The Institute of Electrical and Electronics Engineers Inc.
Region 8 – Europe, Middle East and Africa
Poland Section, Chapter Signal Processing, Circuits and Systems

Audio Engineering Society
Polish Section

PROCEEDINGS OF THE JOINT CONFERENCE:

NTAV/SPA 2012

NEW TRENDS IN AUDIO AND VIDEO SIGNAL PROCESSING: ALGORITHMS, ARCHITECTURES, ARRANGEMENTS AND APPLICATIONS

ŁÓDŹ, POLAND, 27-29 SEPTEMBER 2012

Lodz University of Technology
Institute of Electronics

wydział elektrotechniki elektroniki informatyki i automatyki

PTETiS
Polish Society
of Theoretical and Applied
Electric Engineering

NTAV/SPA 2012

NEW TRENDS IN AUDIO AND VIDEO SIGNAL PROCESSING ALGORITHMS, ARCHITECTURES, ARRANGEMENTS AND APPLICATIONS

CONFERENCE PROCEEDINGS

ŁÓDŹ, 27–29 SEPTEMBER 2012

Conference Organizer

Lodz University of Technology
Faculty of Electrical, Electronic, Computer and Control Engineering
Institute of Electronics, Medical Electronics Division
211/215 Wólczańska Str., B-9 building, 90-924 Łódź, Poland

Co-organizers

Poznan University of Technology
Faculty of Computing Science and Management
Chair of Control and System Engineering
Division of Signal Processing and Electronic Systems
3 Piotrowo Str., 60-965 Poznań, Poland

Wroclaw University of Technology
Institute of Telecommunication, Teleinformatics and Acoustics
Department of Acoustics
27 Wybrzeże Wyspiańskiego Str., 50-370 Wrocław, Poland

Polish Society of Theoretical and Applied Electric Engineering
18/22 Stefanowskiego Str., 90-924 Łódź, Poland

Faculty of Electrical, Electronic, Computer and Control Engineering

The conference was financially supported by the Polish Ministry of Science and Higher Education

