Tunisia (2013)
3. Jahiruddin, M.-A., Dey, L.: A concept-driven biomedical knowledge extraction and visualization framework for conceptualization of text corpora. J. Biomed. Inform. **43**(6), 1020–1035 (2010)
4. Baker, N.C., Hemminger, B.-M.: Mining connections between chemicals, proteins, and diseases extracted from Medline annotations. J. Biomed. Inform. **43**(4), 510–519 (2010)
5. Xu, R., Wang, Q.: Large-scale extraction of accurate drug-disease treatment pairs from biomedical literature for drug repurposing. BMC Bioinform. **14**, 181 (2013)
6. Sperzel, W.-D., Abarbanel, R.-M., Nelson, S.-J., Erlbaum, M.-S., Sherertz, D.-D., Tuttle, M.-S., Olson, N.-E., Fuller, L.-F.: Biomedical database inter-connectivity: an experiment linking MIM, GENBANK, and META-1 via MEDLINE. In: Proceedings of the Annual Symposium on Computer Application in Medical Care, pp. 190–193 (1991)
7. Gall, C., Brahmi, F.A.: Retrieval comparison of EndNote to search MEDLINE (Ovid and PubMed) versus searching them directly. Med. Ref. Serv. Q. **23**(3), 25–32 (2004)
8. Rak, R., Kurgan, L., Reformat, M.: Multi-label associative classification of medical documents from MEDLINE. In: Proceedings Fourth International Conference on Machine Learning and Applications, 15–17 December 2005
9. Al-Mubaid, H., Nguyen, H.-A.: Using MEDLINE as standard corpus for measuring semantic similarity in the biomedical domain. In: Sixth IEEE Symposium on BioInformatics and BioEngineering (BIBE), 16–18 October 2006
10. Yoo, I., Xiaohua, H.: Biomedical ontology MeSH improves document clustering qualify on MEDLINE articles: a comparison study. In: 19th IEEE International Symposium on Computer-Based Medical Systems (CBMS), pp. 577–582 (2006)
11. Plikus, M., Zhang, Z., Chuong, C.-M.: PubFocus: semantic MEDLINE/PubMed citations analytics through integration of controlled biomedical dictionaries and ranking algorithm. BMC Bioinform. **7**(1), 424 (2006)
12. Névéol, A., Shooshan, S.-E., Mork, J.-G., Aronson, A.-R.: Fine-grained indexing of the biomedical literature: MeSH subheading attachment for a MEDLINE indexing tool. In: AMIA Annual Symposium Proceedings, pp. 553–557 (2007)
13. Booth, A., O'Rourke, A.: The value of structured abstracts in information retrieval from MEDLINE. Health Libr. Rev. **14**(3), 157–166 (1997)
14. Kim, J.-J., Pezik, P., Rebholz-Schuhmann, D.: MedEvi: retrieving textual evidence of relations between biomedical concepts from Medline. Bioinformatics **24**(11), 1410–1412 (2008)

15. Humphrey, S.-M., Névéol, A., Gobeil, J., Ruch, P., Darmoni, S.-J., Browne, A.: Comparing a rule-based versus statistical system for automatic categorization of MEDLINE documents according to biomedical specialty. J. Am. Soc. Inf. Sci. Technol. **60**(12), 2530–2539 (2009)
16. Garten, Y., Altman, R.: Pharmspresso: a text mining tool for extraction of pharmacogenomic concepts and relationships from full text. BMC Bioinform. **10**(2), 1–9 (2009)
17. Li, J., Zhu, X., Chen, J.-Y.: Building disease-specific drug-protein connectivity maps from molecular interaction networks and PubMed abstracts. PLoS Comput. Biol. **5**, e1000450 (2009)
18. Chun, H.-W., Kim, J.-D., Choi, Y.-S., Sung, W.-K.: Extracting protein sub-cellular localizations from literature. In: Active Media Technology, pp. 373–382. Springer, Heidelberg (2010)
19. Yeganova, L., Kim, W., Comeau, D.-C., Wilbur, W.-J.: Finding biomedical categories in Medline. J. Biomed. Semant. **3**, S3 (2012)
20. Jung, J.-Y., DeLuca, T.F., Nelson, T.H., Wall, D.P.: A literature search tool for intelligent extraction of disease-associated genes. J. Am. Med. Inform. Assoc.: JAMIA **21**(3), 399–405 (2014)
21. Chun, H.-W., Tsuruoka, Y., Kim, J.-D., Shiba, R., Nagata, N., Hishiki, T., Tsujii, J.: Extraction of gene-disease relations from medline using domain dictionaries and machine learning. In: Pacific Symposium on Biocomputing, pp. 4–15 (2006)
22. Raj, T.-M., Prasanna, S.: Implementation of ML using Naïve Bayes algorithm for identifying disease-treatment relation in bio-science text. Res. J. Appl. Sci. Eng. Technol. **5**, 421–426 (2013)
23. Yepes, A.-J., Prieur-Gaston, E., Neveol, A.: Combining MEDLINE and publisher data to create parallel corpora for the automatic translation of biomedical text. BMC Bioinform. **14**, 146 (2013)
24. Gu, J., Feng, W., Zeng, J., Mamitsuka, H., Zhu, S.: Efficient semisupervised MEDLINE document clustering with MeSH-semantic and global-content constraints. IEEE Trans. Cybern. **43**(4), 1265–1276 (2013)
25. Zhang, Y., Li, D., Tao, C., Shen, F., Liu, H.: An integrative computational approach to identify disease-specific networks from PubMed literature information. In: 2013 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 18–21 December 2013
26. Bchir, A., Karaa, W.-A.: Extraction of drug-disease relations from MEDLINE abstracts. In: 2013 World Congress on Computer and Information Technology (WCCIT), Sousse, Tunisia (2013)
27. Benzarti, S., Karaa, W.-A.: Anno pharma: detection of substances responsible of ADR by annotating and extracting information from MEDLINE abstracts. In Control, 2013 International Conference on Decision and Information Technologies (CoDIT), Hammamet, Tunisia (2013)
28. Kwon, Y., Powelson, S.-E., Wong, H., Ghali, W.-A., Conly, J.-M.: An assessment of the efficacy of searching in biomedical databases beyond MEDLINE in identifying studies for a systematic review on ward closures as an infection control intervention to control outbreaks. Syst. Rev. **3**, 135 (2014)
29. Ben Abacha, A., Zweigenbaum, P.: A hybrid approach for the extraction of semantic relations from MEDLINE abstracts. In: 12th International Conference, CICLing 2011, Tokyo, Japan, 20–26 February 2011
30. Nhung, T.H.-N., Makoto, M., Yoshimasa, T., Takashi, C., Satoshi, T.: Wide-coverage relation extraction from MEDLINE using deep syntax. BMC Bioinform. **16**(4) (2015)
31. Hopcroft, J.-E., Ullman, J.-D.: Introduction to Automata Theory, Languages and Computation. Addison-Wesley Longman Publishing Co., Inc., Boston (1990)

32. Silberztein, M.: Dictionnaires électroniques et analyse automatique de texte: le système INTEX. Masson. Editor. Paris (1993)
33. Dey, N., Karaa, W.-A.: Biomedical Image Analysis and Mining Techniques for Improved Health Outcomes. Advances in Bioinformatics and Biomedical Engineering (ABBE) book series (2015)
34. Karaa, W.-A., Ashour, A.-S., Ben Sassi, D., Roy, P., Kausar, N., Dey, N.: MEDLINE text mining: an enhancement genetic algorithm based approach for document clustering. In: Applications of Intelligent Optimization in Biology and Medicine: Current Trends and Open Problems. Springer (2015). Chap. 10
35. Day, N., Samanta, S., Chakraborty, S., Das, A., Chaudhuri, S.-S., Suri, J.S.: Firefly algorithm for optimization of scaling factors during embedding of manifold medical information: an application in ophthalmology imaging. J. Med. Imaging Health Inform. **4**, 384–394 (2014)
36. Beldjehem, M.: A unified granular fuzzy-neuro min-max relational framework for medical diagnosis. Int. J. Adv. Intell. Paradig. **3**(2), 122–144 (2011)
37. Djebbar, A., Merouani, H.F.: Optimising retrieval phase in CBR through Pearl and JLO algorithms for medical diagnosis. Int. J. Adv. Intell. Paradig. **5**(3), 161–181 (2013)
38. Pal, A.K., Dey, N., Samanta, S., Das, A., Chaudhuri, S.S.: A hybrid reversible watermarking technique for color biomedical images. In: IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), pp. 1–6 (2013)
39. Wu, M., Nakata, M., Sakai, H.: Two rough set-based software tools for analyzing non-deterministic data. Int. J. Rough Sets Data Anal. (IJRSDA) **1**(1), 32–47 (2014)
40. Gonzalez, M.F., Facal, D., Navarro, A.B., Geven, A., Tscheligi, M., Urdaneta, E., Yanguas, J.: Analysis of older users' perceived requests and opportunities with technologies: a scenario-based assessment. In: Pervasive and Ubiquitous Technology Innovations for Ambient Intelligence Environments, p. 40 (2012)
41. Chowdhury, S.R., Ray, R., Dey, N., Chakraborty, S., Karaa, W.B.A., Nath, S.: Effect of demons registration on biomedical content watermarking. In: 2014 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT), pp. 509–514 (2014)
42. de Rooij, A., Broekens, J., Lamers, M.F.: Abstract expressions of affect. Int. J. Synth. Emot. **4**(1) (2013)
43. Rajkumar, M.A.: Continuous review retrial inventory system with impatient customers. Int. J. Comput. Syst. Eng. **1**(3), 193–199 (2013)
44. Kamal, S., Ripon, S.H., Dey, N., Ashour, A.S., Santhi, V.: A MapReduce approach to diminish imbalance parameters for big deoxyribonucleic acid dataset. Comput. Methods Programs Biomed. **131**, 191–206 (2016)

Cluster Analysis for European Neonatal Jaundice

P.K. Nizar Banu[1(✉)], Hala S. Own[2], Teodora Olariu[3], and Iustin Olariu[3]

[1] Department of Computer Science, Christ University, Bengaluru, India
nizarbanu@gmail.com
[2] National Research Institute of Astronomy and Geophysics, Helwan, Egypt
halaown@gmail.com
[3] Vasile Goldis Western University of Arad, Arad, Romania
olariu_teodora@yahoo.com, iustin_olariu@yahoo.com

Abstract. The objective of this paper is to propose and analyze clustering techniques for neonatal jaundice which will help in grouping the babies of similar symptoms. A variety of methods have been introduced in the literature for neonatal jaundice classification and feature selection. As far as we know, clustering techniques are not used for neonatal jaundice data set. This paper studies and proposes clustering techniques such as K-Means, Genetic K-Means and Bat K-Means for jaundice disease. To find the number of clusters elbow method is used. The clusters are validated using RMSE, SI and HI. The experimental results carried out in this paper shows bat k-means clustering performs better than K-means and genetic K-means.

Keywords: Neonatal jaundice · Clusters · K-means · Bat K-means · Genetic K-means

1 Introduction

Relationships and patterns within the medical datasets provide novel medical knowledge due to the accumulation of large quantities of information about patients and their medical conditions (Zhong et al. 2001; Cios et al. 1998; Carlin et al. 1998; Lavrajc et al. 1997). Neural network, Bayesian classifier, genetic algorithms, fuzzy theory and rough sets have been used in medical data analysis (Beligiannis et al. 2006; Adamopoulos et al. 2002; Beligiannis et al. 2005). Each technique has a distinct methodology for addressing problems in its field (Own and Abraham 2012).

Neonatal jaundice is the common problem in newborns (Samar et al. 2009; Chang et al. 2006). It causes yellowing of skin and whitening of eye that is known as sclera. Almost all the infants are jaundice prone (Shrivastava 2013); among them around 60% are term and 40% are preterm infants. Jaundice occurs due to the breakdown of red blood cells; this breakdown process is known as hemolysis. If the cell breakdown rate occurs faster than the usual rate, it increases the level of bilirubin in the body and causes jaundice to the infants. A moderate level (6 mg/dl–9 mg/dl) can be tolerable and

V.E. Balas et al. (eds.), *Soft Computing Applications*, Advances in Intelligent Systems and Computing 633, DOI 10.1007/978-3-319-62521-8_35

easily treated, but delay in the treatment can cause fatal and irreversible brain damage of the newborn. Therefore, early detection of the newborn jaundice is mandatory (Gupta et al. 2015).

Applying clustering methods for any data is a very good start for any kind of analysis on the data. Cluster detection model finds the data records similar to each other. This is basically an undirected data mining technique, since the goal is to find previously unknown similarities in data. Self-similar clusters serves as the basic need for knowing what is in the data and for assuming how to bring out the best use of it. Although the dataset taken for analysis in this paper has classes for categorizing, there are chances of forcing a record to be a part of any of the existing classes. Thus, the hidden unique characteristics of those records may be left unfound. At this state, finding natural grouping of the data is significant. Hence, the study in this paper aims at applying clustering techniques to reveal the unseen similarity of the patient records. A comparative study on evolutionary and clustering techniques that are designed to handle the volatile nature of network data is presented (Sharma and Bhatnagar 2012).

In (Shrivastava 2013), Artificial Neural Networks is applied for the diagnosis of neonatal jaundice. In (Own and Abraham 2012), weighted rough set framework is applied for real time Egyptian neonatal jaundice dataset. First, a weighted attribute reduction algorithm is applied to find the reduct set. Unsupervised Quick Reduct proposed in (Velayutham and Thangavel 2011a), Unsupervised Relative Reduct proposed in (Velayutham and Thangavel 2011b), U-TRS-RelRed proposed in (Inbarani and Banu 2012), is applied for gene expression dataset in (Banu and Inbarani 2013). In (Inbarani et al. 2014), US-PSO-RR and USRR methods are applied for fetal heart rate. Quick Reduct (QR), Entropy Based Reduct (EBR), Unsupervised Tolerance Rough Set based Quick Reduct (U-TRS-QR) are applied for neonatal jaundice dataset in (Banu et al. 2014, 2015). Hard and fuzzy clustering algorithms for thyroid disease dataset are proposed and optimal number of clusters is found using elbow criterion method (Azar et al. 2013). Managing and analysis of medical big data involve many different issues regarding their structure, storage and analysis. For reducing the dimensions, selecting features and classification in such datasets LHNFCSF is presented (Azar and Hassanien 2014). Six different types of SVM such as St-SVM, PSVM, LSVM, NSVM, LPSVM and SSVM is applied to Wisconsin breast cancer dataset and analyzed. The experimental results reveal that SVM classifiers achieve fast, simple and efficient breast cancer diagnosis (Azar and El-Said 2014). Competitive learning networks and unsupervised data clustering methods are used to model the differential grading in childhood autistic rating scale based assessment (Pratap and Kanimozhiselvi 2014). Enhanced swarm based feature selection method for clustering gene expression data is proposed (Banu and Andrews 2015). A new iterative, non-parametric, partitioning clustering algorithm called prime equivalence algorithm is proposed and applied for benchmark datasets (Ashok and Judy 2015).

The rest of the paper is structured as follows. Section 2 presents cluster analysis for jaundice dataset, followed by experimental analysis and discussion in Sect. 3. Finally conclusion and future direction is given in Sect. 4.

2 Cluster Analysis

Cluster analysis is one of the major data mining methods which help to identify the natural grouping in a set of data items. Clustering (Hartigan 1975; Devijver and Kittler 1982; Jain and Dubes 1988) is an unsupervised classification method. It is used when the data available is unlabeled. In clustering, a set of patterns are organized into coherent and contrasted groups or clusters, such that all data in the same group are similar to each other, while data from different clusters are dissimilar (Hartigan 1975). Five different influential factors are studied and identified for the performance improvement of K-Means clustering, (Khan et al. 2013). Almost all the clustering algorithms rely on the number of clusters *'K'*. In this paper, optimal number of clusters is identified using elbow criterion method. The following subsections present a brief description of various clustering techniques proposed for Egyptian neonatal jaundice.

2.1 K-Means

The K-Means (McQueen 1967) algorithm is a popular partitional clustering method due to its easy implementation and rapid convergence. It divides data into predefined number of clusters in order to optimize a predefined criterion. This algorithm iteratively updates the solution in a deterministic manner such that their results are heavily influenced by the choice of initial solution. It is the simplest and most commonly used algorithm that uses a squared error criterion (Jain 2008). Provided with a set of m numeric objects and an integer number K ($K \leq m$), it computes a partition of patterns into K clusters, assigning each object to a cluster with the nearest center. Then the centroid of each cluster is computed and the cluster centers are updated. Objects are reassigned to the clusters and the algorithm is iterated until it converges to a stable solution (Bellazi and Zupan 2007). K-Means algorithm is described in Fig. 1.

Algorithm : **K-Means**
Input : Jaundice dataset J_{mXn}
K – Number of clusters $(K \leq m)$
Output : Clusters $C_1, C_2, \ldots, C_K$ and
Centroid of clusters $Z_1, Z_2, \ldots, Z_K$

Procedure:

1. Choose K initial cluster centers from the set of objects
2. At the p^{th} iterative step, distribute the objects $\{J_i\}$ among the K clusters by computing the distance between cluster center and object using
$$J \in C_j(K) \; if \; \|J - Z_j(K)\| < \|J - Z_i(K)\|$$
3. Compute the new cluster centers such that the sum of the squared distances from all points in cluster to the new cluster center is minimized.
4. Repeat steps 2 and 3 until convergence

Fig. 1. K-Means algorithm

2.2 Genetic K-Means

Genetic algorithm is based on the optimization techniques guided by the principles of evolution and natural genetics having a large amount of implicit parallelism (Kala *et al.* 2010; Sun and Xiong 2009) and on the natural selection process seen in nature (Yannis *et al.* 2008; Whitley 1994). Best fit organism of the current generation is carried to the next generation. Based on the survival of the fittest, a few clusters are selected and biologically inspired operators like mutation is applied on these clusters to produce new generation. Predetermined number of clusters from the search space is taken and iterated till the termination condition is met. Genetic k-means algorithm is presented in Fig. 2.

$$f(C_i) = \frac{1}{\mu} \tag{1}$$

where $\mu = \sum_{i=1}^{K} \mu_i$ and $\mu_i = \sum_{x_j \in z_i} \left\| x_j - z_i \right\|$, where $i = 1, 2, \ldots K$ is number of clusters, x_j are the objects in the K^{th} cluster, Z_i is center of the K^{th} cluster. z_i is represented as follows.

$$z_i = \frac{1}{n_K} \sum_{j=1}^{n_K} x_j \tag{2}$$

where n_K represents the number of objects in cluster. Fittest cluster among the entire population are considered for the next iteration. Proportional selection is used for selecting the best solution from the current population. According to the probability distribution $\{p_1, p_2, \ldots p_K\}$ is defined as follows

$$p_K = \frac{f(C_i)}{\sum_{i=1}^{K} f(C_i)} \tag{3}$$

Algorithm : **Genetic K-Means**
Input : *K* – Number of Clusters; *D* – Dataset containing *n* objects and *m* features; *tmax* - Maximum number of iterations
Output : Set of *K* clusters
Procedure:
Initialize each cluster to contain *K* randomly chosen centroids from the dataset
For *it = 1: tmax*
For each object
Assign each vector to the cluster with the closest centroid
End
Calculate fitness using equation 1
Select a set of objects for reproduction using proportional selection function given in equation 3
Mutate objects in the cluster using probability distribution function given in equation 4
Update Centroids using equation 5
End

Fig. 2. Genetic K-Means algorithm

where $f(C_i)$ denotes fitness of cluster. Mutation operator helps the algorithm to move towards global optimum. Mutation is done using the following probability distribution function as given by Lu *et al.* (2004).

$$C_i = \frac{1.5 * d_{max}(x_j) - d(x_j, z_i) + 0.5}{\sum_{i=1}^{K}(1.5 * d_{max}(x_j) - d(x_j, z_i) + 0.5)} \tag{4}$$

where $d(x_j, z_i)$ is the Euclidean distance between object x_j and the centroid Z_i of the K^{th} cluster, and $d_{max}(x_j) = max_k\{d(x_j, z_i)\}$. If K^{th} cluster is empty then $d(x_j, z_i)$ is defined as 0. Centroid of the cluster is updated using Eq. 5.

$$z_i(K+1) = \frac{1}{N_i}\sum\nolimits_{x \in C_i(K)} x; i = 1, 2, \ldots K \tag{5}$$

where N_i is the number of objects in $C_i(k)$.

2.3 Bat K-Means

Randomly assign objects (bats) to K clusters. This is considered as the centroid of the clusters. Now, select each object and find the distance between the object and the

Algorithm : **Bat K-Means**
Input : Jaundice dataset $J_{mXn,}$ K – Number of clusters $(K \leq m)$, frequency factor Q *and* loudness A
Output : Clusters $C_1, C_2, \ldots, C_K$ and Centroid of clusters $Z_1, Z_2, \ldots, Z_K$

Procedure:

1. Randomly assign K clusters for each of the N bats
2. For each bat, select K objects from S data objects as initial centroids, by taking the mean values of the attribute of the objects within their given clusters
3. Calculate the fitness of the centroid in each bat, and find the best solution that is represented by the total fitness values of centroid in a bat
4. Generate a new solution by adjusting the frequency, updating the velocity and creating new centroid values
5. If random[0..1]>pulse rate R
   5.1 For each bat, select a solution among a set of best solutions from the other bats, and generate a new local solution around the selected best solution, else goto step 9
6. If random[0..1]<A_i and $f(x_i)<f(x_j)$ else goto step 9
7. Accept new solutions, increase R_i and reduce A_i
8. Reassign the clusters
9. Output the best cluster configuration that is represented by the bat that has the greatest fitness
10. Repeat steps 2 to 9 until convergence

Fig. 3. Bat K-Means algorithm

cluster center. Assign the object to the cluster to which it is closer. Update the centroid and calculate the fitness of the cluster. Adjust the frequency, update the velocity and generate the new solution. Output the best cluster configuration. If the termination condition is not met reassign each data points to different clusters and check for new solutions. Bat flies in search of prey, once it finds the prey; it reduces its loudness and increases its pulse emission rate. For our experiments, loudness A and pulse rate is set to 0.5, frequency factor is set between 0 and 0.02. The values of loudness and pulse rate changes based on the iteration. As the loudness usually decreases once a bat finds its prey, while the pulse rate increases, the loudness can be chosen by any specific value. So it is initialized to 0.5. Bat clustering algorithm presented in (Tang et al. 2012) is shown in Fig. 3 and the same is used for our experiments.

3 Experimental Analysis and Discussion

This section discusses on jaundice dataset, experiments carried out, patients grouped into different clusters and performance measures of various clustering algorithms discussed in Sect. 3.

3.1 Dataset Description

We have taken the dataset used in (Own and Abraham 2012). This is a real time dataset collected from the newborns during January to December 2007 in Neonatal Intensive Care Unit in Cairo, Egypt. This dataset consists of 808 patient records with 16 features and 3 classes. Dataset with its description is shown in Table 1. Original dataset consists of null values. For our experimentation, we filled null values with average of the corresponding attribute.

Experiment is conducted for all 808 records, is shown in Table 2.

After clustering results are obtained, it is significant to validate the accuracy of the clusters formed. Cluster validity indices can be used to evaluate the fitness of data partitions produced by a clustering algorithm (Zalik 2011). Performance analysis of K-Means, Genetic K-Means and Bat K-Means clustering algorithms are presented in Tables 3, 4 and 5 respectively. Root mean squared error (RMSE) (Castro et al. 2007), Homogeneity index (HI) and separation index (SI) (Shamir and Sharan 2001) are used to evaluate the clusters formed. Homogeneity index is calculated as the average distance between each vector and the center of the cluster it belongs to. Separation index is calculated as the weighted average distance between cluster centers. HI reflects the compactness of the clusters while SI reflects the overall distance between clusters. Decreasing HI or Increasing SI suggests an improvement in the clustering results.

Performance analysis of K-means clustering is shown in Figs. 4 and 5 presents the performance analysis of genetic K-means clustering. Performance of bat K-means clustering is shown in Fig. 6. It can be observed from the figures bat K-means produces less error for the clusters from 2 to 11, and it is analogous. HI is also analogous compared to other two clustering methods. SI is high for eight clusters. In other two clustering methods, more variation can be seen for 2 to 11 clusters.

Table 1. Neonatal jaundice data and description

S. No	Attribute name	Description
1	Sex	Male or Female
2	Age/day	Postnatal age in days on admission
3	Gest. Age	Gestational age (F = full term, N = near term, P = preterm)
4	Wt/g	Weight in grams on admission
5	Onset of J at day	Postnatal age of patient on the day in which onset of jaundice was occurred
6	Days of adm.	Days of admission in hospital
7	Peak of T bil	Peak of total bilirubin level
8	bil peak at day	Postnatal age of patient on the day in which total bilirubin peak was recorded
9	T bil d of presentation	Total bilirubin level on the day of presentation
10	D bil d of presentation	Direct bilirubin level on the day of presentation
11	T bil 24 h later	Total bilirubin level after 24 h of presentation
12	D bil 24 h later	Direct bilirubin level after 24 h of presentation
13	T bil after 2day	Total bilirubin level after 2 days of presentation
14	D bil after 2day	Direct bilirubin level after 2 days of presentation
15	T bil before disc	Total bilirubin level before discharge from hospital or death
16	D bil before disc	Direct bilirubin level before discharge from hospital or death
17	Pattern (Class attribute)	1. Patient with indirect hyperbilirubinemia 2. Patients with indirect hyperbilirubinemia then changed into direct hyperbilirubinemia 3. Patents with direct hyperbilirubinemia

Table 2. Jaundice dataset and its distribution

S. No	Class name	Class size	Class distribution
1	Indirect hyperbilirubinemia	737	91%
2	Changed from indirect to direct hyperbilirubinemia	41	13%
3	Direct hyperbilirubinemia	25	3%

The performance results disclose that separation index is high for bat-K-means clustering for 8 clusters and homogeneity index is less for genetic k-means with 7 clusters. With respect to RMSE, genetic K-means yields less error value for 2 clusters. Initially, the optimal number of clusters has to be defined. To find the optimal number of clusters, elbow criterion process is used. It is a common rule of thumb to determine what number of clusters should be chosen (Azar et al. 2013). It is applied with random number of clusters like 5, 8, 10, 13 and the results are shown in Fig. 7.

Table 3. Performance analysis of K-Means clustering algorithm

Dataset	Performance analysis of K-Means				
	No. of clusters	RMSE	SI	HI	No. of patients in clusters
Jaundice	2	2.9005	5.0207	6.7283	474, 333
	3	3.7156	4.2234	8.8847	455, 331, 21
	4	2.3903	5.2324	5.5996	193, 126, 399, 89
	5	3.1378	5.0449	7.0998	215, 222, 93, 118, 159
	6	2.6829	2.0886	6.2298	110, 11, 70, 153, 311, 152
	7	2.2739	3.1144	5.2989	135, 255, 118, 12, 11, 255, 21
	8	2.3604	2.1156	5.5303	285, 53, 4, 113, 72, 200, 56, 24
	9	2.0099	5.0793	4.9140	28, 18, 57, 282, 68, 89, 19, 244, 2
	10	1.9763	**5.9057**	4.7683	143, 101, 13, 24, 21, 86, 159, 65, 130, 65
	11	**1.9650**	2.8850	**4.7100**	22, 26, 201, 25, 84, 10, 109, 55, 85, 41, 149

Table 4. Performance analysis of genetic K-Means clustering algorithm

Dataset	Performance analysis of genetic K-Means				
	No. of clusters	RMSE	SI	HI	No. of patients in clusters
Jaundice	2	**0.1757**	3.5844	5.4192	761, 46
	3	0.8045	5.0289	1.7165	30, 474, 303
	4	1.0239	5.9533	2.7198	261, 373, 125, 48
	5	1.1570	1.8597	2.8032	227, 158, 142, 239, 41
	6	1.0458	4.7174	2.5794	46, 223, 74, 131, 256, 77
	7	0.5239	1.4844	**1.3539**	327, 97, 68, 86, 49, 75, 105
	8	0.9991	1.2842	2.2787	100, 79, 49, 7, 96, 257, 154, 65
	9	1.0762	1.6216	2.5839	32, 28, 131, 111, 169, 92, 36, 76, 132
	10	1.0665	5.7918	2.8639	119, 28, 38, 62, 73, 44, 97, 178, 69, 99
	11	0.8599	**6.1532**	2.2883	160, 21, 20, 29, 53, 53, 139, 119, 102, 51, 90

Table 5. Performance analysis of Bat K-Means clustering algorithm

Datasets	Performance analysis of Bat K-Means				
	No. of clusters	RMSE	SI	HI	No. of patients in clusters
Jaundice	2	3.4896	3.2728	8.2725	548, 259
	3	3.3025	4.7250	7.8135	454, 89, 264
	4	3.1843	4.0610	7.4211	364, 100, 271, 72
	5	3.1039	5.8264	7.1671	9, 109, 67, 26, 596
	6	3.1681	6.8589	7.4996	78, 297, 96, 79, 18, 239
	7	3.0130	11.743	7.0553	22, 395, 5, 91, 111, 12, 171
	8	3.0149	**11.987**	6.7833	1, 7, 316, 1, 2, 274, 13, 193
	9	2.9454	8.9884	6.9344	137, 12, 118, 59, 269, 168, 3, 28, 13
	10	**2.7097**	5.9696	**6.1845**	2, 1, 391, 73, 28, 37, 10, 130, 89, 46
	11	3.4164	NAN	7.5313	100, 5, 3, 2, 366, 302, 6, 8, 12, 2, 2

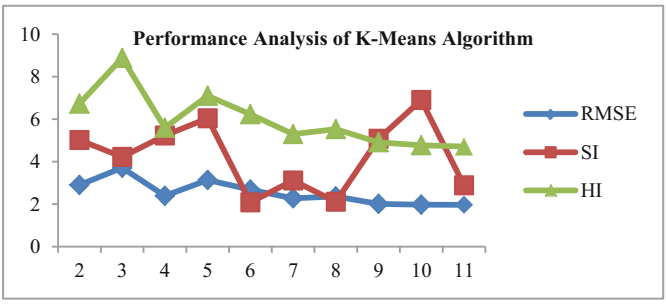


Fig. 4. Performance analysis of K-Means clustering

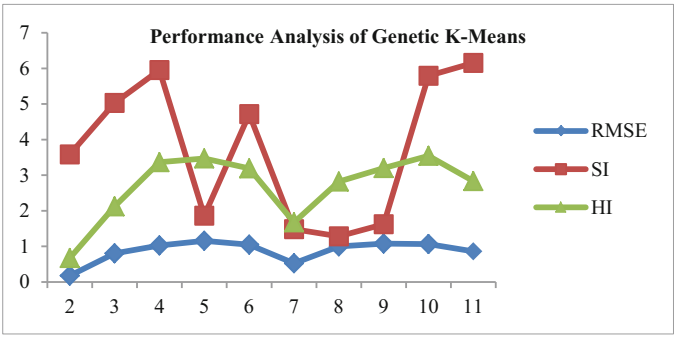


Fig. 5. Performance analysis of genetic K-Means clustering

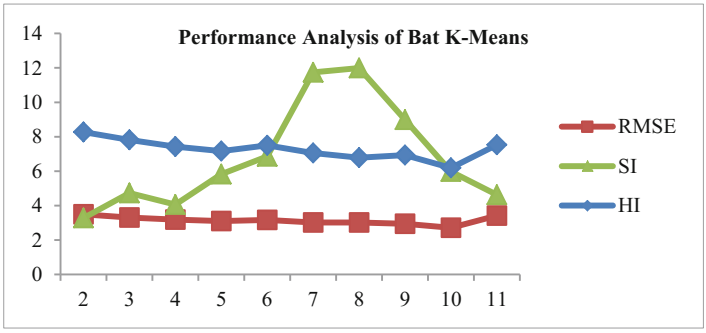


Fig. 6. Performance analysis of Bat K-Means clustering

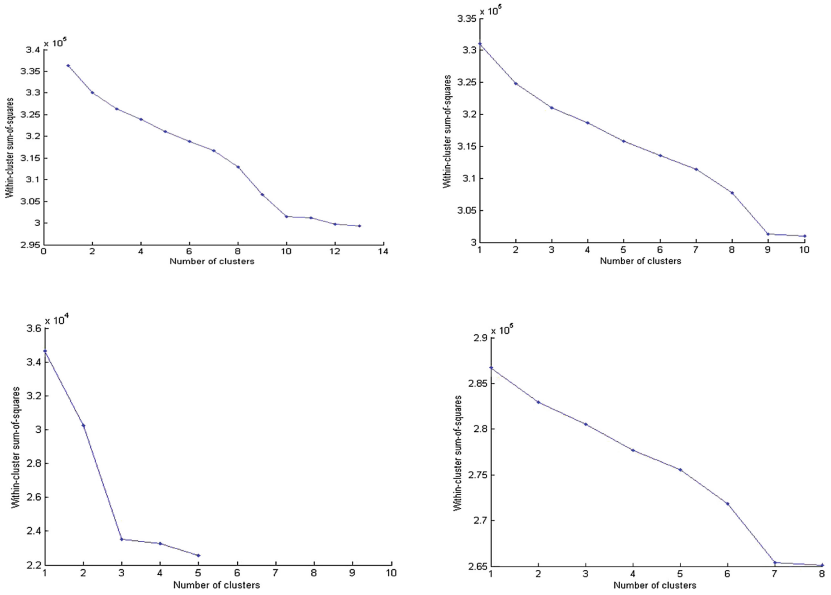


Fig. 7. Optimal number of clusters using elbow criterion method

4 Conclusion

In this study, European neonatal jaundice dataset is clustered using K-Means clustering, Genetic K-Means clustering, and Bat K-Means clustering. Different cluster validity indices such as root mean squared error (RMSE), homogeneity index (HI) and separation index (SI) are used in performances analysis. Several runs are carried out with a different number of clusters between 2 and 11 to establish the optimum number of clusters. Also, to find optimal number of clusters, elbow criterion method is applied. As the number of cluster increase, optimal number of cluster tends to change. According to the validity measures out of 5, 8, 10 clusters, bat k-means gives less error and decreasing homogeneity index for 10 clusters. Also, bat k-means results with high separation score for 10 clusters. This shows bat k-means performs better than the other two algorithms. In future, more efficient methods than elbow criterion for tuning number of clusters can be applied.

References

Adamopoulos, A.V., Anninos, P.A., Likothanassis, S.D., Beligiannis, G.N., Skarlas, L.V., Demiris, E.N., Papadopoulos, D.: Evolutionary self-adaptive multimodel prediction algorithms of the fetal magneto cardiogram. In: Proceedings of the 14th International Conference on Digital Signal Processing (DSP 2002), pp. 1149–1152 (2002)

Jain, A.K.: Data clustering: 50 years beyond K-means. In: Daelemans, W., Goethals, B., Morik, K. (eds.) ECML PKDD 2008. LNCS, vol. 5211, pp. 3–4. Springer, Heidelberg (2008). doi:10.1007/978-3-540-87479-9_3

Pratap, A., Kanimozhiselvi, C.S.: Predictive assessment of autism using unsupervised machine learning models. Int. J. Adv. Intell. Paradigms **6**(2), 113–121 (2014)

Gupta, A., Kumar, A., Khera, P.: Method and model for jaundice prediciton through non-invasive bilirubin detection technique. Int. J. Eng. Res. Technol. **4**(8), 34–38 (2015)

Azar, A.T., El-Said, S.A., Hassanien, A.E.: Fuzzy and hard clustering analysis for thyroid disease. Comput. Methods Programs Biomed. **111**(1), 1–16 (2013)

Beligiannis, G.N., Skarlas, L.V., Likothanassis, S.D., Perdikouri, K.G.: Nonlinear model structure identification of complex biomedical data using a genetic programming-based technique. IEEE Trans. Instrum. Measur. **54**, 2184–2190 (2005)

Beligiannis, G., Hatzilygeroudis, I., Koutsojannis, C., Prentzas, J.: A GA driven intelligent system for medical diagnosis. In: Gabrys, B., Howlett, R.J., Jain, Lakhmi C. (eds.) KES 2006. LNCS, vol. 4251, pp. 968–975. Springer, Heidelberg (2006). doi:10.1007/11892960_116

Carlin, S., Komorowski, J., Ohrn, A.: Rough set analysis of medical datasets: a case of patients with suspected acute appendicitis. In: Proceedings of Workshop on Intelligent Data Analysis in Medicine and Pharmacology (ECAI 1998), Brighton, UK, pp. 18–28 (1998)

Cios, K., Pedrycz, W., Swiniarski, R.: Data Mining Methods for Knowledge Discovery. Kluwer Academic, Boston (1998)

Devijver, P.A., Kittler, J.: Pattern Recognition: A Statistical Approach. Prentice-Hall, London (1982)

de Castro, P.A.D., de Franca, F.O., Ferreira, H.M., Von Zuben, F.J.: Applying biclustering to perform collaborative filtering. In: Proceedings of the Seventh International Conference on Intelligent Systems Design and Applications, pp. 421–426 (2007)

Hannah Inbarani, H., Nizar Banu, P.K.: Unsupervised feature selection using tolerance rough set based relative reduct. In: Proceedings of the International Conference on Advances in Engineering, Science and Management, pp. 326–331 (2012)

Hannah Inbarani, H., Nizar Banu, P.K., Azar, A.T.: Feature selection using swarm-based relative reduct technique for fetal heart rate. Neural Comput. Appl. **24**(7-8), 1–14 (2014)

Hartigan, J.: Clustering Algorithms. Wiley, New York (1975)

Jain, A.K., Dubes, R.C.: Algorithms for Clustering Data. Prentice-Hall, Englewood Cliffs (1988)

Lavrajc, N., Keravnou, E., Zupan, B.: Intelligent Data Analysis in Medicine and Pharmacology. Kluwer Academic, Boston (1997)

Kala, R., Shukla, A., Tiwary, R.: A novel approach to clustering using genetic algorithm. Int. J. Eng. Res. Industr. Appl. **3**(1), 81–88 (2010)

Lu, Y., Lu, S., Fotouhi, F., Deng, Y., Brown, S.: Fast genetic Kmeans algorithm and its application in gene expression data analysis. In: Proceedings of the ACM Symposium on Applied Computing (SAC). Technical report. TR-DB-06-2003 (2004)

McQueen, J.: Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium Mathematical Statistics and Probability, vol. 1, pp. 281–297 (1967)

Nizar Banu, P.K., Andrews, S.: Enhanced swarm based feature selection for clustering gene expression data. Int. J. Appl. Eng. Res. **10**(82), 523–527 (2015)

Nizar Banu, P.K., Hannah Inbarani, H.: A comparative analysis of rough set based intelligent techniques for unsupervised gene selection. Int. J. Syst. Dyn. Appl. **2**(4), 33–46 (2013)

Nizar Banu, P.K., Hannah Inbarani, H., Azar, A.T., Own, H.S., Hassanien, A.E.: Rough set based feature selection for Egyptian neonatal jaundice. Adv. Mach. Learn. Technol. Appl., Commun. Comput. Inf. Sci. **488**, 367–378 (2014). doi:10.1007/978-3-319-13461-1_35

Own, H.S., Abraham, A.: A new weighted rough set framework based classification for Egyptian neonatal jaundice. Appl. Soft Comput. **12**, 999–1005 (2012)

Bellazi, R., Zupan, B.: Towards knowledge-based gene expression data mining. J. Biomed. Inform. **40**, 787–802 (2007)

Shamir, R., Sharan, R.: Algorithmic Approaches to Clustering Gene Expression Data, Current Topics in Computational Biology, pp. 269–300. MIT Press, Boston (2001)

Tang, R., Fong, S., Yang, X.-S., Deb, S.: Integrating nature-inspired optimization algorithms to k-means clustering. In: IEEE Seventh International Conference on Digital Information Management (ICDIM 2012), Macau, pp. 116–123 (2012)

El-Beshbishi, S.N., Sharttuck, K.E., Mohammad, A.A., Petersen, J.R.: Hyperbilirubinemia and transcutaneous bilirubinometry. Clin. Chem. **55**(7), 1280–1287 (2009)

Sharma, S., Bhatnagar, V.: Evolutionary and incremental clustering techniques for analysis of dynamic networks: a comparative study. Int. J. Comput. Syst. Eng. **1**(2), 139–150 (2012)

Shejuti Khan, S.M., Monzurur Rahman, M., Tanim, F., Ahmed, F.: Factors influencing K means algorithm. Int. J. Comput. Syst. Eng. **1**(4), 217–228 (2013)

Shrivastava, S.: Diagnosis of neonatal jaundice using artificial neural networks. Int. Indexed Refereed Res. J. **4**, 43–44 (2013a)

Shrivastava, S.: Diagnosis of neonatal jaundice using artificial neural networks. Int. Indexed Refereed Res. J. **4**, 69–72 (2013b)

Ashok, S., Judy, M.V.: A novel iterative partitioning approach for building prime clusters. Int. J. Adv. Intell. Paradigms **7**(3–4), 313–325 (2015)

Hao-jun, S., Lang-huan, X.: Genetic algorithm based high-dimensional data clustering technique. Proc. Int. Conf. Fuzzy Syst. Knowl. Discov. **1**, 485–489 (2009)

Velayutham, C., Thangavel, K.: Unsupervised quick reduct algorithm using rough set theory. J. Electron. Sci. Technol. **9**(3), 193–201 (2011)

Velayutham, C., Thangavel, K.: Unsupervised feature selection using rough set. In: Proceedings on International Conference-Emerging Trends in Computing, pp. 307–314 (2011b)

Whitley, L.D.: A genetic algorithm tutorial. Stat. Comput. **4**, 65–85 (1994)

Yannis, M., Magdalene, M., Michael, D., Nikolaos, M., Constantin, Z.: A hybrid stochastic genetic–GRASP algorithm for clustering analysis. Int. J. Oper. Res. **8**, 33–46 (2008)

Chang, Y.-H., Hsieh, W.-S., Chou, H.-C., Chen, C.-Y., Jing-Yi, W., Tsao, P.-N.: The effectiveness of a noninvasive transcutaneous bilirubin meter in reducing the need for blood sampling in Taiwanese neonates. Clin. Neonatal. **13**(2), 60–63 (2006)

Zalik, K.R., Zalik, B.: Validity index for clusters of different sizes and densities. Pattern Recogn. Lett. **32**, 221–234 (2011)

Zhong, N., Skowron, A.: Rough sets based knowledge discovery process. Int. J. Appl. Math. Comput. Sci. **11**(3), 603–619 (2001)

Azar, A.T., EI-Said, S.A.: Performance analysis of support vector machines classifiers in breast cancer mammography recognition. Neural Comput. Appl. **24**(5), 1163–1177 (2014)

Azar, A.T., Hassanien, A.E.: Dimensionality reduction of medical big data using neural-fuzzy classifier. Soft Comput. **19**(4), 1115–1127 (2014)

Image, Text and Signal Processing

Nonlinear Fourth-Order Diffusion-Based Model for Image Denoising

Tudor Barbu(✉)

Institute of Computer Science of the Romanian Academy, Iasi, Romania
tudor.barbu@iit.academiaromana-is.ro

Abstract. An effective fourth-order PDE-based scheme for image restoration is proposed in this article. First a novel PDE variational model is described. Then, a nonlinear fourth-order diffusion model is obtained from it. A robust explicit numerical approximation scheme based on the finite-difference method and converging fast to the solution of this PDE is then developed for this differential model. The proposed diffusion restoration scheme provide an effective noise removal that also overcome the unintended effects, as resulting from the performed experiments and method comparison.

Keywords: Fourth-order PDE · Image restoration · Nonlinear diffusion · Numerical approximation algorithm · Variational scheme

1 Introduction

This paper approaches the partial differential (PDE) based image restoration domain, a nonlinear diffusion image denoising model being proposed here. The nonlinear PDE-based have been widely used for image enhancement in the last 25 years [1].

Since they perform the diffusion along the edges and not across them, these PDE models preserve the boundaries very well during the denoising process. Unlike the conventional image filters [2], they overcome successfully the image blurring. Since Perona and Malik introduced their influential anisotropic diffusion algorithm [3], and Rudin, Osher and Fetami developed the well-known Total Variation (TV) Denoising scheme [4], numerous second-order nonlinear diffusion approaches and PDE variational models have been proposed for image restoration [5].

While the second-order PDE-based denoising approaches remove successfully the Gaussian noise and avoid the blurring effect, they often generate the unintended staircase (or *blocky*) effect. Some improved second-order diffusion-based techniques that alleviate this undesired effect have been proposed, but the best staircasing overcoming solution is represented by the nonlinear fourth-order PDE models.

The most popular fourth-order PDE denoising model is the isotropic diffusion scheme elaborated by You and Kaveh [6]. The LLT scheme, elaborated by Lasaker et al. represents another influential restoration technique [7].

We also developed some effective second-order and fourth-order nonlinear diffusion-based restoration methods, in our past works [8–10].

V.E. Balas et al. (eds.), *Soft Computing Applications*, Advances in Intelligent Systems and Computing 633, DOI 10.1007/978-3-319-62521-8_36

In this paper we describe a novel fourth-order PDE-based approach for image enhancement. The proposed nonlinear diffusion model, obtained from a PDE variational scheme, is described in the next section. Then, a robust numerical approximation approach is provided in the third section of this paper.

The denoising experiments and method comparisons that prove our method effectiveness are presented in the fourth section. This article ends with conclusions and a section of references.

2 Nonlinear Fourth-Order PDE-Based Denoising

In this section we obtain a nonlinear diffusion-based image restoration model from a PDE-based variational scheme [5, 11]. So, let us consider the following variational model that minimizes an energy cost functional:

$$u^* = \arg\min_u \int_\Omega \left(\frac{\alpha}{2}\xi_u(\|\Delta u\|) + \frac{\lambda}{2}(u-u_0)^2\right) d\Omega \tag{1}$$

where $\alpha, \lambda \in (0,1)$, u_0 is the observed image, the domain $\Omega \subseteq R^2$ and the regularizer of the model, ξ, represents an increasing function.

Then, a nonlinear PDE model is derived from this variational scheme, by determining the corresponding Euler-Lagrange equation. So, from (1) one gets the next Euler-Lagrange equation:

$$\alpha\nabla^2\left(\frac{\xi_u'(|\Delta u|)}{|\Delta u|}\nabla^2 u\right) + \lambda(u-u_0) = 0 \tag{2}$$

Let us consider $\varphi_u(s) = \frac{1}{s}\frac{\partial \xi_u(s)}{\partial s}$, therefore one obtains:

$$\alpha\nabla^2\left(\varphi_u(|\Delta u|)\nabla^2 u\right) + \lambda(u-u_0) = 0 \tag{3}$$

Then, by applying the gradient descent method and adding some boundary conditions [11], one obtains the following fourth-order PDE model:

$$\begin{cases} \frac{\partial u}{\partial t} + \alpha\Delta\left(\varphi_u\left(\left|\nabla^2 u\right|\right)\Delta u\right) + \lambda(u-u_0) = 0 \\ u(0,x,y) = u_0(x,y), \forall (x,y) \in \Omega \\ u(t,x,y) = 0, \forall (x,y) \in \partial\Omega \end{cases} \tag{4}$$

where the Laplacian $\Delta = \nabla^2$ and $\partial\Omega$ represents the frontier of the domain Ω.

The diffusivity function of this differential model, $\varphi_u : [0,\infty) \to (0,\infty)$, is properly constructed for an effective image denoising, being expressed in the following form:

$$\varphi_u(s) = \frac{\beta}{\left(\frac{s}{v(u)}\right)^k + v(u)\left|\log_{10}\left(\frac{s}{v(u)}\right)^k\right|} \tag{5}$$

with

$$v(u) = \gamma(median\,(|\nabla u|) + \mu(\|\Delta u\|)), \tag{6}$$

where μ returns the average value, *median* computes the median value and the parameters $\beta, k, \gamma \in [1, 3)$.

Thus, the denoised image is determined by solving this partial differential equation. The PDE-based model provided by (1)–(3) is well-posed, admitting an unique and weak solution. This solution, representing the enhanced image, is determined in the next section, where a numerical approximation of the continuous model is proposed.

3 Consistent Numerical Approximation

The described fourth-order PDE model is discretized by applying a finite-difference based numerical approximation scheme [12]. Let us consider a space grid size of h and a time step Δt, the space and time coordinates being quantized as follows:

$$x = ih, y = jh, t = n\Delta t, \forall i \in \{0, .., I\},\ j \in \{0, ., J\}, n \in \{0, .., N\} \tag{7}$$

The equation $\frac{\partial u}{\partial t} + \alpha\Delta\left(\varphi_u\left(\left|\nabla^2 u\right|\right)\Delta u\right) + \lambda(u - u_0) = 0$ is then approximated by using finite differences [12]. The Laplacian of the current image is discretized as:

$$\Delta u_{i,j}^n = \frac{u^n(i+1,j) + u^n(i-1,j) + u^n(i,j+1) + u^n(i,j-1) - 4u^n(i,j)}{h^2} \tag{8}$$

One then computes $\zeta_{i,j}^n = \varphi_u\left(\left|\nabla^2 u_{i,j}^n\right|\right)\Delta u_{i,j}^n$ for $\forall\, n \in \{0, .., N\}$, so

$$\Delta\left(\varphi_u\left(\left|\nabla^2 u_{i,j}^n\right|\right)\Delta u_{i,j}^n\right) = \frac{\zeta_{i+1,j}^n + \zeta_{i-1,j}^n + \zeta_{i,j+1}^n + \zeta_{i,j-1}^n - 4\zeta_{i,j}^n}{h^2} \tag{9}$$

Also, the term $\frac{\partial u}{\partial t} + \lambda(u - u_0)$, is discretized as follows:

$$\frac{u_{i,j}^{n+\Delta t} - u_{i,j}^n}{\Delta t} + \lambda\left(u_{i,j}^n - u_{i,j}^0\right) = \frac{u_{i,j}^{n+\Delta t} + (\lambda\Delta t - 1)u_{i,j}^n - \lambda\Delta t u_{i,j}^0}{\Delta t} \tag{10}$$

We could take the parameter values $h = 1$ and $\Delta t = 1$. So, one obtains the following numerical approximation for the PDE:

$$u^{n+1}(i,j) = (1-\lambda)u^n(i,j) - \alpha(\zeta^n_{i+1,j} - \zeta^n_{i-1,j} - \zeta^n_{i,j+1} + 4\zeta^n_{i,j}) - \lambda u_0(i,j) \tag{11}$$

So, we get an explicit numerical approximation scheme expressed that is consistent to the fourth-order PDE model (1). It also converges fast to the PDE solution, the number of iterations, N, being quite low. This iterative algorithm applies operation (11) on the degraded $[I \times J]$ image, for each $n \in \{0, \ldots, N\}$, until the optimal restoration u^{N+1}, is finally obtained.

4 Experiments

We have tested the nonlinear diffusion-based approach presented here on several hundreds of images corrupted by Gaussian noise. Satisfactory denoising results have been achieved.

The proposed technique reduces considerably the noise, while overcoming the unintended effects, such as blurring, staircasing or speckling, and preserving the boundaries and other important features. Our restoration algorithm executes quite fast, its average runtime being less than 1 s and the number of iterations, N, being low.

The performance of our restoration algorithm is measured by using some image similarity metrics [13], such as the Peak Signal-to-Noise Ratio (PSNR), Norm of the Error Image and Structural Similarity (SSIM) index. Our technique outperforms the conventional two-dimension filters and many linear and nonlinear PDE schemes, producing much better values for these performance measures. We describe here only the PSNR related results. The obtained PSNR values are displayed in Table 1.

Table 1. PSNR values corresponding to several models

Denoising method	PSNR
This fourth-order PDE	26.95 (dB)
Average	25.43 (dB)
Gaussian	25.17 (dB)
Median 2D	25.53 (dB)
Perona-Malik	25.91 (dB)
Tv Denoising	26.17 (dB)
You-Kaveh	26.73 (dB)

Our technique is much better than 2D classic filters, like 2D Gaussian, Average or Median 2D [2], avoiding the blurring, and preserving the details. It outperforms also some nonlinear second- and fourth-order PDE-based schemes, like Perona-Malik model [3], TV Denoising [4] and the You-Kaveh isotropic diffusion algorithm [6], by avoiding the blocky effect and speckle noise, and operating faster than these denoising techniques.

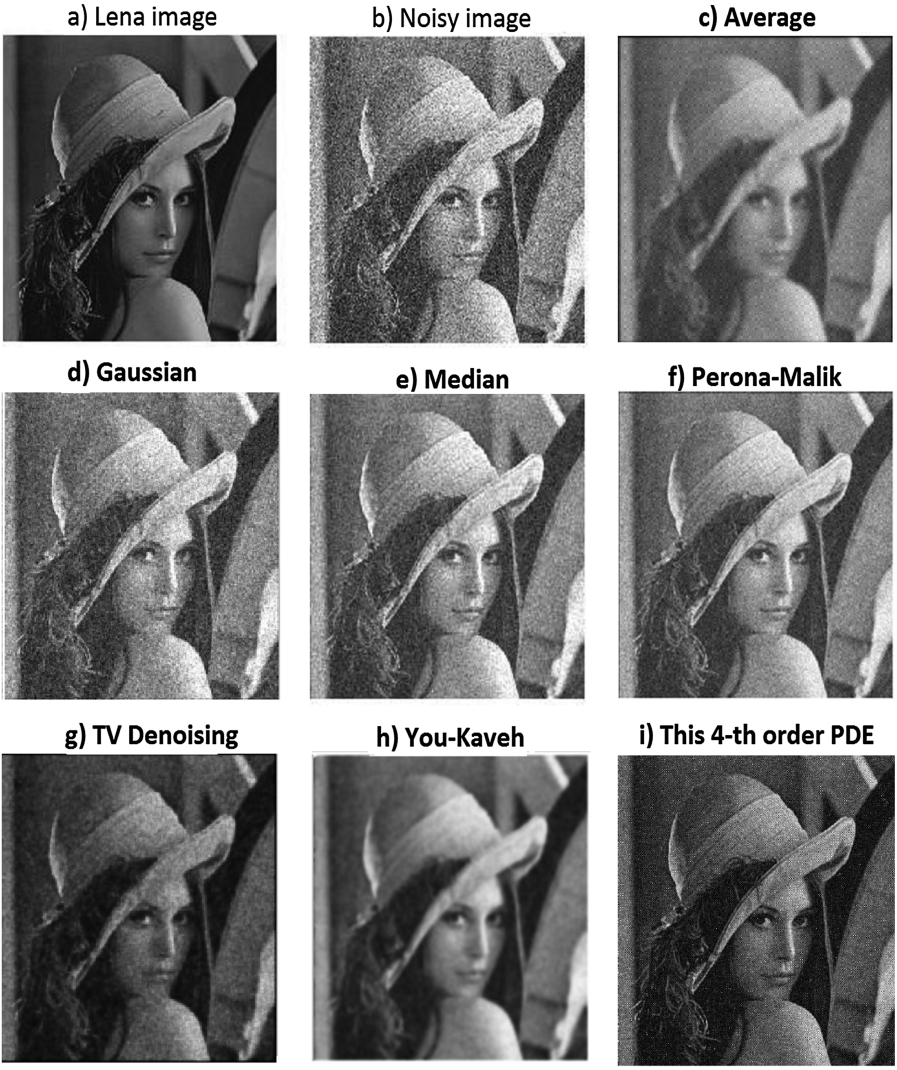


Fig. 1. Denoising results of various filtering methods

The smoothing results produced by the mentioned filtering methods on the *Lena* image are displayed in Fig. 1: (a) – the original image; (b) that image affected by the Gaussian noise with parameters $\mu = 0.21$ and *variance* = 0.02; (c) – Average filtering; (d) – Gaussian filtering; (e) – the Median 2D filter; (f) - Perona-Malik smoothing; (g) - TV Denoising; (h) - You-Kaveh model and (i) – our PDE-based model.

5 Conclusions

A novel nonlinear fourth-order diffusion image denoising model was described in this paper. The proposed PDE-based technique removes successfully the Gaussian noise, while preserving the essential features and avoiding all the undesired effects.

The proposed variational denoising scheme and the fourth-order PDE model achieved from it represent the most important contributions of our article. The explicit and fast-converging finite difference method - based numerical approximation algorithm developed here constitutes another contribution of this paper.

This restoration approach is a better smoothing solution than conventional methods, since it overcomes the unintended effects. Also it outperforms some nonlinear diffusion-based techniques, by alleviating the staircase effect, image blurring and the speckle noise, and also by running much faster than those methods.

While our smoothing method works satisfactory for Gaussian noise removal, it does not have the same efficiency with other noise types. But the proposed denoising model can be further improved or transformed such that to work properly for other types of noise, also.

Thus, we will consider new diffusivity functions for this differential model. We will also try to combine this fourth-order PDE-based scheme to second-order diffusion models, into more effective restoration systems, as part of our future research.

Acknowledgement. The research work described here was supported from the project PN-II-ID-PCE-2011-3-0027, which is financed by the Romanian Minister of Education and Technology.

We also acknowledge the research support of the Institute of Computer Science of the Romanian Academy, Iaşi, ROMANIA.

References

1. Weickert, J.: Anisotropic Diffusion in Image Processing. European Consortium for Mathematics in Industry. B.G. Teubner, Stuttgart (1998)
2. Gonzalez, R., Woods, R.: Digital Image Processing, 2nd edn. Prentice Hall, Upper Saddle River (2001)
3. Perona, P., Malik, J.: Scale-space and edge detection using anisotropic diffusion. In: Proceedings of IEEE Computer Society Workshop on Computer Vision, pp. 16–22, November 1987
4. Rudin, L., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. Phys. D: Nonlinear Phenom. **60**(1), 259–268 (1992)
5. Chan, T., Shen, J., Vese, L.: Variational PDE models in image processing. Not. AMS **50**(1), 14–26 (2003)
6. You, Y.L., Kaveh, M.: Fourth-order partial differential equations for noise removal. IEEE Trans. Image Process. **9**, 1723–1730 (2000)
7. Lysaker, M., Lundervold, A., Tai, X.C.: Noise removal using fourth-order partial differential equation with applications to medical magnetic resonance images in space and time. IEEE Trans. Image Process. **12**, 1579–1590 (2003)

8. Barbu, T.: Robust anisotropic diffusion scheme for image noise removal. Procedia Comput. Sci. **35**, 522–530 (2014). Proceedings of 18th International Conference in Knowledge Based and Intelligent Information & Engineering Systems, KES 2014, 15–17 September, Gdynia, Poland. Elsevier
9. Barbu, T., Favini, A.: Rigorous mathematical investigation of a nonlinear anisotropic diffusion -based image restoration model. Electron. J. Differ. Eqn. **129**, 1–9 (2014)
10. Barbu, T.: PDE-based restoration model using nonlinear second and fourth order diffusions. Proc. Rom. Acad. Ser. A: Math. Phys. Tech. Sci Inf. Sci. **16**(2), 138–146 (2015)
11. Hazawinkel, M.: Variational Calculus. Encyclopedia of Mathematics. Springer, Heidelberg (2001)
12. Johnson, P.: Finite Difference for PDEs. School of Mathematics, University of Manchester (2008)
13. Thung, K.H., Raveendran, P.: A survey of image quality measures. In: Proceedings of the International Conference for Technical Postgraduates (TECHPOS), pp. 1–4 (2009)

On Time-Frequency Image Processing for Change Detection Purposes

Dorel Aiordachioaie(✉)

Electronics and Telecommunication Department, "Dunărea de Jos" University of Galați, 47 Domneasca Street, 80008 Galati, Romania
Dorel.Aiordachioaie@ugal.ro

Abstract. Change detection and diagnosis are important activities in many domains, such as mechanical, aerospace, biomedical and seismic engineering. Process monitoring systems based on vibration measuring and processing use various transforms, for instance time-frequency transforms. Wigner-Ville distribution is used as an example of time-frequency transform, but it generates many interference terms. Some basic properties of the input signal, concerning the number of components, the average time and frequency, the bandwidth and time duration are estimated by computing statistical moments and the Renyi entropy. The paper also presents some results of the evaluation of an automatic system to detect and remove artifacts from time-frequency images. The point is to build a database with blocks, whose size depend on application, with and without interferences (bad blocks), and - by cross-correlation - to detect these blocks in the processed image. In the present state of development, the system is working for some particular structures, concerning the number and the parameters of the analyzed signal components. Adding knowledge about the analyzed signal, the performances can be improved. The results are encouraging and suggest optimization of the proposed method.

Keywords: Time-frequency transform · Image · Detection · Signal processing · Pattern recognition

1 Introduction

Change detection and diagnosis (CDD) are important scientific and technical problems. They are present in the monitoring of some usual vibrational systems, and in many complex pieces of equipment as well. They could be associated with high risk activities, such as: energy production (thermo, hydro, nuclear), crude oil extraction, chemical plants, high power machinery and industrial equipment, etc. (see some reference papers, such as [1–3], with different points of view concerning system theory, signal processing, or system engineering). Many applications on this subject make use of theories based on statistics, [4], which provide the theoretical instruments for solving the problem of early detection. Such a theory is referred to as the local approach [5] and is based on the transformation of the general detection problem into a classical problem of monitoring the mean of a Gaussian vector variable. Another approach is based on the signal-processing paradigm, signals coming from the observed process, most commonly vibration signals. It is also the case of this work.

V.E. Balas et al. (eds.), *Soft Computing Applications*, Advances in Intelligent Systems and Computing 633, DOI 10.1007/978-3-319-62521-8_37

The vibrations analysis and surveillance of the correct functioning of machinery or industrial equipment represent important cases in the detection and diagnosis problems, e.g. [6]. There are two basic approaches in CDD, which can be shortly described as based on quantitative models (using analytical redundancy), see e.g. [7], and qualitative models. Regarding the existing challenges of the CDD domain, there is still a "gap" between the theoretical results and the applications. This is mainly caused by the requirements of some "strong" hypotheses in the existing algorithms, which are not easily verified in practice. A main challenge is the time restriction in the case of real time applications, when low computation effort is possible and reduced order models have to be used, but without reducing the performance of the detection and diagnosis system.

The general method for CDD based on signal processing starts with an acquisition step of the vibration signals and continues with processing steps to define a signature of the monitored equipment, or of an individual component, e.g. a gearbox. The time-frequency distribution of the vibration signals is treated as an image. CDD is performed by interpreting the patterns of the image by the various procedures from pattern recognition engineering field, such as segmentation, feature extraction, and classification.

One of the most simple time-frequency representations is the Wigner-Ville distribution (WVD). The transform is non-linear; therefore any two components at different times or different frequencies are bound to interfere, so the distribution becomes clouded with false images. The WVD is incapable of indicating the true existence for all components at certain time and frequency coordinates, or of representing true causality between any sections of a signal. These interferences cause difficulties in interpreting the distribution, and represent an obstacle to the automatic interpretation of the vibration signature, [8, 9].

In the field of CDD, a research project called VIBROCHANGE is under development, [10]. It supports the development of the toolbox VIBROTOOL and an experimental model, VIBROMOD. VIBROTOOL is a toolbox for the CDD problem, built as a set of programs that compute specific parameters and solve specialized tasks for change detection and diagnosis, and point changing estimation as well. The algorithms make use of classical techniques (pattern recognition, maximum likelihood, model-based techniques, novelty detection, and blind separation) and some of the more recent techniques (multiresolution analysis, soft computing, and information fusion).

The experimental model implements some of the algorithms developed in VIBROTOOL, with simplified versions and requires small computation resources. It is the reason to use and implement some simple time-frequency transforms, such as WVD, even if the global properties are not perfect as stated above. In addition, some software tools are developed in order to assist the operator in processing and reading the time-frequency image. This is the context which supports the considerations of this paper.

The paper is organized as follows: Sect. 2 highlights the basic elements of time-frequency transforms. Section 3 considers some aspects of signal processing, for information extraction from time-frequency images. Section 4 presents the main results obtained by computer simulation, regarding the estimation of the number of the independent components from an image and automatic identification and correction of the interference and real components.

2 Time–Frequency Transforms

If *x(t)* is a continuous (possible complex) signal, the Wigner distribution (WD) of the signal *x(t)* is defined as

$$W_x(t,f) = \int_{-\infty}^{\infty} x\left(t+\frac{\tau}{2}\right)\cdot x^*\left(t-\frac{\tau}{2}\right)\cdot e^{-j\cdot 2\pi\cdot f\cdot\tau}d\tau \tag{1}$$

where * denotes complex conjugation. If *X(f)* is the Fourier transform of *x(t)*, an equivalent definition is

$$W_x(t,f) = \int_{-\infty}^{\infty} X\left(f+\frac{\eta}{2}\right)\cdot X^*\left(f-\frac{\eta}{2}\right)\cdot e^{j\cdot 2\pi\cdot t\cdot\tau}d\eta \tag{2}$$

For the case where *x(t)* is an analytical signal, the Wigner distribution is termed the Wigner-Ville distribution (WVD), [13]. This distribution satisfies a large number of desirable mathematical properties, as described in the specialized literature, e.g. [11] or [12]. In particular, the WVD is always real-valued; it preserves time and frequency shifts and satisfies the marginal properties.

Two properties related to the involved processing signal are presented: (i) the possibility to recover the instantaneous frequency of a signal x from the WVD as its first order moment (or center of gravity) in frequency:

$$f_x(t) = \int_{-\infty}^{\infty} v\cdot W_{x_a}(t,v)dv \Big/ \int_{-\infty}^{\infty} W_{x_a}(t,v)dv \tag{3}$$

where x_a is the analytic signal associated to x; (ii) the group delay of x can be obtained as the first order moment in time of its WVD:

$$t_x(v) = \int_{-\infty}^{\infty} t\cdot W_{x_a}(t,v)dt \Big/ \int_{-\infty}^{\infty} W_{x_a}(t,v)dv \tag{4}$$

The imaginary part $x_I(t)$ is equal to the Hilbert transform of the real part $x_R(t)$ so that

$$x_I(t) = \frac{1}{\pi}\int_{-\infty}^{\infty}\frac{x_R(\tau)}{t-\tau}\cdot d\tau \tag{5}$$

The discrete WVD $W(n,m)$ is defined as

$$W(n,m) = \frac{1}{2N}\sum_{k=0}^{N-1} x(kT)\cdot x^*((n-k)T)\exp\left(-\frac{j\pi\cdot m\cdot(2k-n)}{N}\right) \tag{6}$$

It is easy to verify that $W(n,m)$ is a periodic function of period $2N$ in both time and frequency. In the range $0\le n<2N-1, 0\le m<2N-1$, representing one complete period, the WVD needs only be calculated over the range $0\le n<N-1, 0\le m<N-1$, having an area of one quarter of the area of the complete period.

The definition (1) requires knowledge from −inf to + inf, which can be a problem in practice. By using a windowed observation frame, this is leading to the new distribution:

$$PW_x(t,f) = \int_{-\infty}^{\infty} h(t)\cdot x\left(t+\frac{\tau}{2}\right)\cdot x^*\left(t-\frac{\tau}{2}\right)\cdot e^{-j\cdot 2\pi\cdot f\cdot\tau}d\tau \tag{7}$$

where $h(t)$ is a regular window. This distribution is called the Pseudo Wigner-Ville distribution (PWVD).

The WVD interference terms will be non-zero regardless of the time-frequency distance between the two signal terms. These interference terms are troublesome since they may overlap with auto-terms (signal terms) and thus they make it difficult to visually interpret the TF image.

The rule of interference construction of the WVD can be summarized as follows: two points of the time-frequency plane interfere to create a contribution on a third point, which is located at their geometrical midpoint. The interference terms oscillate perpendicularly to the line joining the two interfering points, with a frequency proportional to the distance between these two points, [13].

Thus, because of their oscillating nature, the interferences will be attenuated in the pseudo-WVD compared to the WVD. The consequence of this improved readability is that many properties of the WVD are lost: the marginal properties, the unitarity, and the frequency-support conservation; the frequency-widths of the auto-terms are increased by this operation, [13].

The interference terms present in any quadratic time-frequency representation, even if they disturb the readability of the representation, contain some information about the analyzed signal. The precise knowledge of their structure and construction rule is useful to interpret the information that they contain, [13]. For instance, the interference terms contain some information about the phase of a signal.

Even if all TDFs tend to the same goal, each representation has to be interpreted differently, according to its own properties. For example, some of them present important interference terms, others are only positive, others are perfectly localized on particular signals, etc. The extraction of information has to be done with care, from the knowledge of these properties.

3 Extraction of Information

A. Statistical Parameters

Depending on the statistic properties of the analyzed signal, especially the property of stationarity, the interpretation of the time-frequency image, which describes the evolution with time of the frequency content of the signal, could be a problem, [14].

Some statistical moments, in time and in frequency, of a time-frequency image, could be computed immediately. A simple way to characterize a signal simultaneously in time and in frequency is to consider its mean localizations and spreading in each of these representations. This can be obtained by considering $|x(t)|^2$ and $|X(f)|^2$ as probability distributions, and looking at their mean values and standard deviations. In the time domain, there is the average

$$t_m = \frac{1}{E_x} \int_{-\infty}^{+\infty} t \cdot |x(t)|^2 \cdot dt \cong \frac{1}{E_x} \sum_{k=1}^{N} t[k] \cdot |x[k]|^2 \cdot Ts \tag{8}$$

and time spreading

$$T^2 = \frac{4\pi}{E_x} \int_{-\infty}^{+\infty} (t - t_m)^2 \cdot |x(t)|^2 \cdot dt \cong \frac{4\pi}{E_x} \sum_{k=1}^{N} (t[k] - tm)^2 \cdot |x[k]|^2 \cdot Ts \tag{9}$$

In the frequency domain, the average frequency is:

$$f_m = \frac{1}{E_f} \int_{-\infty}^{+\infty} f \cdot |X(f)|^2 \cdot df \cong \frac{1}{E_f} \sum_{i=1}^{N} f[i] \cdot |X[i]|^2 \cdot \Delta f \tag{10}$$

and the frequency spreading is:

$$B^2 = \frac{4\pi}{E_f} \int_{-\infty}^{+\infty} (f - f_m)^2 |X(f)|^2 df \cong \frac{4\pi}{E_f} \sum_{i=1}^{N} (f[i] - f_m)^2 \cdot |X(i)|^2 \cdot \Delta f \tag{11}$$

where E_x is the energy of the signal, assumed to be finite (bounded), which respects the equations:

$$E_x = \int_{-\infty}^{+\infty} |x(t)|^2 \cdot dt \cong Ts \cdot \sum_{k=1}^{N} |x[k]|^2 < \infty \tag{12}$$

$$E_f = \int_{-\infty}^{+\infty} |X(f)|^2 \cdot df \cong \Delta f \cdot \sum_{k=1}^{N} |X[k]|^2 < \infty \tag{13}$$

where $|x(t)|^2$ and $|X(f)|^2$ are interpreted as energy densities, respectively in time and in frequency. Then a signal can be characterized in the time-frequency plane by its mean position (t_m, f_m) and a domain of main energy localization whose area is proportional to the time-bandwidth product $T \cdot B$ (always >1). For details, please see [13, 14] or [15]. They describe the averaged positions and spreads in time and in frequency of the signal. For some particular distributions, if the signal is considered in its analytic form, the first order moment in time also corresponds to the instantaneous frequency, and the first order moment in frequency to the group delay of the signal.

B. Renyi Entropies

The Renyi entropy could be used as

$$R_x^{\alpha} = \frac{1}{1-\alpha} \log \left[\int_{-\infty}^{\infty} f^{\alpha}(x) dx \right] \tag{14}$$

where α is the order of information. Third order Renyi information, applied to a time-frequency distribution $C_x(t, f)$, is defined as

$$R_C^{\alpha} = -\frac{1}{2} \log \left[\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} c_x^3(t,f) dt df \right] \tag{15}$$

with

$$c_x(t,f) = C_x(t,f) \Big/ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} C_x(t,f) dt df \tag{16}$$

being the equivalent/associated probability density function. The discrete case, with dT sampling period and dF the resolution on frequency axis, is

$$R_C^{\alpha} = -\frac{1}{2} \log \left[dT \cdot dF \sum_{i=1}^{N} \sum_{j=1}^{N} c_x^3(i,j) \right] \tag{17}$$

and

$$c_x(i,j) = C_x(i,j) \Big/ \left[dT \cdot dF \sum_{i=1}^{N} \sum_{j=1}^{N} C_x(i,j) \right] \tag{18}$$

Some examples and more considerations are available in [16, 17].

Interesting information that one may need to know about an observed non-stationary signal is the number of elementary signals composing this observation. A solution is given by applying an information measure to a time-frequency distribution of the signal. The result produced by this measure is expressed in bits, [13]: if

one elementary signal yields zero bit of information, then two well-separated elementary signals will yield one bit of information (2^1), four well-separated elementary signals will yield two bits of information (2^2), and so on. In particular, it is possible to show that the Renyi information measure provides a good indication of the time separation at which the atoms are essentially resolved, with a better precision than with the time-bandwidth product, [13].

C. Time-Frequency Image Analysis

The proposed solutions in the literature design a decision test (statistic), based on two distinct approaches.

The first method processes the time-frequency image (TFI) by using the correlation between a TFI of the analyzed signal and some two-dimensional templates, with the same or reduced size, constructed using the a priori information available on the signal. The framework of this method is considered in this paper and is described below. Figure 1 introduces the basic structure of an automatic system for the detection and filtering of the interferences of TFI.

The system is working in two steps. In the first one, the system defines a database

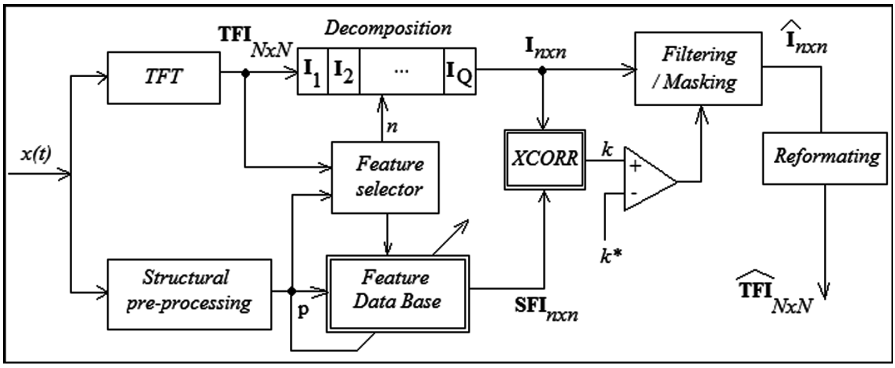


Fig. 1. Structure of the automatic detection and filtering system

with specific features of the processed signal, *x(t)*. The selection process is supervised by a specialized block, the control unit, which is not presented in Fig. 1, for clarity reasons. In the second step, the blocks of the TFI image are compared to the blocks of the database. Once a bad block is detected, it is filtered and/or masked. The recognition of blocks from the image is based on cross correlation.

The structural preprocessing block should estimate the number of components, the information about time and frequency content, and the changes during the analyzed time interval. This information is preserved in a vector **p**.

The TFT block computes the time-frequency transform and generates a 2D signal, which will be considered as an image, **TFI,** of size *NxN*. On rows the frequency, and on columns the distributed time.

The decomposition block reshapes the TFI image, by considering a set of small images, of size *nxn*, in order to recognize and filter the bad blocks. Part of the blocks is stored in a database, as references for good and bad blocks of the TFI signal. The size *n* of the reduced block could be variable, depending on the specific features, which should be detected.

The comparison between two block images is based on cross-correlation. If the correlation coefficient k is greater than an imposed level k^*, then the two blocks are similar, otherwise they are not. Finally, the processed blocks $\hat{\mathbf{I}}$ are assembled by reshaping (reformatting) to generate the filtered time-frequency image.

The second method consists of applying a transform on the time-frequency representation of the analyzed signal, which enhances some characteristics of the signal to be detected, and of applying a test on this new space of decision. An example could be the detection of signals with frequency modulation (variation). If the problem is reduced to the problem of detection and estimation of a line in an image, this can be done by using the Hough transform, dedicated to the detection of lines, [18].

4 Experiments and Results

A. Signal Generation

The test signals are close to the experimentally measured time domain averages of the vibration produced by gears in industrial equipment. The time domain record has $N = 1024$ samples. Possible signals could have a pure sine wave, an amplitude modulated sin wave, and a phase modulated sine wave or Gaussian impulses, as defined by [9]

$$s_1[k] = A_1 \cdot \sin(2\pi \cdot f_1 \cdot k/N) \tag{19}$$

$$s_2[k] = (1 + A_2 \cos(2\pi \cdot f_2 \cdot k/N)) \cdot A_3 \sin(2\pi \cdot f_3 \cdot k/N) \tag{20}$$

$$s_3[k] = A_4 \sin(2\pi \cdot f_4 \cdot k/N + A_5 \sin(2\pi \cdot f_5 \cdot k/N)) \tag{21}$$

$$s_4[k] = A_6 \exp\left(-B_6^2 \cdot (k - \tau_6)^2/N^2\right) \tag{22}$$

with $k = 0, 1, 2, \ldots, N\text{-}1$. Four base simple signals are used in this paper, which are sinusoidal signals modulated by Gaussian impulses, based on equations

$$x_i[k] = A_i \cdot \exp\left(-B_i^2(k - \tau_i)^2 \cdot \sin(2 \cdot \pi \cdot f_i \cdot k/N)\right), i = 1, 2, 3, 4 \tag{23}$$

with

$$\mathbf{A} = [1.5, 0.75, 11.5, 2, 3, 1] \tag{23.a}$$

$$\mathbf{f} = [100, 350, 650, 1000, 0, 0, 0] \tag{23.b}$$

$$\mathbf{B} = [0, 0, 8, 16, 32, 64, 150] \tag{23.c}$$

$$\boldsymbol{\tau} = [0, 0, 0, 100, 300, 600, 900] \tag{23.d}$$

The values of the above vectors were chosen to present clearly the signals and phenomena, but, fortunately, they correspond as order of magnitude to the physical values from the field of the electrical machines.

As case study, ten cases are considered, as they are presented in Table 1. Four individual signals, $\{x_1, x_2, x_3, x_4\}$, plus three combinations of two, $\{x_1 + x_2, x_1 + x_3, x_1 + x_4\}$, plus two combination of three, $\{x_1 + x_2 + x_3, x_1 + x_3 + x_4\}$, and plus one combination of all four base signals, $\{x_1 + x_2 + x_3 + x_4\}$, are considered. Table 1 also presents the possible interference terms, in the *Real* columns, terms which have two indices and a plus sign, e.g. $(+x_{12})$ or $(+x_{13})$.

Figure 2 presents two examples of TFI images, for the cases 5 and 7. For each case, two "correct" disks are presented, which correspond to the real components. In the middle of the real components, there is an interference term, which is an artifact because it is not part of the analyzed signal. The TF transform identifies and extracts correctly the time and frequency content of the input signals.

Table 1. Test signals

No	Case		No	Case	
	Theoretic	Real		Theoretic	Real
1.	x_1	x_1	6.	$x_1 + x_3$	$+x_{13}$
2.	x_2	x_2	7.	$x_1 + x_4$	$+x_{14}$
3.	x_3	x_3	8.	$x_1 + x_2 + x_3$	$+x_{12} + x_{13} + x_{23}$
4.	x_4	x_4	9.	$x_1 + x_3 + x_4$	$+x_{13} + x_{14} + x_{34}$
5.	$x_1 + x_2$	$+x_{12}$	10.	$x_1 + x_2 + x_3 + x_4$	$+x_{12} + x_{13} + x_{14} + + x_{23} + x_{24} + + x_{34}$

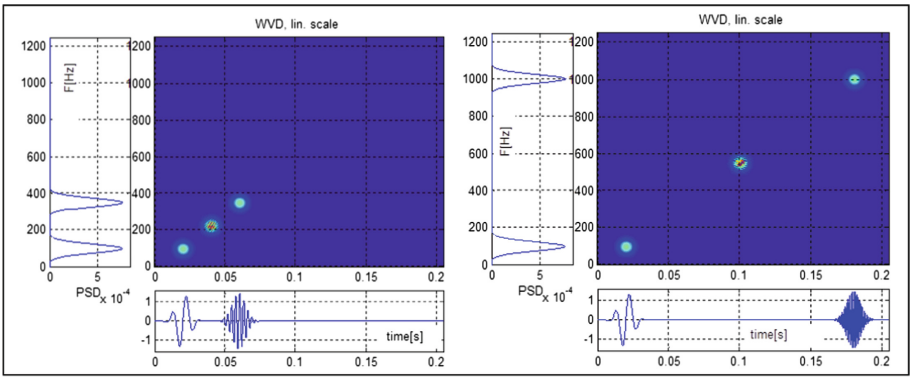


Fig. 2. Time-frequency analysis by WVD transform, cases 5 and 7

Figure 3 presents the results of two different time-frequency distributions, the Wigner-Ville (WVD) and the Pseudo Wigner-Ville (PWVD), as they were defined by Eqs. (1) and (7). The last one decreases the effects of the interference terms. The decrease of the interference terms by slightly changing (e.g. by weighting) the kernel of the transform is not considered here.

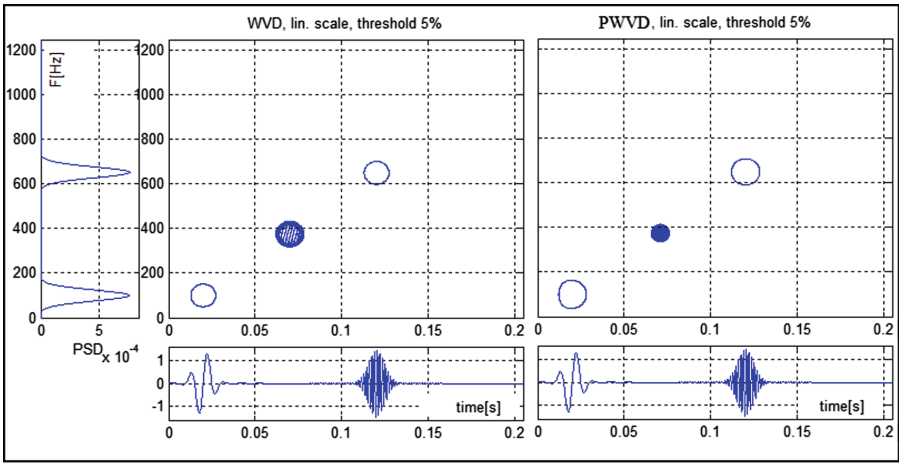


Fig. 3. WVD vs. Pseudo-WVD results, case 6

Starting from these TF images, the processing system should compute some parameters regarding the time width of the signals (impulses, in this study), the frequency content behavior as frequency spectrum and bandwidth. These parameters are considered in the next sub-section. The numerical results are presented in Table 2 for the WV distribution.

Table 2. Average parameters (moments) - for localization and spreading

No	WVD				
	t_m [s]	f_m [Hz]	T [s]	B [Hz]	$B.T$
1.	0.0200	100.0	0.0118	84.6	1.0005
2.	0.0600	350.0	0.0118	84.6	1.0000
3.	0.1200	650.0	0.0118	84.6	1.0000
4.	0.1800	1000.0	0.0118	84.6	1.0000
5.	0.0400	225.0	0.0719	451.1	32.4234
6.	0.0700	375.0	0.1776	978.5	173.8184
7.	0.1000	550.0	0.2838	1597.4	453.4122
8.	0.0667	366.7	0.1462	801.5	117.1482
9.	0.1067	583.4	0.2342	1315.8	308.2299
10.	0.0950	525.0	0.2152	1195.3	257.2468

B. Statistical Measures

The values of t_m and f_m presented in Table 2 are correct for the first four cases, when individual signals were considered. The correctness refers to the physical meaning. When signals with more components are considered, the parameters should be considered as averages. These values are graphically represented in Fig. 4.

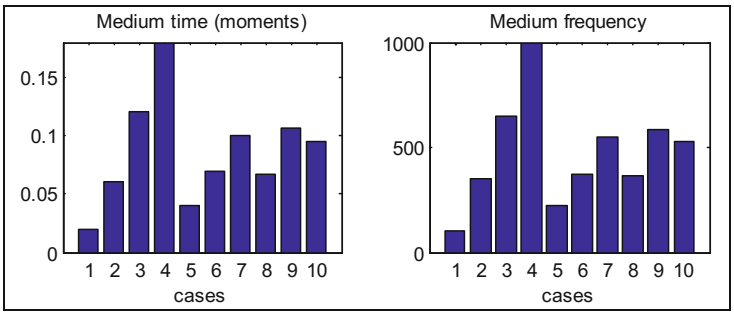


Fig. 4. Graphical representation of parameters t_m and f_m used for localization

For cases where combinations of signals are presented, time and frequency moments, t_m and f_m, are not very relevant. Two examples are presented in Fig. 5 for a single component and, respectively, multicomponent signal. In the case of the single component, t_m and f_m indicate the position of the component in the time-frequency plane.

The parameters T and B indicate the widths of the signal, on time and frequency axes. In the case of multi-component signals, t_m and f_m are averages, and T and B indicate the spreading in time (from the first component to the last one) and frequency (from the lowest to the highest frequency).

The parameters B and T, which indicate the spread, are represented in Fig. 6, and are more appropriate to be used as first descriptor parameters of the multicomponent

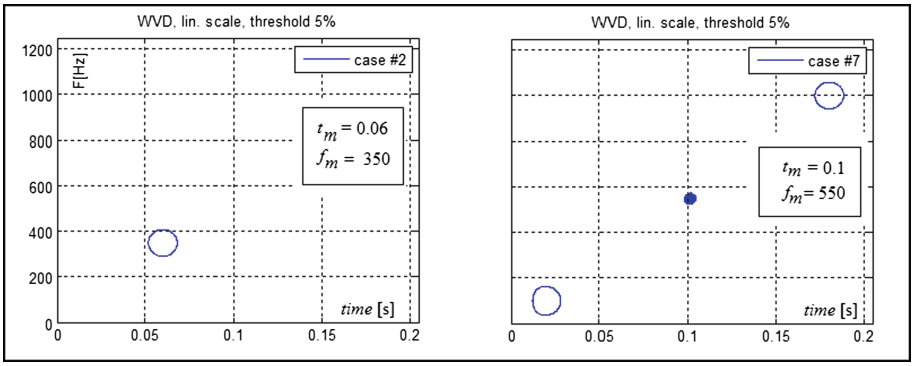


Fig. 5. Average localization in time and frequency for various signals

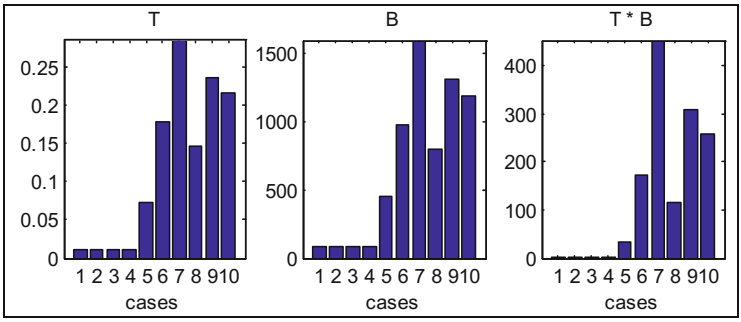


Fig. 6. Representation of parameters B and T, used for spreading information

signals. The maximum value of the time spreading is obtained for case 7, i.e. 0.2838 [s]. This case also generates the maximum bandwidth, of 1597.4 [Hz].

The ratio between t_m and T, or between f_m and B, could be used for a primary description and classification of the analyzed signal. Thus, if $t_m > T$ (or $f_m > B$) then a signal with a single component is more likely to have a component only. Otherwise, the analyzed signal has more components. This statement is verified by the data available in Table 2.

C. Information Measures

Table 3 presents the numerical values of the absolute (*RH*) and differential Renyi entropies (*DRH*). The differential entropy is obtained by subtracting the entropy of the case #1 from the absolute values of all others entropies. Two distributions are considered, namely WVD and PWVD.

Figure 7 shows a comparison between entropies, as solution describing the complexity of the analyzed signal, in all ten considered cases. It is important to notice the superiority of the entropies. If the complexity of the signal is growing (at least as the number of components) then the entropy is rising, too. Moreover, differential entropies

Table 3. Renyi entropies

No	WVD		PWVD	
	RH	DRH	RH	DRH
1.	−0.2012	0.0	−0.0548	0.0
2.	−0.2075	−0.0063	−0.1896	−0.1349
3.	−0.2075	−0.0063	−0.2030	−0.1482
4.	−0.2075	−0.0063	−0.2055	−0.1507
5.	+1.1465	+1.3477	1.2647	+1.3195
6.	+1.1469	+1.3480	0.9987	+1.0535
7.	+1.1469	+1.3481	0.9262	+0.9810
8.	+2.0483	+2.2495	2.0225	+2.0772
9.	+2.0895	+2.2907	1.7881	+1.8429
10.	+2.5921	+2.7933	2.4355	+2.4903

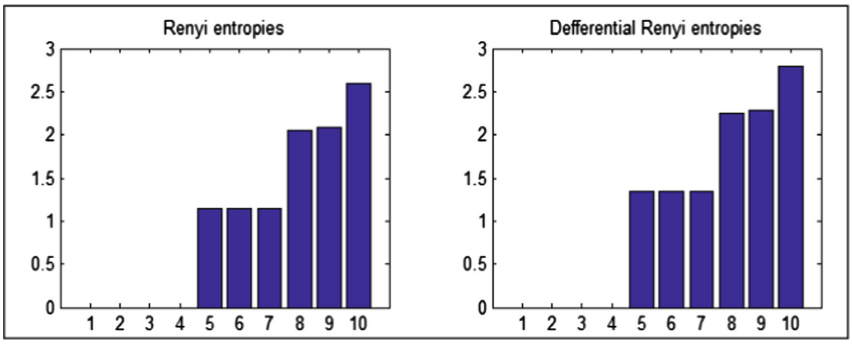


Fig. 7. Evolution of the main descriptors (entropies), information-based

are able to predict the number of the components in the analyzed signal, as it was described in Sect. 3B.

The information obtained from entropies combined with the ratios of statistical parameters, as described above, could drastically improve the results of the signal analysis, both on qualitative (as number of components) and quantitative aspects (localization and width in time and frequency).

D. Detection and Filtering of Interferences

With reference to the system of Fig. 1, let $\mathbf{TI}_{nxn}$ be a block image for testing. Let $\mathbf{bI}_{nxn}$ (bad image) be the block image with artifacts and let $\mathbf{NI}_{nxn}$ (normal image) be the block image with no interferences. The masking decision is based on the decision rule:

$$\begin{aligned} &\text{IF } corr(\mathbf{TI}, \mathbf{bI}) = \max \text{ AND } \quad corr(\mathbf{TI}, \mathbf{NI}) = \min \\ &\quad \text{THEN FILTER } \mathbf{TI} \end{aligned} \tag{24}$$

where *corr(.,.)* is the statistic correlation operator working on square matrices.

Figure 8 presents four block images: two for clean (correct) images, $\mathbf{x}_1$ and $\mathbf{x}_4$, and two block images with interferences, $\mathbf{x}_{12}$ and $\mathbf{x}_{34}$. The interference images have two specific features, which could be used in automatic detection and classification: (i) it contains a lot of parallel ellipses; (ii) The main diagonals are parallel to the segment which connects the centers of the closest neighbored components. This set of features could be used to distinguish the real from false components in time-frequency images. In this case, the self-checking procedure could be used before calling a classifier for recognition purposes.

Figure 9 presents the TFI for case # 8 and the results of the cross-correlation between the TFI and the bad block assigned to $\mathbf{x}_{12}$. The maximum cross-correlation indicates the position of the most similar block, which should be processed. Finally, Fig. 10 shows the effect of removing two bad blocks (block with interferences) from the raw time-frequency image (TFI), for the same case (#8).

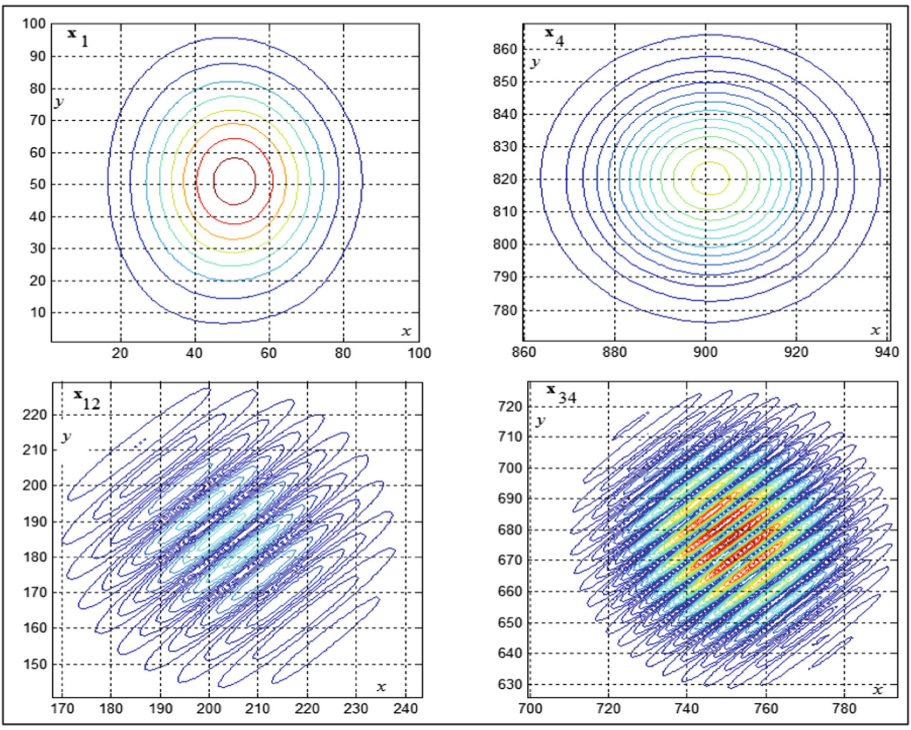


Fig. 8. Samples from the feature image data base: with and free of interference

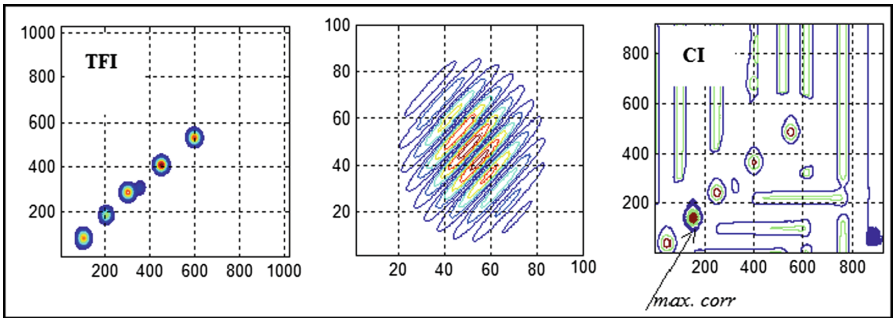


Fig. 9. Raw TFI image and the cross-correlation image between x_{12} and raw TFI

The complete set of experiments can fix the performance of the system. The results are as expected for all cases with double components. For the cases with more components, the results depend on the distance between components. If some components are in the middle, between two or more components, then they are modified and the reconstruction becomes quite difficult with this system.