Ministry of Science and Higher Education
Republic of Poland

Published by:
Lodz University of Technology Press

Edited by:
Paweł Strumiłło, Michał Bujacz, Monika Popielata

Cover design:
Anna Janisz

ISBN 978-83-7283-502-4
NTAV-SPA 2012 Conf ID 21037

Printed by: QuickDruk, 11 Łąkowa Str., Łódź

Audio Engineering Society

IEEE

PTETiS

NTAV/SPA 2012

NEW TRENDS IN AUDIO AND VIDEO SIGNAL PROCESSING ALGORITHMS, ARCHITECTURES, ARRANGEMENTS AND APPLICATIONS

CONFERENCE PROCEEDINGS

ŁÓDŹ, 27–29 SEPTEMBER 2012

Scientific Committee

CHAIRMAN

Prof. Andrzej Materka	Lodz University of Technology, Poland

VICE-CHAIRMAN

Prof. Adam Dąbrowski	Poznan University of Technology, Poland

VICE-CHAIRMAN

Prof. Andrzej Dobrucki	Wroclaw University of Technology, Poland

Prof. Andrzej Bartoszewicz	Lodz University of Technology, Poland
Dr. Michał Bujacz	Lodz University of Technology, Poland
Prof. Ryszard Choraś	University of Technology and Life Sciences Bydgoszcz, Poland
Prof. Zygmunt Ciota	Lodz University of Technology, Poland
Prof. Anthony Davies	King's College London, Great Britain
Prof. Marek Domański	Poznan University of Technology, Poland
Prof. Andrzej Dziech	AGH University of Science and Technology Cracow, Poland
Prof. Alfred Fettweis	Ruhr-Universität Bochum, Germany
Prof. Ewa Hermanowicz	Gdansk University of Technology, Poland
Prof. Lidia Jackowska-Strumiłło	Lodz University of Technology, Poland
Prof. Michał Jacymirski	Lodz University of Technology, Poland
Prof. Jacek Kabziński	Lodz University of Technology, Poland
Prof. Tomasz Kacprzak	Lodz University of Technology, Poland
Prof. Mos Kaveh	President of the IEEE Signal Processing Society
Prof. Piotr Kleczkowski	AGH University of Science and Technology Cracow, Poland
Prof. Christian Kollmitzer	University of Applied Sciences Technikum Wien, Austria
Prof. Bożena Kostek	Gdansk University of Technology, Poland
Prof. Krzysztof Kozłowski	Poznan University of Technology, Poland
Prof. Rolf Kraemer	IHP Microelectronics, Frankfurt/Oder, Germany
Prof. Jacek Kucharski	Lodz University of Technology, Poland
Prof. Zbigniew Kulka	Warsaw University of Technology, Poland
Prof. Józef Modelski	Warsaw University of Technology, Poland
Prof. George Moschytz	Bar-Ilan University, Israel
Prof. Włodzimierz Mosorow	Lodz University of Technology, Poland
Prof. Andrzej Napieralski	Lodz University of Technology, Poland
Prof. Peter Noll	Technische Universität Berlin, Germany
Prof. Antoni Nowakowski	Gdansk University of Technology, Poland
Prof. Maciej Ogorzałek	Jagiellonian University in Krakow, Poland
Prof. Stanisław Osowski	Warsaw University of Technology, Poland
Prof. Aleksander Petrovsky	Bialystok University of Technology, Poland
Prof. Jan Purczyński	West Pomeranian University of Technology, Szczecin, Poland
Prof. Kamisetty R. Rao	University of Texas at Arlington, USA

Conference Organizing Committee

NTAV/SPA 2012

27-29TH SEPTEMBER, 2012, ŁÓDŹ, POLAND

NEW TRENDS IN AUDIO AND VIDEO / SIGNAL PROCESSING ALGORITHMS, ARCHITECTURES, ARRANGEMENTS AND APPLICATIONS

IEEE

CONFERENCE INFORMATION

Conference Record #: 21037

Conference Title: 2012 Joint Conference New Trends In Audio & Video And Signal Processing: Algorithms, Architectures, Arrangements And Applications

Conference Acronym: NTAV/SPA 2012

Conference Dates: 9/27/2012 to 9/29/2012

Location: Lodz University of Technology

City: Lodz

Country: Poland

Exhibits: N **No. of Exhibits:** 0 **Tutorials:** Y **Attendance:** 100 **Producing Publication:** Y **Concentration Banking Info:** No

Conference Scope: DSP theory, algorithms and implementations, Image and video processing, Audio processing, Digital Television and stereovision, Multimedia data compression and editing, data bases, Electroacoustics, Psychoacoustics, Sound mastering, Human computer interfaces, Image synthesis, analysis and recognition, Filter design and implementation, Vision and audio-based diagnosis in medicine and industry, Text-to-speech & speech-to-text algorithms, Artificial intelligence applications in DSP, Distributed and networked DSP systems, Virtual and augmented reality, Biometrics

Conference Keywords: dsp, image processing, audio processing, speech processing, multimedia, audio and visual systems, ntav/spa 2012

Conference Focus: Application, Scientific/Academic

WWW URL: http://www.ntavspa.pl

CFP URL: http://www.ntavspa.pl/__files/call_ntav_spa_2012.pdf

Abstract Submission Date: 5/31/2012

Notification of Acceptance: 7/15/2012

Final Paper Submission Date: 8/15/2012

Expenses: –

Revenue: –

Sponsors

Lodz University of Technology, Poland (Co Sponsor) – 40%

Audio Engineering Society – AES (Co Sponsor) –30%

IEEE Poland Section Signal Processing Chapter (Co Sponsor) –10%

IEEE Poland Section Circuits and Systems Chapter (Co Sponsor) –10%

Poznan University of Technology, Poland (Co Sponsor) –10%

Information Contact:
Julian Balcerek
Piotrowo 3
Poznan University of Technology
Chair of Control and System Engineering
Division of Signal Processing and Electronic Systems
Poznan 60-965
POLAND
Ph : +48 61 665 2833
Fax: +48 61 665 2840
julian.balcerek@put.poznan.pl