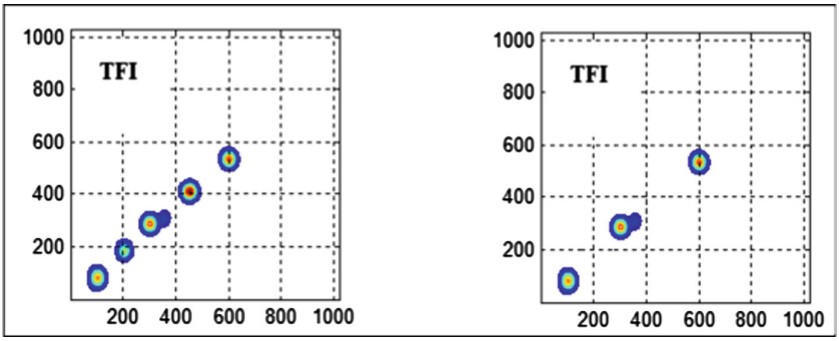


Fig. 10. Raw and filtered TFI images, case 8

5 Conclusions

A system meant to analyze, detect and filter artifacts in images coming from time-frequency transforms was considered and discussed, in the context of change detection and diagnosis of industrial equipment, which generates mechanical vibrations.

The complexity of the input signal is well described by the Renyi entropy, which increases when the complexity of the analyzed signal is growing (number of components, mainly). The average parameters of time and frequency are good descriptors only for mono-components signals. Otherwise, they refer to the average content.

The detection system uses the correlation between small (blocks) images and scanned input images. The raw time-frequency image is decomposed in basic (small) blocks of size $nxn = 100 \times 100$. The decomposition could be static, by scanning the raw image on vertical and horizontal axis, in steps of one or several pixels, or could be dynamic, by changing the size of the scanning steps, based on the content of the image. Sparse blocks will generate higher scanning steps. The dynamic scanning speeds up the analysis of the image.

The interference images have two specific features, which could be used in automatic detection and classification: (i) it contains a lot of parallel ellipses; (ii) The main diagonals are parallel to the segment which connects the centers of the closest neighbored components. This set of features could be used to distinguish the real from false components in time-frequency images.

The system needs a training stage (time) to complete a database with what means good and bad features in the analyzed images. This process is supervised and could be run all the time. The similarity checking between the content of the raw image and the selected features for database is made on cross-correlation basis.

The performance of the system is acceptable for signals with well-separated components, in time and frequency, which is the case of many real faults in the industry of rotating machines. If the components and the artifacts interfere, the system cannot work as expected, and generates errors. It could be involved in estimating the values of the real components, but this aspect is not considered here.

Acknowledgement. The authors thank the Executive Agency for Higher Education, Research, Development and Innovation Funding (UEFISCDI) for its support under Contract PN-II-PT-PCCA-2013-4-0044 (VIBROCHANGE).

References

1. Isermann, R.: Supervision, fault-detection and fault-diagnosis methods. CEP **5**(5), 639–652 (1997)
2. Patton, R.J., Frank, P., Clark, R. (eds.): Fault Diagnosis in Dynamic Systems – Theory and Application. Prentice Hall, Upper Saddle River (1989)
3. Basseville, M., Nikiforov, I.: Detection of Abrupt Changes – Theory and Application. Prentice Hall, Upper Saddle River (1993)
4. Gustafson, F.: Statistical signal processing approaches to fault detection. ARC **31**, 41–54 (2007)
5. Basseville, M.: On-board component fault detection and isolation using the statistical local approach. Publication Int. No. 1122, IRISA (1997)
6. Timusk, M., Lipsett, M., Mechefske, C.K.: Fault detection using transient machine signals. Mech. Syst. SP **22**, 1724–1749 (2008)
7. Venkatasubramanian, V., Rengaswamy, R., Yin, K., Kavuri, S.N.: A review of process fault detection and diagnosis, part I: quantitative model-based methods. Comput. Chem. Eng. **27**, 293–311 (2003)
8. Wang, W.J., McFadden, P.D.: Time-Frequency Domain Analysis of Vibration Signals for Machinery Diagnostics, (III) The Present Power Spectral Density, University of Oxford, Report No. OUEL 1911/92 (1992)
9. McFadden, P.D., Wang, W.: Time-Frequency Domain Analysis of Vibration Signals for Machinery Diagnostics. (I) Introduction to the Wigner-Ville Distribution, University of Oxford, Report No. OUEL 1859/92 (1990)
10. The VIBROCHANGE project, Experimental Model for Change Det. and Diag. of Vibrational Proc. Using Advanced Measuring and Analysis Techn. Model-Based (2015). www.etc.ugal.ro/VIBROCHANGE
11. Cohen, L.: Time-frequency distributions - a review. Proc. IEEE **77**(7), 941–980 (1989)
12. Hlawatsch, F., Boudreaux-Bartels, F.: linear and quadratic time-frequency signal representations. IEEE Sig. Process. Mag. **9**, 21–67 (1992)
13. Flandrin, A.P., Gonçalvès, P., Lemoine, O.: Time-Frequency Toolbox For Use with MATLAB, CNRS (France)/Rice Univ. (USA) (1995–1996)
14. Auger, F.: Representations Temps-Frequence Des Signaux Nonstationnaires: Synthese Et Contributions. Ph.D. thesis, Ecole Centrale de Nantes, France (1991)
15. Flandrin, P.: Temps-frequence, Hermes, Trait des Nouvelles Technologies, serie Traitement du Signal (1993)
16. Popescu, T., Aiordachioaie, D.: Signal segmentation in time-frequency plane using renyi entropy - application in seismic signal processing. In: 2nd International Conference on Control and Fault-Tolerant Systems, SysTol-2013, 9–11 October, Nice, France, pp. 312–317 (2013)
17. Aiordachioaie, D.: Signal segmentation based on direct use of statistical moments and renyi entropy. In: 10th International Conference on Electronics, Computer and Computation (ICECCO 2013), Istanbul, Turkye, pp. 359–362 (2013)
18. Barbarossa, S.: Analysis of multicomponent LFM signals by a combined wigner-hough transform. IEEE Trans. Sig. Process. **43**(6), 1511–1515 (1995)

Adaptive Multi-round Smoothing Based on the Savitzky-Golay Filter

József Dombi[1(✉)] and Adrienn Dineva[2,3]

[1] Institute of Informatics, University of Szeged,
6720 Szeged Árpád tér 2, Szeged, Hungary
dombi@inf.u-szeged.hu
[2] Doctoral School of Applied Informatics and Applied Mathematics,
Óbuda University, 1034 Budapest Bécsi út 96/b, Budapest, Hungary
dineva.adrienn@kvk.uni-obuda.hu
[3] Doctoral School of Computer Science, Department of Information Technologies,
Universitá degli Studi di Milano, Crema Campus, 65 Via Bramante,
26013 Crema (CR), Italy

Abstract. Noise cancellation is the primary issue of the theory and practice of signal processing. The Savitzky-Golay (SG) smoothing and differentiation filter is a well studied simple and efficient technique for noise eliminating problems. In spite of all, only few book on signal processing contain this method. The performance of the classical SG-filter depends on the appropriate setting of the windowlength and the polynomial degree. Thus, the main limitations of the performance of this filter are the most conspicious in processing of signals with high rate of change. In order to evade these deficiencies in this paper we present a new adaptive design to smooth signals based on the Savitzky-Golay algorithm. The here provided method ensures high precision noise removal by iterative multi-round smoothing. The signal approximated by linear regression lines and corrections are made in each step. Also, in each round the parameters are dynamically change due to the results of the previous smoothing. The applicability of this strategy has been validated by simulation results.

Keywords: Savitzky-Golay filter · Adaptive multi-round smoothing · Iterative smoothing · Denoising

1 Introduction

Many areas of signal processing require highly efficient processing methods in order to achieve the desired precision of the result. In a particular class of tasks, for e.g. chemical spectroscopy, smoothing and differentiation is very significant. An ample number of studies have been revealed the details of the smoothing filters. In 1964 a great effort has been devoted to the study of Savitzky and Golay, in which they introduced a particular type of low-pass filter, the so-called digital smoothing polynomial filter (DISPO) or Savitzky-Golay (SG) filter [17].

V.E. Balas et al. (eds.), *Soft Computing Applications*, Advances in Intelligent Systems and Computing 633, DOI 10.1007/978-3-319-62521-8_38

The main advantage of the SG-filter in contrast to the classical filters - that require the characterzation and model of the noise process-, is that both the smoothed signal and the derivatives can be obtained by a simple calculation. Critical reviews and modifications of the original method can be read, for instance in [15,21]. The core of this algorithm is fitting a low degree polynomial in least squares sense on the samples within a sliding window. The new smoothed value of the centerpoint obtained from convolution. There is a rapidly growing literature discussing the properties and improvements of SG filters [1,3,6,7,12,16,24,25]. Also the importance and applicability of a digital smoothing polynomial filter in chemometric algorithms are well established [8,11,22]. The frequency domain properties of SG-filters are addressed in [2,10,18,19]. In [14], the properties of the SG digital differentiator filters and also the issue of the choice of filter length are discussed. Paper [13] concerns the calculation of the filter coefficients for even-numbered data. Also the fractional-order SG differentiators have been investigated, for e.g., by using the Riemann-Lioueville fractional order definition in the SG-filter. For instance, the fractional order derivative can be calculated of corrupted signals as published in [4]. There are several sources and types of noise that may distort the signal, for e.g., eletronic noise, electromagnetic and electrostatic noise, etc. [23]. In the theory of signal processing it is commonly assumed that the noise is an additive white Gaussian noise (AWGN) process. However, in engineering practice often nonstationary, impulsive type disturbances, etc., can degrade the performance of the processing system. Since, for the noise removal issue of signals with a large spectral dynamic or with a high rate of change, the classical SG filtering is an unefficient method. Additionally, the performance depends on the appropriate selection of the polynomial order and the window length. The arbitrary selection of these parameters is difficulty for the users. Usually the Savitzky-Golay filters perform well by using a low order polynomial with long window length or low degree with short window. This latter case needs the repetition of the smoothing. It has also been declared that the performance decreases by applying low order polynomial on higher frequencies. Nonetheless, it is possible to further improve the efficiency. With this goal, in this work we introduce an adaptive smoothing approach based on the SG filtering technique that ensures acceptable performance independent of the type of noise process.

2 Mathematical Background of the Savitzky-Golay Filtering Technique

In this section we briefly outline the premise behind the SavitzkyGolay filtering according to [9]. Let us consider equally spaced input data of n$\{x_j; y_j\}$, $j = 1, \ldots, n$. The smoothed values derives from convolution, given by

$$g_i = \sum_{i=-m}^{m} c_i y_{k+i}, \tag{1}$$

where the window length $M = 2m+1$, $i = -m, \ldots, \lambda, \ldots, m$, and λ denotes the index of the centerpoint. The k^{th} order polynomial P can be written as

$$P = a_0 + a_1(x - x_\lambda) + a_2(x - x_\lambda)^2 + \ldots + a_k(x - x_\lambda)^k \tag{2}$$

The aim is to calculate the coefficients of Eq. (2) by minimizing the fitting error in the least squares sense. The Jacobian matrix is as follows

$$J = \frac{\partial P}{\partial a} \tag{3}$$

The polynomial at $x = x_\lambda$ is a_0, hence in order to evaluate the polynomial in the window we have to solve a system of M equations which can be written in matrix form

$$J \cdot a = y \tag{4}$$

$$\begin{pmatrix} 1 & (x_{\lambda-m} - x_\lambda) & \cdots & (x_{\lambda-m} - x_\lambda)^k \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & (x_{\lambda+m} - x_\lambda) & \cdots & (x_{\lambda+m} - x_\lambda)^k \end{pmatrix} \times \begin{pmatrix} a_0 \\ \vdots \\ \vdots \\ \vdots \\ a_k \end{pmatrix} = \begin{pmatrix} y_{\lambda-m} \\ \vdots \\ \vdots \\ \vdots \\ y_{\lambda+m} \end{pmatrix}$$

The coefficients are found from the normal equation in the following writing

$$J^T(Ja) = (J^T J)a \tag{5}$$

so

$$a = (J^T J)^{-1}(J^T y). \tag{6}$$

Since

$$P(x_\lambda) = a_0 = (J^T J)^{-1}(J^T y), \tag{7}$$

by replacing y with a unit vector in Eq. (6) the c_0 coefficient can be calculated as

$$c_j = \sum_{i=1}^{k+1} |(J^T J)^{-1}|_{0i} J_{ij}. \tag{8}$$

With a size of $(2m+1)$x$(k+1)$ the G matrix of the convolution coefficients

$$G = J(J^T J) = [g_0, g_1, \ldots, g_j]. \tag{9}$$

Figure 1 demonstrates the performance of the classical SG-filter. It can be observed that the smoothing is not precise. To overcome this problem, the following section will present an adaptive strategy (Table 1).

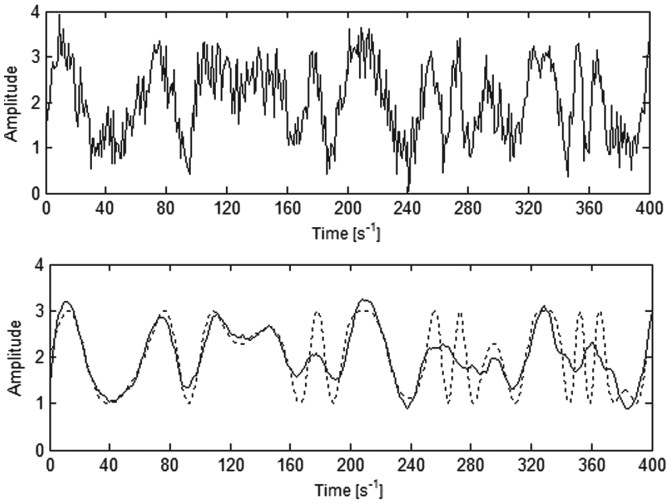


Fig. 1. Performance of the SG filter. Upper chart: signal with contaminating noise. Lower chart: dotted line - original signal, solid line - smoothed signal, $k = 3$, $M = 35$

Table 1. Some SG coefficients. $M = 2m + 1$ is the window length and k denotes the polynomial degree

Savitzky-Golay coefficients		
M	**k**	**Coefficients**
2*9	2	-0.0909 0.0606 0.1688 0.2338 0.2554 0.2338 0.1688 0.0606 -0.0909
	4	0.0350 -0.1282 0.0699 0.3147 0.4172 0.3147 0.0699 -0.1282 0.0350
2*11	3	-0.0839 0.0210 0.1026 0.1608 0.1958 0.2075 0.1958 0.1608 0.1026 0.0210 -0.0839
	5	0.0420 -0.1049 -0.0233 0.1399 0.2797 0.3333 0.2797 0.1399 -0.0233 -0.1049 0.0420

3 Adaptive Multi-round Smoothing Based-On the SG Filtering Technique

3.1 Multi-round Smoothing and Correction

This new adaptive strategy aims setting automatically the suitable polynomial order and window length at the different frequency components of the signal. Hence, it is possible to avoid the undershoots and preserve the peaks that could be important from different data analysis aspects. Since we perform in the time domain, this method provides efficient results independent of the type of contaminating noise. At first, the classical Savitzky-Golay filtering is performed. Assuming that only the corrupted signal is available, this step serves for revealing the peaks, hence the window length and degree of the polynomial may be arbitrary. After the first smoothing, the coordinates of the local minimum and maximum points can be obtained. From now on, we can also define the d distace vector which contains the number of samples between two neighboring points

of local minima and maxima. Then, the next step is the separation of the high- and low frequency components using the bordering points and setting the proper parameters for the smoothing. The window should match the scale of the signal and the polynomial degree should vary by depending on the framesize and frequency. Since the next fuzzy relation can be defined between the section length;

$$F(d_{max} >> \bar{d_R}) = \frac{1}{1 + e^{-(\delta_{max} - \bar{d_R})}} \in [0, 1] \tag{10}$$

where $\bar{\delta_R}$ stands for the average length of the sections in the current R parts of the signal, while $\delta_{max} = max(d)$ in the observed signal. If $g(d_{max}, \bar{d_R}) = 1$, the current part of the signal contains high freqency componenst. Hence, the following rules are applied:

$$\begin{aligned}
&\text{if } 1 > g(d_{max}, \bar{d_R}) > 0.9 \text{ then } k = 5, M = nint(0.3\bar{\delta_R}) \\
&\text{if } 0.89 > g(d_{max}, \bar{d_R}) > 0.75 \text{ then } k = 4, M = nint(0.5\bar{\delta_R}) \\
&\text{if } 0.75 > g(d_{max}, \bar{d_R}) > 0.45 \text{ then } k = 3, M = nint(\bar{\delta_R}) \\
&\text{if } 0.44 > g(d_{max}, \bar{d_R}) > 0.2 \text{ then } k = 2, M = nint(0.5R_n) \\
&\text{else } k = 1, M = nint(0.8R_n),
\end{aligned} \tag{11}$$

where R_n is the total number of samples of the R part, and we can assign the k and M values to each R part of the signal. The values for the bounds have been determined according to the forlmula 2^k modified by experimental results.

3.2 Approximation Using Modified Shephard Method for Corretion

The correction carried out with taking the linear approximation of the obtained signal. Then, it is extracted from the smoothed one. This step reveals the higher deviation, thus the next smoothing procedure can be modified accordint to its result. As we have the coordinates of the local minimum and maximum points and the vector d, we can easily fit a regression line on the points between two local extrema. In this case, the ending and starting points of two consecutive lines do not necessary follow so as to match at the same value. In order to ensure the continuous joining of the lines we can perform this step by appling the Lagrange-multiplicator method given by

$$\sum_{x_i \in [x_1, x_2]} (m_1 x^{(i)} + b_1 - y^{(i)})^2 + \sum_{x_i \in [x_1, x_2]} (m_2 x^{(i)} + b_2 - y^{(i)})^2 \Rightarrow \min, \tag{12}$$

with the following constraints:

$$m_1 x_2 + b_1 - m_2 x_2 + b_2. \tag{13}$$

However, in some cases the peaks can contain the information of interest. Therefore, the form of the peak or valley should be processed with special care. To addres this issue, a modified Shepard - method can be applied. Let us consider

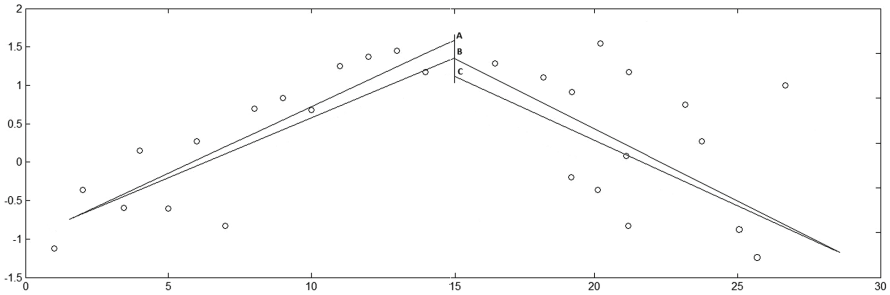


Fig. 2. Illustration of the problem of joining of the regression lines

the points around the local extrema in radius r. The new values are calculated by weighting according to the neighboring points distance. There are several variations of the Shepard method [20], now let us consider the *GMS* (*Groundwater Modeling System*) form below

$$w_i = \frac{(\frac{d-d_i}{dd_i})^2}{\sum_{i=1}^{n}(\frac{d-d_i}{dd_i})^2}. \tag{14}$$

Equation 14 can be trasformed into

$$w_i = \frac{(\frac{1-u_i}{u_i})^2}{\sum_{j=1}^{n}(\frac{u-u_j}{u_j})^2}, \tag{15}$$

in which $u_i(x) = \frac{d^i(x)}{d(x)}$. Now, using the similarity between the form of Eq. 15 and the *Dombi* operator [5] we can define the following new parametric weighting function:

$$w_i = \frac{1}{1 + (\frac{u_i}{1-u_i})^2 \sum_{j=1}(\frac{1-u_j}{u_j})^\lambda} \tag{16}$$

in which the setting of λ and radius r (where it performs) have the effect on the smoothness of the result (Fig. 2).

4 Simulation Results

The performance of the proposed method have been tested on a noisy signal (see, Fig. 3). The simulation has been carried out by using Matlab 8. Figure 4 shows the approximated signal after the first round. In Fig. 5 the resulted and the original signal can be seen after two rounds. It can be observed that the applied technique can efficiently recover the signal.

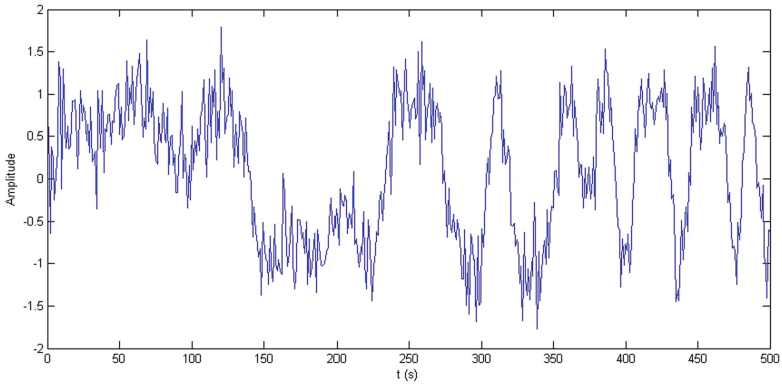


Fig. 3. The corrupted signal

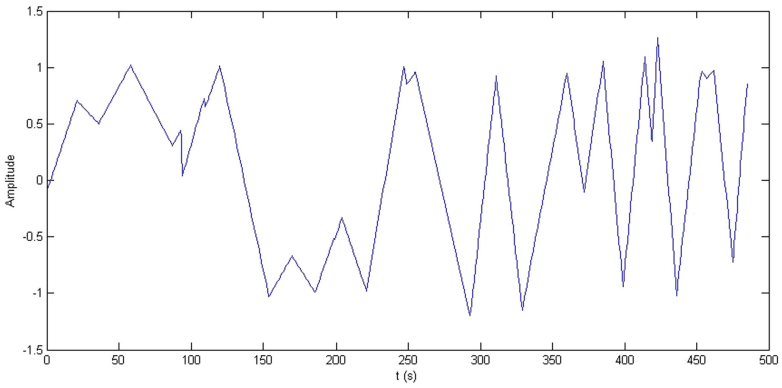


Fig. 4. The approximation of the signal after the first round.

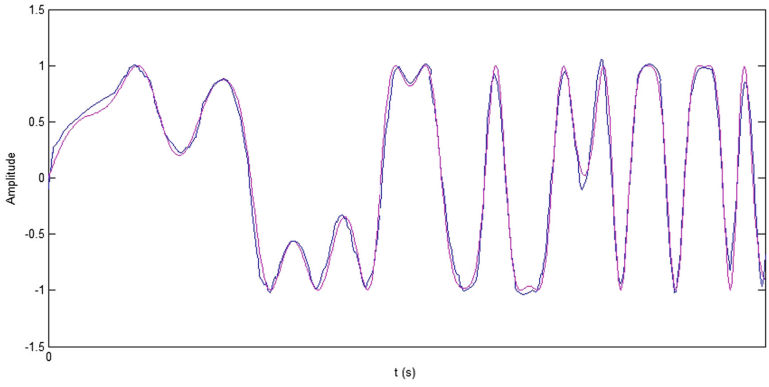


Fig. 5. The recovered (blue) and the original (magenta) signal.

5 Conclusions

In this paper a new adaptive smooting strategy has been introduced based on the Savitzky-Golay filtering technique. The proposed method allows to evade the main difficulties of the original SG filter by automatically setting the smoothing parameters. Furthermore, for the precise reconstruction of the signal a multi round correction has been applied using the linear approximation of the signal. For the reconstruction of the peaks and valleys that may contain the important information, a new weighting function has been introduced with the combination of the GMS method and the *Dombi* operator. The applicability and efficiency of this new strategy have been validated by simulation results.

Acknowledgement. This work has been sponsored by the Hungarian National Scientific Research Fund (OTKA 105846). The authors also thankfully acknowledge the support of the Doctoral School of Applied Informatics and Applied Mathematics of Obuda University.

References

1. Ahnert, K., Abel, M.: Numerical differentiation of experimental data local versus global methods. Comput. Phys. Commun. **177**, 764–774 (2007)
2. Bromba, M.U.A., Ziegler, H.: Application hints for Savitzky-Golay digital smoothing filters. Anal. Chem. **53**(11), 1583–1586 (1981)
3. Browne, M., Mayer, N., Cutmore, T.: A multiscale filter for adaptive smoothing. Digit. Sig. Proc. **17**, 69–75 (2007)
4. Chen, D., et al.: Digital fractional order Savitzky-Golay differentiator. IEEE Trans. Circ. Syst. **58**(11), 758–762 (2011)
5. Dombi, J.: Towards a general class of operators for fuzzy systems. IEEE Trans. Fuzzy Syst. **16**(2), 477–484 (2008)
6. Edwards, T., Willson, P.: Digital least squares smoothing of spectra. Appl. Spectrosc. **28**, 541–545 (1974)
7. Engel, J., et al.: Breaking with trend in pre-processing? Trends Anal. Chem. **2013**, 96–106 (2013)
8. Finta, C., Teppola, P.J., Juuti, M., Toivanen, P.: Feasibility of parallel algorithms and graphics processing unit acceleration in chemical imaging. Chemometr. Intell. Lab. Syst. **127**, 132–138 (2013)
9. Flannery, B., Press, H., Teukolsky, A., Vetterling, W.: Numerical Recipes in C: The Art of Scientific Computing, 2nd edn. Cambridge University Press, New York (1992)
10. Hamming, R.W.: Digital Filters, 3rd edn. Prentice-Hall, Englewood Cliffs (1989). Chap. 3
11. Komsta, L.: A comparative study on several algorithms for denoising of thin layer densitograms. Anal. Chim. Acta **641**, 52–58 (2009)
12. Krishnan, R., Seelamantula, C.: On the selection of optimum Savitzky-Golay filters. IEEE Trans. Sig. Process. **61**(2), 380–391 (2013)
13. Luo, J., Ying, K., Bai, J.: Savitzky-Golay smoothing and differentiation filter for even number data. Sig. Process. **85**(7), 1429–1434 (2005)

14. Luo, J., Ying, K., He, P., Bai, J.: Properties of Savitzky-Golay digital differentiators. Digit. Sig. Proc. **15**(2), 122–136 (2005)
15. Madden, H.: Comments of Savitzky-Golay convolution method for least squares fit smoothing and differentiation of digital data. Anal. Chem. **50**, 1383–1386 (1978)
16. Persson, P.O., Strang, G.: Smoothing by Savitzky-Golay and legendre filters. In: Rosenthal, J., Gilliam, D.S. (eds.) Mathematical Systems Theory in Biology, Communications, Computation, and Finance, vol. 134, pp. 301–315. Springer, New York (2003)
17. Savitzky, A., Golay, M.J.: Smoothing and differentiation of data by simplified least squares procedures. Anal. Chem. **36**(8), 1627–1639 (1964)
18. Schafer, R.W.: On the frequency-domain properties of Savitzky-Golay filter. In: Proceedings of the the 2011 DSP/SPE Workshop, Sedona, AZ, pp. 54–59 (2011)
19. Schafer, R.W.: What is a Savitzky-Golay filter? (lecture notes). IEEE Sig. Process. Mag. **28**(4), 111–117 (2011). doi:10.1109/MSP.2011.941097
20. Shepard, D.: A two-dimensional interpolation function for irregularly-spaced data. In: Proceedings of the 1968 ACM National Conference, pp. 517–524 (1968)
21. Steiner, J., Termonia, Y., Deltour, J.: Comments on smoothing and differentiation of data by simplified least squares procedura. Anal. Chem. **4**, 1906–1909 (1972)
22. Tong, P., Du, Y., Zheng, K., Wu, T., Wang, J.: Improvement of nir model by fractional order Savitzky-Golay derivation (FOSGD) coupled with wavelength selection. Chemometr. Intell. Lab. Syst. **143**, 40–48 (2015)
23. Vaseghi, S.V.: Advanced Digital Signal Processing and Noise Reduction, 4th edn. Wiley, Hoboken (2008)
24. Wayt, H.J., Khan, T.R.: Integrated Savitzky-Gola filter from inverse Taylor series approach. In: Proceedings of the 2007 15th International Conference on Digital Signal Processing (DSP 2007), pp. 375–378. IEEE (2007)
25. Zhao, A., Tang, X., Zhang, Z.: The parameters optimization selection of Savitzky-Golay filter and its application in smoothing pretreatment for FTIR spectra. In: Proceedings of the IEEE 9th Conference on Industrial Electronics and Applications (ICIEA), pp. 516–521 (2014)

Point Cloud Processing with the Combination of Fuzzy Information Measure and Wavelets

Adrienn Dineva[1,2(✉)], Annamária R. Várkonyi-Kóczy[3], Vincenzo Piuri[4], and József K. Tar[5]

[1] Doctoral School of Applied Informatics and Applied Mathematics, Óbuda University, Bécsi út 96/b, Budapest 1034, Hungary
dineva.adrienn@kvk.uni-obuda.hu

[2] Doctoral School of Computer Science, Department of Information Technologies, Universitá degli Studi di Milano, Crema Campus, 65 Via Bramante, 26013 Crema (CR), Italy

[3] Department of Mathematics and Informatics, J. Selye University, Komarno, Slovakia
koczya@mail.ujs.sk

[4] Department of Information Technologies, Universitá degli Studi di Milano, Crema, Italy
vincenzo.piuri@unimi.it

[5] Antal Bejczy Center for Intelligent Robotics (ABC iRob), Óbuda University, Budapest, Hungary
tar.jozsef@nik.uni-obuda.hu

Abstract. Processing of remotely sensed point clouds is a crucial issue for many applications, including robots operating autonomously in real world environments, etc. The pre-processing is almost an important step which includes operations such as remove of systematic errors, filtering, feature detection and extraction. In this paper we introduce a new pre-processing strategy with the combination of fuzzy information measure and wavelets that allows precise feature extraction, denoising and additionally effective compression of the point cloud. The suitable setting of the applied wavelet and parameters have a great impact on the result and depends on the aim of preprocessing that usually is a challenge for the users. In order to address this problem the fuzzy information measure has been proposed that supports the adaptive setting of the parameters. Additionally a fuzzy performance measure has been introduced, that can reduce the complexity of the procedure. Simulation results validate the efficiency and applicability of this method.

Keywords: Point cloud processing · Fuzzy information measure · Signal processing · Denoising

1 Introduction

Recently, there is a growing interest in several areas of engineering practice for the automatic processing methods of three-dimensional remotely sensed models.

V.E. Balas et al. (eds.), *Soft Computing Applications*, Advances in Intelligent Systems and Computing 633, DOI 10.1007/978-3-319-62521-8_39

Most of the 3D scanning devices produce a set of data points of the environment usually defined with their Descartes (x, y, z) coordinates. The resulted data set can contain faulty sensor values of the environment, disturbing noises, etc.,. Since the sampling of the raw point cloud is insufficient for most applications the main steps of processing includes the outlier (extreme values) removal, noise cancellation and smoothig, surface reconstruction and feature extraction. Additionally, depending on the specific requirements of the application further post processing techniques can be used [10].

Classical approaches, for instance linear filtering, can remove the noise, but with weak feature localization ability. To overcome this deficieny nonlinear filters have been proposed [8,16]. Among the classical signal processing techniques, wavelet-based noise reduction has a major potential application to filter data, due to its excellent feature localization property. It reveals the information at a level of detail, which is not available with Fourier-based methods [4]. The so-called wavelet shrinkage employs nonlinear soft thresholding functions in the wavelet domain [6]. Also, the fast discrete wavelet signal transform (DWT) algorithms can fulfil time-critical needs. However, the selection of the appropriate wavelet and number of resolution levels, threshold function, etc. is still a challenging task. The point cloud may contain several objects, artifacts with different surface complexity thus the different parts need the appropriate shrinkage method. This paper introduces the fuzzy information measure for supporting the adaptive shrinkage procedure selection. The here described method conserns outlier removal and feature extraction. The complexity of an object is determined by the number of edges.

In order to conserve precisely the shape that may contain the information of interest with simultaneously filtering out the noise, the objects with high complexity are processed with higher levels of resolution and appropriate threshold function. While the other parts can be smoothed. Furthermore, processing all of the points can be time-consuming, thus the fuzzy performance measure proposed for selecting the minimum amount of data that is necessary for obtaining the desired result. This serves also a basis for efficient data compression. The applicability and efficiency of this method have been validated via simulation results.

2 Point Cloud Processing with the Combination of Fuzzy Information Measure and Wavelets

2.1 Principles of Discrete Wavelet Shrinkage

The point cloud contains points that are returned from the terrain objects, including ground, buildings, bridges, vehicles, trees, and other non-ground features. For many applications it is important to detect, separated, or removed the artifacts in order to extract the required information [17]. Many mathematical tools have been derived for other applications need the precise reconstruction of the artifacts or involve searching for the minimum of a continuous function or

a sizing optimization process [11]. Thus, the pre-processing is almost important step which includes operations such as remove of systematic errors, filtering, feature detection and extraction, etc.

The utilization and theory of wavelets is well established, for e.g. see [4]. The discrete wavelet transform analyses the signal at different frequency scales with different resolutions by reducing the signal into approximate and detail information. For removing noise, wavelet shrinkage employs nonlinear soft thresholding functions in the wavelet domain. The popularity of this nonparametric method is due to the excellent localization and feature extracting behavior. However several threshold estimators exist, it is still a challenging task to select the appropriate shrinkage method that fits to the type of signal and contaminating noise and further, is robust against impulse type noises [6]. The other major issue in noise reduction is minimizing the effects of extreme values or elements that deviate from the observation pattern (outliers) (see, for e.g. [2]).

The first step of wavelet shrinkage is the decomposition of signal into the wavelet (detail and approximate) coefficients, as described in [9]. The basic idea behind wavelet shrinkage is setting the value of these coefficients with small magnitude to zero (hard thresholding), or setting their value determined by a lambda (λ) threshold level [7]. After, the reconstruction is carried out by performing the inverse discrete wavelet transform (IDWT). Generally, the shrinkage methods construct nonlinear threshold functions that relies on some statistical considerations. For instance, the smoothness-adaptive method (SureShrink) [7] is proposed to threshold each dyadic resolution level using the principle of Steins Unbiased Estimate of Risk [12], while the universal bound thresholding rule provides results with low computational complexity. The rule of the latter is defined, as follows [7],

$$\nu = \sigma_{MAD}\sqrt{2}\log s_j \tag{1}$$

where $\sigma_{MAD} = \frac{median(\omega_j)}{0.6745}$ denotes the absolute median deviation. The specific choice of the wavelet function, decomposition level, and thresholding rule allows to construct many different shrinkage procedures.

2.2 The Concept of Fuzzy Information Measure

The key issue of image processing, image understanding, data storage, information re-trieval, etc. is what carries the information in images or scenes. However, a wide range of possible concepts replies to this question, containing statistical approaches. Another, more sophisticated approach may also consider the elimination of the additive noises. This raises another question: What is noise? The main problem of answering this questions that noise is an ill-defined category, because noise is usually dependent on the situation and on the objective of the processing [15]. What is characteristic or useful information in one application can be noise in another one. Based on these consideration, the amount of information in an image or point cloud strongly related to the number and complexity of the objects in it [15]. The most characteristics of the objects are their boundaries, i.e. the edges that can be extracted. Since the boundary edges

of the objects contain the primary information about the object shape, these serve as base for classification, object recognition, etc. According to the concept of *Fuzzy Information Measure* revealed in [14,15], also the point clouds content information can be represented by the characteristic features, like boundary edges. Thus, the amount of information in the 3D map is proportional to the number of characteristic boundary lines.

2.3 Fuzzy Information Measure for Feature Extraction

There are also some studies proposing edge detecting procedures for 3D point clouds [3]. Generally, the precise construction of a 3D edge map of the environment from remotely sensed data can be limited. The method proposed further extends our previously approach [5]. It has been shown that applying robust fitting on the wavelet coefficients can efficiently denoise the data. However, objects with high complexity need shrinking with higher resolution levels and performing the fitting with higher order polynomials for preserving the details. While, the other regions that contain unimportant or sparse information can be smoothed. For separating complex objects the number and distance of the most characteristic edges are counted that carries the information. For this reason, it is not necessary to construct the full edge map, just defining the areas where the density of these edges are high. For properly setting the parameters of the wavelet resolution and smoothing the fuzzy information measure, the approach supports this clustering task. As has been described, for e.g. in [3], the local neighborhood of the point is used to test whether a point lies on a potential edge or not. If there is a depth discontinuity in the lidar scan at the point is at a maximum, then that point is part of a contour edge. According to this concept, the fuzzy-edgeness is determined by the depth discontinuity. The discontinuity is determined in the wavelet domain. The $\mu(r) = 1$ fuzzy membership value belongs to a region that carries the highest information. The fuzzy information measure of each region is determined as follows,

$$\mu(r_i) = \frac{1}{1 + e^{(n_{r_{max}} - n_{r_i})}} \tag{2}$$

in which $n_{r_{max}}$ denotes the number of edges in the r_{max} region that contains the maximum number of edges, while n_{r_i} is the number of edges in the i^{th} region. After, the areas with a high number of edges are at first decomposed with *symlet* wavelets using six levels. This step serves for the separation of the noise and the parts of the signal that carries the important information. After this, the robust fitting is performed on the approximate coefficients. Then, the iterative reweighting of the atoms is performed based on their residuals and using a *trisquare* function in a short running window, according to [3]. At last, the inverse discrete wavelet transform (IDWT) is performed. The same procedure holds for the regions with lower measure, but with applying daubechies wavelets with two levels of decomposition and for the fitting a bisquare function within a long sliding window.

2.4 Fuzzy Performance Measure

In engineering practice, the compression ratio measures the relative decrease in complexity of data in a raw form comparing it to the complexity of the compressed form after processing. However, several measures have been applied [13]. In the presented approach the robust fitting performed on the wavelet coefficients can be time consuming if we calculate the weights and new values

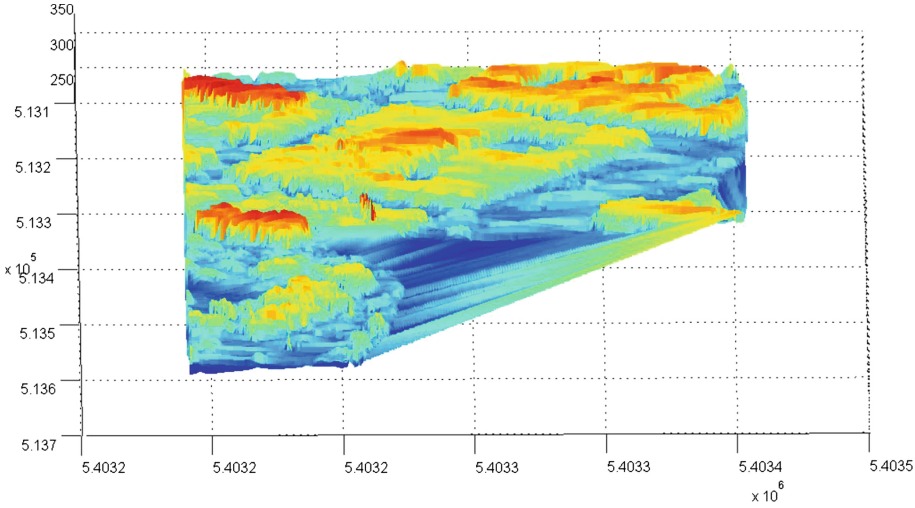


Fig. 1. The raw test data (available at [1]) with additive white Gaussian noise (AWGN).

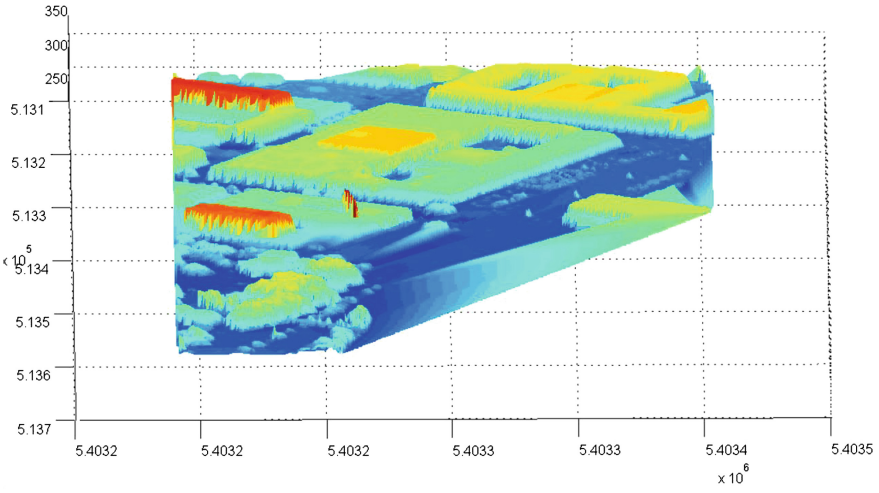


Fig. 2. Result of the denoising using wavelet shrinkage and fuzzy information measure. The raw data is available at [1].

for all the points. Once a point selected for evaluation it is not necessary for calculating again the weights in its vicinity. For selecting the points for the procedure the evaluated fuzzy information measure of the region can be used. In region r_i the running window size determined by the d_{max} distance of the edges. In the regions $mu(r_i) > 0.8$ the 85% of the points are selected for evaluation, while within the regions $0.8 > mu(r_i) > 0.6$ the 50% of the points are enough. These boundaries can be set according to experimental results.

3 Simulation Results

The proposed approach has been tested on a lidar data available at [1]. The data set contains 65536 points. In Fig. 1 the raw data can be seen that has been affected by noise. Figure 2 shows the result of the denoising with the combined approach. It can be seen, that the here described procedure ensures acceptable performance.

4 Conclusions

In this paper a novel concept has been proposed for denoising and feature extracting of remotely sensed point clouds. The here described method relies on the combination of fuzzy information measure and wavelets that allows precise feature extraction, denoising and additionally effective compression of the point cloud. The suitable setting of the applied wavelet and parameters have a great impact on the result and depends on the aim of preprocessing that usually is a challenge for the users. In order to address this problem the fuzzy information measure has been proposed that supports the adaptive setting of the parameters. However, the evaluation of all the points in a large data set is time consuming. For this reason, a fuzzy performance measure has been introduced, that can reduce the complexity of the procedure by reducing the number of points due to the separated regions fuzzy information measure values. Simulation results validate the efficiency and applicability of this method.

Acknowledgements. This work has been sponsored by the Hungarian National Scientific Research Fund (OTKA 105846). This publication is also the partial result of the Research & Development Operational Programme for the project "Modernisation and Improvement of Technical Infrastructure for Research and Development of J. Selye University in the Fields of Nanotechnology and Intelligent Space", ITMS 26210120042, co-funded by the European Regional Development Fund.

References

1. ISPRS test on extracting DEMs from point clouds: a comparison of existing automatic filters. http://www.itc.nl/isprswgiii-3/filtertest/
2. Alvera-Azcárate, A.: Outlier detection in satellite data using spatial coherence. Remote Sens. Environ. **118**, 84–91 (2012)

3. Borges, P., Zlot, R., Bosse, M., Nuske, S., Tews, A.: Vision-based localiztaion using edge map extracted from 3D laser range data. In: IEEE International Conference on Robotics and Automation, pp. 4902–4909, May 3–8, Anchorage, Alaska, USA (2010)
4. Daubechies, I.: The wavelet transform, time frequency localizlocal and signal analysis. IEEE Trans. Inf. Theory **36**, 961–1005 (1990)
5. Dineva, A., Várkonyi-Kóczy, A.R., Tar, J.K.: Improved Denoising with Robust Fitting in the Wavelet Transform Domain. In: Technological Innovation for Cloud-Based Engineering System. IFIP Advances in Information and Communication Technology, vol. 450, pp. 179–187. Springer, Cham (2015)
6. Donoho, D.: Denoising by soft-thresholding. IEEE Trans. Inf. Theory **41**(3), 613–627 (1995)
7. Donoho, D., Johnstone, M.: Adapting to unknown smoothness via wavelet shrinkage. J. Am. Stat. Assoc. **90**, 1200–1224 (1995)
8. Hamming, R.W.: Digital Filters, 3rd edn. Prentice-Hall, Englewood Cliffs (1989)
9. Mallat, S.: Multiresolution approximations and wavelet orthonormal bases of l^2. Trans. Am. Math. Soc. **315**, 69–78 (1989)
10. Orfanidis, S.J.: Introduction to Signal Processing. Prentice-Hall Inc, Upper Saddle River (1995). ISBN 0-13-209172-0
11. Park, J.Y., Han, S.: Application of artificial bee colony algorithm to topology optimization for dynamic stiffness problems. Comput. Math. Appl. **66**, 1879–1891 (2013)
12. Stein, C.M.: Estimation of the mean of a multivariate normal distribution. Ann. Stat. **9**(6), 1135–1151 (1981)
13. Teolis, A.: Computational Signal Processing with Wavelets. Birkhauser, Boston (1998)
14. Várkonyi-Kóczy, A.R., Hancsicska, S., Bukor, J.: Fuzzy information measure for improving HDR imaging. In: Shabbazova, S. (ed) In: Proceedings of the 4th World Conference on Soft Computing, WConSC 2014, pp. 491–497 (2014)
15. Várkonyi-Kóczy, A.R., Tóth, J.: A fuzzy information measure for image quality improvement. In: Nagatsu, M. (ed) In: Proceedings of the 14th International Conference on Global Research and Education in Intelligent Systems, Interacademia 2015, pp. 30–37 (2015)
16. Vaseghi, S.V.: Advanced Digital Signal Processing and Noise Reduction, 4th edn. Wiley, Hoboken (2008)
17. Vosselmann, G., Maas, H.G. (eds.): Airborne and Terrestrial Laser Scanning. Whittles Publishing, Dunbeath (2010)

Significance of Glottal Closure Instants Detection Algorithms in Vocal Emotion Conversion

Susmitha Vekkot and Shikha Tripathi(✉)

Department of Electronics and Communication Engineering,
Amrita School of Engineering, Amrita Vishwa Vidyapeetham,
Amrita University, Bengaluru, India
{v_susmitha, t_shikha}@blr.amrita.edu

Abstract. The objective of this work is to explore the significance of efficient glottal activity detection for inter-emotion conversion. Performance of popular glottal epoch detection algorithms like Dynamic Projected Phase-Slope Algorithm (DYPSA), Speech Event Detection using Residual Excitation And a Mean-based Signal (SEDREAMS) and Zero Frequency Filtering (ZFF) are compared in the context of vocal emotion conversion. Existing conversion approaches deal with synthesis/conversion from neutral to different emotions. In this work, we have demonstrated the efficacy of determining the conversion parameters based on statistical values derived from multiple emotions and using them for inter-emotion conversion in Indian context. Pitch modification is effected by using transformation scales derived from both male and female speakers in IIT Kharagpur-Simulated Emotion Speech Corpus. Three archetypal emotions viz. anger, fear and happiness were generated using pitch and amplitude modification algorithm. Analysis of statistical parameters for pitch after conversion revealed that anger gives good subjective and objective similarity while characteristics of fear and happiness are most challenging to synthesise. Also, use of male voice for synthesis gave better intelligibility. Glottal activity detection by ZFF gave results with least error for median pitch. The results from this study indicated that for emotions with overlapping characteristics like surprise and happiness, inter-emotion conversion can be a better choice than conversion from neutral.