Technical Program Chair:
Paweł Strumiłło
211/215 Wolczanska Str., B-9 building
Lodz University of Technology
Lodz 90-924
POLAND
Ph : +48 42 631 2646
Fax: +48 42 636 2238
pawel.strumillo@p.lodz.pl

Conference Chair:
Andrzej Materka
211/215 Wolczanska Str., B-9 building
Technical University of Lodz
Lodz 90-924
POLAND
Ph : +48 42 631 2644
Fax: +48 42 636 2238
andrzej.materka@p.lodz.pl

Publication Chair:
Tomasz Marciniak
Piotrowo 3
Poznan University of Technology
Chair of Control and System Engineering
Division of Signal Processing and Electronic Systems
Poznan 60-965
POLAND
Ph : +48 61 665 2836
Fax: +48 61 665 2840
tomasz.marciniak@put.poznan.pl

Treasurer:
Zdzisława Sobańska
211/215 Wolczanska Str., B-9 building
Lodz University of Technology
Lodz 90-924
POLAND
Ph : +48 42 631 2621
Fax: +48 42 636 2238
zdzislawa.sobanska@p.lodz.pl

Info Schedule Submitted by
Adam Dąbrowski
+48 61 665 2831
adam.dabrowski@put.poznan.pl

Entered in database: 4/16/2012

Program summary

	THURSDAY 27.09.2012	FRIDAY 28.09.2012	SATURDAY 29.09.2012	
09:00–09:45		Plenary Lecture II prof. Andrzej Czyżewski	Plenary Lecture III prof. Władysław Skarbek	09:00–09:45
10:00–10:15		SESSION 4 Human-Computer Interaction	SESSION 7 Stereovision and Video Coding	10:00–10:15
10:15	OPENING			10:15
10:30–11:00	Plenary Lecture I prof. Arvid Lundervold	COFFEE	COFFEE	10:30–11:00
11:15–12:15	SESSION 1 Image Processing I	SESSION 5 Image Processing II and Biomedical Applications	SESSION 8 DSP, Hardware and Applications	11:15–12:15
12:30–12:45	COFFEE			12:30–12:45
13:00			CLOSING	13:00
13:15–14:00	SESSION 2 Audio Processing I	LUNCH	LUNCH	13:15–14:00
14:15–15:00	LUNCH	SESSION 6 Audio Processing III		14:15–15:00
15:15–15:45	Hardware-software platform for measurements and processing of signals from National Instruments Sponsor Presentation by Wojciech Sommer	COFFEE		15:15–15:45
16:00–16:15	SESSION 3 Audio Processing II	CITY TOUR		16:00–16:15
16:30–16:45	COFFEE			16:30–16:45
17:00–17:45	SESSION 3 Audio Processing II – cont.			17:00–17:45
18:00–18:45				18:00–18:45
19:00–20:45	BANQUET (WHITE FACTORY)	PUB (IRISH PUB)		19:00–20:45
21:00				21:00

All sessions will be held at the Institute of Electronics lecture hall 416b, "Lodex" Building B9, 211/215 Wólczańska Str., Łódź
Lunches will be served at the university cafeteria, 3a Politechniki ave.

Dear Conference Participants!

Welcome to the NTAV/SPA 2012 conference held in Lodz. This year's scientific event joins two successful series of conferences: the New Trends in Audio and Video (NTAV) conference – organized first by the Wrocław University of Technology in 1994 and then biannually by various major technical universities in Poland, and the IEEE Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA) conference – first held in 1993 and then organized yearly since 1999 by the Poznań University of Technology. Credit for the initiation, sustained support and organization of the NTAV and SPA conferences goes to professors Andrzej Dobrucki and Adam Dąbrowski, respectively.

NTAV/SPA 2012 is the second time these two conferences are joined as one event, the first time being in 2008 at the Poznań University of Technology. The aim of the NTAV/SPA 2008 was to gather researchers interested in various signal processing studies (including audio, video and image analysis). Similarly, the aim of this year's event, held four years later, is to bring together colleagues from a broad range of research subjects related to audio, image and video processing, digital signal processing theory, techniques and various applications including human-computer interaction systems and medicine.

An excellent introduction to the conference's broad scope is given by the three invited lectures prepared by distinguished scientists from Norway and Poland. Professor Arvid Lundervold from the University of Bergen presents an important image diagnostic technique known as Functional Magnetic Resonance Imaging (fMRI) and highlights its applications in basic and clinical neurosciences. The lecture by professor Andrzej Czyżewski from the Gdańsk University of Technology is devoted to the dynamically developing field of human-computer interaction and its novel applications in education and aids for persons with disabilities. Finally, professor Władysław Skarbek from the Warsaw University of Technology addresses the appealing concept of literal programming, i.e. the methodology of programming in high level languages resembling natural language more than program code.

The 42 regular papers accepted for NTAV/SPA 2012 were reviewed by two (or in some cases three) reviewers – members of the Scientific Committee. The contributions have been grouped into the following thematic fields: signal processing theory and algorithms, audio processing and acoustics, image processing and analysis, and finally human-computer interaction. All the accepted submissions printed in the conference proceedings will be indexed in the IEEE Xplore database. Selected papers pointed out by the Scientific Committee, after suitable extension, are invited for a publication in peer reviewed journals. It is worth noting here that, thanks to the support of the Audio Engineering Society, the abstracts of the NTAV/SPA 2012 papers devoted to audio and acoustics have been printed in the recent special issue of the Archives of Acoustics, a magazine published by the Polish Academy of Sciences.

The NTAV/SPA 2012 Organizing Committee acknowledges the financial support of the Ministry of Science and Higher Education. In this respect, special thanks are also due to the Dean of the Faculty of Electrical, Electronic, Computer and Control Engineering prof. Sławomir Wiak and the Dean-Elect prof. Sławomir Hausman. The contribution of the Polish Section of the Audio Engineering Society and the Polish Association of Theoretical and Applied Electrotechnics to the organization activities of the conference is also highly appreciated. Finally, sincere thanks are extended to all conference participants who submitted their work to the NTAV/SPA 2012.

We do hope that a wide scope of this joint conference will offer a unique forum for stimulating interdisciplinary discussion on new trends of all aspects of signal, image and video processing techniques and their widening important applications.

We wish you an enjoyable stay in Łódź!

On behalf of the Organizing Committee
Paweł Strumiłło

Contents

SESSION 5:
IMAGE PROCESSING AND BIOMEDICAL APPLICATIONS

New Type of Sensor for Heart Rhythm Monitoring

Wojciech Surtel PhD., Marcin Maciejewski MSc
Institute of Electronics, Lublin University of Technology
Nadbystrzycka 38a 20-618 Lublin
m.maciejewski@pollub.pl

Teresa Małecka-Massalska M.D PhD.
Department of Human Physiology
Medical University of Lublin
Radziwiłłowska 11, 20-080 Lublin

ABSTRACT — **The paper proposes a new type of sensor for heart rhythm monitoring. It is less uncomfortable than classic ECG monitors and provides almost real-time information to the medical team, which is useful in monitoring patients with chronic heart conditions. It also can prove useful in cases when the patient suffers from skin conditions by minimizing contact surface. by utilizing the capabilities of modern mobile devices, whose processing power can also be used to detect abnormal heart operation by sensor signal analysis, which can be later used to inform the medical team. During testing, a prototype was designed, built and tested. A proper application for mobile devices was developed, and trials on a group of volunteers were performed. The device prototype proved to provide sufficient quality signal to be used as a heart monitor.**