Keywords: Glottal Closure Instants · DYPSA · SEDREAMS · ZFF · Inter-emotion conversion

1 Introduction

Emotions form a vital part of human communication. An emotion can be defined as that state of mind which conveys the non-linguistic information from the speaker. Emotions can be expressed through multiple modalities like facial expressions, hand gestures or speech. Vocal emotion forms the non-linguistic part of expression and has a range of variations depending on speaker gender, state of mind, physiology and language. Due to this, the recognition and analysis of human emotions has always been a challenging task. Vocal emotions form part of the para-linguistic information conveyed

V.E. Balas et al. (eds.), *Soft Computing Applications*, Advances in Intelligent Systems and Computing 633, DOI 10.1007/978-3-319-62521-8_40

in the form of spoken language. With the advent of robotic communication technology, human-machine interaction has been gaining immense momentum during the last decade. Effective portrayal of emotions by robots requires that the parameters characterising the synthesis of emotions are estimated, analysed and manipulated at segmental, supra-segmental and sub-segmental levels.

In Fants source-filter model of speech production, the source signal produced at lungs and manipulated by glottal constrictions is linearly filtered through the vocal tract. The operation can be compared to convolution using a glottal source and filter. The vocal tract shapes the characteristics of the vibrations emitted from the source. The resulting sound is emitted to the surrounding air through lips [1].

Glottal Closure Instants (GCIs) are defined as instances of significant excitation of resonances of vocal tract filter which corresponds to high energy in the glottal signal during voiced segment of speech and the onset points of frication in case of unvoiced speech [2]. Accurate detection of GCIs can aid in applications such as prosody modification, speech coding/compression and speaker recognition [3]. The essence of speech production starts from glottal flow and hence a reliable estimate of parameters contributing to emotion impression in speech has to start from accurate determination of dynamics of glottal flow. Modulation characteristics of airflow can be studied by analysing and comparing the glottal waves across multiple emotions. Different algorithms have been proposed for determination of GCIs of voiced speech. These algorithms follow different approaches. Majority of them use Linear Prediction residual (LPR) for epoch extraction which can be used as the best alternative for glottal wave/its derivative [4, 5].

Fewer techniques are available for directly estimating glottal instants from speech. Researchers have been interested in developing accurate methods for glottal activity detection from the last few years. One of the important algorithms for GCI detection is done by Naylor in [6], popularly known as Dynamic Projected Phase-Slope Algorithm (DYPSA). A breakthrough algorithm for accurate detection of GCIs using Zero Frequency Filtered Signal (ZFF) was proposed in [8, 9]. One of the prominent works on glottal activity detection was carried out by Sturme by using multiscale analysis through wavelet transforms [10]. The authors computed absolute maxima along various scales and generated a tree representation. The GCIs were located at the top of the tree branches. A critical evaluation of the various algorithms for detecting GA regions was done in [11]. Important algorithms like DYPSA [6], SEDREAMS [7], ZFF [8] and wavelet based Lines Of Maximum Amplitude (LOMA) [10] were compared using data from speakers of CMU-ARCTIC database. SEDREAMS was found to have better accuracy in terms of identification rates. Another useful technique for detecting GOIs is by estimating the Hilbert transform of the LP residual of speech [12]. Here, GOIs are estimated from difference EGG by picking secondary peaks. Performance evaluation showed an identification rate of 99.91%.

In this paper we focus on the significance of glottal activity detection in conversion across multiple emotions. The rest of the paper is organised as follows: Sect. 2 describes the algorithms used to obtain the GCIs in the paper, Sect. 3 describes the implementation of inter-emotion conversion using the various epoch detection algorithms while Sect. 4 describes the results and objective evaluation using statistical values and comparison of error percentage. Finally, the paper is concluded in Sect. 5 with insights into the challenges involved and future scope of work.

2 Algorithms for Glottal Epoch Detection

This section throws light on the algorithms used for epoch detection from speech signals.

2.1 Dynamic Projected Phase-Slope Algorithm (DYPSA)

The algorithm described in [6] uses a modification of [13] in which Group Delay function is used to determine instants of significant excitation of glottis. As the name implies, DYPSA algorithm works on the slope of unwrapped phase of Short-Time Fourier transformed Linear Prediction (LP) residual, which is known as phase slope in Group Delay Algorithm [6]. The peak GCI candidates are identified from the positive going and projected zero-crossings of the above function. Finally, a dynamic programming approach is used to select the optimum candidates for glottal closure by minimising an objective function.

2.2 Speech Event Detection Using Residual Excitation and a Mean-Based Signal (SEDREAMS)

Speech Event Detection using Residual Excitation And a Mean-based Signal (SEDREAMS) uses a mean-based signal for epoch detection [7]. This signal is obtained as given in Eq. 1:

$$y_m = \frac{1}{2N+1} \sum_{m=-N}^{N} w(m)\, x\,(n+m) \tag{1}$$

where x(n) is speech signal and w(m) is the window used.

The GCI intervals extracted from the above signal may not give accurate detection. Hence mean based minima are found out and intervals between 2 successive minima (starting from mean based minima till 0.35 times local pitch period) are taken to represent the relative positions of GCIs [7]. For an accurate estimation, the local maxima within the above interval is found out which refines the actual GCI positions.

2.3 Zero Frequency Filtering

ZFF algorithm locates epochs in speech by computing the zero frequency filtered signal [8]. Initially, the speech signal is pre-processed to remove any irregularities in recording:

$$d(n) = s(n) - s(n-1) \tag{2}$$

d(n) is known as difference speech. This is filtered through a cascade of two ideal zero-frequency resonators:

$$y(n) = \sum_{k=1}^{4} a_k y(n-k) + d(n)$$
$$\{a_1, a_2, a_3, a_4\} = \{-4, 6, -4, 1\} \tag{3}$$

For removing the trend in y(n), the local mean subtraction is done at each sample to get the zero frequency filtered signal, $y_{ZFF}(n)$.

$$y_{ZFF}(n) = y(n) - \frac{1}{2N+1} \sum_{m=-N}^{N} y(n+m) \tag{4}$$

where 2N + 1 is the window length used. Epoch locations can be estimated by locating the positive zero-crossings of $y_{ZFF}(n)$.

3 Implementation

3.1 Database Used

The database used for experimentation is IITKGP-SESC (Indian Institute of Technology Kharagpur Simulated Emotion Speech Corpus) which is recorded using 10 (5 male and 5 female) experienced professional artists from All India Radio (AIR) Vijayawada, India. For analysing the emotions 15 emotionally neutral Telugu sentences are considered in which each artist speaks with 8 emotions viz. neutral, anger, happy, fear, compassion, sarcasm, surprise and disgust. Each emotion has 1500 utterances, thus making the total utterances to 12000, lasting for around 7 h [14]. The speech signal was sampled at 16 kHz.

For the current study, we have considered 4 basic emotions in IIT-KGP database viz. neutral, anger, happy and fear from male and female artists. For each of the emotions, statistical values i.e. median, mean and standard deviation of pitch are calculated from 10 different utterances. For each speech sample, GCIs are located using each of the above discussed algorithms and used for inter-emotion conversion using the statistical values computed. Inter-emotion conversion among the above 4 emotions is carried out and results are tabulated. The statistical values for pitch of the emotions considered in this paper for male and female speakers are given in Table 1.

3.2 Pitch Transformation Scales

Using the statistical values obtained above, the transformation scales for conversion across different emotions have been obtained using Eq. 5:

$$Pitchfactor(\%) = \frac{T \arg etpitch - Originalpitch}{Originalpitch} \times 100 \tag{5}$$

For analysis, the percentage change is chosen by taking the median value of each of the pitch parameters in the database. This is done to overcome the changes in the mean

Table 1. Gender-wise statistical pitch values obtained for IIT-KGP SESC database

Gender	Emotion	Median	Mean	Standard deviation
Male	Neutral	187.38	186.83	40.09
	Anger	272.72	264.64	50.77
	Fear	249.76	248.67	28.27
	Happy	219.56	216.87	42.83
	Surprise	252.23	251.36	56.85
Female	Neutral	303.69	310.19	80.4
	Anger	370.50	381.52	95.96
	Fear	309.62	310.6	52.72
	Happy	316.5	338.07	94.8
	Surprise	355.99	361.34	85.85

value due to few irregularities in parameter values. Median of a data distribution gives a more accurate estimation as it does not take into account the irregularities. Also, the median and mean differ by less than 5% [15]. The percentage change to be incorporated in original emotions for conversion in each case is given in Table 2. Applying the pitch transformation scales below, the variations in pitch has been effected corresponding to each emotion. After pitch scaling, the resulting speech sample is subjected to amplitude scaling by multiplying the intermediate signal with the amplitude tier of the target emotion. Thus pitch and amplitude scaling for each emotion is performed. The statistical parameters are recalculated after conversion. The results are compared with the median values for target emotion as we have derived the transform scales based on median value.

3.3 Inter-emotion Conversion

Implementation of emotional conversion is carried out in this paper as a two-stage process. The first stage involves incorporating the transformation scales obtained from Table 2 in pitch modification algorithm. The first stage of the conversion process makes use of dynamic pitch modification algorithm depicted by Govind et al. [16]. This method has been chosen based on the efficiency in preserving the prosodic variations for high arousal emotions.

Stage I

1. Estimate the instants of significant excitation/glottal closure using the algorithm chosen for GCI detection.
2. Get the epoch intervals as difference between successive GCIs.
3. Modification factor, m_q for every q^{th} interval is expressed as an integer ratio A_q/B_q. The initial value for modification factor is the transformation scale value derived from Table 2.
4. m_q corresponding to each emotion is derived based on transform scale Eq. 5.
5. New epoch instants are obtained by equally dividing the q^{th} and $q + A^{th}$ location into B_q intervals.

Table 2. Transformation scales obtained for inter-emotion conversion

Gender	Original emotion	Target emotion	Change (%)
Male	Neutral	Anger	45.54
		Fear	33.28
		Happy	17.17
		Surprise	34.60
	Anger	Neutral	−31.29
		Fear	−8.43
		Happy	−19.49
		Surprise	−7.52
	Fear	Neutral	−24.97
		Anger	9.2
		Happy	−12.1
		Surprise	0.99
	Happy	Neutral	−14.65
		Anger	24.22
		Fear	13.76
		Surprise	14.88
	Surprise	Neutral	−25.71
		Anger	8.13
		Fear	−0.99
		Happy	−12.95
Female	Neutral	Anger	23.00
		Fear	1.95
		Happy	4.20
		Surprise	17.22
	Anger	Neutral	−18.03
		Fear	−16.40
		Happy	−14.57
		Surprise	−3.92
	Fear	Neutral	1.91
		Anger	19.67
		Happy	2.22
		Surprise	14.98
	Happy	Neutral	−4.05
		Anger	17.06
		Fear	−2.18
		Surprise	12.47
	Surprise	Neutral	−14.69
		Anger	4.08
		Fear	−13.03
		Happy	−11.09

6. The modified epoch locations are derived from modified instants computed in step 5 of the algorithm.
7. Similarly, the modified instants for each successive interval are calculated by incrementing q^{th} location value by 1.
8. Repeat steps 3–7 until end location to generate new instants sequence.
9. Modified instants are concatenated and new pitch anchor points generated.
10. Speech waveform is reconstructed by copying the speech samples existing in each epoch interval to the adjacent modified epoch interval in the modified instants sequence obtained [17].

Figure 1 plots the target and synthesized happy speech with their pitch contours and spectrograms after Stage I (pitch scaling using transformation scale obtained in each case) of the algorithm.

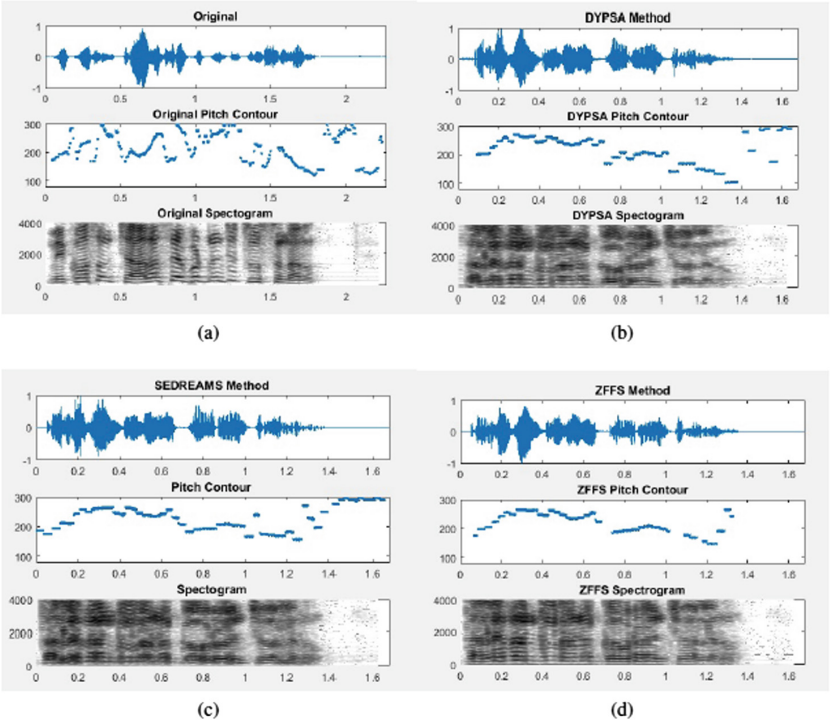


Fig. 1. Results obtained after pitch modification for neutral to happy conversion with GCI estimation using different algorithms. (a) Original happy speech segment, pitch contour and spectrogram. (b) Converted happy, pitch contour and spectrogram with GCI estimation using DYPSA. (c) Converted happy speech, pitch contour and spectrogram with GCI estimation using SEDREAMS. (d) Converted happy speech, pitch contour and spectrogram with GCI estimation using ZFF

The second stage involves intensity modification by imposition of amplitude changes of target emotion into the pitch modified signal from the first stage. The algorithmic steps involved in conversion are listed below:

Stage II

1. The pitch modified speech signal is saved as.wav file and extracted for further processing.
2. The target emotion speech file is processed to separate the intensity tier.
3. The amplitude variations in target emotion are imposed on pitch modified speech by multiplying the extracted amplitude tier with the pitch modified speech signal.

4 Results and Objective Evaluation

The statistical parameters are again calculated after conversion. The results are compared with the median values for original emotion as we have derived the transform scales based on median value.

Table 3. Comparison between statistical values for synthesised emotion and target emotion based on GCI algorithms

GCI algorithm	Gender	Statistical value	Source emotion					Target values	Target emotion
			Neutral	Fear	Happy	Surprise	Anger		
DYPSA	Male	**Median**	**282.77**	**278.6**	**276.17**	**299.42**	**–**	**272.7**	Anger
		Mean	293.44	272.21	252.75	265.53	–	264.64	
		Std. Dev	58.83	37.87	62.63	78.92	–	50.77	
	Female	**Median**	**289.54**	**371.80**	**343.67**	**398.95**	**–**	**357.61**	
		Mean	294.3	367.37	359.42	363.23	–	351.68	
		Std. Dev	83.50	48.78	58.94	84.59	–	75.84	
	Male	**Median**	**212.87**	**–**	**215.59**	**214.88**	**193.26**	**249.76**	Fear
		Mean	200.46	–	207.83	188.13	197.26	248.67	
		Std. Dev	56.78	–	30.45	49.09	45.8	28.27	
	Female	**Median**	**236.15**	**–**	**244.72**	**297.73**	**245.20**	**290.15**	
		Mean	234.68	–	239.18	284.2	294.05	284.79	
		Std. Dev	78.25	–	32.00	60.04	115.4	51.05	
	Male	**Median**	**222.08**	**227.36**	**–**	**255.70**	**213.55**	**219.56**	Happy
		Mean	209.4	213.87	–	254.17	208.44	216.86	
		Std. Dev	64.26	38.97	–	38.46	22.18	42.83	
	Female	**Median**	**298.27**	**233.68**	**–**	**310.00**	**300.00**	**305.33**	
		Mean	313.74	234.94	–	304.40	295.00	327.57	
		Std. Dev	79.64	71.75	–	38.56	92.48	72.3	
SEDREAMS	Male	**Median**	**281.17**	**312.85**	**291.37**	**299.40**	**–**	**272.70**	Anger
		Mean	265.24	305.00	286.82	298.33	–	264.64	
		Std. Dev	47.96	24.50	31.46	51.86	–	50.77	
	Female	**Median**	**331.86**	**398.38**	**332.46**	**467.2**	**–**	**357.61**	
		Mean	326.53	392.21	334.19	467.56	–	351.68	
		Std. Dev	62.80	22.20	86.32	10.75	–	75.84	

(*continued*)

Table 3. (*continued*)

GCI algorithm	Gender	Statistical value	Source emotion					Target values	Target emotion
			Neutral	Fear	Happy	Surprise	Anger		
	Male	**Median**	**216.78**	**–**		**235.42**		**249.76**	Fear
		Mean	227.35	–	240.48	233.54	218.16	248.67	
		Std. Dev	36.52	–	38.90	43.00	31.60	28.27	
	Female	**Median**	**309.82**	**–**	**252.65**	**303.47**	**203.58**	**290.15**	
		Mean	296.54	–	250.22	291.24	238.99	284.79	
		Std. Dev	62.03	–	11.28	46.48	92.53	51.05	
	Male	**Median**	**228.13**	**219.02**	**–**	**238.61**	**204.4**	**219.56**	Happy
		Mean	237.36	210.34	–	235.36	195.27	216.86	
		Std. Dev	39.00	29.30	–	37.73	40.75	42.83	
	Female	**Median**	**309.86**	**338.50**	**–**	**310.24**	**193.00**	**305.33**	
		Mean	330.85	328.57	–	306.89	216.3	327.57	
		Std. Dev	89.00	74.76	–	20.77	70.04	72.3	
ZFF	Male	**Median**	**263.86**	**272.87**	**270.59**	**236.24**	**–**	**272.70**	Anger
		Mean	268.33	269.92	261.30	246.00	–	264.64	
		Std. Dev	47.87	33.97	36.36	52.77	–	50.77	
	Female	**Median**	**372.15**	**373.88**	**351.78**	**279.82**	**–**	**357.61**	
		Mean	361.91	377.50	371.05	323.48	–	351.68	
		Std. Dev	64.27	55.70	47.34	89.97	–	75.84	
	Male	**Median**	**223.62**	**–**	**230.60**	**234.74**	**220.44**	**249.76**	Fear
		Mean	232.98	–	224.68	235.80	215.23	248.67	
		Std. Dev	37.48	–	29.59	43.60	32.98	28.27	
	Female	**Median**	**271.68**	**–**	**320.90**	**298.12**	**248.00**	**290.15**	
		Mean	265.67	–	309.97	301.77	245.82	284.79	
		Std. Dev	37.12	–	94.00	78.35	60.62	51.05	
	Male	**Median**	**224.90**	**221.30**	**–**	**242.76**	**202.97**	**219.56**	Happy
		Mean	233.23	219.74	–	225.99	202.27	216.86	
		Std. Dev	39.86	18.81	–	45.20	19.62	42.83	
	Female	**Median**	**301.60**	**281.10**	**–**	**302.49**	**260.70**	**305.33**	
		Mean	293.08	297.46	–	316.09	278.07	327.57	
		Std. Dev	60.50	42.12	–	67.77	44.96	72.3	

The results for converted (synthesized) emotion are tabulated in Table 3 and are compared with target emotion in each case. The percentage error rates have been calculated for each GCI algorithm based inter-emotion conversion and the results have been listed in Table 4.

A graphical comparison between the algorithms has been drawn in Fig. 2.

As we have used the scaling factor taking into account the median values, a comparison has been drawn based on median values.

Table 4. Comparison of error (%) for inter-emotion conversion based on GCI algorithms

Gender	GCI algorithm	Error (%) when target emotion is:		
		Anger	Fear	Happy
Male	DYPSA	4.23	16.25	4.60
	SEDREAMS	8.60	6.85	1.36
	ZFF	4.33	8.97	1.56
Female	DYPSA	1.85	11.78	6.49
	SEDREAMS	6.95	7.85	5.70
	ZFF	3.69	1.89	6.18

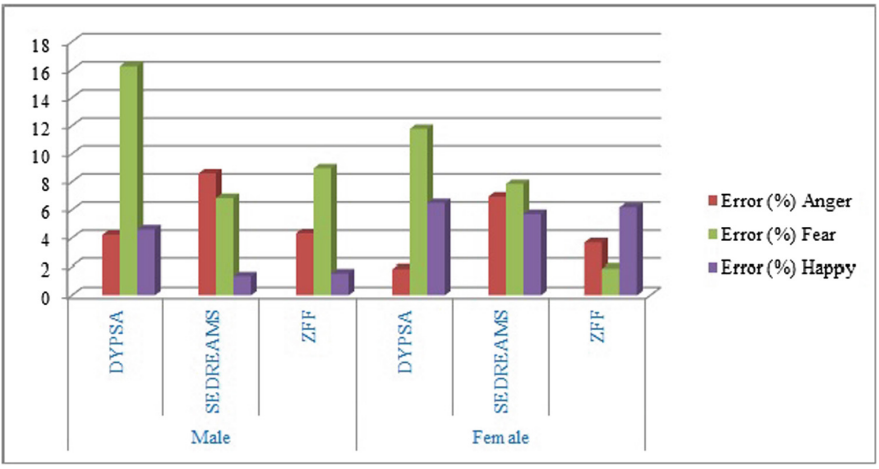


Fig. 2. Comparison of various GCI algorithms based on percentage error. The values are calculated based on median of anger, fear and happy emotions for both male and female speakers

From the experiments, some observations have been made:

- Median value forms a better estimate for probabilistic study of various emotions as it is free from irregularities.
- Conversion efficiency is higher for anger target in both neutral-emotion and inter-emotion cases. This is because anger is mainly categorized by higher pitch and intensity. The variations done as part of this work also focussed on pitch and intensity scaling. Anger gave best results for perception also.
- Most deviation in results was obtained for conversion to/from fear emotion. Fear emotion is characterised with large variance and lower energy with the addition of a panic or anxiety component, thus giving rise to a breathy quality to voice. Mere pitch, duration and amplitude variations are unable to incorporate the harmonics and voice quality of fear.

- Conversion to/from happy also yielded average perception results. But the pitch contour of the synthesised emotion was found to be matching with original happy to some extent as shown by Fig. 1. Though the algorithm succeeded in bringing the pitch and intensity to the required level, it still failed to bring in the smooth rippling kind of inflexions normally found in happy speech.
- From the error comparison Table 4, the minimum percentage error is obtained for emotion conversion in case of male speaker. As far as inter-emotion conversion is concerned, though the trends are fluctuating for the various algorithms we can see that an optimal choice of ZFF for epoch estimation gives acceptable error values. Also, the conversion error is maximum for fear case in both male and female speakers which is rightly correlated by experimental findings and perception.
- The average computation time required for execution of conversion algorithm from neutral to emotion and between emotions were computed for each GCI algorithm and it was found that they are closely similar (around 220–260 μs). This shows that no restriction exists while choosing the mode of conversion (neutral - emotion/emotion - emotion) and either can be chosen depending on interrelationship between source and target. For instance, conversion from surprise to happy yielded good objective and perceptive results in all cases (Table 3). This is due to the fact that surprise is depicted in the database as "pleasant".
- We found that the conversion efficiency depends on expressivity of database to a large extent. Emotions like happy and fear can be better distinguished from angry and sadness if the speaker displays the emotion with characteristic precision. In some cases, the enacted database fails to bring in the complex inundations and voice modulations of emotions while spontaneous recording can bring in a much perceivable quality to the target.

5 Conclusions

The experiments on inter-emotion conversion gave acceptable results for pitch variations. Conversion to anger yielded best objective and perceptive results while fear and happiness conversions failed to bring in the dynamics of emotion involved. For emotions like happiness and surprise, the synthesised emotion was better perceived when converted from each other than from neutral emotion. However, though the computation time required for inter-emotion conversion using various GCI detection algorithms were similar, ZFF gave best results in terms of conversion error rate (%). Studies proved that in addition to incorporating variations in prosody, determining the voice quality parameters in emotion conversion and embedding these at proper instances are instrumental in making the target more expressive and intelligible. The challenge lies in suitably inserting the parameters thus determined without distorting the spectral characteristics of normal speech and further, making a generalized framework for emotion conversion which is database, speaker and language independent.

Acknowledgements. We would like to thank Dr. Govind. D, Amrita Vishwa Vidyapeetham, Coimbatore for providing us with IIT-KGP SESC database used in this research.

References

1. Fant, G.: Acoustic Theory of Speech Production. Mouton, The Hague (1960)
2. Murthy, K.S.R., Yegnanarayana, B.: Epoch extraction from speech signals. IEEE Trans. Audio Speech Lang. Proc. **16**(8), 1602–1614 (2008)
3. Rao, K.S., Yegnanarayana, B.: Prosody modification using instants of significant excitation. IEEE Trans. Audio Speech Lang. Proc. **14**, 972–980 (2006)
4. Ananthapadmanabha, T.V., Yegnanarayana, B.: Epoch extraction from linear prediction residual for identification of closed glottis interval. IEEE Trans. Acoust. Speech Signal Proc. **ASSP-27**, 309–319 (1979)
5. Prasanna, S.R.M., Govind, D.: Analysis of excitation source information in emotional speech. In: Proceedings of the Interspeech, Makuhari, Chiba, pp. 781–784 (2010)
6. Naylor, P., Kounoudes, A., Gudnason, J., Brookes, M.: Estimation of glottal closure instants in voiced speech using the dypsa algorithm. IEEE Trans. Audio Speech Lang. Proc. **15**(1), 34–43 (2007)
7. Drugman, T., Dutoit, T.: Glottal closure and opening instant detection from speech signals. In: Proceedings of Interspeech, 6–10 September 2009, pp. 2891–2894 (2009)
8. Yegnanarayana, B., Murthy, K.S.R.: Event-based instantaneous fundamental frequency estimation from speech signals. IEEE Trans. Audio Speech Lang. Proc. **17**(4), 614–624 (2009)
9. Murthy, K.S.R., Yegnanarayana, B., Joseph, A.: Characterisation of glottal activity from speech signals. IEEE Signal Process. Lett. **16**(6), 469–472 (2009)
10. Sturmel. N., d'Alessandro, C., Rigaud, F.: Glottal closure instant detection using lines of maximum amplitudes (LOMA) of the wavelet transform. In: Proceedings of ICASSP, pp. 4517–4520 (2009)
11. Cabral, J.P., Kane, J., Gobl, C., Carson-Berndsen, J.: Evaluation of glottal epoch detection algorithms on different voice types. In: Proceedings of the Interspeech, Florence, Italy, pp. 1989–1992 (2011)
12. Ramesh, K., Prasanna, S.R.M., Govind, D.: Detection of glottal opening instants using hilbert envelope. In: Proceedings of the Interspeech, Lyon, France, pp. 44–48 (2013)
13. Smits, R., Yegnanarayana, B.: Determination of instants of significant excitation in speech using group delay function. IEEE Trans. Audio Speech Lang. Proc. **3**, 325–333 (1995)
14. Koolagudi, S.G., Maity, S., Vuppala, A.K., Chakrabarti, S., Rao, K.S.: IITKGP-SESC: speech database for emotion analysis. In: Proceedings of IC3, Noida, pp. 485–492 (2009)
15. Akanksh, B., Vekkot, S., Tripathi, S.: Inter-conversion of emotions in speech using TDPSOLA. In: Proceedings of the Advances in Signal Processing and Intelligent Recognition Systems, SIRS 2015, pp. 367–378. Springer, Trivandrum (2015)
16. Govind, D., Joy, T.T.: Improving the flexibility of dynamic prosody modification using instants of significant excitation. Circ. Syst. Sig. Process. **35**(7), 2518–2543 (2016). doi:10.1007/s00034-015-0159-5
17. Govind, D., Prasanna, S.R.M.: Dynamic prosody modification using zero frequency filtered signal. Int. J. Speech Technol. **16**(1), 41–54 (2013). doi:10.1007/s10772-012-9155-3

Analysis of Image Compression Approaches Using Wavelet Transform and Kohonen's Network

Mourad Rahali(✉), Habiba Loukil, and Mohamed Salim Bouhlel

Sciences and Technologies of Image and Telecommunications,
High Institute of Biotechnology, University of Sfax, Sfax, Tunisia
rahali.mourad@yahoo.fr, loukil_habiba@voila.fr,
medsalim.bouhlel@enis.rnu.tn

Abstract. Since digital images require a large space on the storage devices and the network bandwidth, many compression methods have been used to solve this problem. Actually, these methods have, more or less, good results in terms of compression ratio and the quality of the reconstructed images. There are two main types of compression: the lossless compression which is based on the scalar quantization and the lossy compression which rests on the vector quantization. Among the vector quantization algorithms, we can cite the Kohonen's network. To improve the compression result, we add a pre-processing phase. This phase is performed on the image before applying the Kohonen's network of compression. Such a phase is the wavelet transform. Indeed, this paper is meant to study and model an approach to image compression by using the wavelet transform and Kohonen's network. The compression settings for the approach to the model are based on the quality metrics rwPSNR and MSSIM.

Keywords: Image compression · Kohonen's networks · Wavelet transform · Learning algorithm

1 Introduction

Digital images pose a problem of transmission time and storage volume. The compression techniques can reduce the space needed to represent a certain amount of information. The compression techniques are divided into two main categories. First, the lossy compression in which some of the information in the original image is lost. Second, the lossless compression exploits the information redundancy in the image to reduce its size. The lossy compression methods [1] are more likely to achieve higher compression ratio than those obtained by the lossless methods [2, 5]. The methods of image compression by neural networks [3] yield acceptable results. Yet, these methods have a limit on the compression ratio and the reconstructed image quality.

To improve the reconstructed image quality, we combine the discrete wavelet transform (DWT) [11] and the quantization by Kohonen's networks [4]. Thereafter, we use the Huffman coding to encode the quantized values [6, 14]. In this paper, we are interested in the study of an approach to image compression through the use of the wavelet transform and Kohonen's network. We will, in particular, detail the learning

V.E. Balas et al. (eds.), *Soft Computing Applications*, Advances in Intelligent Systems and Computing 633, DOI 10.1007/978-3-319-62521-8_41

process of image compression and evaluate the compression result with a new quality metric rwPSNR "relative weighted PSNR". To develop and improve the assessment, we will use another quality metric; namely, the MSSIM "Means Structural SIMilarity". Next, we test the image compression by using the wavelet and Kohonen's network together. Lastly, we make a comparison by using Kohonen's network only without the wavelet transform and the second comparison between the proposed approach and various compression methods.

2 Proposed Approach

2.1 Image Compression

Image compression is carried out through the following steps:

- Apply a wavelet transform [9, 22] to an original image depending on the decomposition level and the wavelet type.
- Decompose the image into blocks according to a block size (for example 2×2, 4×4, 8×8 or 16×16).
- Search the codebook for each block and the code word with a minimum distance from the block. The index of the selected word is added to the index vector that represents the compressed image.
- Code the index vector by a Huffman coding [14].
- Save the index vectors coded for use during decompression.

The Fig. 1 depicts the steps of compression.

Fig. 1. Image compression steps

2.2 Learning Phase

In fact, learning [7, 8] is deemed to be a very important step to compress images by the neural network. The goal is to construct codebooks to be used during compression. The learning process is described in Fig. 2.

The first step of learning is the wavelet transform of an original image to obtain four sub-images: an approximation image and three detail-images (Fig. 4) in different resolutions depending on the decomposition level and the wavelet choice. The second step is to decompose the four sub-images in blocks according to the block sizes (2×2, 4×4, 8×8 or 16×16). The blocks are arranged in linear vectors to be presented in the Self-Organizing Map (SOM) [4, 16] one after the other. The third step is to adjust the weight-coupling according to an index vector. The weights obtained at the end of learning represent the codebook which will be used for compression. The codebook obtained depends on several settings such as the choice of the learning image, the type

of the wavelet transform, the decomposition level, the block size and the size of the self-organizing map. Therefore, we should create several codebooks to improve the compression.

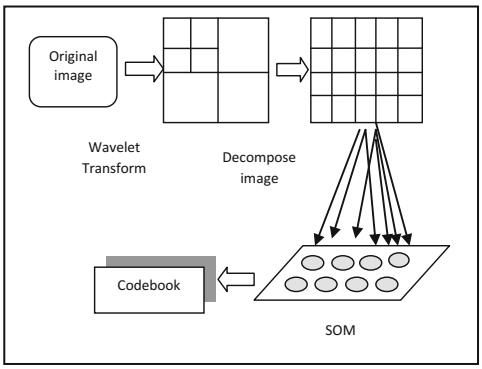


Fig. 2. Learning phase

2.3 Image Decompression

Image decompression is realized throughout these steps:

- Replace each code by the corresponding index to obtain the index vector. This is the decoding step.
- Find the three detail-images and the approximation image by replacing each element of the index vector by the corresponding block in the codebook. In order to improve the reconstructed image quality, in our approach, we keep the approximation image un-indexed by the self organization map.
- The inverse transform is applied to the sub-images obtained after de-quantization to display the reconstructed image.

The Fig. 3 described the decompression steps.

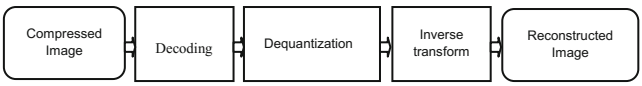


Fig. 3. Image decompression steps

The same codebook is used during both compression and decompression.

2.4 Kohonen's Network Algorithm

Kohonen's network algorithm [4, 10] follows these steps:

- Find the winning neuron of the competition

$$d(X, w_c) \leq d(X, w_i), \forall i \neq c \tag{1}$$

where, X is input vector, w_c is weight vector of the winning neuron c and w_i is weight vector of the neuron i.

- Update weight w_i

$$w_i(t+1) = w_i(t) + h(c, i, t) * [X - w_i(t)] \tag{2}$$

where, w_i is the weight vector of the neuron i in instant t and h is a function defined by:

$$h(c, i, t) = \begin{cases} \alpha(t), i \in N(c, t) \\ 0, \text{ else if} \end{cases} with \ \alpha(t) \in [0, 1] \tag{3}$$

The function h defined the extent of the correction to the winning neuron c and its neighborhood.

In instant t, the neighbors of winning neuron c are determined by the function $N(c, t)$. The final neighbors of a neuron consist of the neuron itself. The function $h(c, i, t)$ assigns the same correction $\alpha(t)$ for all neurons belonging to the neighbors of the winning neuron at instant t [18].

2.5 Image Pretreatment Using Wavelet Transform

The two-dimension-wavelet transform [9, 10] is adopted in our approach. Figure 4 shows the image division into sub-images for the case of seven-band decomposition. The sub-image LL2 represents the lowest frequency of the original image. As a matter of fact, the restoration of the wavelet coefficients in the sub-image LL2 will directly affect the image quality.

The sub-images HH1, LH1, HL1, HH2, LH2, and HL2 contain detail information of the edge, the outline and the vein of the image at different decomposition layers. Concretely, the sub-image HL2 and the sub-image HL1 denote the image-coefficient at the vertical edge after the first and the second layers wavelet decomposition. The sub-image LH2 and the sub-image LH1 indicate the image coefficients at the horizontal edge. The sub-image HH1 and the sub-image HH2 signify the image coefficients on the cross edge.

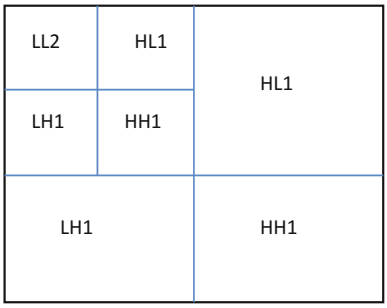


Fig. 4. Decomposition on the frequency by wavelet

3 Objective Assessments: The Quality Index

To improve our method, we use tow quality metric: rwPSNR and MSSIM.

3.1 The Relative Weighted Peak Signal to Noise Ratio rwPSNR

The PSNR quantifies an intensity of the distortion. It does not adjust to the dynamic characteristics of the image. Indeed, the deterioration is more visible in less textured zones (weak variance) and is less visible in more textured zones (stronger variance) [20]. Accordingly, we take the variance of the picture into consideration. Hence, it increases when the variance is high and decreases in the opposite case. We will have a new definition of the MSE.

Let X = {xij | I = 1, …, M; j = 1, …, N}and Y = {yij | I = 1, …, M; j = 1, …, N} be the original image and the test image, respectively. The wMSE is given as:

$$wMSE = \frac{1}{MN}\sum_{m=0}^{M-1}\sum_{n=0}^{N-1}\left(\frac{\left|\left(x_{m,n} - y_{m,n}\right)\right|}{1 + Var(M,N)}\right)^2 \tag{4}$$

Where Var(M, N) is the test image variance. The human eyes do not have an equal sensitivity across the different intensities. In fact, there is a threshold of sensitivity that must be exceeded before an increase of the intensity so that it can be detected. So, as a complement to the wPSNR, we introduce our rwPSNR [12] "relative weighted PSNR" which takes account of the relative difference of the image-gray levels because the noticeable difference of the two stimuli is roughly proportional to the intensity of the stimulus. Actually, an error between two pixels of two images can not translate the same error deviation between two pixels of two other images with the same intensity difference. Indeed, if the intensity difference (10) between the pixels is 10 and 20, it remains numerically the same as that between a pair of pixel values 110 and 120. However, the perception differs on the visual plan. In the first case, the error is quantified at 100% (20 to 10). But, in the second case, the error is quantifiable at 10% (120-110). Therefore, one has to think about the necessity of introducing the relative difference notion in the calculation of the wPSNR from which rwPSNR is derived. So, we have a new definition of the MSE noted as rwMSE "relative weighted Mean Square Error" which takes account of the variance and the image intensity. Our rwMSE is defined as follows:

Let X = {x | i = 1, …, M; j = 1, …, N} and Y = {y | i = 1, …, M; j = 1, …, N} respectively be the original image and the test image. The rwMSE is given as:

$$rwMSE = \frac{1}{MN}\sum_{m=0}^{M-1}\sum_{n=0}^{N-1}\left(2 * \frac{|(x-y)/(x+y)|}{1 + Var(M,N)}\right)^2 \tag{5}$$

The expression of our relative weighted peak signal to noise ratio is given by:

$$rwPSNR = 10 * \log_{10}\left(\frac{x_{\max}^2}{rwMSE}\right) \tag{6}$$

3.2 The Structural Similarity Means MSSIM

The new image quality measurement design is based on the assumption that the human visual system is highly adapted to extract the structural information of the visual field. The measurement of the structural information change can provide a good approximation of the distortion of the perceived image. The error sensitivity approach estimates the perceived errors to quantify the image degradation [21]; whereas, the new philosophy considers the degradation of the image as the perceived changes in the structural information. The luminosity of the surface of the observed object is the product of illumination and reflection. But the structures of the objects in the scene are independent of illumination [23]. Therefore, to explore the structural information in an image, we have to eliminate the influence of illumination. Therefore, the structural information in an image is defined as the attributes that represent the structures of the objects. Since luminosity and contrast may vary across the scene, we use luminosity and the local contrasts in our definition.

The system breaks the task of similarity measurement into three comparisons: Luminosity $L(x, y)$, Contrast $C(x, y)$ and Structure $S(x, y)$. The combination of the three comparisons determines the structural similarity index (SSIM) [13].

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{7}$$

When applying, only one total quality measurement of the whole image is required; whence, a means SSIM index (MSSIM) to assess the overall quality of the image is determined.

$$MSSIM(X, Y) = \frac{1}{M}\sum_{i=1}^{M} SSIM(x_i, y_i) \tag{8}$$

4 Subjective Assessments

The objective assessment is insufficient to assess the visual quality of the compression methods. The subjective assessment analyzes the degradation of visual quality of images. Also, the human eye can judge the compression quality to compare the result between compression methods [19]. In our work, we compare tow compression methods with and without wavelet transform.

4.1 Quantitative Assessments of Results

In our work, we change the compression parameters: the decomposition level (j) of the wavelet, the input block size (BS), the wavelet type (haar, Coiflets, Daubechies, Symlets, …) and the size of the self organizing map (SOM). To evaluate the performance of our approach in image compression, we use the following measures: bits per pixel (Nbpp), the means square error (MSE), the relative weighted peak signal to the noise ratio (rwPSNR) and the structural similarity means (MSSIM) and we compare our approach to image compression without using the wavelet transform (Tables 1 and 2).

Table 1. With wavelet transform

Image	Nbpp	rwPSNR	MSE	MSSIM	Parameters
Lena	4.60	66.98	41.60	0.954	J = 1; BS = 4; SOM = 256
	4.17	66.87	43.43	0.946	J = 1; BS = 4; SOM = 64
	3.02	62.94	66.88	0.933	J = 1; BS = 16; SOM = 256
	1.94	61.06	175.21	0.785	J = 2; BS = 16; SOM = 256
	0.59	60.76	208.31	0.732	J = 2; BS = 256; SOM = 16
	0.34	58.06	502.25	0.526	J = 3; BS = 64; SOM = 16
Cameraman	4.71	61.75	90.08	0.932	J = 1; BS = 4; SOM = 256
	3.90	61.47	93.58	0.928	J = 1, BS = 4, SOM = 64
	2.79	60.04	114.65	0.899	J = 1, BS = 16, SOM = 256
	2.32	59.82	120.43	0.892	J = 1, BS = 16, SOM = 64
	1.82	59.20	135.9	0.859	J = 1, BS = 16, SOM = 4
	1.29	58.08	385.75	0.724	J = 2, BS = 16, SOM = 64
	0.57	57.72	403.24	0.664	J = 2, BS = 256, SOM = 64
Barbara	4.14	62.82	155.29	0.899	J = 1, BS = 4, SOM = 64
	2.96	61.18	204.50	0.857	J = 1, BS = 16, SOM = 256
	2.42	60.49	246.09	0.813	J = 1, BS = 4, SOM = 4
	1.83	60.13	258.69	0.770	J = 1, BS = 256, SOM = 4
	1.36	57.60	394.86	0.650	J = 2, BS = 16, SOM = 64
	0.59	57.68	469.14	0.526	J = 2, BS = 64, SOM = 4
	0.21	56.42	624.43	0.369	J = 3, BS = 256, SOM = 4

Table 2. Without wavelet transform

Image	Nbpp	rwPSNR	MSE	MSSIM	Parameters
Lena	3.92	63.45	92.43	0.938	BS = 4; SOM = 256
	2.54	61.14	133.26	0.869	BS = 4; SOM = 64
	1.91	60.70	151.82	0.845	BS = 16; SOM = 64
	1.66	59.73	175.75	0.753	BS = 16; SOM = 16
	0.82	61.03	248.99	0.738	BS = 64; SOM = 16
	0.31	57.47	566.33	0.486	BS = 256; SOM = 16
Cameraman	3.72	60.43	120.62	0.898	BS = 4; SOM = 256
	2.23	59.86	177.74	0.863	BS = 4; SOM = 64
	2.01	56.04	253.46	0.833	BS = 16; SOM = 64
	1.39	57.21	211.84	0.739	BS = 16; SOM = 16
	0.75	57.65	415.16	0.703	BS = 64; SOM = 16
	0.42	55.79	454.63	0.661	BS = 256; SOM = 64
	0.39	55.65	530.23	0.502	BS = 256; SOM = 16
Barbara	3.23	61.53	156.83	0.865	BS = 4; SOM = 256
	2.99	60.09	181.62	0.837	BS = 4; SOM = 64
	2.36	57.56	265.42	0.792	BS = 16; SOM = 64
	1.56	56.61	316.59	0.687	BS = 16; SOM = 16
	0.75	58.62	366.85	0.643	BS = 64; SOM = 16
	0.47	55.61	507.30	0.541	BS = 256; SOM = 64
	0.33	54.83	563.21	0.364	BS = 256; SOM = 16

Table 3 shows an objective assessment of the relative weighted peak signal to the noise ratio (rwPSNR) depending on the number of bits per pixel (Nbpp). The blue curve represents the rwPSNR of the compressed images using the discrete wavelet transform and Kohonen's network. The green curve represents the rwPSNR of the compressed images using only Kohonen's network without the wavelet transform. Thus, we see very well that the image quality calculated by our metric (rwPSNR) has been improved by the compression approach using the wavelet transform compared to the other approach without wavelet transform. Table 4 shows an objective assessment of the structural similarity means (MSSIM) depending on the number of bits per pixel (Nbpp). The blue curve represents the MSSIM of the compressed images using the discrete wavelet transform and Kohonen's network. The green curve represents the MSSIM of the compressed images using only Kohonen's network without the wavelet transform.