KEYWORDS — heart, ecg, piezoelectric sensor, heart, monitoring

I. INTRODUCTION

Modern day science provides humanity with a vast array of methods of communication and data transfer. They range from instant messaging and voice transmission to video conferences and file sharing. Those methods relay on various means of both wired and wireless transmission to relay information between nodes. This allowed for fast development in all fields of knowledge and birth of many new sciences, like telemedicine. It's a relatively young but a rapid expanding mixture of medicine, bioinformatics, telecommunication, networking and computer science. It uses miniature sensors to gather biometrical data, low power highly efficient microprocessors for initial data management and communication interfaces of various types to allow for remote tele-diagnostics, patient monitoring and data gathering. Telemedical devices can be used to construct local networks in or outside of medical facilities and allow the patient mobility while maintaining 24/7 surveillance and health monitoring which, in many cases, means greatly reducing the arrival time of medical teams and greater chance of successful recovery after experiencing a critical condition.

II. CONSTRUCTION

Mobile recorder usually consists of a sensor of a certain type, a microprocessor for initial data management and storage and a wireless interface for communication (fig. 1). Often, the device utilizes more than one type of sensor for monitoring of certain specific conditions, like a heart rate , blood pressure and ECG sensor in for use in monitoring of patients with heart problems. Not only the device is usually responsible for initial preprocessing and normalization of data but also it sometimes uses advanced compression algorithms to reduce the amount of data sent through the interface or decision algorithms that allow for advanced determining of the patient's state. Sometimes, an internal GPS positioning module provides additional location information to the medical response team, which in turn results in faster patient localization, or, coupled with acceleration or location sensor, state of consciousness. The figure below shows the simplified model of teleinformatic system indicating the placement of patient's wireless recorder within the system (fig. 2).

III. LOCAL SENSOR NETWORKS

The human body is a special kind of environment to consider while placing sensors. A great benefit to using wireless autonomous sensors is easy networking [6]. By applying a multifunctional sensor network it is possible to perform a variety of tasks and simultaneously gather different types of information. This sometimes cannot be done by a single device due to the nature of the measurements. In medical condition monitoring many sensors require precise placement to perform accurate analysis of vital signs, for example an

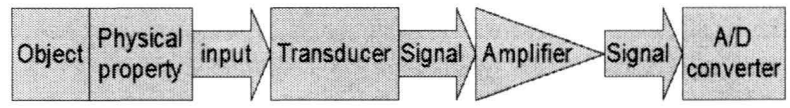


Figure 1. Block diagram showing example sensor input

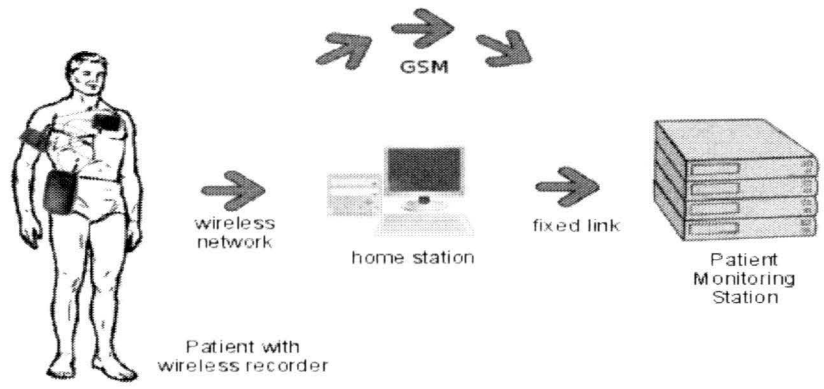


Figure 2. Example biomedical monitoring system idea

ECG sensor gives accurate results when situated on the torso, while a blood pressure monitor performs well when placed around a wrist. This makes using a mesh of sensors highly beneficial (fig. 3). Another important factor is pre-processing

data. Initial assessment of patient condition allows for decrease in the amount of information sent through the interfaces[7]. One central node responsible for information gathering can decide when to relay information to the medical team. This decreases the central server's information overload and response time.