Table 3. Curves of rwPSNR = f(Nbpp)

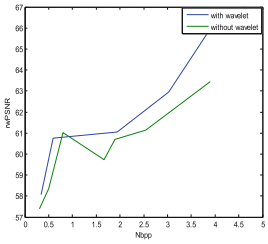
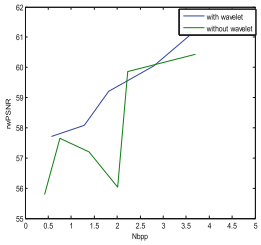
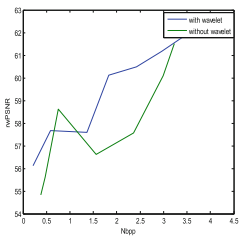
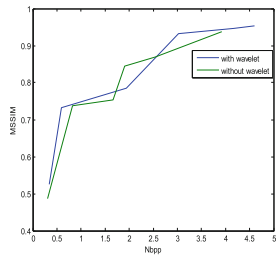
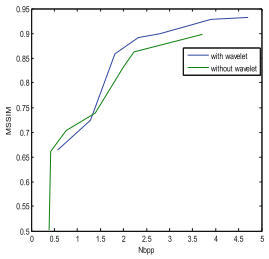
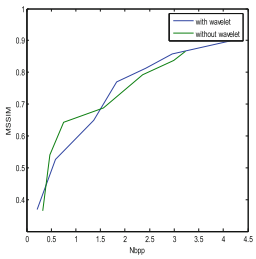
Lena	Cameraman	Barbara
		

Table 4. Curves of MSSIM = f(Nbpp)

Lena	Cameraman	Barbara
		

4.2 Assessment of the Images Visual Quality

The Figs. 5, 6, 7, 8, 9 and 10 provide the assessment of the visual quality of three images (Lena, Cameraman and Barbara) which are compressed by our approach.

(a) MSE=41.60, rwPSNR=66.98, mssim = 0,954

(b)MSE=175.21, rwPSNR=61.06, mssim= 0,785

(c)MSE=502.25, rwPSNR=58.06, mssim= 0,526

Fig. 5. The visual quality With wavelet transform of Lena

(a')MSE=92.43, rwPSNR=63.45, mssim= 0.938

(b')MSE=175.75, rwPSNR=59.73, mssim= 0.753

(c')MSE=566.33, rwPSNR=57.47, mssim= 0.486

Fig. 6. The visual quality without wavelet transform of Lena

(a)MSE=93.58, rwPSNR=61.47, mssim= 0,928

(b)MSE=135.972, rwPSNR=59.20, mssim= 0.859

(c)MSE=403.24, rwPSNR=57.72, mssim= 0,664

Fig. 7. The visual quality with wavelet transform of Cameraman

(a')MSE=120.62, rwPSNR=60.43, mssim= 0.898

(b')MSE=211.84, rwPSNR=57.21, mssim= 0.739

(c')MSE=530.23, rwPSNR=55.65, mssim= 0.502

Fig. 8. The visual quality without wavelet transform of Cameraman

(a)MSE=246.09, rwPSNR=60.49, mssim=0,813

(b)MSE=258.69, rwPSNR=60.13, mssim=0,770

(c)MSE=469.14, rwPSNR=57.68, mssim=0,526

Fig. 9. The visual quality with wavelet transform of Barbara

(a')MSE=265.42, rwPSNR= 57.56, mssim= 0.792

(b')MSE=316.59, rwPSNR=56.61, mssim= 0.753

(c')MSE=507.3, rwPSNR=55.61, mssim= 0.541

Fig. 10. The visual quality without wavelet transform of Barbara

4.3 Comparison Between Our Approach and Various Methods

To evaluate our result, we use the standard image quality metrics for comparisons with other methods. The standard metrics is Peak Signal to Noise Ratio (PSNR) and Compression Ratio (CR). Tables 5 and 6 show the comparison between the proposed approach and various methods [10, 15, 17]. The methods were applied on standard images 256 * 256 (Lena and Cameraman): Neural Network one level, Neural Network two level [15], standard and modified Self-Organization Map SOM [10] and wavelet transform combined by Vector Quantization VQ [17].

Table 5. Comparison results for Lena image

Methods	Nbpp	CR	PSNR
Proposed method	1.39	82.59	25.18
	1.8	78	27.39
Neural Network one level	2.32	71	27.3
Network neuron two levels	1.49	81.3	24.7
Modified SOM	1.84	77	22.15
Standard SOM	1.7	79	19.15
Wavelet transform and Vector Quantization	3.2	60	26.2

Table 6. Comparison results for Cameraman image

Methods	Nbpp	CR	PSNR
Proposed method with wavelet transform	1.77	77.8	25.44
Modified SOM	2.32	71	23.65
Standard SOM	1.79	77.6	20.67

4.4 Simulation and Results

Our work is based on the comparison of different compression methods. The performance evaluation using various image quality metrics like PSNR, rwPSNR, MSE and MSSIM indicates that the best method is the wavelet transform and kohonen's network. After comparing the two compression methods with and without wavelet transform, we can show the importance of using the wavelet transform in compression. In fact, the wavelet transform allows reducing the entropy of the image and separating its details to improve the quality of the reconstructed image according to the number of bits per pixel (Tables 3 and 4). The reconstructed image quality is acceptable; i.e. for rwPSNR it is more than 59 and for MSSIM it is between 0.7 and 1. The block effect is remarkable if the block size (BS) is higher than 256 and the degradation of the visual quality if the level of decomposition of the wavelet (J) is superior or equal to 2.

For performance evaluation, the proposed method is compared with various methods [10, 15, 17]. The comparison is done using various measures such as PSNR, the number of bits per pixel Nbpp and compression ratio CR. From Tables 5 and 6, we can be deduced that the PSNR according of bits per pixel compression ratio of our

method is better than of the some compression methods. From Table 5, the PSNR of our method is 25.18 and the CR is 82.59% but the PSNR of NNs with two levels [15] is 24.7 and CR is 81.3%. So, the PSNR and CR are better in proposed approach

5 Future Scope and Conclusion

In this paper, we use an image compression approach based on the wavelet transform and the vector quantization by Kohonen's network and the Huffman coding. We use two metrics to assess the reconstructed image quality: rwPSNR and MSSIM. We compare our method with other compression methods. The comparison is done using various standard measures such as PSNR and CR, it can be recognized that the PSNR according to CR of our method is better than some of the methods. To improve the reconstructed image quality and the compression ratio, we will add a phase of pre-treatment before the wavelet transform. We will use Weber-Fechner law and the A-law, which are used in the field of telephony, to quantify an image through the semi-logarithmic method.

References

1. Koff, D., Shulman, H.: An overview of digital compression of medical images: can we use lossy image compression in radiology? Can. Assoc. Radiol. J. **57**(4), 211–217 (2006)
2. Zhang, N., Wu, X.: Lossless compression of color mosaic images. IEEE Trans. Image Process. **15**(6), 1379–1388 (2006)
3. Kotyk, T., Ashour, A.S., Chakraborty, S., Balas, V.E.: Apoptosis analysis in classification paradigm: a neural network based approach. In: Healthy World Conference 2015 - A Healthy World for a Happy Life, At Kakinada (AP) India, September 2015
4. Kohonen, T.: The self-organizing map. Proc. IEEE **78**(9), 1464–1480 (1990)
5. Mtimet, J., Benzarti, F., Amiri, H.: Lossless binary image coding using hybrid coding method. In: 6th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications, SETIT 2012, Sousse-Tunisia, 21–24 March 2012, pp. 448–452 (2012)
6. Pujar, J., Kadlaskar, L.: A new lossless method of image compression and decompression using Huffman coding techniques. J. Theor. Appl. Inf. Technol. **15**(1/2), 18–23 (2010)
7. Ribés, J., Simon, B., Macq, B.: Combined Kohonen neural networks and discrete cosine transform method for iterated transformation theory. Signal Process.: Image Commun. **16**(7), 643–656 (2001)
8. Hore, S., Bhattacharya, T., Dey, N., Hassanien, A.E., Banerjee, A., Chaudhuri, S.R.: A real time dactylology based feature extractrion for selective image encryption and artificial neural network. In: Image Feature Detectors and Descriptors. Studies in Computational Intelligence, vol. 630, pp. 203–226, February 2016
9. Raghuwanshi, P., Jain, A.: A review of image compression based on wavelet transform function and structure optimization technique. Int. J. Comput. Technol. Appl. **4**, 527–532 (2013)
10. Boopathi, G.: An image compression approach using wavelet transform and modified self organizing map. Int. J. Comput. Appl. **28**(2), 323–330 (2011)

11. Lewis, A.S., Knowles, G.: Image compression using the 2-D wavelet transform. IEEE Trans. Image Process. **1**(2), 244–250 (1992)
12. Loukil, H., Kacem, M.H., Bouhlel, M.S.: A new image quality metric using system visual human characteristics. Int. J. Comput. Appl., **60**(6) (2012)
13. Brooks, A.C., Zhao, X., Pappas, T.N.: Structural similarity quality metrics in a coding context: exploring the space of realistic distortions. IEEE Trans. Image Process. **17**(8), 1261–1273 (2008)
14. Mathur, M.K., Loonker, S., Saxena, D.: Lossless Huffman coding technique for image compression and reconstruction using binary trees. Int. J. Comput. Technol. Appl. **3**(1), 76–79 (2012)
15. Alshehri, S.A.: Neural network technique for image compression. IET Image Process., 1–5 (2015)
16. Zribi, M., Boujelbene, Y., Abdelkafi, I., Feki, R.: The "self-organizing maps of Kohonen in the medical classification". In: 6th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications, SETIT 2012, Sousse-Tunisia, 21–24 March 2012, pp. 852–856 (2012)
17. Wang, H., Lu, L., Que, D., Luo, X.: Image compression based on wavelet transform and vector quantization. In: International Conference on Machine Learning and Cybernetics, Beijing, November 2002
18. Samanta, S., Ahmed, S.S., Salem, M.A., Chaudhuri, S.: Haralick features based automated glaucoma classification using back propagation neural network. In: International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA), At Bhubaneswar, vol. 327, pp. 351–358 (2014)
19. Sudhakar, R., Karthiga, M.R., Jayaraman, S.: Image compression using coding of wavelet coefficients? A survey. ICGST-GVIP J. **5**(6), 25–38 (2005)
20. Chaouch, D.E., Foitih, Z.A., Khelfi, M.F.: A sliding mode based control of 2DoF robot manipulator using neural network. In: 6th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications, SETIT 2012, Sousse-Tunisia, 21–24 March 2012, pp. 906-911 (2012)
21. Mili, F., Hamdi, M.: Comparative study of expansion functions for evolutionary hybrid functional link artificial neural networks for data mining and classification. Int. J. Hum. Mach. Interact. (IJHMI) **1**(2), 44–56 (2014)
22. Dey, N., Mishra, G., Nandi, B., Pal, M., Das, A., Chaudhuri, S.: Wavelet based watermarked normal and abnormal heart sound identification using spectrogram analysis. In: Computational Intelligence & Computing Research (ICCIC), pp. 1–7, December 2012
23. Espartinez, A.S.: Self-fulfillment and participation of the human person in and through actions. Int. J. Hum. Mach. Interact. (IJHMI) **2**(2), 28–31 (2015)

Application of Higham and Savitzky-Golay Filters to Nuclear Spectra

Vansha Kher, Garvita Khajuria, Purnendu Prabhat(✉), Meena Kohli, and Vinod K. Madan

Model Institute of Engineering and Technology, Jammu, India
vansha26may@gmail.com, garvitakhajuria@gmail.com, klvkmadan@gmail.com, {purnendu.cse,meena.ece}@mietjammu.in

Abstract. A nuclear spectrum is a Type II digital signal. It is processed to extract the qualitative and the quantitative information about the radioisotopes. Noise reduction is usually a step in processing a spectrum, and smoothing filters are employed for this. Perhaps the most widely used filters for the spectral filter-ing since the last 50 years have been Savitzky-Golay (S-G) filters. Incidentally S-G filters are not well known to the digital signal processing community. In this paper the authors have employed rather unknown high fidelity Higham filters used by actuaries of the 19th century for filtering real observed nuclear spectra, and compared their performance with that of an optimal S-G filter. It was observed that the performance of the Higham filters and S-G filters compared favorably amongst each other. This paper describes theory of S-G and Higham filters, their application to nuclear spectra, and presents the results. Higham filters have potential for more applications. There is a scope of exploring more treasure of the past, and merging tools from various disciplines, for more interdisciplinary research.

Keywords: Nuclear spectral processing · Type I and Type II signals · FIR filters · Savitzky-Golay filters · Higham filters

1 Introduction

A nuclear spectrum e.g. a gamma-ray spectrum is processed to extract the qualitative and the quantitative information about the radioisotopes emitting gamma photons. In a typical observed nuclear spectrum over 90% of the information is lost thus greatly reducing the information capacity of a nuclear spectrometer. Processing of a spectrum usually involves employing a smoothing filter to reduce noise. It is desired that the filter should retain spectroscopic information content of the spectrum while attenuating noise. It is sometimes assumed that smoothing is manipulation of data without clear theoretical justification. Digital signal processing (DSP) gives a clear theoretical justification of a smoothing operation. In the frequency domain the signal containing the desired information gets concentrated in the low-frequency region while distortions extend up to π rad.

The separation of signal and distortions appears most clear in the frequency domain. In any other domain like sequency domain, the clarity of separation gets obscured. A smoothing filter is thus adequately justified [1–8].

V.E. Balas et al. (eds.), *Soft Computing Applications*, Advances in Intelligent Systems and Computing 633, DOI 10.1007/978-3-319-62521-8_42

Perhaps the most widely used smoothing filters for the nuclear spectra since the last 50 years have been due to Savitzky and Golay [8]. These are linear phase, noncausal, FIR smoothing filters, also known as polynomial smoothing filters, or least squares smoothing filters. They have been widely used but not well known to the digital signal processing (DSP) community, and "only one out of about 20 had heard of them" [9]. One of the authors (VKM) interacted with both communities, DSP experts as well as nuclear scientists, and had similar experience, and was comforted to read lecture notes of Prof. Schafer [9], a highly respected DSP expert. Earlier the author happened to contribute to a bandwidth matching smoothing criterion for the S-G filters [3]. To bridge the gap between the two communities, he had contributed to a new classification of digital signals viz. Type I and Type II based on the fundamental problems of aliasing and quantization noise. The classification has widened the role of DSP to many disciplines [1, 4, 5].

In the larger perspective DSP techniques to nuclear spectral processing have hardly been used, though the benefits of DSP are widely accepted in diverse applications of science and technology. The least squares peak fitting functions entirely dominate the literature of this field, and these methods fall to a large extent within the category of "art-forms" rather than of "direct methods" [5]. Though DSP methods have not been widely used for the processing of nuclear spectra, they hold a good potential. Only a few researchers have reported the use of DSP techniques in this area. DSP methods for spectral processing, however, proved to be advantageous. They have enhanced the understanding of nuclear spectra while helping in the refinement, extension, and consolidation of various algorithms. The methods have further helped in developing new algorithms especially suited to the spectra and are extremely simple, easy, and fast [1–3, 5, 6].

In the 19th century FIR filters were employed for actuarial data processing [10–12]. The authors have explored two linear phase, high fidelity, FIR Higham filters from that era to filter real observed gamma spectra, and compared their performance with an optimal S-G filter for filtering the same spectra. The performance of the three filters was found favourable with respect to each other.

To the authors knowledge Higham filters have neither been used in electrical engineering nor in nuclear spectrometry. Besides nuclear spectrometry, the filters hold a potential for their application to infrared spectrometry, analytical chemistry, heart rate monitoring using an accelerometer, speech and image processing etc. and that needs to be explored.

This paper describes Type I and Type II signals, smoothing filters viz. S-G and Higham, and their application to filtering real observed nuclear spectra.

2 Type I and Type II Signals

A digital signal can be represented as a mathematical function:

$$y(x) = A \sin(2\pi fx + \theta) \tag{1}$$

where both y(x) and its argument, x, are quantized. The quantity A is the amplitude, f is the frequency and θ is the phase angle.

To widen the scope of DSP, a new classification of digital signals was given by Madan and Sinha [4, 5]. It is based on the fundamental problems of aliasing and quantization noise. Aliasing is a phenomenon where higher-frequency components take on the identity of lower-frequency components because of low sampling rate. It is addressed along the abscissa. Quantization error is the difference between the actual value of the signal and its quantized value. The noise power associated with the quantization of signals is called quantization noise.

Type I digital signals are those where the problems of aliasing and quantization noise are addressed along the abscissa and the ordinate respectively. In Type II digital signals both the problems are addressed along the abscissa.

DSP methods, widely used for processing Type I signals, are generally not used for processing Type II signals. The methods have, however, numerous advantages for processing Type II signals, e.g. nuclear spectra and age returns [4, 5].

The classification of signals arises because of the acquisition process of signals. While acquiring a DSP signals, both Type I and Type II, one has to ensure that aliasing should not be there, and quantization noise should be below a threshold. Once signals have been acquired, DSP techniques like filtering, deconvolution, Fourier domain processing etc. can be employed to process the signals.

It may be mentioned that DSP techniques have demonstrated advantages, both theoretically and in implementation, for processing Type II signals like nuclear spectra and age returns. These methods, however, have hardly been employed.

Figure 1 describes generation of a Type I digital signal like a speech signal, while Fig. 2 describes generation of a Type II digital signal like a nuclear spectrum. Figure 1 is well known to electrical engineers while Fig. 2 is not well known to the engineers.

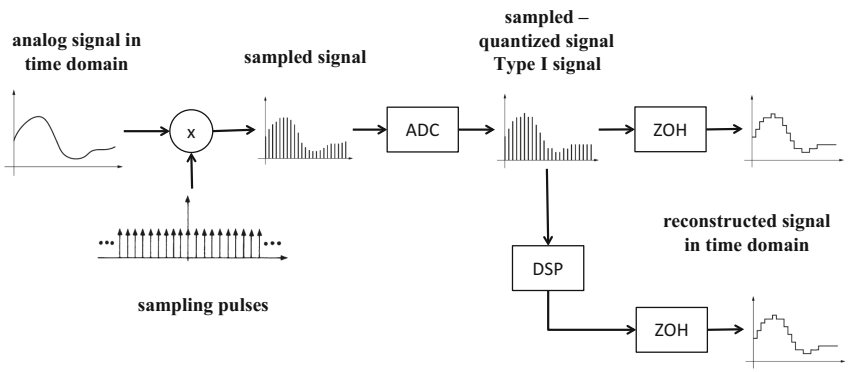


Fig. 1. Acquisition and processing of a Type I signal

The Fig. 2 is described in more detail. Sampling instant is considered as the time when a given pulse arising out of nuclear radiation is detected, and immediately after the process of sampling, the pulse is quantized. Since the output of the nuclear analog to digital converter (NADC) will continue to be a time domain signal, as a matter of

fact it is deliberately rotated 90° as the amplitude of each quantized pulse (or code, or energy channel in nuclear parlance) is quantified as the abscissa of histogram where all the pulses having assigned a particular code (or amplitude or energy) are counted in the same channel. It may be mentioned that a NADC has more stringent specification compared to those of commercial available ADC chips from manufacturers [1, 2].

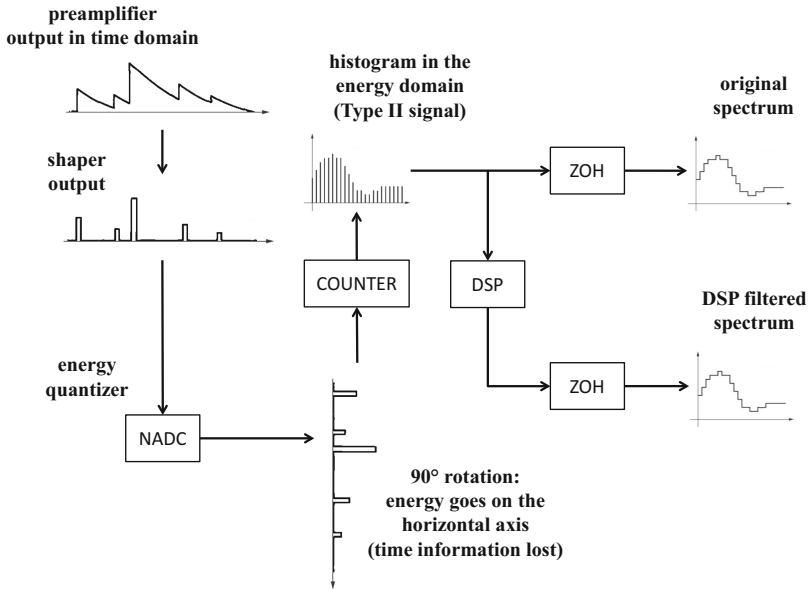


Fig. 2. Acquisition and processing of a Type II signal

Let's go through the signal theory for the two different situations.

Figure 1 depicts electrical waveform in a common digital signal processing chain. Assume x(t) as the input waveform which is sampled at uniform intervals of time T. It is assumed that T is a constant irrespective of the variations in an analog signal, and due to jitter, which is assumed to be insignificant.

The sampled signal $x_s(t)$ is given as:

$$x_s(t) = \sum_{n=-\infty}^{+\infty} x(nT)\delta(t - nT) \tag{2}$$

where, $\delta(t - nT)$ is a unit impulse at time instant t = nT. In Fig. 1 quantization is done by a commercial ADC. In Fig. 2 it is done by a NADC, and the digital signals are condensed into a histogram thus generating a nuclear spectrum. The histogram is generated by employing the reconstruction filter. It is a zero order hold (ZOH) filter. It helps in maintaining the same level of the signal in a spectral channel.

ZOH is a low pass filter with a transfer function $G_T(s)$ which is given by:

$$G_T(\mathrm{s}) = \frac{1}{s}\left(1 - e^{-sT}\right) \tag{3}$$

In Fig. 2, the second filter in the box DSP is a smoothing filter.

3 Gamma-Ray Spectrometer

Figure 3 describes block diagram of a typical gamma-ray spectrometer. A gamma ray source emits gamma rays and they are detected by a HPGe detector. The charge collected by the detector, and then processed by the preamplifier produces sharp rising and pulses which decay with a large time constant. The pulses are shaped into Gaussian like pulses by the spectroscopy amplifier and they are condensed into histogram by a multichannel analyser (MCA) thus generating a nuclear spectrum. The spectrum is filtered and then further processing like background removal, deconvolution etc. is carried out to eventually extract the information about the radioisotoes emitting the nuclear radiation [1, 2].

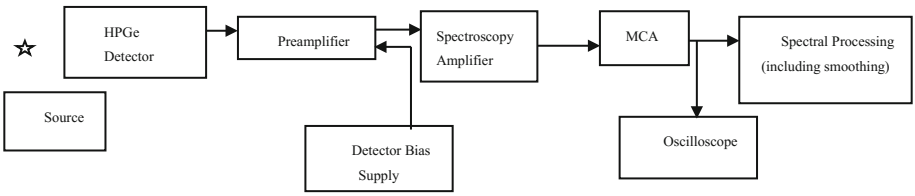


Fig. 3. Block diagram of a gamma-ray spectrometer

4 Savitzky-Golay and Higham FIR Smoothing Filters

In this section S-G and two Higham FIR filters are described. Both filters are non-causal, linear phase (symmetric) filters. S-G filters have been widely used and are easily found in the open literature albeit not in most DSP textbooks. Most DSP experts have not even heard of S-G filters [9]. As far Higham FIR filters, the authors have neither seen their application in electrical engineering nor in nuclear field. Described below is the theory of S-G and Higham filters.

4.1 Savitzky-Golay Filters

The S-G FIR smoothing filters, are also known as polynomial smoothing, or least-squares smoothing filters. It is a happy coincidence that the least squares polynomial fitting turns out to be FIR symmetric smoothing filters. S-G filters preserve high-frequency content representing peaks in a nuclear spectrum while attenuating noise.

Savitzky and Golay put a polynomial to a set of input samples and then at the single point, evaluated the resulting polynomial with least squares fitting [8, 9]. Mathematically it proved to be a discrete convolution having fixed symmetric finite impulse response. The smoothing filter obtained by the least squares method helps in reduction of noise while preserving the shape and height of the nuclear spectral peaks.

Consider for the moment the group of (2M + 1) samples of the signal x[n] centered at n = 0, we obtain the coefficients, a_k, of a polynomial, $p(n)$, that minimizes

$$p(n) = \sum\nolimits_{k=0}^{N} a_k n^k \tag{4}$$

the mean-squared approximation error ε_N for the group of input samples centered on n = 0.

$$\begin{aligned} \varepsilon_N &= \sum\nolimits_{n=-M}^{+M} (p(n) - x[n])^2 \\ &= \sum_{n=-M}^{+M} \left(\sum_{K=0}^{N} a_k n^k - x[n] \right)^2 \end{aligned} \tag{5}$$

The analysis is the same for any other group of (2M + 1) input samples. We will refer to M as the "half width" of the approximation interval.

In S-G filter, the fitted polynomial to the sampled signal is similar to a fixed linear combination of the local set of input samples; the set of (2M + 1) input samples which are within the approximation interval are combined by a fixed set of weighting coefficients that can be calculated once for a given polynomial of order N, and remains same for another approximation interval of length 2M + 1. The polynomial coefficients, a_k, represent impulse response h[n] of the FIR filter. Thus the output samples can be obtained by a discrete convolution method of the form:

$$\begin{aligned} y[n] &= \sum_{m=-M}^{+M} h[m]x[n-m] \\ &= \sum_{m=n-M}^{n+M} h[n-m]x[m] \end{aligned} \tag{6}$$

4.2 Higham Filters

Higham proposed filters that have high fidelity and are simple to implement. We briefly describe here Higham filters using the notation used by the then actuaries. Let n point boxcar average be represented as:

$$y_i = \frac{1}{n}[n]x_i \tag{7}$$

where x_i is the input signal and y_i is the output signal. The operator [n] is defined as:

$$[n]x_i = x_{i-m} + x_{i-m+1} + \ldots + x_{i+m-1} + x_{i+m} \tag{8}$$

Higham filters are depicted in Fig. 4. It has input signal x_i and the output signal as y_i. In operator form, it can be represented as:

$$A_T = \{aA_1 - bA_2\} \tag{9}$$

where A_1 and A_2 are both summation filters, and a and b are scalars. Higham designed a number of filters including the following two. The authors could not find the detailed derivations of the Higham filters in the literature they had access to. The derivations were derived by the authors by hand using the Eqs. 10 and 11, and let these filters be called as Higham Filter 1 and Higham Filter 2 respectively. Both the filters resulted in 17 points smoothing filters [10–15].

$$y_{i=} \frac{1}{125}[2[5]^3[4][2] - 3[5]^4]x_i \tag{10}$$

$$y_i = \left[\frac{64}{10000}[5]^4 - \frac{4}{300}[5]^2[9]\right]x_i \tag{11}$$

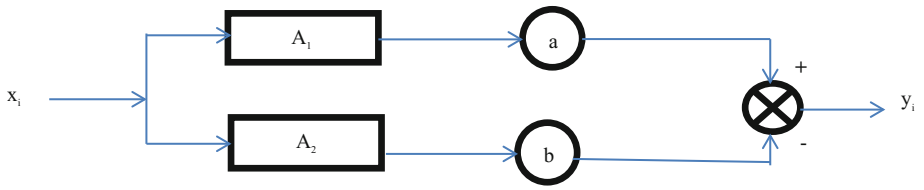


Fig. 4. Representation of a Higham filter

5 Experimental Results

The real observed gamma-ray spectra were used in this paper. The spectra were taken from two earlier publications by one of the authors [3, 7]. These two spectral data sets were acquired using two different nuclear spectrometers. S-G, Higham Filter 1, and Higham Filter 2 filters were employed to filter the spectra. An optimal smoothing interval of 17 was chosen for all the smoothing filters [3]. It may be mentioned that in the earlier two publications [3, 7] employed only S-G filters to the spectra. The effect of smoothing on the spectra was investigated by employing all the three filters.

The investigation included visual inspection, observing monotonicity in spectral peaks viz. it should be monotonically increasing before and monotonically decreasing after the peak position, change in full width at half maximum (FWHM), change in percentage peak areas using difference between sum of counts in the peak region of raw and smoothed spectra, and sum of the squared values of difference between the raw and

smoothed spectral counts in the peak region, and it is called error squared values. The percentage change in peak area used the following formula:

$$\text{Percentage Change in Peak Area} = 100\frac{\text{Raw spectral area} - \text{smoothed spectral area}}{\text{raw spectral area}}$$

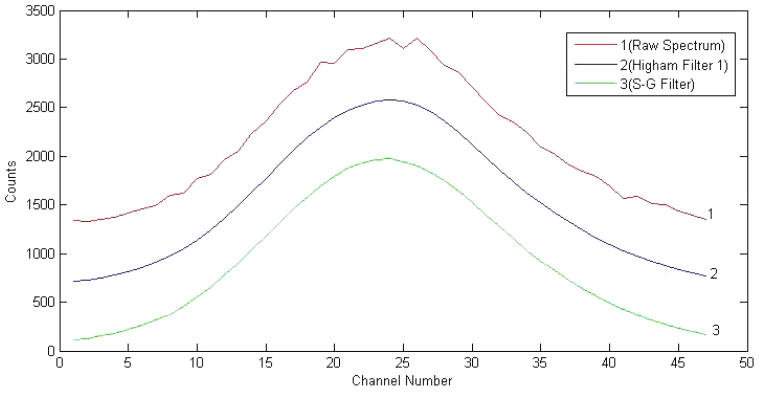


Fig. 5. Raw and smoothed spectra, shifted vertically arbitrarily, using Higham Filter 1 and S-G filter

Figure 5 depicts the raw, and the smoothed spectra obtained by employing the Higham Filter 1, and the S-G filter to one set of data, while Fig. 6 depicts the raw, and the smoothed spectra obtained by employing the Higham Filter 2, and the S-G filter to the other set of data. The visual inspection of the smoothed spectra employing Higham Filter 1, Higham Filter 2, and S-G on both the observed spectra indicated that smoothing effects were pronounced. The spectra were shifted vertically arbitrarily to get clarity of display in the figures. The channel number and counts in the figures correspond to energy and intensity respectively of nuclear radiation.

From the printout of the filtered spectral data it was observed that the monotonicity as mentioned above was met in all the smoothed spectra. The changes in FWHM for all the smoothed spectral peaks were found by plotting the raw, and the smoothed spectra. It was observed that the change in FWHM was on the order of 0.1 channels for all the smoothed spectra. It is negligible. This indicates all the filters preserved high-frequency content representing peaks in a nuclear spectrum while attenuating noise.

The Tables 1 and 2 depict sum of error squared values, and percentage change in peak areas respectively between the raw and the smoothed spectra.

From the Table 1, it is observed that the sum of the error squared values for Higham filters were little better those for the S-G filters. It is observed from the Table 2 that the percentage change in peak areas by employing Higham and S-G filters were within 0.02% and hence negligible.

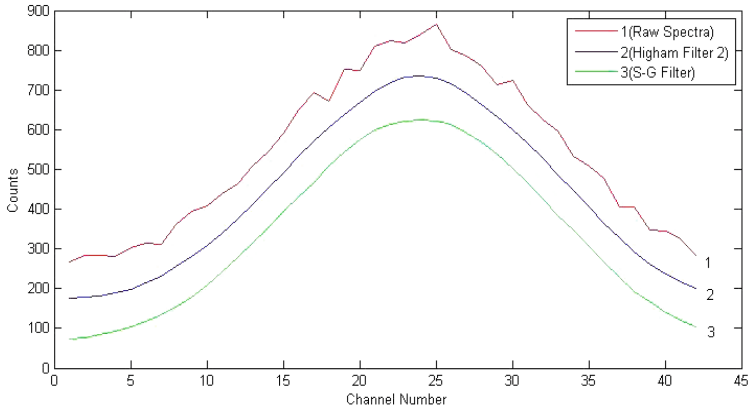


Fig. 6. Raw and smoothed spectra, shifted vertically arbitrarily, using Higham Filter 2 and S-G filter

Table 1. Sum of error squared values

Filter type	Spectrum 1	Spectrum 2
Higham Filter 1	9474	39531
Higham Filter 2	9286	38639
S-G filter	11588	49805

Table 2. Percentage change in peak areas

Filter type	Spectrum 1	Spectrum 2
Higham Filter 1	−0.0158%	−0.0093%
Higham Filter 2	−0.0191%	−0.0119%
S-G filter	−0.0126%	−0.0177%

However the relative error squared values depicted in Table 1, and the relative changes in percentage peak areas depicted in Table 2 can be considered as within the experimental limits.

Hence from the above performance metrics it is concluded that the smoothing performance of all the three filters was favorable with respect to each other.

Higham filters have potential for their application to many areas and that should be explored.

6 Conclusion

For nuclear spectral filtering, perhaps the most widely used filters since the last 50 years have been S-G filters. However the filters are not well known to the DSP community, and "only one out of about 20 had heard of them" [7].

The authors have employed rather unknown Higham filters from the 19th century for filtering of real observed nuclear spectra. To the authors knowledge these filters

have neither been employed in electrical engineering nor for nuclear applications. The performance of the Higham filters and S-G filter compared favorably with respect to each other.

This paper demonstrated an interdisciplinary approach to nuclear spectral filtering. It has merged modern and 19th century tools from disciplines with encouraging application to nuclear spectrometry. Higham filters have potential for their application to infrared spectrometry, analytical chemistry, heart rate monitoring using an accelerometer, speech and image processing etc. and that needs to be explored.

There is a scope of exploring more treasure of the past, and there is a need, perhaps more than ever, for merging of tools from various disciplines for the interdisciplinary research to be effective and fruitful.

Acknowledgments. The authors would like to thank Director, MIET Prof. Ankur Gupta for his encouragement. This paper is respectfully dedicated to Prof. S.C. Dutta Roy.

References

1. Madan, V., Balas, M., Staderini, E.: Novel original digital signal processing techniques of nuclear spectra: a review. In: Kulczycki, P., Szafran, B. (eds.) Information Technology and Computational Physics (2016, designated for inclusion). Special Issue
2. Madan, V., Balas, M.: Nuclear spectrometry. Presented at the Congress of Information Technology, Computational and Experimental Physics (CITCEP 2015), Kraków, Poland, 18–20 December 2015, pp 194–198 (2015)
3. Kekre, H.B., Madan, V.K., Bairi, B.R.: A fourier transform method for the selection of a smoothing interval. Nucl. Instr. Meth. **A279**, 596–598 (1989)
4. Madan, V.K., Sinha, R.K.: Digital smoothing of census data employing fourier transforms. Comput. Stat. Data Anal. **20**, 285–294 (1995)
5. Kekre, H.B., Madan, V.K.: Frequency domain and sequency domain filtering of nuclear spectral data. Nucl. Instr. Meth. **A245**, 542–546 (1986)
6. Madan, V.K., Abani, M.C., Bairi, B.R.: A digital processing method for analysis of complex gamma spectra. Nucl. Instr. Meth. **A343**, 616–622 (1994)
7. Madan, V.K., Sachdev, M.S.: A simpler method to reduce noise in uranium gamma spectra. Presented at the International Conference Radiation Safety in Uranium Mining, Saskatoon, Sask., Canada, 25–28 May (1992)
8. Savitzky, A., Golay, M.: Smoothing and differentiation of data by simplified least squares procedures. Anal. Chem. **36**, 1627–1639 (1964)
9. Schafer, R.W.: What is a Savitzky Golay filter? IEEE Sig. Process. Mag. **28**(34), 111–117 (2011)
10. Higham, J.A.: On the adjustment of mortality tables. J. Inst. Actuar. **23**, 335–352 (1882)
11. Higham, J.A.: On the adjustment of mortality tables. J. Inst. Actuar. **24**, 44–51 (1883)
12. Higham, J.A.: On the graduation of mortality tables. J. Inst. Actuar. **25**, 15–24 (1884)
13. Finlaison, J.: Report on the Law of Mortality of the Government Life Annuitants. The House of Commons, London (1829)
14. Wilkinson, R.H.: Early history of the high-fidelity finite-impulse-response filter. In: IEE Proceedings, vol. 133, pt. G, no. 5, October 1986
15. Madan, V.K.: Digital signal processing: from antiquity to emerging paradigms with emphasis on nuclear spectrometry. ASET Colloquium, Tata Institute of Fundamental Research, Mumbai, 4 October 2013. http://www.tifr.res.in/∼aset/

The Efficiency of Perceptual Quality Metrics 3D Meshes Based on the Human Visual System

Nessrine Elloumi(✉), Habiba Loukil Hadj Kacem, and Med Salim Bouhlel

Research Unit: Sciences and Technologies of Image and Telecommunications, Higher Institute of Biotechnology, University of Sfax, Sfax, Tunisia
elloumi nessrine@gmail.com, loukil_habiba@net-c.com, medsalim.bouhlel@enis.rnu.tn

Abstract. The representation of content as a 3D mesh is a very emerging technology. These three-dimensional meshes can be a scan of objects, characters or 3D scenes. Mesh quality is a determining factor in treatment of effectiveness, accuracy of results and rendering quality. You can show users these 3D meshes with a texture on the 3D mesh surface. The estimated quality by an observer is a very complex task related to the complexity of the Human Visual System (HVS). In this paper we present the efficiency of perceptual quality metrics 3D meshes based on the human visual system.

Keywords: Perceptual quality · 3D metric · 3D meshes · HVS

1 Introduction

For several years, researchers in image [1] processing and computer graphics have worked on the definition of an effective objective metric, able to predict the perceived quality of images, videos and 3D scenes [2]. To design such a system, it was necessary to study and determine the structures of visual content. This task is more complex when we look at 3D content represented in the form of triangular meshes.

These 3D meshes are generally discrete packets of a non- uniform way to represent objects and scenes. Also, many operation can be performed on this 3D meshes [3], for example If we want to protect the 3D content we can use watermarking mesh algorithms for security purposes. It is also possible to apply a compression scheme to transmit on a low bandwidth network [4, 5] to reduce the size of 3D meshes [6]. Other operations on 3D meshes can be used, among which we mention: the quantification, enameling or deformation. These algorithms usually introduce inevitable distortion on the geometry sufaces mesh. It is therefore important to measure and evaluate the effect of these distortions that affect the quality triangular meshes.

2 The Perceptual Quality 3D Meshes

The study of perceptual quality is an important task since most visual data are for a final human observer. Measuring the perceptual quality can be carried out using measurements made by observers (subjective measures) or measurements from

V.E. Balas et al. (eds.), *Soft Computing Applications*, Advances in Intelligent Systems and Computing 633, DOI 10.1007/978-3-319-62521-8_43

algorithmic processes (objective measures). For each observer, quality can have a different definition of personal criteria. The main objective of measuring the perceptual quality is to analyze the behavior of the HVS [7] and quantify the quality as perceived by observers [8]. Human Vision System is based on two mechanisms: low-level mechanisms that affect the biophysical structure of the sensory system and high-level mechanisms related to human cognitive system.

3 Objective Measures of Quality Meshes

The estimate of the perceptual quality is an important element in assessing the meshes quality after treatments. The objective of the perceptual quality metrics is to judge the quality of a mesh based on human perception characteristics [9]. To do this, there are different approaches the top-down approach considers the HVS as a black box and tries to imitate the behavior of the system from the perspective of inputs/outputs and the bottom-up approach based on simulation and imitation of each component of HVS. To develop a 3D metrics [10], researchers can use two alternatives: either they use existing metrics of perceptual quality 2D images [11–13] on two-dimensional projections of 3D meshes (metric-based images and videos) or they use develop metrics based-models that exploit the term mesh geometry and connectivity signal to evaluate the quality [14]. To estimate the quality of 3D meshes, other classification of the quality metrics is possible.

*** Metrics "Full – Reference":** The original model is available in full, in which case we will try to quantify the existing difference between the original image and the degraded version.

*** Metrics to "Reduced Reference":** The original image is available in its entirety, but is represented by a vector containing a reduced set of attributes. In this case, an attempt to quantify the difference existing between the vectors associated to the original image and the vector resulting from a distorted version.

*** Metrics "No – Reference":** The original image is not available. In this case, it is possible to work with a priori on the degradation types that may be encountered, or, conversely, by making no assumption about possible distortions. Many of these metrics exploit the proprieties of the human visual system. To compare two 2D images or 3D meshes, it is common to use the signal to noise ratio measurements or geometric distances. These geometrical measurements do not include the properties of human vision system. From the literature it is approved that the metric based on the HVS are more appropriate than others [15].

4 The Human Visual System

The human visual system is a complex sensory system and not yet fully mastered. Nevertheless, it may be considered an information transcription system turned into usable data by the brain (Fig. 1). A conversion step can capture the received information and decoded into signals by the brain.

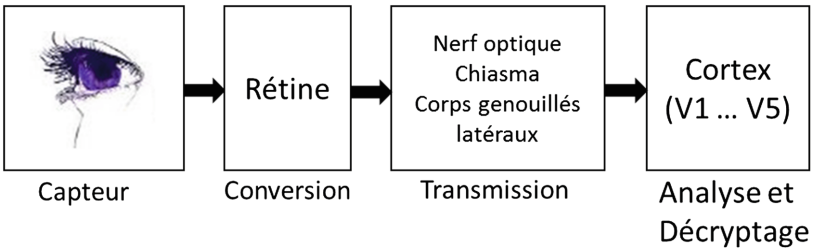


Fig. 1. Human visual system: simplified

It is the role of the eye that converts light energy into sensory signals and then transmits them to the visual cortex via the optic nerves and geniculate nucleus. The visual cortex decrypts and processes the information. To better understand the flow of information and operation of the HVS, we recall the essential elements that are involved in the process of gathering and processing of visual information (Capture, conversion, transmission and processing of information).

4.1 Sensor

- **The eye:** Which an optical system whose primary role is to converge the light signals to the conversion zone and transmit the decoded form in the brain. It comprises several important elements including:
- **The cornea:** is a convex spherical surface which separated the eye from the external environment. Its primary role is to focus the light rays received to the retina.
- **Iris:** Adapts the light intensity by varying its opening. So acts as an optical diaphragm.
- **The crystalline:** Allows to redirect the light flow to the retina where it's arranged into photoreceptor cells (Cones and rods). It plays the role of variable focus lens.

4.2 Conversion

- **The retina:** It converts the light signal captured by the receiving photocells into electrical signals which are then transmitted to the visual cortex via the optic nerve.
- **The retina:** is composed essentially of two types of receptors.
- **Cones:** Which are responsive to detail.
- **Rods:** Which are sensitive to low luminance (blurred vision and course) and intervene rather at night vision (monochrome).

4.3 Transmission

- **The optic nerve:** it transports information from the retina to the visual cortex, through the chiasm and lateral geniculation body.

– **The chiasm:** the direction Information goes through from the right eye and the left eye, chiasm's role is to transmit the information received to the lateral geniculation body.

4.4 Analysis and Decryption

– **The visual cortex:** His major role is the decrypting (or decoding) and the analyzing of the received signals.

5 The Human Visual System and the Perception of Visual Quality

Most of the existing metrics follow the approach Top – Down to study the human visual system and imitate its behavior.

The existing metrics has for objective to maximize the correlation of the results (profits) of prediction with the subjective scores. The subjective evaluation of the meshing quality is established by means of observers through psychometric experiments. Many databases were developed to have measures of the real structure of the 3D meshes perceptual quality. There are two important properties of the HVS that we have to consider in order to measure the perceptual quality of the 3D meshing [16].

5.1 Contrast Sensitivity Function

The function of spatial contrast sensitivity (notes CSF) models the sensitivity of the human visual system to spatial frequencies of the visual stimulus. The first Research Campbell - Robson on CSF have shown that Works with the HVS selectivity on spatial frequencies, sensitivity achieved its Maximum Value around four cycles per degree of visual angle [17].

Figure 2(a) shows that the pixel intensity is modulated by a horizontal sine function. When the spatial frequency of the pixel increases in a logarithmical way (on the horizontal axis), the contrast increases from top to bottom. Although the change contrast is the same for all frequencies, we observe that the bars appear to be highest in the middle of the image following the shape of the sensitivity function in Fig. 2(b). This effect is not made by the image, it's rather made by the property of frequency selectivity of the human vision system.

In the context of dynamic visual content (2D videos, dynamic meshes, etc.), the CSF must be modulated by stimuli speed. This is due to the variability of the contrast sensitivity depending on the speed of the stimulus movement. Since 1977, Kelly has presented an experimental study for modulating the function of contrast sensitivity affected by the test conditions (see Fig. 3(a)). In 1998, Daly has developed a dynamic model of CSF through a time function of CSF tends to move to lower frequencies. These studies were performed for achromatic data. Other studies have integrated color effect on the CSF: The contrast sensitivity (Fig. 3(b)). With increasing speed, the CSF

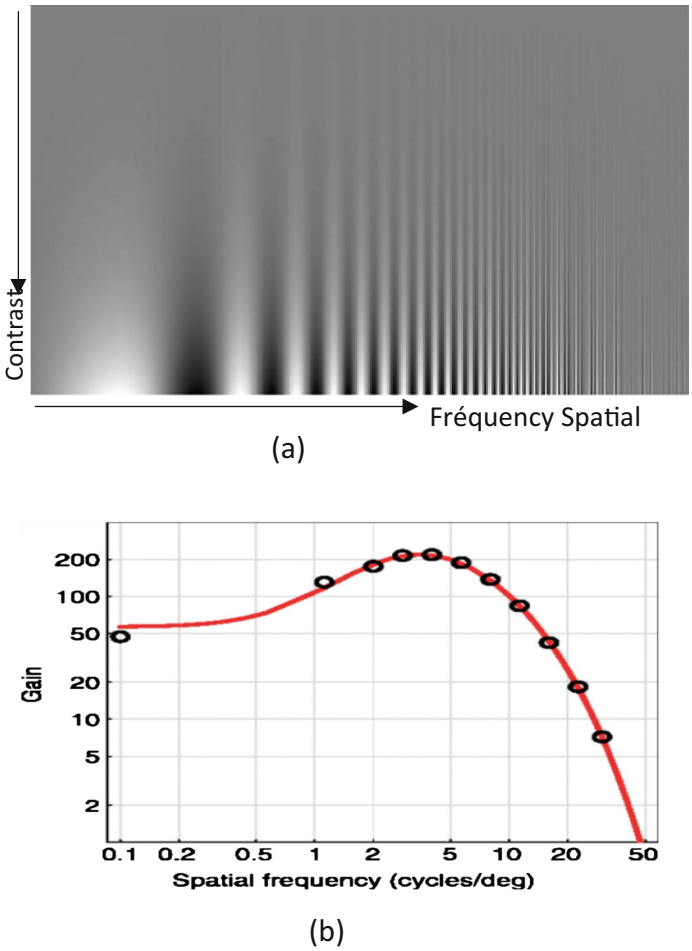


Fig. 2. The spectral properties of the HVS: (A) - graph illustrating Campbell- Robson contrast sensitivity (CSF), (b) - curve of HVS sensitivity as a function of spatial frequency.

curve tends to shift towards low frequencies. These studies were realized to achromatic data [18].