IV. HEART PROBLEMS

Heart problems have been the cause of deaths for as long as the human race exists. Due to the sudden nature and subtle signs preceding heart failures accurate and timed diagnosis is extremely important [4]. Nowadays, it's possible to monitor the heart condition in patients using electrocardiography by attaching medical ECG recorders[3]. Nevertheless, these devices can't provide timely information to the doctor, because the data can only be read once in a while, meaning it can be too late. The answer to this problem is telemedical monitoring, which can ensure 24/7 patient care by heart analysis[1][2]. In this paper a new device utilizing an innovative method to perform heart monitoring is proposed. Standard ECG monitors rely on electrodes attached to the skin. Those systems require multiple nodes, which as a result can leave the patient in a serious discomfort. It's also a known fact that some patients suffer from skin rash and allergy, which can make the use of electrodes impossible.

V. DEVICE

The proposed method utilizes new type of piezoelectric-quartz sensor to monitor heart condition. It was necessary to optimize the sensor for low-frequency signals [5]. It requires just one device attached to the skin, which in turn reduces discomfort and makes it less visible. By using a remote connection with a mobile device, it is not only possible to record the signal, but also to perform signal analysis in real time using the phone's processing power (fig. 4).

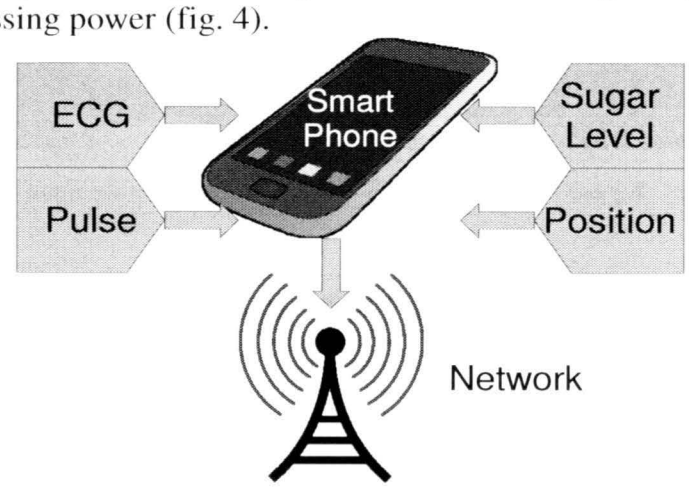


Figure 3. Sensor network example with one central device

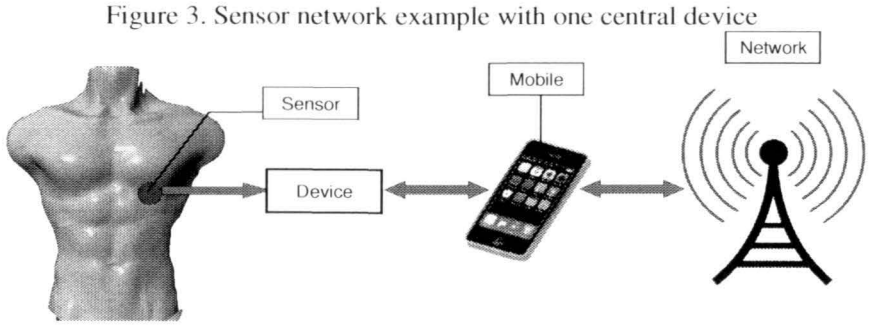


Figure 4. Simplified diagram of the system.

By applying decision mechanism based on the pre-calculated signal parameters it is possible to detect symptoms of heart failure, and in turn to inform the patient and medical staff of the probability of serious life-threatening situation. It is also possible to implement an external or internal motion sensor to detect patient position, which can detect the event of falling or fainting. Also, a GPS module can provide geographical data for the ambulance. Simplified diagram is show in fig. 5. The device consists of five key components: the transducer, an amplifier, A/D converter, interface, and a control system.