5.2 The Masking Effect

Masking is the effect of modification of the component bordures visibility in a multi-media content (masked signal) by the presence of another component (masking signal). The magnitude of this effect is measured by the change of the masked signal visibility with or without the presence of the masking signal [19]. Masking may intervene in the same frequency band (in-band masking) or between different frequency bands (inter-band masking). This type of mask in literature, called the entropy masking effect [20]. There is another type of contrast masking named masking contrast connected to the

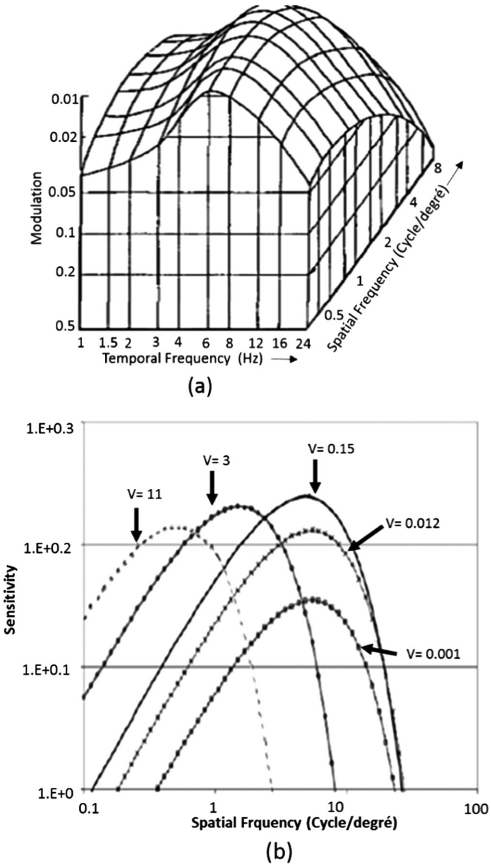


Fig. 3. Kelly Illustration (a) - for the presentation of the combination of the effect of spatial and temporal frequency on contrast sensitivity, and (b) - the CSF illustration depending on the speed. The time function of CSF in (b) is calculated from an empirical equation. The speeds V are measured in degree/second.

change of the surface visibility as a function of the contrast values. The spatial masking effect is often linked to the concept of surface roughness for 3D meshes [21].

5.3 The Perceptual Quality 3D Mesh

To measure the distance between two 3D meshes [22] many 3D metrics exist. We represent in the following sheet, the perceptual metrics developed for 3D meshes [23].

5.3.1 The 3D Watermarking Perception Metric

Corsini et al. have developed a new metric of quality [24]. Their approach, called 3DWPM (3D Watermarking Perception Metric), is based on the calculation of the perceptual distance between two meshes resting on surface roughness. This measured 3DWPM distance between two meshes M_1 and M_2 are identified by:

$$3\mathrm{DWPM}(\mathrm{M1},\mathrm{M2}) = \log\left(\frac{\rho(M2-\rho(M1)}{\rho(M1)} + K\right) - \log(\mathrm{k}) \tag{1}$$

where $\rho(\mathrm{M}_1)$ and $\rho(\mathrm{M}_2)$ measure the overall roughness of the two meshes and k is a constant varied digital. Two variants of 3DWPM were developed using descriptors of roughness differences. The first descriptor of roughness is used to 3DWPM1, is inspired by that of Lavoué [25]. The roughness measurement is calculated through measuring the dihedral angles between the normal facets in vicinity. The normal facet of a smooth surface does not vary greatly. In contrast, on textured regions (rough) these normal vary more significantly. A Multi - scale analysis of these entities is considered [26] to evaluate dihedral angles using the direct vicinity (1 ring) and the extended vicinity (1 ring, 2 rings, etc.). The second roughness measurement adopted by Corsini et al. is 3DWPM2, it is based on estimating of the roughness of surfaces by Wang et al. [27]. Their approach is based on the comparison of a mesh and smoothed versions of the same mesh. Smooth regions correspond to small differences while the rough areas have more significant differences.

5.3.2 The Mesh Structural Distortion Measure Metric

The MSDM (Mesh Structural Distortion Measure) uses the amplitude of the average curvature of the 3D mesh surface to quantify the perceptual distortion. This metric has been improved recently under the name MSDM2 by integrating multi-scale analyses [25]. In 2006, Lavoué et al. have introduced a structured distortion measurement called Structural Mesh Distortion Measure (MSDM) [28]. This quality measure was inspired by the quality measurement of 2D pictures SSIM (Structural SIMilarity index) introduced by Corsini et al. [24].

The MSDM is based on the statistical difference of the average amplitude curves to measure the perceptual difference of two meshes. The average curvatures ($\mathrm{CM}_{\mathrm{si}}$) are calculated for each as the average of the minimum and maximum curvatures:

$$\mathrm{CMsi} = \frac{(\mathrm{CM}_{\mathrm{si}}) = |Cmin, si| + |Cmax, si|}{2} \tag{2}$$

The amplitude of minimum and maximum curvatures ($|Cmin, si|$ and $|Cmax, si|$) are deducted from the fair values of the curvature tensor. The approximation of the curvature tensor used, is that introduced by Cohen-Steiner et al. on spatial vicinity defined by the geodesic disc resulting from the projection of a sphere of radius h on the surface of mesh. The average and standard deviation of the average curvature in a spatial window w containing n peaks, denoted respectively μ_w *and* σ_w are defined as:

$$\mu_w = \frac{1}{n}\sum\nolimits_{si\in w} CMsi, \tag{3}$$

$$\sigma_w = \sqrt{\frac{1}{n}\sum\nolimits_{si\in w} (CMsi - \mu_w)^2}, \tag{4}$$

The covariance between the curvatures of two window s $w1$ *and* $w2$ of the compared meshes M1 and M2 is defined as:

$$\sigma_{w_1 w_2} = \frac{\sigma^{w_1}_{w_1 w_2} + \sigma^{w_2}_{w_1 w_2}}{2} \tag{5}$$

where $\sigma^{w_1}_{w_1 w_2}$ defines the covariance calculated on the window $w1$:

$$\sigma^{w_1}_{w_1 w_2} = \frac{1}{n} \sum\nolimits_{S_{1i} \in w_1} \left(\left(CM^{w_1}_{S_{1i}} \right) - \mu_{w_1} \right) \left(\left(CM^{w_2}_{S2i} \right) - \mu_{w_2} \right) \tag{6}$$

where $S2i$ is the nearest summit to $S1i$ on $S1i$ the covariance σ^{w2}_{w1w2} on the window w_2 is calculated in the same way. To calculate the overall distance between two meshes, MSDM based on local distances, we note MSDML:

$$MSDML = (\alpha * L(s)^a + \beta * C(s)^a + \gamma * M(s)^a)^{\frac{1}{a}} \tag{7}$$

With α, β, γ and a are scalars for combining different quantities. The parameters L, C, M represent respectively the differences bends, contrasts and structures measured by:

$$L(s) = \frac{||\mu_{w1} - \mu_{w2}||}{MAX(\mu_{w1}, \mu_{w2})} \tag{8}$$

$$\mathbf{C}(\mathbf{w1}, \mathbf{w2}) = \frac{||\boldsymbol{\sigma}_{\mathbf{w1}} - \boldsymbol{\sigma}_{\mathbf{w2}}||}{\mathbf{MAX}(\boldsymbol{\sigma}_{\mathbf{w1}}, \boldsymbol{\sigma}_{\mathbf{w2}})} \tag{9}$$

$$M(w1, w2) = \frac{||\sigma_{w1}\sigma_{w2} - \sigma_{w1}\sigma_{w2}||}{\sigma_{w1}\sigma_{w2}} \tag{10}$$

where w represents the number of local windows on mesh surfaces to compare, and a parameter that has the same value as in the Eq. 7. An improved version of MSDM named MSDM2 was proposed in 2011 by integrating a multi-scale analysis [29]. MSDM2 also compares two meshes that do not share the same connectivity through a matching step through the structure data AABB tree implemented in the CGAL [30]. We note that to establish a measure of perceptual symmetrical distance, MSDM2 calculates the average of the two distances unidirectional (respectively M_1 to M_2 and M_2 to M_1). MSDM and MSDM2 are based only on statistics of the curvatures magnitudes.

5.3.3 The Fast Mesh Perceptual Distance Metric

In 2012, a new metric for estimating the perceptual quality was introduced par Corsini et al. [31]. It is entitled Fast Mesh Perceptual Distance (FMPD). The type of this metric is reduced-reference, it is based on the comparison of overall roughness measurements calculated on the two grids to compare. The roughness descriptor retained in FMPD is

derived from Laplacian of the Gaussian curvature. For each summit *Si* the Gaussian curvature (noted CG_i) is established by the following equation:

$$CG_i = \left| 2\pi - \sum_{j \epsilon N_i(F)} \alpha_j \right| \tag{11}$$

or *Ni* (*F*) represents all vicinity facets at the summit *Si* and the j angle αj represent the crossing facet with the actual summit. The local roughness, calculated on a top *si* is approximated by the quantity:

$$RL_i = \left| CG_i + \frac{\sum_{j \in N_i(S)} D_{ij} CG_j}{D_{ii}} \right| \tag{12}$$

where $N_i(s)$ is the set of neighboring vertices aware summit, the matrix D it's the array of discrete Laplace established by:

$$D_{ij} = \frac{\cot(\beta_{ij}) + \cot(\beta_{ij})}{2} pourj \epsilon N_i(s) \tag{13}$$

$$D_{ii} = -\sum_j D_{ij} \tag{14}$$

where βij and $\beta ij'$ are the opposite angles in which the peak relies *si* to *sj*. Figures 4(b) and (d) represent respectively the color card of the roughness measurement on the surface of the Armadillo and Venus meshes. The warm colors match at high RL values, cold colors correspond to low values of RL. We note that this descriptor is able to properly evaluate and classify the surfaces of meshes into smooth and rough areas regions. FMPD in the local roughness measurements are modulated by a power function [32] to establish local values of roughness RLF_i in each peak s_i. This modulation is performed to capture the spatial masking effect on the mesh surfaces.

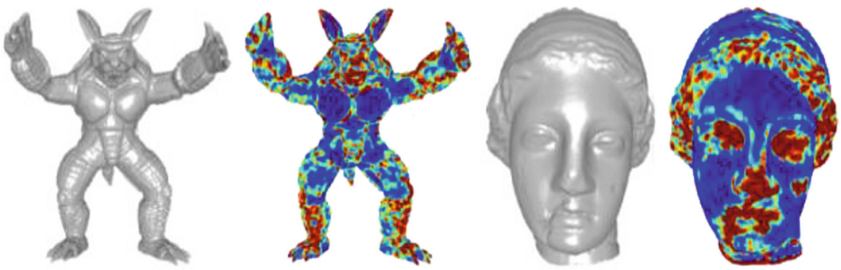


Fig. 4. (a) - and (c) - respectively present the 3D meshes of Armadillo and Venus, (b) - and (d) - present colors of the cards from the respective surfaces roughness Armadillo of mesh and mesh Venus.

The overall roughness is subsequently calculated by a steady sum of the local measurements:

$$\mathrm{RG} = \frac{\sum_i RLF_i a_i}{\sum_i a_i} \tag{15}$$

where a_i is a coefficient connected to the area of the incident facets at the summit s_i the measure of the perceptual distance FMPD is established as the difference between the values of global roughness RG_1 and RG_2 of two meshes M_1 and M_2 to compare:

$$FMPD = c|RG_1 - RG_2| \tag{16}$$

With c as a scalar set a scale parameter for the FMPD in the interval distances [0; 1].

5.3.4 The Dihedral Angle Mesh Error Metric

Recently, Vasa and Rus have developed a new metric of perceptual quality for static 3D mesh based on the dihedral angles [33]. This metric, named DAME (Dihedral Angle Mesh Error) offers a compromise between the complexity and the prediction efficiency. On each pair of triangles (t1 = $\{S_1S_2S_3\}$; t2 = $\{S_3S_2S_4\}$) neighbors of the mesh, the dihedral angle oriented are calculated use this expression:

$$D_{t1t2=\arccos(n1,n2)*sgn(n1.(s4-s3))} \tag{17}$$

where n_1 and n_2 represent respectively the surface normal associated with t1 and t2 triangles. Measurement of DAME perceptual distance is calculated as follows:

$$\mathrm{DAME} = \frac{1}{||\Omega||} \tag{18}$$

$$\sum\nolimits_{\{t1,t2\}\in\Omega} \left\| D_{t1t2} - \overline{D_{t2t2}} \right\| \cdot m_{t1t2}.(w_1 + w_2) \tag{19}$$

where D_{t1t2} and $\overline{D}_{t1t2}$ respectively represent the measurements of the dihedral angles of the two meshes to compare, m and w represent respectively a masking modeling function and a visibility coefficient.

6 Conclusion

In this paper, we have studied the importance of perceptual quality 3D meshes compared to the objective measures meshes. We have also enumerated the different existing approaches which are based on the metrics of the perceptual quality 3D meshes. We have also detailed various properties of the HVS, such as the CSF and the masking effect. It is important to consider these properties for guiding quality measures development process perceptual 3D meshes. These existing perceptual objective

metrics mainly considers the spatial masking effect to establish quality metrics collected for 3D meshes. Finally we have itemized metrics which are based on the HVS properties.

References

1. Lecuire, V., Duran-Faundez, C., Krommenacker, N.: Energy-efficient image transmission in sensor networks. Int. J. Sens. Netw. **4**(1/2), 37–47 (2008)
2. Salehpour, M., Behrad, A.: 3D face reconstruction by KLT feature extraction and model consistency match refining and growing. In: 2012 6th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT), Sousse, pp. 297–302 (2012)
3. Chan, A.T., Gamino, A.: Integration of assistive technologies into 3D simulations: an exploratory study. In: Information Technology: New Generations, Advances in Intelligent Systems and Computing, vol. 448, pp. 425–437. Springer (2016). ISBN 978-3-319-18295-7
4. El-Bendary, M.A.M., El-Tokhy, M., Kazemian, H.B.: Efficient image transmission over low-power IEEE802.15.1 network over correlated fading channels. In: The 6th International Conferences: Sciences of Electronics, Technologies of Information and Telecommunications "SETIT 2012", Mars 2012, Sousse-Tunisie, IEEE Conferences, pp. 563–567 (2012). doi:10.1109/SETIT.2012.6481973
5. Abderrahim, Z., Techini, E., Bouhlel, M.S.: State of the art: compression of 3D meshes. Int. J. Comput. Trends Technol. (IJCTT) **4**(6), 765–770 (2012)
6. Tang, H., Joshi, N., Kapoor, A.: Learning a blind measure of perceptual image quality. In: Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition, pp. 305–312 (2011)
7. Hemanth, D.J., Balas, V.E., Anitha, J.: Hybrid neuro-fuzzy approaches for abnormality detection in retinal images. In: Proceedings of the 6th International Workshop Soft Computing Applications, SOFA 2014, Timisoara, Romania, 24–26 July 2014, pp. 295–305 (2014)
8. El-Bendary, M.A.M., El-Tokhy, M., Shawki, F., Abd-El-Samie, F.E.: Studying the throughput efficiency of JPEG image transmission over mobile IEEE 802.15.1 network using EDR packets. In: The 6th International Conferences: Sciences of Electronics, Technologies of Information and Telecommunications "SETIT 2012", Mars 2012, Sousse-Tunisie, IEEE Conferences, pp. 573–577 (2012). doi:10.1109/SETIT.2012.6481975
9. Escribano-Barreno, J., García-Muñoz, J.: Integrated metrics handling in open source software quality management platforms. In: Information Technology: New Generations. Advances in Intelligent Systems and Computing, vol. 448, pp. 509–518. Springer (2016). ISBN 978-3-319-18295-7
10. Triki, N., Kallel, M., Bouhlel, M.S.: Imaging and HMI, fondations and complementarities. In: The 6th International Conferences: Sciences of Electronics, Technologies of Information and Telecommunications, SETIT 2012, Mars 2012, Sousse-Tunisie, IEEE Conferences, pp. 25–29 (2012). doi:10.1109/SETIT.2012.6481884
11. Nandi, D., Ashour, A.S., Samanta, S., Chakraborty, S., Salem, M.A., Dey, N.: Principal component analysis in medical image processing: a study. Int. J. Image Min. **1**(1), 65–86 (2015)
12. Cho, J.-W., Prost, R., Jung, H.-Y.: An oblivious watermarking for 3-D polygonal meshes using distribution of vertex norms. IEEE Trans. Sig. Process. **55**(1), 142–155 (2007)

13. Wang, K., Lavoué, G., Denis, F., Baskurt, A.: Robustand blind mesh watermarking based on volume moments. Comput. Graph. **35**(1), 1–19 (2011)
14. Abderrahim, Z., Techini, E., Bouhlel, M.S.: Progressive compression of 3D objects with an adaptive quantization. Int. J. Comput. Sci. Issues (IJCSI) **10**(2), 504–511 (2013)
15. Campbell, F.-W., Robson, J.-G.: Application of Fourier analysis to the visibility of gratings. J. Physiol. **197**, 551–566 (1968)
16. Hirai, K., Tsumura, N., Nakaguchi, T., Miyake, Y., Tominaga, S.: Spatio-velocity contrast sensitivity functions and video quality assessment. In: International Symposium on Intelligent Signal Processing and Communication Systems, pp. 1–4 (2010)
17. Ninassi, A., Meur, O.L., Le Callet, P., Barba, D.: On the performance of human visual system based image quality assessment metric using wavelet Domain. In: Proceedings of the SPIE Human Vision and Electronic Imaging (2008)
18. Ninassi, A., Meur, O.L., Le Callet, P., Barba, D.: Which semi-local visual masking model for wavelet based image quality metric In: Proceedings of IEEE International Conference on Image Processing, pp. 1180–1183 (2008)
19. Fernandez-Maloigne, C., Larabi, M.-C., Bringier, B., Richard, N.: Spatio temporal characteristics of the human color perception for digital quality assessment. In: Proceedings of International Symposium on Signals, Circuits and Systems, pp. 203–206 (2005)
20. Kelly, D.-H.: Visual contrast sensitivity. Opt. Acta: Int. J. Opt. **24**(2), 107–129 (1977)
21. Daly, S.-J.: Engineering observations from spatio velocity and spatio temporal visual models. In: Proceedings of the SPIE Human Vision and Electronic Imaging III, vol. 3299, pp. 180–191 (1998)
22. Wang, Z., Bovik, A.-C.: Modern Image Quality Assessment. Morgan & Claypool, San Rafael (2006)
23. Rogowitz, B.-E., Rushmeier, H.-E.: Are image quality metrics adequate to evaluate the quality of geometric objects. In: Proceedings of SPIE Human Vision and Electronic Imaging, pp. 340–348 (2001)
24. Corsini, M., Drelie Gelasca, E., Ebrahimi, T., Barni, M.: Watermarked 3-D mesh quality assessment. IEEE Trans. Multimedia **9**(2), 247–256 (2007)
25. Lavoué, G.: A multiscale metric for 3D mesh visual quality assessment. Comput. Graph. Forum **30**(5), 1427–1437 (2011)
26. Lavoué, G., Drelie Gelasca, E., Dupont, F., Baskurt, A., Ebrahimi, T.: Perceptually driven 3D distance metrics with application to watermarking. In: Proceedings of SPIE Electronic Imaging (2006)
27. Wang, Z., Bovik, A.-C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)
28. Alliez, P., Tayeb, S., Wormser, C.: 3D fast intersection and distance computation (AABB tree). In: CGAL User and Reference Manual (2012)
29. Wu, J.-H., Hu, S.-M., Tai, C.-L., Sun, J.-G.: An effective feature-preserving mesh simplification scheme based on face constriction. In: Pacific Conference on Computer Graphics and Applications, pp. 12–21 (2001)
30. Gelasca, E.-D., Ebrahimi, T.: Objective evaluation of the perceptual quality of 3D watermarking. In: IEEE International Conference on Image Processing, pp. 241–244 (2005)
31. Corsini, M., Gelasca, E.-D., Ebrahimi, T.: A multi-scale roughness metric for 3D watermarking quality assessment. In: Workshop on Image Analysis for Multimedia Interactive Services (2005)
32. Wang, K., Torkhani, F., Montanvert, A.: A fast roughness-based approach to the assessment of 3D mesh visual quality. Comput. Graph. **36**(7), 808–818 (2012)
33. Vása, L., Rus, J.: Dihedral angle mesh error: a fast perception correlated distortion measure for fixed connectivity triangle meshes. Comput. Graph. Forum **31**(5), 1715–1724 (2012)

Fireworks Algorithm Based Image Registration

Silviu-Ioan Bejinariu[1(✉)], Hariton Costin[1,2], Florin Rotaru[1], Ramona Luca[1], Cristina Diana Niţă[1], and Camelia Lazăr[1]

[1] Institute of Computer Science, Romanian Academy Iaşi Branch, Iasi, Romania
{silviu.bejinariu, florin.rotaru, ramona.luca, cristina.nita, camelia.lazar}@iit.academiaromana-is.ro, hncostin@mail.umfiasi.ro
[2] Faculty of Medical Bioengineering, "Grigore T. Popa" University of Medicine and Pharmacy, Iasi, Romania

Abstract. In the Image Processing (IP) domain, optimization algorithms have to be applied in many cases. Nature-inspired heuristics allow obtaining near optimal solutions using lower computing resources. In this paper the Fireworks Algorithm (FWA) behavior is studied for Image Registration (IR) problems. The IR results accuracy is analyzed for different types of images, mainly in case of pixel based registration using the Normalized Mutual Information. FWA is compared to Particle Swarming (PSO), Cuckoo Search (CSA) and Genetic Algorithms (GA) in terms of results accuracy and number of objective function evaluations required to obtain the optimal geometric transform parameters. Because the pixel based IR may fail in case of images containing graphic drawings, a features based IR approach is proposed for this class of images. Comparing to other nature inspired algorithms, FWA performances are close to those of PSO and CSA in terms of accuracy. Considering the required computing time, that is determined by the number of cost function evaluations, FWA is little slower than PSO and much faster than CSA and GA.

Keywords: Fireworks Algorithm · Optimization · Image registration

1 Introduction

Image registration (IR) is the process of geometric overlaying or alignment of two or more images of the same scene for subsequent common processing [1]. The goal of an IR procedure is to approximate the parameters of the geometric transform to be applied to a *source* image such that it is aligned to a *model* image. The overlaying quality is evaluated using a similarity criterion which must be maximized using an optimization algorithm. In case of a large number of parameters, or of complex similarity measures, the optimal parameters may be difficult to obtain. Based on the strategies of live beings to survive, find food or avoid obstacles, or modeling other natural phenomena, the nature-inspired metaheuristics offer the possibility to obtain faster a near optimal solution [2].

V.E. Balas et al. (eds.), *Soft Computing Applications*, Advances in Intelligent Systems and Computing 633, DOI 10.1007/978-3-319-62521-8_44

A relatively new nature inspired algorithm is the Fireworks Algorithm (FWA) proposed in [3] and deeply analyzed in [4]. A detailed study and some enhancements of the operators (strategies) used in FWA are presented in [5]. Other versions of FWA are: an adaptive version of FWA proposed in [6] and a cooperative strategy among fireworks in [7]. Multi-objective versions of FWA were also developed [8]. In most papers the algorithm's performances are studied using benchmark functions.

The IR issue was addressed in [9–11] using Bacterial Foraging Optimization and in [12] using Bat and Cuckoo Search algorithms. In [13] the behavior of Bacterial Foraging Optimization, Particle Swarming, Multi-swarming and Firefly algorithms in multi-threshold image segmentation was studied. In [14] single and multi-objective FWA convergence was analyzed using test functions.

The paper is organized as follows. In the second section FWA is briefly described. The third section describes the IR problem and the experiments made using different types of images, including an analysis of an IR failure case. In the fourth section, the accuracy of FWA results is compared to that of the results obtained using the Particle Swarming, Cuckoo Search and Genetic Algorithms as optimization procedure. In the fifth section, a features based IR approach is shortly described. It allows the registration of images that contain graphic drawings, that fails in case of pixel based registration. The last section concludes the paper.

2 Fireworks Algorithm

The Fireworks Algorithm (FWA) belongs to the Swarm Intelligence algorithms category and it is inspired from the fireworks explosion simulation [3, 4]. As in other swarming algorithms, the possible solutions are encoded as position of individuals that evolve in the problem domain. The evolution is simulated by selection of sparks resulted in the fireworks explosion process. The objective to optimize is evaluated in all the reached positions and the best obtained position is considered as being the problem solution [14].

FWA was initially developed for single-objective optimization problems, in which the parameter value that minimize (or maximize) a fitness function have to be determined:

$$\min_{x \in S} f(x) \tag{1}$$

where S is the problem domain (constraints).

FWA is an iterative algorithm which starts with a set of individuals (fireworks) randomly placed in the problem domain. For each iteration (*epoch*) of FWA, a variable number of descendants (*sparks*) is generated for each firework. The descendants are evaluated and some of them became fireworks in the next evolution step (epoch).

Let's consider $X = \{x_i, i = 1, \ldots N\}$ a set of N fireworks, and $x_i = \left(x_i^1, \ldots, x_i^d\right)$ their positions in the problem domain, where d is the dimension of the problem.

The general structure of the algorithm is described below [4]:

Generate X= $\{x_i, i = 1, \ldots N\}$ a randomly placed set of fireworks
Compute fitness for each firework x_i
while the stop condition is not met
  for each firework $x_i \in X$
    Compute S_i, the number of sparks for each firework
    Compute A_i, the sparks amplitudes for each firework
  end for
  Generate regular sparks
  Evaluate sparks and check for the best solution
  Generate Gaussian sparks
  Evaluate sparks and check for the best solution
  Select the next generation: the best, the worst and other $N - 2$ random selected sparks
end while

Each step of the FWA loop involves strategies that are important for the algorithm's performances. The fireworks explosion which ensures the quality of the next fireworks generation is characterized by strength, amplitude and displacement.

The explosion strength refers to the number of sparks S_i of each firework. It depends on the current firework quality, and it is computed as:

$$S_i = m * \frac{f_{worst} - f(x_i) + \varepsilon}{\sum_1^n (f_{worst} - f(x_i)) + \varepsilon} \tag{2}$$

where m is the total number of sparks, f_{worst} is the worst fitness value of the current fireworks set, $f(x_i)$ is the fitness value of the i^{th} firework and ε is a constant used to avoid zero-division operations. To keep the number of sparks in a reasonable interval, minimum and maximum values that depend on m are imposed for S_i.

The amplitude A_i of the i^{th} firework sparks is used to avoid the convergence to local solutions. The amplitude is higher as the distance to the best firework increases and it is computed as:

$$A_i = \overline{A} * \frac{f(x_i) - f_{best} + \varepsilon}{\sum_1^n (f(x_i) - f_{best}) + \varepsilon} \tag{3}$$

where $\overline{A}$ is the desired sum of amplitudes, f_{best} is the best fitness value in the current fireworks set.

Two types of sparks are created in the i^{th} firework explosion: regular and Gaussian. Regular sparks are obtained from the parent firework by adding a uniform random number in the interval $(-A_i, A_i)$ to some randomly chosen coordinates.

$$x_i^k = x_i^k + U \tag{4}$$

where k are the randomly chosen coordinates and U is the uniform random value.

The Gaussian sparks assure the diversity of the next fireworks generation by altering some randomly chosen coordinates of some randomly chosen fireworks by random numbers with Gaussian distribution:

$$x_i^k = x_i^k * g \tag{5}$$

where k are the randomly chosen coordinates and g are random numbers.

If the generated sparks are placed outside the problem domain they must be re-positioned inside.

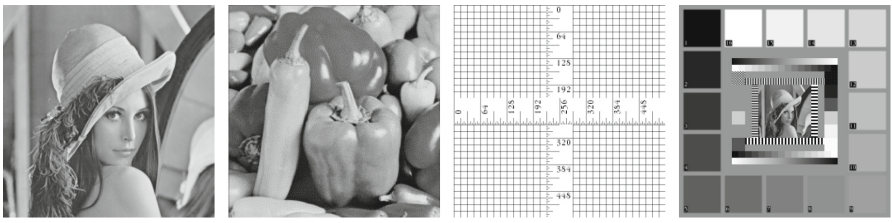
The fitness function is evaluated in the all new sparks and the next generation fireworks population is selected. The best, the worst and the other $N - 2$ randomly chosen fireworks replace the population of fireworks.

3 FWA Based Image Registration

Image registration is the process of geometric overlaying or alignment of two or more images of the same scene taken at different times, from different viewpoints, and/or by different sensors [1]. There are two different approaches of IR: pixel level and features level registration. In both cases, IR is an optimization procedure because it requires to approximate the parameters of a geometric transform which applied to a source image minimize a similarity measure computed between the transformed source image and a *model* image.

In the FWA based IR the possible solutions are encoded as positions of the fireworks and the objective to optimize is the Normalized Mutual Information (NMI). The fireworks evolve in a multidimensional space whose dimension is equal to the number of parameters in the geometric transform to be approximated. In the next paragraphs the Normalized Cross-Correlation (NCC) is also used to evaluate the results, but it is not used as optimization criteria.

The FWA based IR procedure was evaluated using the images available in the 'Miscellaneous' volume of USC-SIPI database [15]. All used images were resized to 256×256 resolution, and color depth was reduced to 256 gray levels if not already so. In the following paragraphs the results obtained using four model images from this database are presented. Two of these images contain photos (Fig. 1a and b), one contains graphics (Fig. 1c) and the last one is mixed (Fig. 1d).



a. 'Lena' model image b. 'Peppers' model image c. 'Ruler' model image d. 'Testpat' model image

Fig. 1. Original test images, from [15]

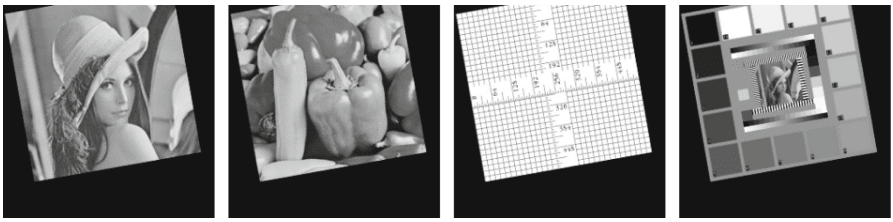
The source images were artificially generated by applying to model images an affine transform with the following parameters: rotation by $\theta = 10^\circ$ against the rotation center $c_x = 30, c_y = 30$ and scaling by the factor $s = 0.8$. The *computed* inverse affine transform has the following parameters: rotation by $\theta' = -10^\circ$ against the same center and scaling factor $s' = 1.250$.

FWA was applied to register the source images (Fig. 2) to model images (Fig. 1). The parameters of the algorithm were chosen as follows:

- fitness function: NMI,
- number of epochs = 250,
- number of fireworks = 5,
- maximum amplitude = 40,
- regular sparks factor = 10 and
- number of Gaussian sparks = 5.

These values of FWA parameters were chosen accordingly to those proposed in [4] and they were validated by our experiments.

Because optimal value of NMI is unknown, there was no stop criterion other than the number of epochs. It must be noticed that the fitness function evaluation requires to apply the approximated geometric transform that correspond to the firework position and then to compute the similarity measure between the model image and transformed source image. This evaluation is time consuming; in fact more than 90% of IR execution time is spent in the fitness function evaluation [10].



a. 'Lena' source image b. 'Peppers' source image c. 'Ruler' source image d. 'Testpat' source image

Fig. 2. Source images obtained by applying the geometric transform to model images

Even if the number of individuals (fireworks) and iteration (epochs) are reduced comparing to other nature-inspired IR procedures, the results are quite good.

The final results evaluation is performed using two different similarity measures: NMI and NCC. The columns named *Expected* contain the NMI/NCC values computed between the model image and source image transformed using the *computed* inverse transform. The columns named *Approximated* contain the NMI/NCC values computed between the model image and source image transformed using the *approximated* inverse transform. The column '#evals' contains the total number of cost function evaluations, column 'Best eval' the number of the evaluation in which the best fitness was reached and the last column contains the execution time in seconds – not so relevant while it depends on computer's performances.

By analyzing the values in Tables 1 and 2, the following conclusions can be drawn:

- In most cases the approximated value of the similarity measure is greater than the expected value. This can be explained by the two geometric transforms applied to the image which means that pixels values were interpolated two times.
- The parameters of the approximated inverse transform (Table 2) are very close to those of the computed inverse.
- In case of 'Ruler' image the registration process failed. The cause is discussed below.
- The best value was obtained enough late in the iterative process, which means that better solutions can be obtained by increasing the number of iterations. Regarding the number of epochs and number of fireworks used in FWA it must be said that a reduced number of fireworks (3 or 4) requires a considerably increase of the number of epochs in order to obtain similar results. By increasing the number of epochs and keeping constant the number of fireworks (=5), the results are improved but not significantly while the processing time greatly increases.
- The application was tested on an Intel Core i3, 2.4 GHz, 6 GB RAM based computer. The execution time is presented for the case of parallel execution. Even if not present in the Table 1, the sequential execution time is about 30% greater with the observation that other applications were running in the same time.

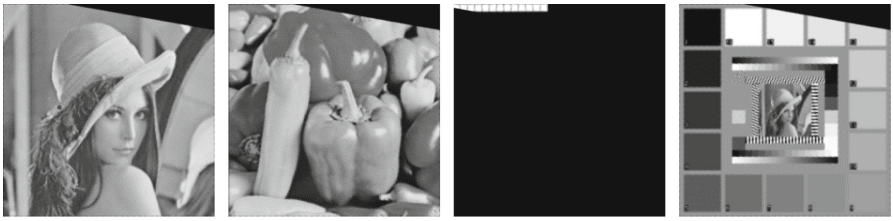
Table 1. Results evaluation of IR process

Image	NMI		NCC		# evals	Best eval	Time (sec)
	Expected	Approximated	Expected	Approximated			
Lena	1.3252	1.3253	0.9765	0.9766	14369	14062	27.300
Peppers	1.3463	1.3466	0.9675	0.9675	14390	14329	27.986
Ruler	1.1147	1.0492	0.9699	0.3511	14109	13867	26.567
Testpat	1.5014	1.5014	0.9525	0.9525	14349	14223	29.422

Table 2. Parameters of the approximated inverse transform

Image	c'_x	c'_y	θ'	s'
Lena	30.1959	30.0245	−9.9944	1.2504
Peppers	30.0360	30.0073	−10.0032	1.2500
Ruler	40.7941	48.6515	−9.9937	1.2519
Testpat	29.9996	30.0161	−10.0012	1.2500

The registered images are depicted in Fig. 3. The failure of 'Ruler' image (which contains graphics) registration is also visible (Fig. 3c).

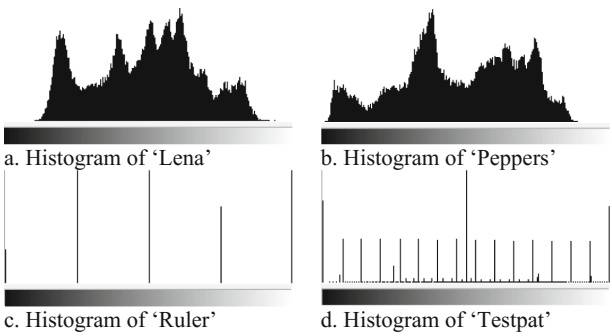


a. 'Lena' registered image | b. 'Peppers' registered image | c. 'Ruler' registered image | d. 'Testpat' registered image

Fig. 3. Registered images obtained by applying the approximated inverse transform

3.1 IR Failure Analysis

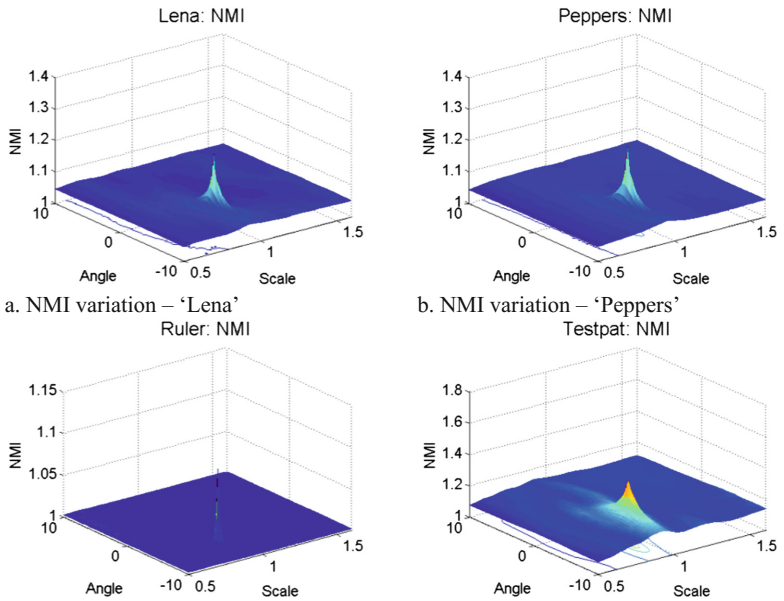
As it was seen above, the IR of 'Ruler' image failed. Let's analyze the histograms of the four images (Fig. 4). Only 5 levels of gray are used in the 'Ruler' image, most of them corresponding to the image background. By applying the affine transform to the model, most pixels have the same or close values to those in the original, leading to small variations of the similarity measure. This is visible in Fig. 5 which contains for all images the variation of NMI depending on the rotation angle and scale factor, keeping the rotation center fixed and equal to the values determined in the computed inverse transform. The peak of NMI surface is visible for all images, but the variation is steep close to the maximum value in case of 'Ruler' while the growth is smoother for the other images. It must be specified also that these surfaces have about 400 local maximum values in case of 'Lena' and 'Peppers', 122 local maximum in case of 'Testpat' and about 1040 maximum values in case of 'Ruler'. Considering the fact that during evolution only 5 fireworks are used, it is understandable why the algorithm converges to a local solution. In fact, because the initial positions are randomly chosen, for repeated executions of IR procedure on 'Testpat' image, in about 10% of cases the algorithm leads to the correct solution.



a. Histogram of 'Lena' | b. Histogram of 'Peppers'

c. Histogram of 'Ruler' | d. Histogram of 'Testpat'

Fig. 4. Histograms of the model images



a. NMI variation – 'Lena'
b. NMI variation – 'Peppers'
c. NMI variation – 'Ruler'
d. NMI variation – 'Testpat'

Fig. 5. Normalized mutual information variation

The easiest solution to solve these failures is to change the algorithm's parameters. The attempt to increase the number of iterations (epochs) to 500 or 1000 did not lead to the expected result, because it depends on the spatial distribution of the initial population. By increasing the number of fireworks in the initial population to 10 or 15, the percent of cases in which the algorithm leads to the correct solution has increased to about 25%. In our opinion, both these parameters must be increased, even if this solution is not feasible because the execution time increases also.

Another solution of this problem is the usage of features based IR. Some results are shortly described later in the fifth section.

3.2 Image Registration of Noise Affected Images

The behavior of the FWA based IR procedure was studied also for noise affected images. The source images were warped using the same affine transform described in the previous section and then they were degraded using Gaussian or Salt & Pepper noise. The degraded source images were registered to the original model images. Image degradation is evaluated using the Signal-To-Noise ratio (Table 3). The objective evaluation of IR results is presented in Tables 4 and 5 and the processed images are presented in Fig. 6.

The conclusion that can be drawn is that the registration process is not affected by the degradation of images, but depends on the similarity measure used as objective function in the optimization. In fact this demonstrates that the procedure can be applied for multimodal images.

Table 3. SNR value for degraded source images

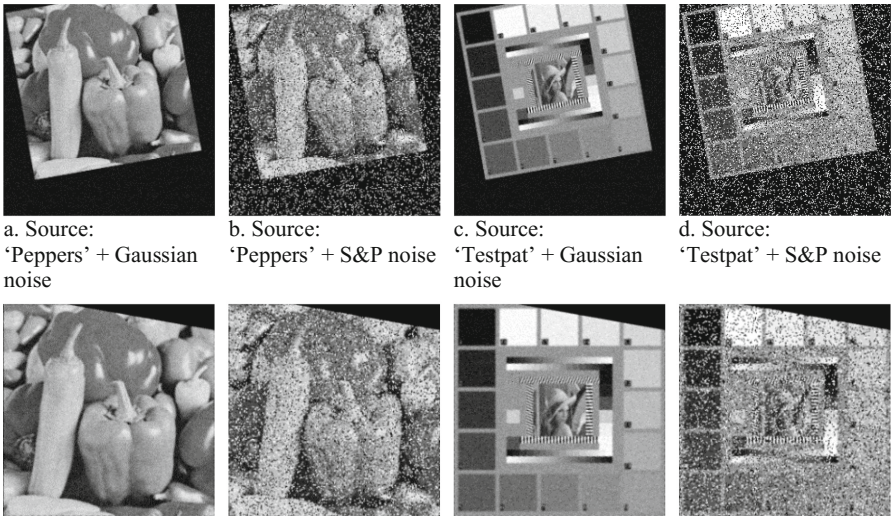
Model image	SNR (dB)	
	S&P noise	Gaussian noise
Peppers	2.46	21.19
Testpat	2.72	21.72

Table 4. Evaluation of IR of degraded images

Image	Noise	NMI		NCC		# evals	Best eval	Time (sec)
		Expected	Computed	Expected	Computed			
Peppers	Gaussian	1.2081	1.2081	0.9662	0.9661	14289	13812	29.890
	S&P	1.1160	1.1166	0.9069	0.9070	14491	14390	33.291
Testpat	Gaussian	1.2526	1.2533	0.9513	0.9517	14516	14497	24.258
	S&P	1.1407	1.1410	0.8921	0.8919	14408	13962	26.801

Table 5. Computed parameters of the inverse transform

Image	Noise	cx	cy	Angle	Scale
Peppers	Gaussian	29.8557	29.9494	−10.0009	1.2498
	S&P	30.0255	30.0632	−10.0169	1.2504
Testpat	Gaussian	30.5535	30.0995	−10.0220	1.2513
	S&P	29.9779	30.0260	−9.9906	1.2497



a. Source: 'Peppers' + Gaussian noise
b. Source: 'Peppers' + S&P noise
c. Source: 'Testpat' + Gaussian noise
d. Source: 'Testpat' + S&P noise
e. Registered: 'Peppers' + Gaussian noise
f. Registered: 'Peppers' + S&P noise
g. Registered: 'Testpat' + Gaussian noise
h. Registered: 'Testpat' + S&P noise

Fig. 6. IR results obtained using degraded source images

4 FWA Versus Other Optimization Algorithms

The accuracy of the image registration results is compared to that obtained using other optimization algorithms: Particle Swarming (PSO), Cuckoo Search (CSA) and Genetic Algorithm (GA).

The PSO implementation described in [13] was customized for IR with the following parameters: number of particles – 100, number of iterations – 100, inertia weight – 0.729, local weight – 1.4945 and global weight – 1.4945. The weights are those proposed in [16]. Our experiments demonstrated that even small variations of these values decrease the results accuracy.

The CSA based IR procedure [12] was used with the following parameters: number of nests – 25, number of iterations – 1000 and discovery rate – 0.25.

For GA based IR an implementation based on real encoding (each chromosome is characterized by real values representing the geometric transform parameters) was used [17]. Discrete, average and simplex crossover operators were applied with different probabilities. The GA was applied as in [17] using the following parameters: number of generations – 500, number of chromosomes – 1500, probability of discrete crossover – 0.05, probability of average crossover – 0.15, probability of simplex crossover – 0.2 and mutation rate – 0.2.

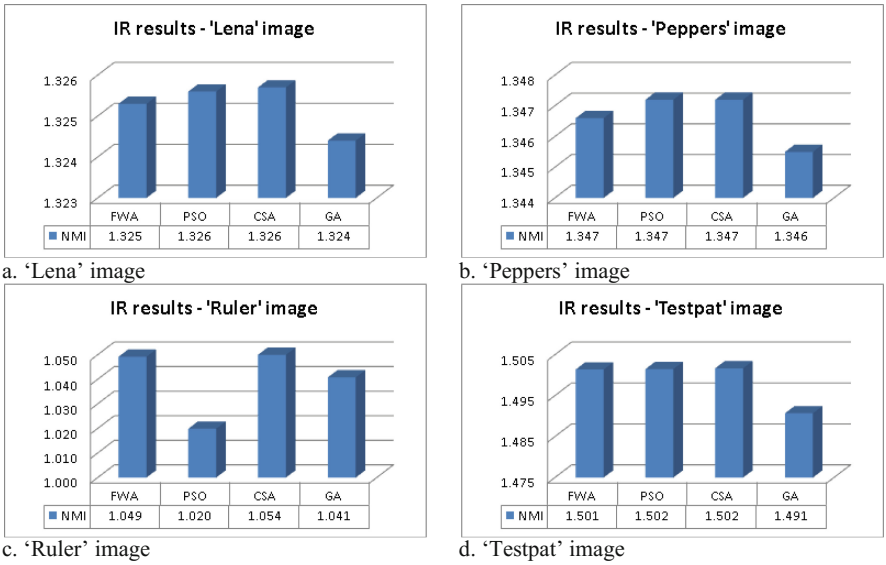
The IR procedure was applied using all these optimization algorithms on the same images as in the previous section. The obtained results are described in Table 6. Both NMI and NCC values are presented but NMI was used as objective function to be maximized. The maximum values of NMI are embossed for each image case.

Table 6. Evaluation of IR results using FWA, PSO, CSA and GA as optimization algorithm

Image			FWA		PSO		CSA		GA	
	Expected		Computed		Computed		Computed		Computed	
	NMI	NCC	NMI	NCC	NMI	NCC	NMI	NCC	NMI	NCC
Lena	1.3252	0.9765	1.3253	0.9766	1.3256	0.9766	**1.3257**	0.9766	1.3244	0.9765
Peppers	1.3463	0.9675	1.3466	0.9675	**1.3472**	0.9676	**1.3472**	0.9676	1.3455	0.9675
Ruler	1.1147	0.9699	1.0492	0.3511	1.0199	0.3184	**1.0535**	0.8995	1.0408	0.2925
Testpat	1.5014	0.9525	1.5014	0.9525	1.5015	0.9525	**1.5017**	0.9526	1.4907	0.9523

The same results are graphically presented for each test image in Fig. 7. The values presented in Fig. 7 have been rounded to three decimal digits. For this reason, in some graphs the values are equal but the corresponding bars have different heights. A brief analysis of these results leads to the following conclusions:

- First of all, it must be noticed that in case of 'Ruler' test image the registration fails also in case of PSO, CSA and GA optimization,
- All four optimization algorithms produce enough good results for three test images: 'Lena', 'Peppers' and 'Testpat'.
- CSA offers the best results for all four images. In fact, also for 'Ruler' test image, FWA, PSO and GA produce a result similar to that presented in Fig. 3c, while CSA



a. 'Lena' image
b. 'Peppers' image
c. 'Ruler' image
d. 'Testpat' image

Fig. 7. Evaluation of pixel based IR results obtained using FWA, PSO, CSA and GA for each test image

tends to approximate the inverse geometric transform. The 'scale' and 'angle' parameters are enough well approximated, only the rotation center coordinates are pretty much shifted from the expected values. It is possible that an increased number of iterations and/or an increased number of host nests allow CSA to obtain a better solution.

- The results obtained by using GA in optimization are less accurate than the other three algorithms results.
- The accuracy of results obtained by using PSO and FWA is good even if lower than CSA results accuracy.

The PSO, CSA and GA based IR procedures were applied also for the images altered by Gaussian and Salt & Pepper noise with quite similar results.

A complete analysis of the registration results have to consider also the processing time which is determined by the number of objective function evaluations. As it was already noticed in the previous section, in case of pixel based IR applications, more than 90% of the processing time is spent in these evaluations, so, reducing them may be important at least for real time applications. In Fig. 8, the average number of objective function evaluations is presented. On the other hand, the number of evaluations *required* to obtain the best solution is important given the fact that this class of algorithms usually obtain the best solution during the evolution but not necessary in the last iteration.

As is depicted in Fig. 8, FWA and PSO use a lower number of cost function evaluations, while GA uses the greatest number while its results are less accurate. Concerning the number of cost function evaluations *required* to obtain the best solution,

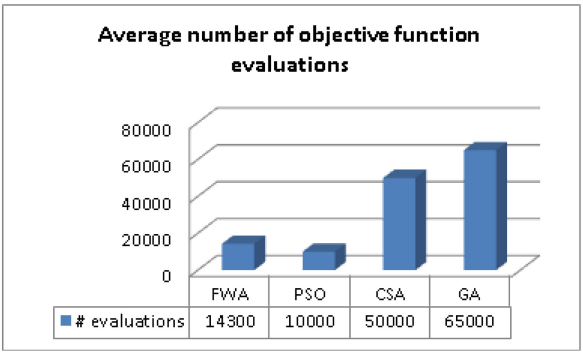


Fig. 8. Average number of objective function evaluations

it must be noticed that in case of FWA, PSO and CSA it is close to the total number of evaluations. This leads to the conclusion that more precise solutions can be obtained by increasing the number of iterations and/or individuals used in the evolutionary process. A different situation is encountered in case of GA. Its best solutions are obtained after about 8–9000 cost function evaluations, i.e. about one eight of the total number which is lower than in case of the other algorithms. This means that GA optimization represents a good choice if less accurate solutions are acceptable.

As a conclusion, CSA can be used to obtain more precise solutions; GA can be used to faster obtain an acceptable solution; FWA and PSO is the proper choice if good solutions have to be also quickly obtained.

5 Pixel Based Versus Features Based IR

We still have to find a solution to register images that contain graphic drawings, like the 'Ruler' test image. A good solution is the usage of features based IR. In this case, distinctive and stable features (points, lines, contours, regions) have to be detected in both model and source images. A features based IR procedure using bio-inspired computing is presented in [17]. As features the key-points determined using the Scale Invariant Feature Transform (SIFT) [18] are used.

To evaluate images similarity, the correspondences between features present in both images have to be identified. The Euclidean distances between SIFT descriptors are computed for all key-points pairs detected in the model and source images. For each key-point descriptor in the model image, the distances to the key-points descriptors in the source image are sorted in ascendant order. If the smallest distance is less than a specific percent (usually 30%) of the second distance then a match is established between the key-point in the model image and the key-point in the source image [17].

The registration procedures are similar excepting the objective function used in optimization. In case of features based IR the Euclidean distance between the positions of the common features in the two images is used for images similarity evaluation. It must be noticed also that in this case the geometric transform corresponding to each

possible solution have to be applied only for the key-points coordinates while in the pixel based IR, the entire source image have to be transformed. This explains why the features based IR is much faster than pixel based IR. In fact, increasing the number of iterations and/or individuals does not have a great impact on the processing time.

However, features based IR has one great disadvantage. For multimodal or some noise affected images the correspondences between stable features in the two images can't be established.

Trying to register images that contain graphic drawings ('Ruler' image), the four optimization algorithms were applied for features based IR using the SIFT key-points. A number of 2607 stable features were determined in the model image and 2525 stable features were determined in the source image. From these only for 60 pairs of key-points the correspondence was established and they were used in registration.

The image registration succeeded for all algorithms. As depicted in Fig. 9, the best registration accuracy is obtained when FWA was used as optimization algorithm but also the others produce enough good results. Concerning the number of objective function evaluations it is similar to that of pixel based IR, because the parameters chosen for the optimization algorithms were the same. Anyway this is less important because the IR is very fast. As comparison, the duration of IR procedure applied for the 'Lena' image and FWA usage in optimization, on a system equipped with an Intel Core i5 3.1 GHz processor and 4 GB RAM, is shown in Table 7.

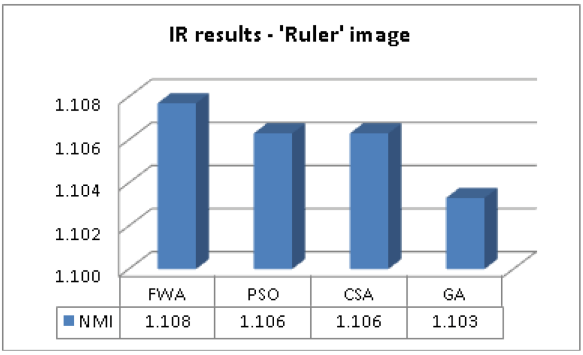


Fig. 9. Evaluation of features based IR results using FWA, PSO, CSA and GA for 'Ruler' test image

Table 7. IR duration (seconds) for 'Lena' image and FWA optimization

	Pixel based IR	Features based IR
Parallel	14.72	0.04
Sequential	20.13	0.05

The features based IR was applied also for the images degraded using Gaussian and Salt & Pepper noise. In the first case the features based IR succeeded as well even if fewer correspondences were found in the two images. In the second case, the

registration failed because no correspondences were established between the features detected in the two images.

6 Conclusions

In this paper, the behavior of the Fireworks optimization algorithm for image registration problems is analyzed. Both pixel based and features based image registration were addressed for different types of images: pictures, graphics and a combination of these.

In most cases the optimization algorithm is chosen and its parameters are tuned depending on the problem to be solved. FWA is compared to other nature inspired evolutionary algorithms: Particle Swarming, Cuckoo Search and Genetic algorithms. Concerning the pixel based IR, for which the Normalized Mutual Information was used as objective function, FWA may be a choice if enough good registration results have to be obtained in a reasonable execution time because the accuracy of its results is almost as good as for PSO and CSA algorithms and better than those of GA usage. Regarding the number of objective function evaluations that is proportional to the execution time, in the studied cases FWA is almost as fast as PSO is and much faster than CSA and GA. Similar results were obtained for images degraded by noise.

The pixel based registration may fail if the images to be registered contain graphics. For this reason a features based IR approach is also analyzed. It is based on the SIFT features and FWA offers more accurate results than PSO, CSA and GA optimization algorithms. The features based IR has the disadvantage that it might fail for multi-modal or noise affected images, because the correspondences between the stable features detected in the two images can't be found.

The results presented in this paper were obtained using FWA, PSO, CSA and GA implementations in C++ [9–14]. The parallel implementation uses the computing power of multi-core processors and was developed using the parallel computing support of the Microsoft Visual Studio 2015 framework. The OpenCV open-source library was used for images manipulation. The graphs presented in Fig. 5 were drawn using a Matlab application.

This research will be continued by the analysis of other nature-inspired optimization algorithms and their usage in image processing applications.

References

1. Zitova, B., Flusser, J.: Image registration methods: a survey. Image Vis. Comput. **21**, 977–1000 (2003). Elsevier
2. Yang, X.-S.: Nature-Inspired Optimization Algorithms. Elsevier Inc., Amsterdam (2014)
3. Tan, Y., Zhu, Y.: Fireworks algorithm for optimization. In: Tan, Y., Shi, Y., Tan, K.C. (eds.) ICSI 2010, Part I. LNCS, vol. 6145, pp. 355–364 (2010)
4. Tan, Y.: Fireworks Algorithm A Novel Swarm Intelligence Optimization Method. Springer, Heidelberg (2015)

5. Zheng, S., Janecek, A., Tan, Y.: Enhanced fireworks algorithm. In: Proceedings of 2013 IEEE Congress on Evolutionary Computation, Cancún, México, pp. 2069–2077 (2013)
6. Li, J., Zheng, S., Tan, Y.: Adaptive fireworks algorithm. In: Proceedings of 2014 IEEE Congress on Evolutionary Computation (CEC), pp. 3214–3221 (2014)
7. Zheng, S., Li, J., Janecek, A., Tan, Y.: A cooperative framework for fireworks algorithm. In: IEEE/ACM Transactions on Computational Biology and Bioinformatics, pp. 1–13 (2015)
8. Liu, L., Zheng, S., Tan, Y.: S-metric based multi-objective fireworks algorithm. In: 2015 IEEE Congress on Evolutionary Computation (CEC), pp. 1257–1264 (2015)
9. Costin, H., Bejinariu, S.I.: Medical image registration by means of a bio-inspired optimization strategy. Comput. Sci. J. Moldova **20,** 2(59), 178–202 (2012)
10. Bejinariu, S.-I.: Image registration using bacterial foraging optimization algorithm on multi-core processors. In: 4th International Symposium on Electrical and Electronics Engineering (ISEEE), Galaţi, România (2013)
11. Bejinariu, S.-I., Costin, H., Rotaru, F., Luca, R., Niţă, C.: Social behavior in bacterial foraging optimization algorithm for image registration. In: Proceedings of the 18th International Conference on System Theory, Control and Computing, Sinaia, Romania, pp. 330–334 (2014)
12. Bejinariu, S.-I., Costin, H., Rotaru, F., Luca, R., Nita, C.D.: Image processing by means of some bio-inspired optimization algorithms. In: Proceedings of the IEEE 5th International Conference on E-Health and Bioengineering – "EHB 2015", Iasi, Romania, pp. 1–4 (2015)
13. Bejinariu, S.-I., Costin, H., Rotaru, F., Luca, R., Nita, C.D.: Automatic multi-threshold image segmentation using metaheuristic algorithms. In: 2015 International Symposium on Signals, Circuits and Systems (ISSCS), Iasi, Romania, pp. 1–4 (2015)
14. Bejinariu, S.-I., Costin, H., Rotaru, F., Luca, R., Nita, C.D.: Fireworks algorithm based single and multi-objective optimization. Paper Submitted to Buletinul Institutului Politehnic din Iasi, Sectia Automatica si Calculatoare (2016)
15. University of Southern California, USC-SIPI Image Database. http://sipi.usc.edu/database/database.php?volume=misc. Accessed 15 Mar 2016
16. Pedersen, M.E.H.: Good parameters for particle swarm optimization. Hvass Laboratories, Technical report no. HL100 (2010)
17. Bejinariu, S.-I., Costin, H., Rotaru, F., Luca, R., Niţă, C., Lazăr, C.: Parallel processing and bio-inspired computing for biomedical image registration. Comput. Sci. J. Moldova **22**, 2(65), 253–277 (2014). Invited Article
18. Lowe, D.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**(2), 91–110 (2004)

Discrete Wavelet Transforms for PET Image Reduction/Expansion (wavREPro)

Hedi Amri[1(✉)], Malek Gargouri[1], Med Karim Abdmouleh[1], Ali Khalfallah[1], Bertrand David[2], and Med Salim Bouhlel[1]

[1] Research Unit: Sciences and Technologies of Image and Telecommunications, Higher Institute of Biotechnology, University of Sfax, Sfax, Tunisia
{hedi.amri,malek.gargouri}@setit.rnu.tn, medkarim.abdmouleh@isggb.rnu.tn, ali.khalfallah@enetcom.rnu.tn, medsalim.bouhlel@enis.rnu.tn

[2] LIRIS Laboratory, UMR 5205 CNRS Central School of Lyon, University of Lyon, 36, av Guy de Collongue, 69134, Lyon-Ecully Cedex, France
bertrand.david@ec-lyon.fr

Abstract. The large volume of medical images remains a major problem for their archiving and transmission. In this context, we propose a novel protocol wavREPro that aims to minimize the image size before its storage and then to enlarge it at reception. This process has been achieved by exploiting the Discrete Wavelet Transforms (DWT) namely Haar, Le Gall (5/3) and Daubechies (9/7) wavelets. These tools represent the image in the multi-resolution domain that follows the human psycho-visual system. Therefore, the reduced image is none other than the approximation of the image. Its magnification is carried out by either cancelling the details (wavREProZ) or estimating them (wavREProED) using the DWT^{-1} on the reduced image. Our experiments have been conducted on a PET (Positron Emission Tomography) medical image database and the results have been presented for the three well-known color spaces RGB, HSV and YCbCr. The reported results have promoted the wavREProZ application with the Haar wavelets on RGB images since it achieved maximum fidelity between the original and reduced then enlarged images. The good performance of this approach encourages its adoption to display images on screens having different sizes.

Keywords: Telemedicine · Archiving · Transmission · Medical images · Discrete Wavelet Transforms · Image reduction · Image expansion

1 Introduction

The reduction and expansion of digital images are current, useful and even necessary operations in many applications related to the imaging domain especially in the medical field [7,14]. In fact, the big size of medical images [4,24] make their transmission and archiving quite difficult due to the limitation of both the hard drive capacity and the transfer rate [10]. The only possible solution to overcome this problem is to reduce the size of these files by compression techniques

V.E. Balas et al. (eds.), *Soft Computing Applications*, Advances in Intelligent Systems and Computing 633, DOI 10.1007/978-3-319-62521-8_45

such as JPEG [11], JPEG 2000 [34], Fractal coding methods [6], Region of Interest Coding Techniques [12,17], Lossless dynamic and adaptive compression [9], low-complexity compression [29] and genetic algorithms [33] and REPro (Reduction/Expansion Protocol) [2]. In this work, we are interested in REPro [11] that compresses images by reducing their definition (pixel number). In this context, various methods have been developed such as the decimation by square-square mesh [15], the decimation from a square mesh to a staggered mesh [28] and the decimation from a staggered mesh to a square mesh. On the other hand, the enlargement process is required to access to the image details. It can also be used to increase the resolution of the image in the same spatial support to improve the visual comfort of the viewer. In the literature, different techniques have been utilized for enlarging color images [27] namely the zero-padding [26], the polynomial nearest neighbor [32], the linear, quadratic and cubic interpolations and the cardinal spline transformation [13].

In our research, a novel approach has been developed to reduce and enlarge PET (Positron Emission Tomography) images. We have exploited the image multiresolution representation using Discrete Wavelet Transforms (DWTs) [8, 16,19,25,30] in the reduction and expansion phases.

This paper is organized as follows. Section 2 is reserved to the presentation of our new approach wavREPro and its two versions. In this part, we also present an overview about the different used tools which are the DWTs and the image quality metrics. Section 3 presents our experimental results and their assessments. In the last section (Sect. 4), we summarize the main contributions of our work.

2 Proposed Approach

2.1 Medical Context

Telemedicine is an effective solution to remedy the lack of medical specialists and equipment. Indeed, thanks to this technology, it is possible to exchange information having different sizes and natures between a tele-staff. This fact requires the adaptation of the image size to the screen size mainly when displaying medical images [20]. However, these images are usually large which can slow the communication. Furthermore, to display these images, they must also be adapted to the resolution of the screens [31]. In this context, REPro permits the reduction of the image definition that speeds its transmission. On the other hand, this protocol can enlarge the transferred image to ensure more visual comfort or it may further reduce the image definition to be adaptable to the low resolutions of some display devices such as tablets and smart phones (Fig. 1).

2.2 Methodology

To reduce or to enlarge the image, we have chosen to transform the image to the multiresolution domain to benefit from its simultaneous spatial and frequency locations. This process is carried out by the use of the Discrete Wavelet Transforms (Fig. 2).

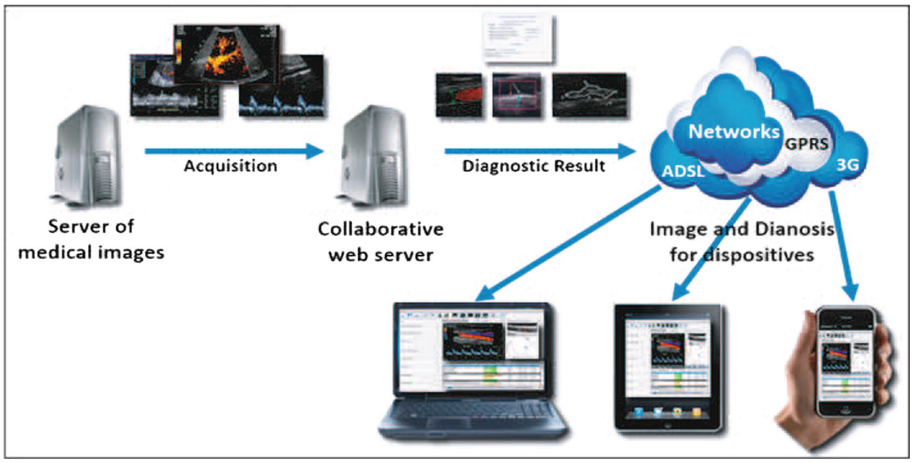


Fig. 1. Image transmission process using REPro

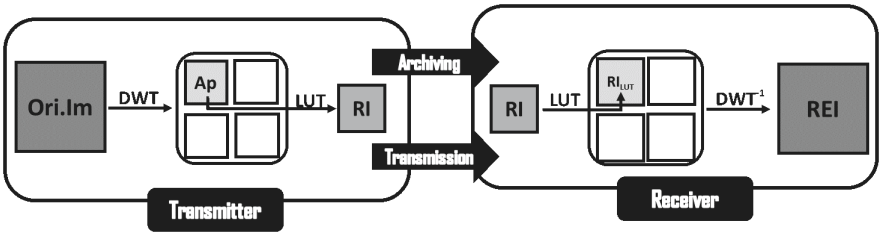


Fig. 2. Protocol wavREPro

Where "ori.Im" is the original image, DWT is the Discrete Wavelet Transforms, LUT is the Look-Up Table, RI is the reduced image, RILUT is the approximation space related to the applied DWT, REI is the reduced then enlarged image.

2.2.1 Used Wavelet Transforms

Wavelets are interesting tools as they provide a simultaneous frequency and spatial locations of the image. This feature following the human psychovisual model allows wavelets to be exploited in various image processing applications such as compression (JPEG 2000), filtering [3], segmentation [5] and watermarking [19]. To apply a DWT on an image, we simply utilized it on lines then on columns. The original image was first decomposed into an approximation Ap_0 and 3 detail sub images called HD_0, VD_0 and DD_0 which denoted Horizontal Details, Vertical Details and Diagonal details, respectively [23]. The recursive application of wavelets yielded the representation of the image in the multiresolution domain at level n as shown in Fig. 3. The level n denoted the number of iterations of the application of wavelets.

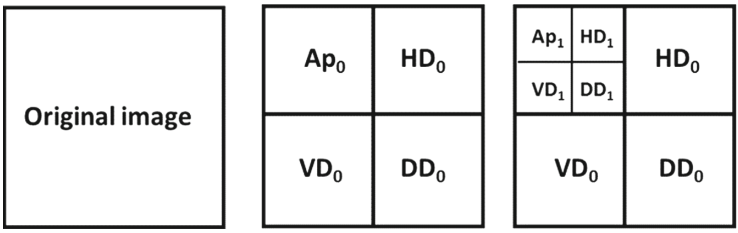


Fig. 3. Representation of an image decomposition: one level and two level

In this work, three well-known transforms were exploited. In fact, the Haar wavelet was used for its simple implementation [22]. Moreover, we utilized Le Gall 5/3 and Daubechies 9/7 wavelets due to their frequent usages in the JPEG 2000 compression standard.

The equations of Haar, 5/3 et 9/7 DWTs are depicted in Eqs. 1, 2 and 3, respectively.

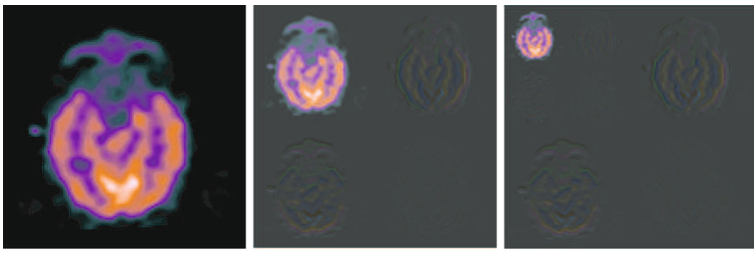
$$\begin{cases} s[n] = \frac{1}{2}(s_0[n] + d_0[n]) \\ d[n] = \frac{1}{2}(s_0[n] - d_0[n]) \end{cases} \tag{1}$$

$$\begin{cases} d[n] = d_0[n] - \lfloor \frac{1}{2}(s_0[n+1] + s_0[n]) \rfloor \\ s[n] = s_0[n] + \lfloor \frac{1}{4}(d[n] + d[n-1]) + \frac{1}{2} \rfloor \end{cases} \tag{2}$$

$$\begin{cases} d[n] = d_0[n] + \lfloor \frac{1}{16}((s_0[n+2] + s_0[n-1]) - 9(s_0[n+1] + s_0[n])) + \frac{1}{2} \rfloor \\ s[n] = s_0[n] + \lfloor \frac{1}{4}(d[n] + d[n-1]) + \frac{1}{2} \rfloor \end{cases} \tag{3}$$

where s_0 and d_0 are the initial approximation and details and s and d are the new approximation and details.

Figure 4 illustrated the image decompositions in one level and two level after the application of the 5/3 wavelet on an original color image (Fig. 4a).



(a) (b) (c)

Fig. 4. Image decomposition using the 5/3 wavelet: (a) original image, (b) decomposition in level 1, (c) decomposition in level 2

2.2.2 The WavREPro Protocol

The wavREPro protocol (or "Reduction/Expansion Protocol based on wavelets") exploited the DWTs not only to reduce the image but also to enlarge it. Two versions of this protocol were proposed (Fig. 5). On the one hand, we obtained wavREProZ when the wavREPro protocol used Zero details. On the other hand, wavREProED represented wavREPro using enlarged details. These two approaches used the same reduction process that involved DWTs. This aimed to get the approximation on which we applied a Look-Up Table (LUT) [1] transform before its transmission to respect the image coding standards. We explain below the expansion process of wavREProZ and wavREProED.

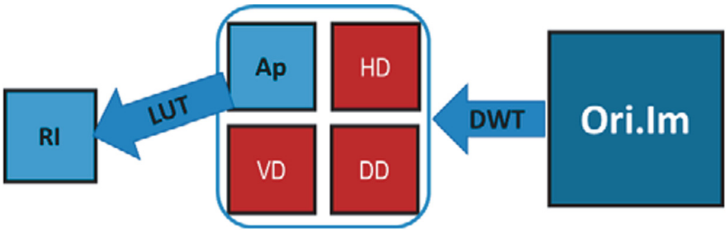


Fig. 5. Image reduction by wavREPro

2.2.2.1 The wavREProZ Protocol

When the user received the reduced image (RI), it applied on it a color palette (Look-Up Table: LUT). This table was based on linear functions to make the values in accordance with the approximation space related to the applied DWT. The resulting matrix was called RILUT. We associated to this matrix 3 zero matrices having the same dimensions as RILUT. These matrices were considered as matrices of Horizontal Details (HD = 0), Vertical Details (VD = 0) and Diagonal Details (DD = 0).

Finally, DWT^{-1} was applied on the set {RILUT, HD = 0, VD = 0, DD = 0} to get the reduced then enlarged image using wavREProZ and we named it REI-wavZ (Fig. 6).

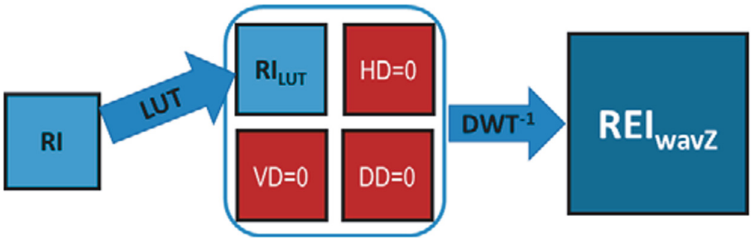


Fig. 6. Expansion of the reduced image using wavREProZ

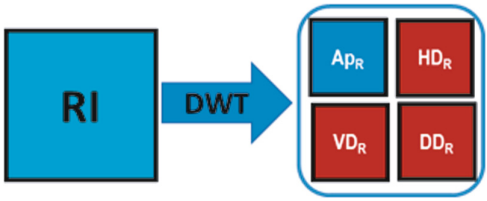


Fig. 7. Decomposition of the reduced image using DWT

2.2.2.2 The wavREProED Protocol

This protocol tried to find the vertical (VD), horizontal (HD) and diagonal (DD) details lost in the reduction phase. This process was carried out in two phases. In fact, the DWT was applied on the reduced image as illustrated in Fig. 7.

Then, the DWT^{-1} is carried out on each of the Horizontal Details HDR, Vertical Details VDR and Diagonal Details DDR of the reduced mage (RI). To compute one of these components, the other components should be cancelled as described in Fig. 8. Thus, to estimate a detail component, we have enlarged its counter-parts obtained from the application of the DWT on the reduced image while respecting the frequency location of these details.

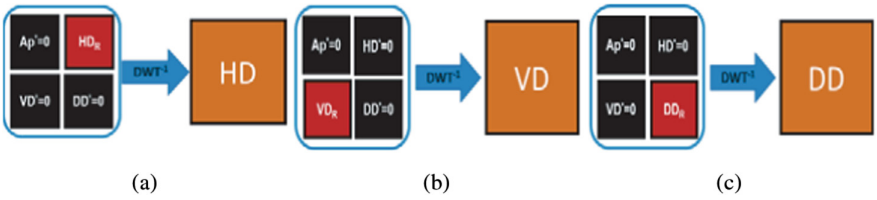


Fig. 8. Computation of the different details using wavREProED: (a) computation of HD, (b) computation of VD, (c) computation of DD

Finally, to get the image enlarged by the wavREProED protocol, we have adapted the reduced image RI to the approximation space related to the used DWT to obtain the same image RILUT used by wavREProZ. The magnified image (REI-wavED) is obtained by applying the DWT^{-1} on the set {RILUT, HD, VD, DD} as illustrated in Fig. 9.

2.2.3 Image Quality Metrics

The image quality metrics were considered in our experiments to assess the level of distortion introduced by the reduction-expansion process. In this work, we used the Peak Signal to Noise Ratio (PSNR) and the Structural Similarity Index Metric (SSIM) to get an objective evaluation of the performance of the proposed approaches.

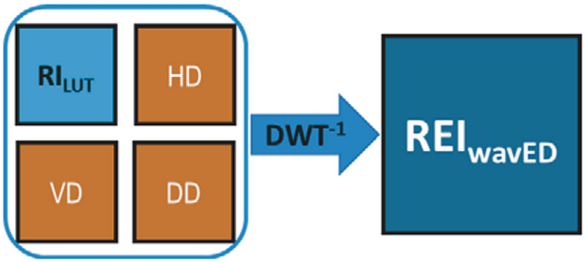


Fig. 9. Expansion of the reduced image using wavREProED

2.2.3.1 Structural SIMilarity
The Structural SIMilarity (SSIM) index is a method for measuring the similarity between two images where we focus on the structure image fidelity. The SSIM [18] is expressed by Eq. 4.

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}$$
$$C_1 = (K_1 L)^2,\ C_2 = (K_2 L)^2 \tag{4}$$

where μ_x and μ_x denote the average values of the images X and Y respectively, σ_x and σ_y are the standard deviations (the square root of variance) of X and Y σ_{xy} is the covariance of X and Y. L is the dynamic range of the pixel values (255 for the images encoded on 8 bits) and C_1 and C_2 are small positive constants (K_1 and $K_2 \ll 1$).

At each pixel (i, j), a local SSIM (i, j) index is defined by evaluating the average, the standard deviation and the covariance on a local neighbor around that pixel (i, j). The overall quality of the image is measured by the Means SSIM index (MSSIM) which is expressed by Eq. 5.

$$MSSIM = \frac{1}{M}\sum_i \sum_j SSIM(i,j) \tag{5}$$

where M is the total number of local SSIM indices.

2.2.3.2 Peak Signal Noise Ratio
The PSNR quantifies the image distortion in decibel (dB) [21]. It is a function of the Mean Square Error (MSE). Equations 6 and 7 describe the MSE and PSNR metrics, respectively.

$$MSE = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\left(I_{ij} - I^*_{ij}\right)^2}{nm} \tag{6}$$

$$PSNR\ (I, I^*) = 10\log 10(X\,\max^2 / MSE(I, I^*)) \tag{7}$$

where I and I^* are images having the same size $n \times m$.

When a PSNR is upper than or equal to 30 dB, it reflects an acceptable image degradation. From 60 dB, it represents an imperceptible image distortion.

2.2.3.3 Selection Rate

The quantification of the distortion introduced by two methods on an image database was generally carried out according results provided by $PSNR_{average}$ and $SSIM_{average}$.

However, if the $PSNR_{average}$ of the first approach is higher than that of the second method, this fact does not prove that the first approach introduces less distortion on each image of the database.

To get a more objective assessment, we introduced the concept of selection rate. Hence, the selection rate of the method 1 was simply the percentage of the images of the database where this method introduced less distortion compared to the second method.

Considering this definition, we obtained selection rates based on PSNR and SSIM to evaluate the performances of the approaches on our image database.

3 Experiments and Discussion

We proposed to evaluate the performance of our two approaches wavREProZ and wavREproED and to analyze the impact of each color space (RGB, HSV, YCbCr) and each adopted wavelet on the reduction-expansion process. Our experiments were conducted on 30 PET images.

Initially, we were interested in assessing the performance of each wavREPro version when the images were represented in the RGB color space. The reported results in Fig. 10 and Table 1 obviously showed that the adoption of the Haar wavelet in wavREPro yielded two different performances n terms of PSNR and SSIM. Indeed, the use of this wavelet in the wavREProZ process, hence its name wavHaarZ, ensured the best degree of similarity between the original and reduced then enlarged images. However, the same wavelet caused the most image quality degradation when it was used in the wavREProED method, hence its name wavHaarED.

In addition, we noted that PSNRs obtained after the application of 5/3 and 9/7 wavelets on PET images were very close (Fig. 10a). However, the results shown in Fig. 10b promoted the utilization of the 9/7 wavelet followed by the 5/3 wavelet in terms of SSIM.

Thus, the Haar wavelet improved the performance of wavREProZ while it was preferable to use wavREProED with the 9/7 or 5/3 wavelets (Table 1).

Where wav5/3Z and wav5/3ED denoted the 5/3 wavelets applied in wavREProZ and wavREProED, respectively. Similarly, wav9/7Z and wav9/7ED represented the 9/7 wavelets applied in wavREProZ and wavREProED, respectively.

The results provided by Table 2 confirmed the values of PSNR and SSIM. Indeed, wavHaarZ was selected as the best technique of reduction-expansion for all the PET images of the database when the color space was RGB.

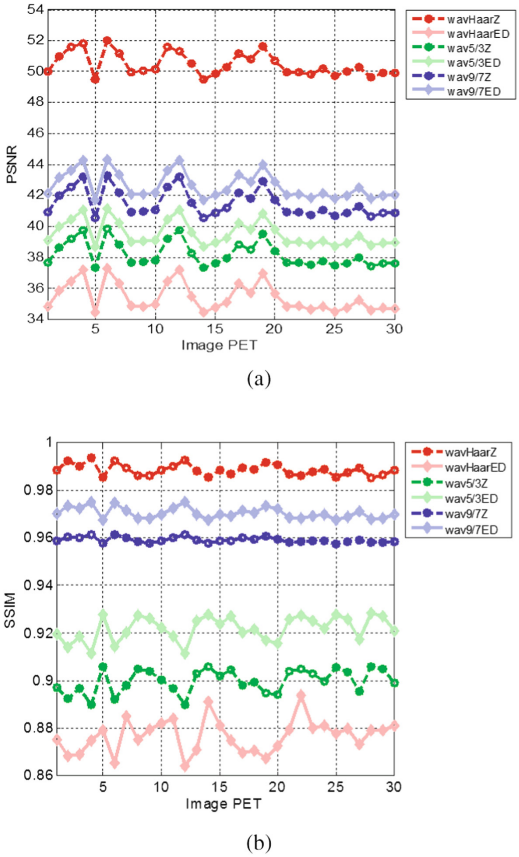


Fig. 10. Effects of the wavREPro application on PET images represented in RGB color space: (a) Quality assessment of the RGB images in terms of PSNR, (b) Quality assessment of the RGB images in terms of SSIM

In order to evaluate the performances of wavREProZ and wavREProED and the impact of the choice of wavelets on the HSV (Hue Saturation Value) image quality, we converted the test images to this color space. Then, these images were reduced then enlarged. Finally, they were converted back to the RGB space. We exploited PSNR, SSIM and selection rates to illustrate the effect of the reduction operation followed by the image magnification process.

The results presented in Fig. 11 and Table 3 proved again that wavReproZ based on Haar wavelet transforms performed better in terms of PSNR and SSIM.

Similarly to the results attained in the RGB color space, the Haar wavelet had the greatest performance when we adopted it in wavREProZ.

But, contrary to the results provided in the RGB color space, wavREProED based on wavHaarED seemed more efficient than wavREProED based on the 9/7 wavelets. Indeed, wavHaarED performed better than wav9/7ED for 20% of the database images in terms of PSNR and reached even 26.6% of this set in terms of SSIM (Table 4).

Table 1. Quality evaluation of the RGB images after the application of wavREPro.

	wavHaarZ	wavHaarED	wav5/3Z	wav5/3ED	wav9/7Z	wav9/7ED
$PSNR_{average}$	50.47	35.42	38.21	39.54	41.51	42.63
$SSIM_{average}$	0.989	0.877	0.904	0.922	0.959	0.97

Table 2. Selection rates of wavelets applied on RGB images.

	wavHaarZ	wavHaarED	wav5/3Z	wav5/3ED	wav9/7Z	wav9/7ED
PSNR	100%	0.0%	0.0%	0.0%	0.0%	0.0%
SSIM	100%	0.0%	0.0%	0.0%	0.0%	0.0%

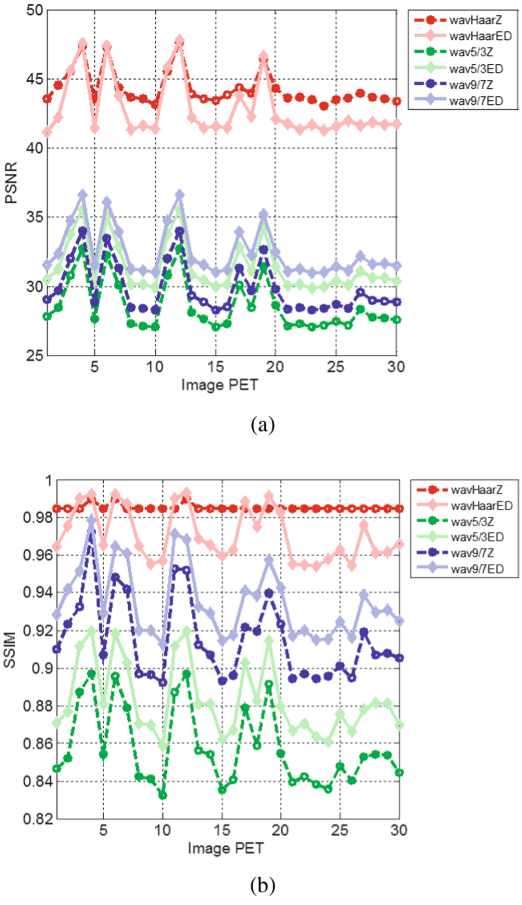


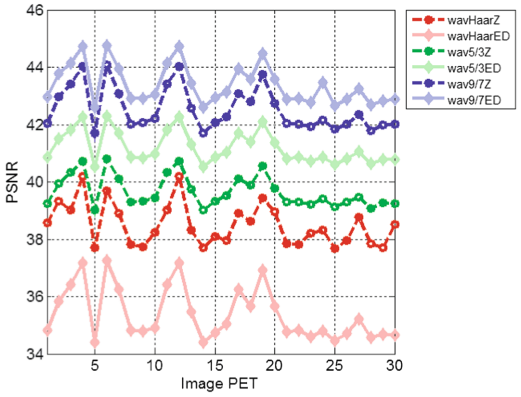
Fig. 11. Effects of the wavREPro application on PET images represented in HSV color space: (a) Quality assessment of the HSV images in terms of PSNR, (b) Quality assessment of the HSV images in terms of SSIM

Table 3. Quality evaluation of the HSV images after the application of wavREPro.

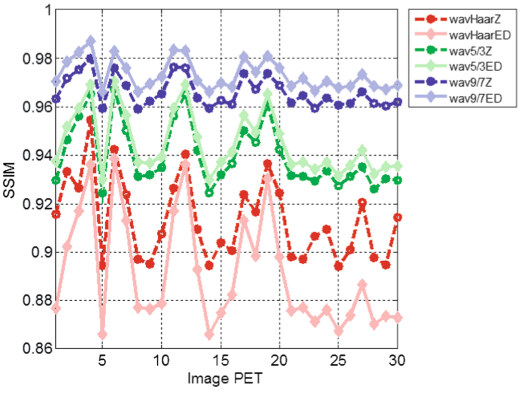
	wavHaarZ	wavHaarED	wav5/3Z	wav5/3ED	wav9/7Z	wav9/7ED
$PSNR_{average}$	44.30	42.85	28.62	31.46	29.84	32.51
$SSIM_{average}$	0.985	0.971	0.858	0.883	0.915	0.935

Table 4. Selection rates of wavelets applied on HSV images.

	wavHaarZ	wavHaarED	wav5/3Z	wav5/3ED	wav9/7Z	wav9/7ED
PSNR	80%	20%	0.0%	0.0%	0.0%	0.0%
SSIM	73.33%	26.67%	0.0%	0.0%	0.0%	0.0%



(a)



(b)

Fig. 12. Effects of the wavREPro application on PET images represented in YCbCr color space: (a) Quality assessment of the YCbCr images in terms of PSNR, (b) Quality assessment of the YCbCr images in terms of SSIM

We also carried out our experiments on our test images that we converted to the YCbCr color space. Figure 12 illustrated the degree of resemblance between the original and reduced then enlarged images. The performances of the different methods were summarized in Table 5. According to Fig. 11a, we clearly noticed that PSNRs are higher than 30 dB for all the applied methods. In other words, there was a good conservation of the content of all the images.

Table 5. Quality evaluation of the YCbCr images after the application of wavREPro.

	wavHaarZ	wavHaarED	wav5/3Z	wav5/3ED	wav9/7Z	wav9/7ED
$PSNR_{average}$	38.50	35.39	39.67	41.20	42.54	43.38
$SSIM_{average}$	0.913	0.891	0.940	0.945	0.966	0.973

On the other hand, we noted that PSNR and SSIM reached great values when wav9/7ED was adopted (Table 5). Moreover, the 5/3 wavelet provided the second best performance followed by the Haar wavelet. Therefore, wav9/7ED was obviously selected as the best technique since it ensured the most image content preservation (Table 6).

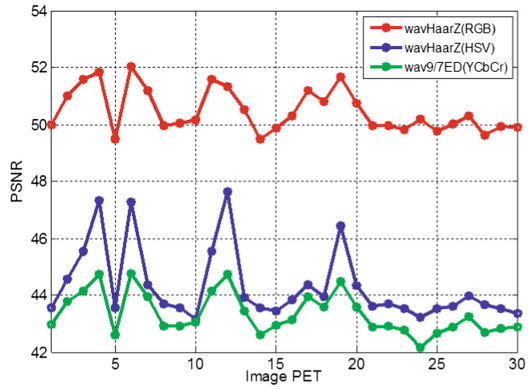
Table 6. Selection rates of wavelets applied on YCbCr images.

	wavHaarZ	wavHaarED	wav5/3Z	wav5/3ED	wav9/7Z	wav9/7ED
PSNR	0.0%	0.0%	0.0%	0.0%	0.0%	100%
SSIM	0.0%	0.0%	0.0%	0.0%	0.0%	100%

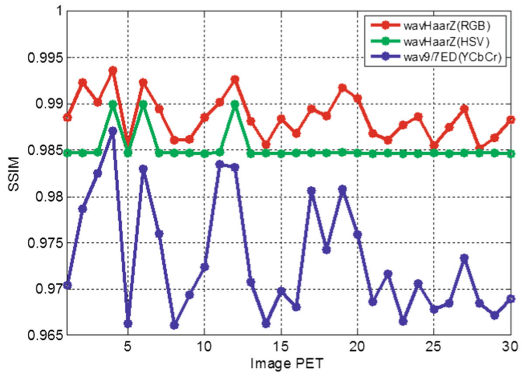
Finally, we noted that the Haar wavelet always gave the best results when it was included in wavREProZ whereas the 9/7 and 5/3 wavelets performed better when used in the wavREProED protocol.

To highlight the impact of the space color choice on the wavREPro performance, we compared the best wavREPro approaches selected by each color space. This leaded us to compare the performances of wavHaarZ applied in RGB and HSV color spaces to those of wav9/7ED utilized in the YCbCr color space.

Figure 13 and Tables 7 and 8 showed the performance assessments of methods in terms of PSNR and SSIM. We noted that RGB represented the best color space to adopt and that is why it was selected for all images (Fig. 13 and Tables 9 and 10).



(a)



(b)

Fig. 13. Effects of the wavREPro application on PET images: (a) Quality assessment of the images in terms of PSNR, (b) Quality assessment of the images in terms of SSIM

Table 7. Performance assessment of selected wavelets applied on TEP images in terms of $PSNR_{average}$.

wavHaarZ (RGB)	wavHaarZ (HSV)	Ond9/7ED (YCbCr)
50.47	44.30	43.38

Table 8. Performance assessment of selected wavelets applied on TEP images in terms of $SSIM_{average}$.

wavHaarZ (RGB)	wavHaarZ (HSV)	wav9/7ED (YCbCr)
0.989	0.985	0.973

Table 9. Selection of the best wavelet according to PSNR.

wavHaarZ (RGB)	wavHaarZ (HSV)	wav9/7ED (YCbCr)
100%	0.0%	0.0%

Table 10. Selection of the best wavelet according to SSIM.

wavHaarZ (RGB)	wavHaarZ (HSV)	Wav9/7ED (YCbCr)
100%	0.0%	0.0%

4 Conclusion

In this paper, we presented a new approach of reduction and enlargement called wavREPro that exploited the image representation in the multiresolution domain. This field provided simultaneously the spatial and frequency locations of the image and followed the human psycho-visual model. The transition to this domain was achieved using the Discrete Wavelet Transforms. In this context, we used three types of DWTs namely Haar, Le Gall (5/3) and Daubechies (9/7) wavelets. These methods allowed the reduction of the image before its storage or transfer then its enlargement before its display on the screen.

The minimized image was none other than the approximation that was the result of the DWT application on the image to transfer. To enlarge the image, we proposed two versions of wavREPro which added details to the obtained approximation and then applied the DWT^{-1} on the 4 components. The wavREPoZ version canceled details while the wavREProED version estimated the details of the image transformed to the multiresolution domain.

The performances of these different approaches were assessed in different color spaces namely RGB, HSV and YCbCr. The obtained results promoted the use of the Haar wavelet, which canceled the details in the image magnification process. We also noted that it was appropriate to exploit the RGB space color for the application of wavREPro on PET images.

In addition, it was obviously noted that the Haar wavelet was more effective when used in wavREProZ while wavREProED achieved good results when the 5/3 and 9/7 wavelets were utilized.

This work has been exploited to reduce images before their storage or transfer. In addition, our approach has ensured their adaption to the different types of screens before their display. This can be carried out by resizing the image to fit appropriately the screen size.

References

1. Abdmouleh, M.K., Khalfallah, A., Bouhlel, M.S.: A chaotic cryptosystem for color image with dynamic look-up table. In: Proceedings of the 7th International Conference on Image and Signal Processing (ICISP 2016), pp. 91–100. Springer International Publishing, Trois-Rivieres (2016)

2. Amri, H., Hanna, F., Lapayre, J.C., Khalfallah, A., Bouhlel, M.S.: REPRO: a new reduction/expansion protocol to increase the performance of image transmission in medical telediagnosis platforms. Biomed. Eng.: Appl. Basis Commun. **27**(6), 1550054 (2015)
3. Aramendi, E., Irusta, U., Ayala, U., Naas, H., Kramer-Johansen, J., Eftestol, T.: Filtering mechanical chest compression artefacts from out-of-hospital cardiac arrest data. Resuscitation **98**, 41–47 (2016)
4. Aribi, W., Khalfallah, A., Bouhlel, M.S., Elkadri, N.: Evaluation of image fusion techniques in nuclear medicine. In: 2012 6th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT), pp. 875–880 (2012)
5. Bai, X., Jin, J.S., Feng, D.: Segmentation-based multilayer diagnosis lossless medical image compression. In: Proceedings of the Pan-Sydney Area Workshop on Visual Information Processing, VIP 2005, pp. 9–14 (2004)
6. Bhavani, S., Thanushkodi, K.G.: Comparison of fractal coding methods for medical image compression. IET Image Process. **7**(7), 686–693 (2013)
7. Bonnans, V., Humbert, P., Pazart, L., Marzani, F., Lapayre, J.C., Lang, C.: Collaborative platform for skin cancer screening and associated optical fibered probe for diagnosis. In: Sin'Fran 2009, Singaporean-French IPAL Symposium, pp. 44–55 (2009)
8. Boudjelal, A., Messali, Z., Boubchir, L., Chetih, N.: Nonparametric Bayesian estimation structures in the wavelet domain of multiple noisy image copies. In: 2012 6th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT), pp. 495–501 (2012)
9. Castiglione, A., Pizzolante, R., De Santis, A., Carpentieri, B., Castiglione, A., Palmieri, F.: Cloud-based adaptive compression and secure management services for 3D healthcare data. Future Gener. Comput. Syst. **43**(44), 120–134 (2015)
10. Chaabouni, I., Fourati, W., Bouhlel, M.S.: Using ROI with ISOM compression to medical image. IJCVR **6**(1/2), 65–76 (2016)
11. Ciznicki, M., Kurowski, K., Plaza, A.: Graphics processing unit implementation of JPEG2000 for hyperspectral image compression. J. Appl. Remote Sens. **6**(1), 061507-1–061507-14 (2012)
12. Doukas, C., Maglogiannis, I.: Region of interest coding techniques for medical image compression. IEEE Eng. Med. Biol. Mag. **26**(5), 29–35 (2007)
13. Fahmy, M.F., Fahmy, G., Fahmy, O.F.: B-spline wavelets for signal denoising and image compression. Signal Image Video Process. **5**(2), 141–153 (2011)
14. Fuin, D., Garcia, E., Guyennet, H., Lapayre, J.C.: Collaborative interactions for medical e-Diagnosis. HPCN Int. J. High-Perform. Comput. Netw. **5**(3), 189–197 (2008)
15. Gotchev, A.P., Egiazarian, K.O., Marchokov, G., Saramki, T.: A near least squares method for image decimation. ICIP **2**, 929–932 (2003)
16. Hassen, W., Mbainaibeye, J., Olivier, C., Amiri, H.: New video encoder based on wavelet coefficients and motion compensation. In: 2012 6th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT), pp. 440–447 (2012)
17. Hernndez-Cabronero, M., Blanes, I., Pinho, A.J., Marcellin, M.W., Serra-Sagrist, J.: Progressive lossy-to-lossless compression of dna microarray images. IEEE Signal Process. Lett. **23**(5), 698–702 (2016)
18. Joshi, Y.G., Loo, J., Shah, P., Rahman, S., Chang, Y.C.: A novel low complexity local hybrid pseudo-SSIM-SATD distortion metric towards perceptual rate control. In: 2013 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB), pp. 1–6 (2013)

19. Kammoun, F., Khalfallah, A., Bouhlel, M.S.: New scheme of digital watermarking using an adaptive embedding strength applied on multiresolution filed by 9/7 wavelet. Int. J. Imaging Syst. Technol. **16**(6), 249–257 (2006)
20. Kassab, R., Lapayre, J.C., Aupet, J.B., Marzani, F., Pieralli, C.: Scars collaborative telediagnosis platform using adaptive image flow. Integr. Comput.-Aided Eng. ICAE **1**(20), 3–14 (2013)
21. Amri, H., Khalfallah, A., Lapayre, J.C., Bouhlel, M.S.: Watermarking for improving the reduction-expansion process of medical images (WREPro). Int. J. Imaging Robot.™ **16**(3), 124–139 (2016)
22. Khalfallah, A., Bouhlel, M.S.: The impact of the concept of the family relative signatures on the non-blind watermarking in the multiresolution domain using 9/7 and 5/3 wavelets. Int. J. Inf. Commun. Technol. **4**(3–4), 111–118 (2011)
23. Khalfallah, A., Brahim, K., Olivier, C.: A new approach to encrypted semi-blend watermarking in multi resolution field by 9/7 wavelet. In: IEEE International Conference of E-Medical Systems E-Medisys 2007. Frs, Morocco (2007)
24. Lazrag, H., Naceur, M.S.: Despeckling of intravascular ultrasound images using curvelet transform. In: 2012 6th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT), pp. 365–369 (2012)
25. Lazrag, H., Naceur, M.S.: Wavelet filters analysis for speckle reduction in intravascular ultrasound images. In: 2012 6th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT), pp. 375–379 (2012)
26. Liu, S., Wang, Q., Liu, G.: A versatile method of discrete convolution and FFT (DC-FFT) for contact analyses. Wear **243**(1/2), 101–111 (2000)
27. Nini, B.: Projection based permutation of color images. In: 2012 6th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT), pp. 460–467 (2012)
28. Petraglia, A., Pereira, J., Barugui, F.: Low-sensitivity IIR switched-capacitor decimation filters using delay lines with staggered T/H stages. IEEE Proc. - Circuits Devices Syst. **153**(5), 193–198 (2006)
29. Pizzolante, R., Carpentieri, B., Castiglione, A.: A secure low complexity approach for compression and transmission of 3-D medical images. In: Proceedings of the 2013 Eighth International Conference on Broadband and Wireless Computing, Communication and Applications, BWCCA 2013, pp. 387–392. IEEE Computer Society (2013)
30. Saidani, T., Atri, M., Said, Y., Tourki, R.: Real time FPGA acceleration for discrete wavelet transform of the 5/3 filter for JPEG 2000. In: 2012 6th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT), pp. 393–399 (2012)
31. Sanchez Santana, M.A., Aupet, J.B., Betbeder, M.L., Lapayre, J.C., Camarena, A.: A tool for telediagnosis of cardiovascular diseases in a collaborative and adaptive approach. J. Univers. Comput. Sci. JUCS **19**(9), 1275–1294 (2013)
32. Schafer, R.W., Rabiner, L.R.: A digital signal processing approach to interpolation. Proc. IEEE **61**(6), 692–702 (1973)
33. Shih, F.Y., Wu, Y.T.: Robust watermarking and compression for medical images based on genetic algorithms. Inf. Sci. **175**(3), 200–216 (2005)
34. Wu, X., Li, Y., Liu, K., Wang, K., Wang, L.: Massive parallel implementation of JPEG2000 decoding algorithm with multi-GPUS. In: Proceedings of the SPIE, vol. 9124, pp. 91240S–91240S-6 (2014)

Tips on Texture Evaluation with Dual Tree Complex Wavelet Transform

Anca Ignat[1(✉)] and Mihaela Luca[2]

[1] Faculty of Computer Science, University "Alexandru Ioan Cuza" of Iaşi, Iaşi, Romania
ancai@info.uaic.ro
[2] Institute of Computer Science, Romanian Academy, Iaşi, Romania
mihaela.luca@iit.academiaromana-is.ro

Abstract. Dual Tree Complex Wavelet Transform (DTCWT) is a practical tool offering the possibility of extracting characteristics from images, a transformation which has good directional selectivity, is approximately shift invariant and it is computationally efficient. DTCWT may be employed in different stages of image processing, including multiscale texture featuring. We are measuring texture similarities on characteristics vectors by applying cosine, Pearson correlation coefficient, entropy-based and histogram related distances and we tested these measures on slightly rotated samples from Brodatz texture album. Comparisons and observations are made on the methods that we employed.

Keywords: Dual Tree Complex Wavelet Transform · Feature extraction · Similarity measures · Texture

1 Introduction

Understanding images is a complex process which implies training during long periods of time in order to use high amounts of accumulated knowledge. To recognize images similar to those that we are acquainted to, we are scaling them and situating them in environmental contexts, in order to place the objects from the image in an appropriate scenario. As human beings we are merely evaluating colors, textures, lines, shapes, shades and lights, materials and spatial positioning, while a computer needs a large variety of mathematical models attempting to represent the cognitive processes in order to obtain comparable results. Intelligent agent architectures aim to adapt to each special task apparently simple for human eye-brain chain which in fact is a very complex paradigm. Image features are extracted and decomposed in hierarchical stages, color and texture being characterized by complex mathematical methods.

Texture characterization is important in many domains, among them being the medical imaging evaluation and interpretation, industrial applications, remote sensing, etc. [1–4]. With the actual amazing increase of the internet image databases it is essential in automatic object retrieval to have a good texture identification procedure. The texture featuring and analysis with a relatively new mathematical tool, the Dual Tree Complex Wavelet Transform DTCWT [5, 6], is discussed in our paper. Note that the DTCWT is abbreviated with an italic *C* to differ from the Continuous Wavelet Transform.

V.E. Balas et al. (eds.), *Soft Computing Applications*, Advances in Intelligent Systems and Computing 633, DOI 10.1007/978-3-319-62521-8_46

We used the classic Brodatz album with texture images [7] and some images from our experimental database. Texture features were computed using the six directional coefficients provided by DTCWT applied on all these texture images and their rotated versions, as we will further explain. In order to identify a certain texture in an area of an image, we apply several distances (cosine, Pearson correlation coefficient, entropy-based and histogram related) on the extracted features.

2 Disadvantages of Wavelet Transforms

The main disadvantages of (real) wavelet transform [8, 9] are: oscillations around singularities (1), shift variance (2), aliasing (3) and the lack of directionality (4). These issues can be overcome by using Complex Wavelets. Yet, it is difficult to compute complex wavelets satisfying the perfect reconstruction property. In [5, 6], Kingsbury introduced the Dual Tree Complex Wavelet Transform (DTCWT) which is a new way of computing complex wavelets with perfect reconstruction with a 2^d:1 redundancy for d dimensional signals.

Important aspects about computing DTCWT, its properties and possible applications were developed in [9–11].

3 Discussion on Dual Tree Complex Wavelet Transform

The decimated Discrete Wavelet Transform [12] has two major inconveniences: shift variance (small shifts in the input signal can generate big variations in the wavelet coefficients) and reduced directional selectivity (especially in diagonal directions). To overcome these limitations one can use complex wavelets. The difficulty in computing complex wavelets is given by the problem of designing perfectly reconstructed filter banks beyond the first level of decomposition. The problem was solved by Kingsbury [5, 6]. He developed the Dual Tree Complex Wavelet Transform, which provides approximate shift invariance, good directional selectivity, uses short linear phase filters with perfect reconstruction and has efficient computation costs (O(N) for 1D-signals). The toll to be paid for these properties is increased redundancy (2:1 for 1D-signals, 4:1 for 2D signals, 8:1 for 3D, and so on) but still it is much less than the undecimated wavelet transform.

In order to obtain approximate shift invariance for the decimated DWT, one must double the sample rate at each level of decomposition; this is equivalent with having two parallel fully decimated trees and also the filters at level 1 must be delayed by one sample. After level 1, to reestablish the uniform interval sampling, one must work with odd length filter on one tree and even length filter on the other. In order to have a better symmetry between the trees odd and even filters must alternate when changing the decomposition level. The output obtained from the two parallel trees can be interpreted as the real and the imaginary parts of the complex wavelet coefficients. In order to compute CWT for 2D signals, each level of decomposition contains two stages: one consists in filtering the rows and the second stage is to filter the result along columns.

When applied to digital images the DTCWT yields at each level of decomposition six bandpasses sub-images of complex coefficients. These sub-images provide strong directional selectivity at angles $\pm 15^{o}$, $\pm 45^{o}$, and $\pm 75^{o}$.

For feature extraction the energy, the mean, the standard deviation, the entropy of the six complex directional coefficients (or their amplitude) yielded by the DTCWT are usually employed: in [13, 14] they use energy and standard deviation, [15] works with the mean, the standard deviation of the amplitude of the sub-band images, and [16] employs the mean and the entropy. In [17] features were extracted using a threshold for counting the significant elements of the DTCWT coefficients. In [8, 18, 19], rotation invariant features were designed using DTCWT.

In our paper, we study the influence of rotation and of changing the brightness of images (by applying a γ transformation) over the DTCWT coefficients. As features we use the histograms of the DTCWT directional sub-images. For comparisons, we employed ten similarity measures for probability density functions presented in [20] and for classification we used the minimum-distance classifier and SVM (Support Vector Machine) [21].

4 Texture Features Selected with DTCWT

Due to the exponentially growing amount of digital images, each new aspect has to be taken into account to facilitate the automatic image classification and clustering.

In order to group the photos in classes depicting the same scene, it is possible to extract by means of segmentation some texture samples from each image. We might classify the image to a certain class by taking into account the information provided by the retrieved dominant textures.

In order to identify textures, we extract feature vectors and we match the texture from the image with textures from a database. We tested some similarity measures in order to compare the textures of the test images with those from the database. The class in which we introduce the test image is the one that gives most matches. We study in this paper the effect of rotation on classification process using DT*C*WT for feature extraction and we proposed some similarity measures. As an example, Fig. 1 is presenting an image from our town which contains as main textures bark, grass and pavement (Fig. 2). In Fig. 3 we have the same scene but it was taken by rotating the photo camera with an angle about 15–20°. We manually extracted three samples of textures from both of the images. The selected textures are depicted in Figs. 2 and 4. The samples represent the same three textures from both images (bark, grass, and pavement) but they are not cropped from the same spot in order to catch different illuminations of the scene.

Alike visual detection of natural environments, we note that analysing the directions of important lines, shades, colours, we may detect, on different scales, different properties that help us to identify a texture. Because of its good directional selectivity, we considered the DT*C*WT as an appropriate tool for texture characterization.

We have considered a database containing 12 grayscale texture images from the Brodatz album (Fig. 5). We uniformly preprocessed these images (the original and the rotated ones) in order to obtain a size of 64×64.

Fig. 1. Image of a sidewalk in Iasi, Romania

Fig. 2. (a), (b) Selected textures: bark, grass, pavement

Fig. 3. Rotated scene of the original image from Fig. 1

Figure 6(a) and (b) are presenting the six directional DTCWT sub-band images $S_{j,1}$ for the texture samples from Fig. 4.

Fig. 4. Bark, grass, pavement samples from Fig. 3

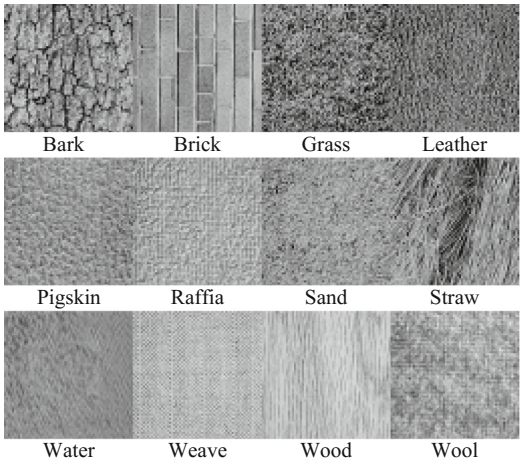


Fig. 5. Brodatz textures database – 12 selected samples for our tests

For the DTCWT we used the following indexes for the six sub-images: $j = 1$ for +15º, $j = 2$ for +45º, $j = 3$ for +75º, $j = 4$ for −75º, $j = 5$ for −45º, and $j = 6$ for −15º. We tested different levels of wavelet decomposition ($D = 1, 2, 3, 4$) thus obtaining $2 \times 6 \times D$ directional coefficients.

The provided sub-band images are complex:

$$W_{j,d} = V_{j,d} + iU_{j,d}, j = 1, 2, \ldots, 6. \tag{1}$$

The DTCWT provides the $\{V_{j,d}, U_{j,d}\}$ pairs for $j = 1,..., 6$, $d = 1,..., D$, along with the real lowpass image from the last decomposition level. For our computations we employed only the amplitude $S_{j,d}$ of the $\{V_{j,d}, U_{j,d}\}$ coefficients, i.e.:

$$S_{j,d} = |W_{j,d}| = \sqrt{V_{j,d}^2 + U_{j,d}^2}, j = 1, \ldots, 6, d = 1, \ldots, D \tag{2}$$

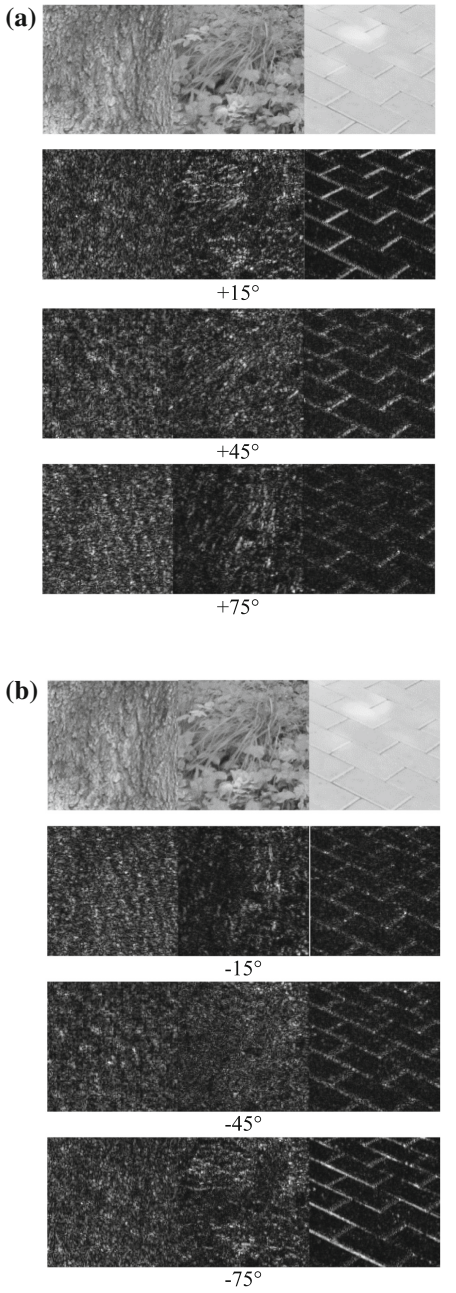


Fig. 6. (a) Selected textures: bark, grass, pavement 2 and the DT*C*WT amplitudes in the three directions: 15˚, 45˚, 75˚. (b) Selected textures: bark, grass, pavement 2 and the DT*C*WT amplitudes in the three directions: −15˚, −45˚, −75˚, $d = 1$

We computed the $\{S_{j,d}; j = 1, \ldots, 6, d = 1, \ldots, D\}$ sub-band images for the normalized texture images (with intensity values in *[0, 1]*).

We multiplied the $S_{j,d}$ by 100, rounded to the nearest integer and computed the histogram, thus obtaining six 100–dimensional feature vectors for each wavelet decomposition level. In order to compare these histograms we tested some of the distances described in [20] (five from the χ^2 family, five from the Shannon's entropy family, the cosine and the Pearson correlation coefficient). We selected five that gave the best results.

Let p and q be two histograms of size $m = 100$ we want to compare. We used the following distances in order to compare the histograms:

$$d_{\cos}(p, q) = 1 - \cos(p, q) \tag{3}$$

$$\cos(p, q) = \frac{\sum_{i=1}^{m} p_i q_i}{\sqrt{\sum_{i=1}^{m} p_i^2} \sqrt{\sum_{i=1}^{m} q_i^2}} \tag{4}$$

$$d_{corr}(p, q) = \frac{1 - r(p, q)}{2} \tag{5}$$

$$r(p, q) = \frac{\mathrm{cov}(p, q)}{\sigma_p \sigma_q} \tag{6}$$

where by *cov(p,q)* we denoted the covariance of the random variables p, q, and σ_p, σ_q are the standard deviations of p and q respectively.

From the Shannon's entropy family we used Topsøe, Jeffreys and Jensen difference distances:

$$d_{Top}(p, q) = \sum_{i=1}^{m} \left[p_i \log \frac{2p_i}{(p_i + q_i)} + q_i \log \frac{2q_i}{(p_i + q_i)} \right] \tag{7}$$

$$d_J(p, q) = \sum_{i=1}^{m} (p_i - q_i) \log \frac{p_i}{q_i} \tag{8}$$

$$d_{Jd}(p, q) = \sum_{i=1}^{m} \left[\frac{p_i \log p_i + q_i \log q_i}{2} - \frac{p_i + q_i}{2} \log \frac{p_i + q_i}{2} \right] \tag{9}$$

and from the χ^2 family we employed the squared χ^2:

$$d_{SqChi}(p, q) = \sum_{i=1}^{m} \frac{(p_i - q_i)^2}{p_i + q_i} \tag{10}$$

We employed these similarity measures in the following way:

$$d_{hist}(X,Y) = \sum_{d=1}^{D} \sum_{j=1}^{6} d_*(hist(S_{X,j,d}), hist(S_{Y,j,d})) \tag{11}$$

where d_* is one of the five distances described above.

Figure 7 depicts the histograms of the DTCWT coefficients in the direction of 75° (histograms of S_3) for the Sand, Water and Wood textures. As can be seen, these histograms can be used as a separation measure between images.

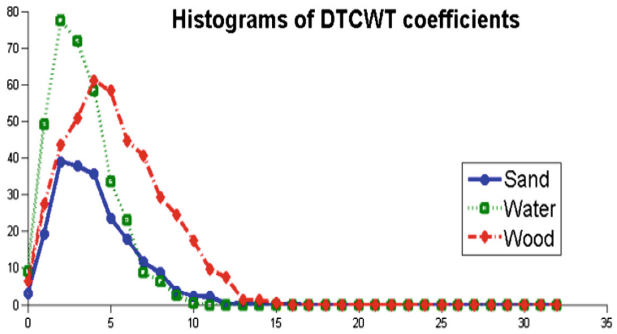


Fig. 7. Histogram of the DTCWT coefficients for the 75° direction for Sand, Water and Wood original textures

For testing the features and the classifiers we applied two types of transformations: a geometrical one (i.e. rotations with different rotation angles) and γ intensity transforms. We used the following γ transformation [22]:

$$T(r) = r^{\gamma}. \tag{12}$$

Using DTCWT, the above presented similarity measures and minimum distance classifiers or SVM we seek to retrieve for each transformed image the original texture that produced it.

5 Results and Discussions

For the numerical tests we employed MATLAB. To compute the DTCWT we adapted [23] the software package from Professor Nick Kingsbury's web page [24].

As mentioned above, we tested the extracted features and the classifiers on 12 textures selected from the Brodatz album ([7], Fig. 5).

We consider as classifiers the minimum distance classifier (using the distances previously introduced, i.e. (3), (5), (7)–(10)) and SVM. All the texture images are of size 64×64.

Four types of classes representing the textures were considered. The first type has classes containing only one texture, the original one. The rotated and the γ transformed ones are used separately for tests. The last three types of classes contain more than one image.

The original Brodatz album contains images of size 512×512. Using these images and resizing we computed rotated and γ transformed textures of the same size (512×512). For each such texture we extracted 64 (8·8) non-overlapping sub-images of size 64×64. Each class that represents a texture is formed by putting together these 192 (3·64) images. We consider three situations: in the first two cases we treat independently the rotated and the γ transformed images and in the third one the classes contain a mixture of all types of images (base, rotated, γ transformed).

In the rotation case the class is formed only with rotated sub-images and those extracted from the original texture. In the γ transformed situation the class is formed only with original and γ transformed images.

We formed the training set by choosing randomly 50% of the class content (50% of each type of image: original, rotated, γ transformed), the other ones were employed for classification purpose.

We have rotated the texture images with the following angles: 5°, 10°, 15° and 30°. In Fig. 8 are depicted the Raffia and the Wood textures and their rotated versions.

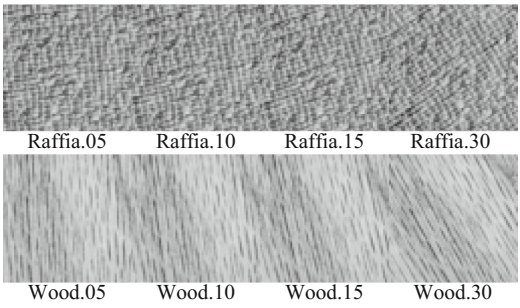


Fig. 8. Brodatz textures database – 2 rotated samples for tests

We applied also several γ transformations to the original Brodatz. In Fig. 9. we show for Grass, Wood and Sand different γ-transformed variants of these textures. In this paper, we considered values for γ in *[0.5, 2]*.

In the case of classes with only one element we used the following five distances: cosine, correlation, Topsøe, Jensen difference, and squared χ^2.

For the rotation angle of 5°, 10° and 15° we get 100% identification rate for all five considered distances. For 30° rotated textures the cosine distance provides 3 errors, the correlation distance 4 errors, the other three distances have each 2 errors.

The comparisons between the Raffia and Wood rotated textures and the original ones using the five similarity measures described applied to the histogram of the DT sub-band images are shown in Tables 1, 2, 3 and 4. The misidentifications appear for the 30° rotation angle.

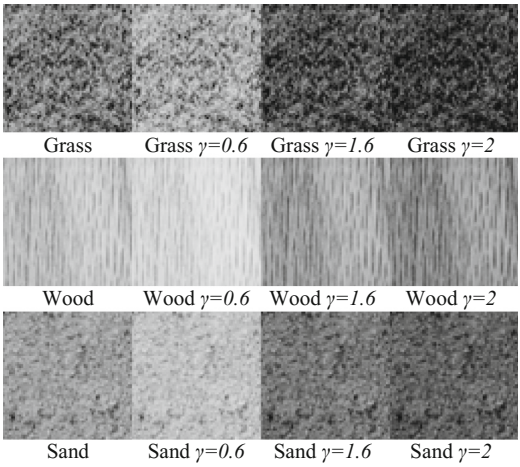


Fig. 9. γ-transformed Brodatz textures – Grass, Wood and Sand

Table 1. Cosine and correlation distances - rotated Raffia images

	Raffia.05		Raffia.10		Raffia.15		Raffia.30	
	d_{cos}	d_{corr}	d_{cos}	d_{corr}	d_{cos}	d_{corr}	d_{cos}	d_{corr}
Bark.00	0.295	0.170	0.285	0.163	0.280	0.161	0.288	0.167
Brick.00	0.333	0.181	0.347	0.189	0.342	0.186	0.390	0.212
Grass.00	0.328	0.183	0.323	0.180	0.324	0.182	0.297	0.167
Leather.00	0.155	0.081	0.138	0.072	0.125	0.065	**0.126**	**0.066**
Pigskin.00	0.160	0.085	0.165	0.088	0.175	0.094	0.188	0.101
Raffia.00	**0.045**	**0.021**	**0.073**	**0.037**	**0.080**	**0.040**	0.138	0.072
Sand.00	0.143	0.076	0.131	0.069	0.122	0.064	0.150	0.080
Straw.00	0.181	0.099	0.208	0.114	0.193	0.106	0.220	0.122
Water.00	0.281	0.147	0.286	0.149	0.284	0.149	0.326	0.171
Weave.00	0.228	0.119	0.241	0.126	0.204	0.106	0.208	0.109
Wood.00	0.342	0.178	0.346	0.180	0.337	0.175	0.407	0.213
Wool.00	0.136	0.072	0.141	0.075	0.145	0.077	0.184	0.098

The cosine and the correlation distances make the same mistake (Leather) and the other three distances misidentifies Raffia with Sand. Both mistakes are understandable from the visual point of view taking into account that we use a directional instrument.

The same thing happens with the Wood texture, i.e., the cosine and the correlation distances provide Weave as classification result and the other three classify Wood as Water.

We repeated the same classification procedure as we used for the rotated textures for the γ transformed ones. We get 100% identification rate for $\gamma \in [0.6, 1.7]$.

In Table 5 are the classification results for Grass and Sand with $\gamma = 0.6$ and in Table 6 are the comparisons results for Grass and Wood with $\gamma = 2$.

Table 2. Topsøe, Jensen difference and squared χ^2 distances - rotated Raffia images

	Raffia.05			Raffia.10			Raffia.15			Raffia.30		
	d_T	d_{Jd}	d_{Sq}	d_T	d_{Jd}	d_{Sq}	d_T	d_{Jd}	d_{Sq}	d_T	d_{Jd}	d_{Sq}
Bark.00	12.7	6.4	21.0	12.5	6.2	20.6	12.3	6.2	20.5	13.3	6.6	22.0
Brick.00	29.0	14.5	48.8	29.3	14.6	49.2	29.5	14.8	49.7	32.7	16.3	53.9
Grass.00	18.6	9.3	30.9	18.6	9.3	30.9	18.3	9.2	30.7	17.9	9.0	30.1
Leather.00	12.0	6.0	20.4	11.3	5.7	19.4	11.0	5.5	18.9	10.5	5.3	18.3
Pigskin.00	9.8	4.9	17.8	9.9	5.0	17.8	9.8	4.9	17.6	10.9	5.4	19.4
Raffia.00	**1.5**	**0.7**	**2.7**	**2.3**	**1.1**	**4.2**	**3.0**	**1.5**	**5.6**	8.6	4.3	15.3
Sand.00	6.9	3.5	12.8	6.7	3.3	12.4	6.5	3.2	12.0	**6.9**	**3.4**	**12.6**
Straw.00	18.2	9.1	30.3	18.0	9.0	30.0	17.8	8.9	29.8	18.0	9.0	30.0
Water.00	30.8	15.4	50.8	31.2	15.6	51.2	31.4	15.7	51.5	33.9	17.0	55.1
Weave.00	16.5	8.3	27.3	16.5	8.2	27.1	15.6	7.8	25.8	12.7	6.4	21.4
Wood.00	29.2	14.6	46.5	29.8	14.9	47.3	30.4	15.2	48.4	35.7	17.9	56.4
Wool.00	9.5	4.7	16.9	9.6	4.8	16.9	9.8	4.9	17.4	12.6	6.3	21.9

Table 3. Cosine and correlation distances - rotated Wood images

	Wood.05		Wood.10		Wood.15		Wood.30	
	d_{cos}	d_{corr}	d_{cos}	d_{corr}	d_{cos}	d_{corr}	d_{cos}	d_{corr}
Bark.00	0.605	0.333	0.609	0.336	0.590	0.326	0.509	0.283
Brick.00	0.192	0.104	0.190	0.102	0.203	0.110	0.272	0.146
Grass.00	0.619	0.336	0.624	0.339	0.616	0.336	0.544	0.297
Leather.00	0.447	0.235	0.437	0.230	0.425	0.224	0.350	0.185
Pigskin.00	0.448	0.236	0.430	0.227	0.406	0.214	0.335	0.177
Raffia.00	0.350	0.182	0.346	0.180	0.336	0.175	0.353	0.185
Sand.00	0.421	0.221	0.421	0.222	0.400	0.211	0.329	0.174
Straw.00	0.409	0.220	0.400	0.214	0.385	0.207	0.355	0.192
Water.00	0.140	0.070	0.129	0.065	0.112	0.056	0.174	0.088
Weave.00	0.318	0.164	0.320	0.165	0.306	0.158	**0.232**	**0.119**
Wood.00	**0.045**	**0.020**	**0.053**	**0.024**	**0.082**	**0.040**	0.239	0.123
Wool.00	0.390	0.205	0.384	0.203	0.361	0.191	0.339	0.180

The five distances behave in the same manner as for the rotated textures: the cosine and the correlation misidentify the same texture, and Topsøe, Jensen difference and squared χ^2 make the same mistake.

We also studied the influence of histogram equalization of the texture images before computing the DTCWT.

When we use histogram equalization, we obtain better results for the γ-transformed images i.e. the interval with zero errors in the classification process extends to $\gamma \in [0.05, 5.4]$. On the contrary when this procedure is performed on the rotated images the results are worse. In the case of histogram equalization for rotated textures classification the results are: the cosine and correlation distances have 1 error each for the 5°, 10°, 15° rotation angles and 4 respectively 5 misidentifications for 30°.

Table 4. Topsøe, Jensen difference and squared χ^2 distances - rotated wood images

	Wood.05			Wood.10			Wood.15			Wood.30		
	d_T	d_{Jd}	d_{Sq}	d_T	d_{Jd}	d_{Sq}	d_T	d_{Jd}	d_{Sq}	d_T	d_{Jd}	d_{Sq}
Bark.00	37.6	18.8	59.3	37.9	19.0	59.8	37.9	19.0	60.2	33.5	16.7	53.5
Brick.00	19.7	9.9	34.0	20.0	10.0	34.2	19.8	9.9	33.8	25.8	12.9	42.5
Grass.00	43.1	21.6	68.1	43.7	21.8	68.9	44.0	22.0	69.8	39.9	20.0	63.9
Leather.00	33.2	16.6	52.9	33.1	16.5	52.7	33.1	16.5	53.1	27.8	13.9	45.5
Pigskin.00	37.8	18.9	59.7	37.1	18.6	58.7	36.1	18.1	57.3	26.4	13.2	42.5
Raffia.00	30.0	15.0	47.6	30.3	15.2	48.0	30.3	15.2	48.2	31.8	15.9	51.4
Sand.00	35.4	17.7	56.2	35.3	17.6	56.0	34.8	17.4	55.4	26.3	13.1	42.7
Straw.00	28.1	14.0	45.8	28.8	14.4	46.8	29.3	14.6	47.8	28.4	14.2	45.7
Water.00	12.8	6.4	22.1	12.4	6.2	21.4	11.8	5.9	20.3	**19.3**	**9.7**	**32.1**
Weave.00	40.1	20.1	63.2	40.0	20.0	63.0	38.9	19.5	61.5	27.7	13.8	44.8
Wood.00	**1.9**	**0.9**	**3.4**	**2.9**	**1.5**	**5.3**	**4.5**	**2.3**	**8.2**	22.8	11.4	36.1
Wool.00	30.4	15.2	48.5	30.0	15.0	47.9	29.5	14.7	47.2	24.7	12.3	40.3

Table 5. Grass and Sand gamma transformed images comparisons ($\gamma = 0.6$)

	Grass					Sand				
	d_{cos}	d_{corr}	d_T	d_{Jdiff}	d_{Sq}	d_{cos}	d_{corr}	d_T	d_{Jdiff}	d_{Sq}
Bark	0.183	0.112	6.7	3.4	11.7	0.338	0.194	18.5	9.3	31.5
Brick	0.572	0.318	34.6	17.3	57.5	0.289	0.155	23.4	11.7	39.1
Grass	**0.082**	**0.047**	**1.9**	**0.9**	**3.3**	0.399	0.223	28.6	14.3	48.8
Leather	0.193	0.109	6.9	3.4	11.9	0.202	0.108	17.9	9.0	31.2
Pigskin	0.268	0.148	22.6	11.3	38.0	0.138	0.073	**3.6**	**1.8**	**6.3**
Raffia	0.287	0.159	16.8	8.4	27.9	0.135	0.070	11.1	5.6	20.7
Sand	0.251	0.141	15.8	7.9	27.6	0.118	0.061	4.0	2.0	7.4
Straw	0.240	0.136	13.6	6.8	23.8	0.223	0.123	21.6	10.8	36.4
Water	0.566	0.310	49.5	24.8	79.1	0.207	0.106	22.3	11.1	36.5
Weave	0.417	0.231	18.9	9.4	31.0	0.180	0.092	13.1	6.5	22.1
Wood	0.589	0.318	41.5	20.8	65.4	0.364	0.190	35.5	17.8	57.2
Wool	0.305	0.170	26.7	13.3	44.1	**0.114**	**0.060**	3.9	1.9	6.8

The Topsøe, Jensen difference and squared χ^2 distances have 6 errors each for the 30° rotated textures and no misidentification for the smaller rotation angles.

For the cases when the classes have more than one texture we used as classifiers SVM and minimum distance with cosine, Jeffreys, Topsøe, Jensen difference, and squared χ^2.

When the rotations and γ-transformations were treated separately, each class in the training set has 64 elements, 32 from the original texture and 32 from the transformed one. The test set contains a total of 768 images, 64 from each texture.

The rotation angles are the same as previously ($\theta \in \{5°, 10°, 15°, 30°\}$) and the values for γ are in $\{0.5, 0.6, 0.8, 0.8^{-1}, 0.6^{-1}, 2\}$. We observe from Table 7 that as θ increases usually the classification results improve, one of the poorest results is for

Table 6. Grass and Wood gamma transformed images comparisons ($\gamma = 2$)

	Grass					Wood				
	d_{cos}	d_{corr}	d_T	d_{Jdiff}	d_{Sq}	d_{cos}	d_{corr}	d_T	d_{Jdiff}	d_{Sq}
Bark	0.200	0.119	7.1	3.5	11.7	0.526	0.293	32.1	16.0	51.0
Brick	0.468	0.256	30.7	15.4	51.5	0.187	0.104	17.4	8.7	31.1
Grass	0.179	0.102	8.6	4.3	15.4	0.530	0.288	36.3	18.1	57.2
Leather	0.138	0.075	**5.1**	**2.5**	**9.1**	0.349	0.183	26.8	13.4	42.9
Pigskin	0.146	0.078	10.3	5.1	17.9	0.386	0.205	41.0	20.5	64.5
Raffia	0.171	0.092	8.3	4.2	14.3	0.286	0.150	30.6	15.3	48.1
Sand	**0.134**	**0.072**	5.3	2.7	9.7	0.384	0.204	37.3	18.7	59.7
Straw	0.205	0.115	13.3	6.7	23.1	0.291	0.156	18.7	9.3	30.4
Water	0.418	0.224	36.5	18.2	58.9	0.162	0.085	23.1	11.5	38.6
Weave	0.310	0.168	11.9	6.0	19.9	0.344	0.180	43.2	21.6	67.7
Wood	0.494	0.262	34.5	17.2	54.4	**0.112**	**0.057**	**6.9**	**3.5**	**12.3**
Wool	0.184	0.099	13.4	6.7	22.7	0.320	0.170	33.7	16.9	53.7

Table 7. Classification results for rotated case

	Rotation angles			
	$\theta = 5°$	$\theta = 10°$	$\theta = 15°$	$\theta = 30°$
SVM	96.75% (25)	96.09% (30)	95.31% (36)	95.44% (35)
Jeffreys	91.54% (65)	92.06% (61)	92.32% (59)	92.84% (55)
Topsøe	93.88% (47)	93.75% (48)	93.75% (48)	94.79% (40)
Jensen	93.88% (47)	93.75% (48)	93.75% (48)	94.79% (40)
Squared χ^2	93.62% (49)	93.62% (49)	93.75% (48)	94.66% (41)
Cos	85.29% (113)	84.64% (118)	85.94% (108)	85.16% (114)

$\theta = 5°$, the DT*C*WT provides better separation of textures if the rotation angle is greater than 10°. Best results are obtained with the SVM and the Topsøe, Jensen difference distances.

When applying γ transformations, we considered pairs $(g, 1/g)$, $g<1$. As was expected as g approaches *1* the results are getting better. Comparing the results provided by each pair $(g, 1/g), g<1$ one remarks that we get better results for γ values greater than *1* which provide better contrast than $\gamma < 1$. Again, with the SVM one gets the best results and Topsøe, Jensen difference, and squared χ^2 distances come next, making the same number of mistakes (Table 8).

We also analysed cases when all the textures were put together: first, the original and the rotated textures using all angles, second the original and all the γ-transformed images, and the third was when all the images representing a texture were present in class content.

In the first situation, each class has 160 elements and 1920 texture were tested. In the second case the classes have each 352 elements and 4224 textures were classified.

In the last case, the classes contain 480 images each and 5760 texture were employed for classification. The results are in Table 9.

Table 8. Classification results for γ transformation case

	γ Transformations					
	$\gamma = 0.5$	$\gamma = 0.6$	$\gamma = 0.8$	$\gamma = 0.8^{-1}$	$\gamma = 0.6^{-1}$	$\gamma = 2$
SVM	95.96% (31)	97.66% (18)	99.74% (2)	99.74% (2)	97.92% (16)	96.22% (29)
Jeffreys	95.70% (33)	96.62% (26)	96.62% (26)	100% (0)	97.14% (22)	95.31% (36)
Topsøe	96.62% (26)	97.79% (17)	100% (0)	99.87% (1)	96.75% (25)	95.83% (32)
Jensen	96.62% (26)	97.79% (17)	100% (0)	99.87% (1)	96.75% (25)	95.83% (32)
Squared χ^2	96.35% (28)	97.79% (17)	100% (0)	99.87% (1)	97.01% (23)	95.83% (32)
cos	87.89% (93)	91.54% (65)	95.44% (35)	98.57% (11)	94.01% (46)	91.78% (63)

Table 9. Classification results for rotations, γ -transformed considered together

	Rotations	γ-transformed	All
SVM	97.19% (54)	99.55% (19)	98.23% (102)
Jeffreys	95.21% (92)	99.12% (37)	98.19% (104)
Topsøe	95.99% (77)	99.12% (37)	98.72% (74)
Jensen	95.99% (77)	99.12% (37)	98.72% (74)
Squared χ^2	96.41% (69)	99.17% (35)	98.82% (68)
cos	90.42% (184)	96.97% (128)	95.24% (274)

In this case the squared χ^2 minimum distance classifier gives in average the best results and then the SVM.

If we consider all the cases with classes represented by multiple textures, we computed the average classification rate: the best results are obtained with SVM (with a total of 399 errors), then follows the squared χ^2 distance (460 errors), the Topsøe, Jensen difference distances make 472 errors each, Jeffreys yields 616 misidentifications and the cosine distance makes a total of 1352 errors.

For the situation of class content with mixture of all types of images we computed for each texture the total number of misidentifications, the results are summarized in Table 10. It is surprising that a texture such as Brick which has good saliency, with

Table 10. Number of misidentifications

	Rotations	γ-transformed	All
Bark	64	85	161
Brick	169	27	146
Grass	21	56	61
Leather	8	19	33
Pigskin	19	5	23
Raffia	13	2	6
Sand	121	26	113
Straw	30	25	21
Water	0	0	0
Weave	51	24	76
Wood	3	0	4
Wool	54	24	52

very distinct visual and directional characteristics has one of the highest misidentification rate and the Water texture which visually is very similar with Sand or Pigskin has no errors in classification.

We tested the cosine, correlation, Topsøe, Jensen difference and squared χ^2 histogram similarity measures upon the six real world textures from Figs. 2 and 4. There are three classes of textures (bark, grass, and pavement) each class having two samples. We computed for all these textures the similarity measures to the other five images. In Table 11 we marked 'T' (true) when the minimum distance was achieved for the image in the same class and 'F' (false) otherwise. The DT*C*WT -based similarity measures provided the correct result in all the cases (the DT*C*WT column in Table 11).

Table 11. Comparisons for real world images – distances applied to the histograms of the DTCWT coefficients and of the images

	Cosine		Correlation		Topsøe		Jensen		Squared χ^2	
	DTCWT	Img	DTCWT	Img	DTCWT	Img	DTCWT	Img	DTCWT	Img
Bark	T	F	T	F	T	F	T	F	T	F
Grass	T	T	T	T	T	T	T	T	T	T
Pavement	T	F	T	F	T	T	T	T	T	T

We compare the classification obtained applying the five distances to the histogram of the DT*C*WT coefficients to the same classification performed when computing the similarity of the histogram of the textures (the Img column in Table 11). We obtained the correct texture when comparing the histogram of the rotated textures with the not rotated ones, in almost all cases, as expected. The cosine and the correlation distances had 1 misidentification each for the 30° case.

6 Conclusions

In this paper, we studied texture identification in digital images with a new method. We employed the Dual Tree Complex Wavelet Transform (DTCWT) for texture feature extraction and identification using various similarity measures. The DT*C*WT provides a set of six complex selective directional coefficients, for which we employed their magnitudes $\{S_{j,d};\ j = 1,...,6,\ d = 1,..,D\}$. We computed the histograms of these coefficients and classified 12 textures from the Brodatz album using SVM or minimum distance classifiers employing six similarity measures: cosine and Pearson correlation coefficient, Jeffreys, Topsøe, Jensen difference and squared χ^2. Our main focus was on the influence of rotation and γ-transformation on the capacity of texture recognition. We analysed several cases: classes with only one texture or with multiple textures, the rotations and γ-transformations were treated separately or together. We also tested the proposed method on real world images with good identification results. Our further research will concentrate on testing the influence of the image size, the noise and also on other types of image transformations (blur, shear, etc.) beside rotations and

γ-transformations on the capacity of the presented methods for texture identification. These procedures will be tested also on image segmentation and indexing based on texture recognition.

References

1. Petrou, M., Sevilla, P.G.: Image Processing: Dealing With Texture. Wiley, England (2006)
2. Mirmehdi, M., Xie, X., Suri, J. (eds.) Handbook of Texture Analysis. Imperial College Press, London (2008)
3. Wang, X., Georganas, N.D., Petriu, E.M.: Fabric texture analysis using computer vision techniques. IEEE Trans. Instrum. Measur. **60**(1), 44–56 (2011)
4. Basile, T.M.A., Caponetti, L., Castellano, G., Sforza, G.: A texture-based image processing approach for the description of human oocyte cytoplasm. IEEE Trans. Instrum. Measur. **59** (10), 2591–2601 (2010)
5. Kingsbury, N.G.: The dual-tree complex wavelet transform: a new technique for shift invariance and directional filters. In: Proceedings of the 8th IEEE DSP Workshop, Paper no. 86, Utah, 9–12 August 1998
6. Kingsbury, N.G.: The dual-tree complex wavelet transform: a new efficient tool for image restoration and enhancement. In: Proceedings of the European Signal Processing Conference, Rhodes, September 1998, pp. 319–322 (1998)
7. Brodatz, P.: Textures: a photographic album for artists and designers, Dover, New York, USA (2014). http://www.ux.uis.no/~tranden/brodatz.html
8. Hill, P.R., Bull, D.R., Canagarajah, C.N.: Rotationally invariant texture features using the dual-tree complex wavelet transform. In: IEEE Proceedings of the International Conference on Image Processing (ICIP), Vancouver, Canada, vol. 3, pp. 901–904 (2000)
9. Selesnick, I.W., Baraniuk, R.G., Kingsbury, N.G.: The dual tree complex wavelet transform DTCWT. IEEE Sig. Process. Mag. **22**(6), 123–151 (2005)
10. Kingsbury, N.G.: Complex wavelets for shift invariant analysis and filtering of signals. J. Appl. Comput. Harmonic Anal. **3**, 234–253 (2001)
11. Kingsbury, N.G.: Design of Q-shift complex wavelets for image processing using frequency domain energy minimization. In: Proceedings of the IEEE Conference on Image Processing, Barcelona, 15–17 September (2003). Paper 1199
12. Mallat, S.G.: A theory for multiresolution signal decomposition: the wavelet representation. IEEE Trans. Pattern Anal. Mach. Intell. **11**(7), 674–693 (1989)
13. Hatipoglu, S., Mitra, S.K., Kingsbury, N.G.: Image texture description using complex wavelet transform. In: Proceedings of the IEEE International Conference on Image Processing, Vancouver, BC, Canada, September 2000, vol. 2, pp. 530–533 (2000)
14. Wang, H.-Z., He, X.-H., Zai, W.-J.: Texture image retrieval using dual-tree complex wavelet transform. In: Proceedings of the 2007 International Conference on Wavelet Analysis and Pattern Recognition, Beijing, China, 2–4 November 2007, pp. 230–234 (2007)
15. Mumtaz, A., Gilani, M.S.A., Hameed, K., Jameel, T.: Enhancing performance of image retrieval systems using dual tree complex wavelet transform and support vector machines. J. Comput. Inf. Technol. (CIT) **16**(1), 57–68 (2008)
16. Celik, T., Tjahadi, T.: Multiscale texture classification using dual-tree complex wavelet transform. Pattern Recogn. Lett. **30**, 331–339 (2009)
17. Hatipoglu, S., Mitra, S.K., Kingsbury, N.: Texture classification using dual-tree complex wavelet transform. IEEE Image Process. Appl. **465**, 344–347 (1999)

18. Lo, E.H.S., Pickering, M., Frater, M., Arnold, J.: Scale and rotation invariant texture features from the dual-tree complex wavelet transform. In: IEEE Proceedings of the International Conference on Image Processing (ICIP), 24–27 October, Singapore, vol. 1, pp. 227–230 (2004)
19. Chen, S., Shang, Y., Mao, B., Lian, Q.: Rotation invariant texture classification algorithm based on DT-CWT and SVM. In: Liu, D., Fei, S., Hou, Z., Zhang, H., Sun, C. (eds.) Proceedings of the 4th International Symposium on Neural Networks Advances in Neural Networks ISNN 2007, pp. 454–460. Springer, Heidelberg (2007).
20. Cha, S.-H.: Comprehensive survey on distance/similarity measures between probability density functions. Int. J. Math. Models Methods Appl. Sci. **1**(4), 300–307 (2007)
21. Xu, R., WunschII, D.C.: Clustering. IEEE Press/Wiley, Hoboken (2009)
22. Gonzales, R.C., Woods, R.E.: Digital Image Processing, 3rd ed. Chap. 3, pp. 132–137. Prentice Hall, Upper Saddle River (2008)
23. Costin (Luca), M., Ignat, A.: Pitfalls in using dual tree complex wavelet transform for texture featuring: a discussion. In: IEEE WISP 2011 - 7th IEEE International Symposium on Intelligent Signal Processing, Floriana, Malta, 19–21 September 2011, pp. 110–115 (2011)
24. Kingsbury's, N.G.: http://www-sigproc.eng.cam.ac.uk/Main/NGK

Author Index

V.E. Balas et al. (eds.), *Soft Computing Applications*, Advances in Intelligent Systems and Computing 633, DOI 10.1007/978-3-319-62521-8