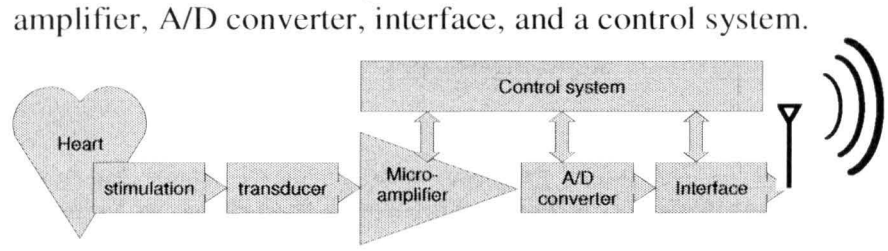


Figure 5. General device operation

VI. SENSOR

The system utilizes a piezoelectric quartz sensor for heart monitoring. Sensors of this type are widely used in industrial applications or in motorization [8]. The sensor has limited use in medical industry, mostly limited to stethoscopes. The piezoelectric sensors are able to pick up low frequency signals from physical objects which makes them ideal for heart rhythm monitoring applications. Also, low cost, small dimensions and insignificant weight allows for integrating in mobile telemedical systems [9]. In this work we used one of these stethoscope sensor modules. Simplified diagram of the sensor is shown in fig. 6. The sensor is optimized for picking up vibrations in the normal plane, at the same time providing inhibition for high-frequency vibration. It operates from voltages as low as 2.7 V and can give rail-to-rail output. Simplified block diagram of the electronic circuit is shown in fig. 7.

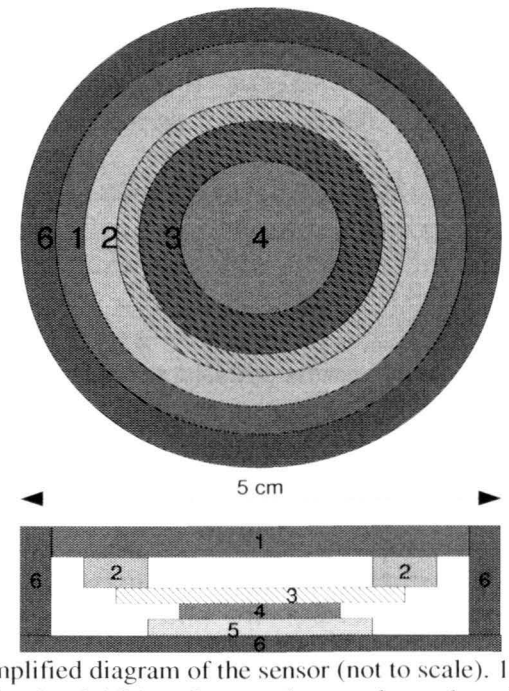


Figure 6. Simplified diagram of the sensor (not to scale). 1 – chassis. 2 – support and vibration inhibitor. 3 – metal sensor base plate. 4 – piezoelectric quartz film. 5 – rubber diagonal tension inhibitor. 6 – outer membrane. The size of the base plate for the quartz film does not exceed 2 cm.

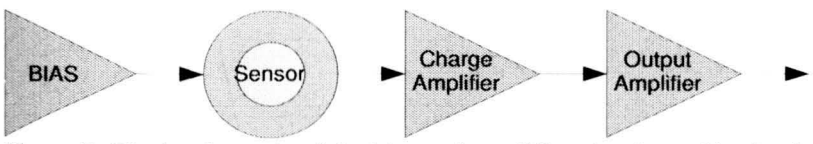


Figure 7. Block schematic of the bias and amplifier circuit used in the device.

VII. COMMUNICATION PROTOCOL

To be able to communicate with the smartphone application, the device utilizes a wireless Bluetooth interface. The device has a unique ID which is required to connect through the appropriate API methods in the smartphone communication layer. After connection, initial configuration is performed, including transfer speed, sampling speed and device status negotiation. This allows the device to enter standby mode, waiting for the order to start taking samples. After receiving the order, the device starts adding consequent signal samples to an internal queue and awaits the order to send the queue to the smartphone. In case of transmission error, it is possible to resend the last communication packet with the queue values. This allows for error correction. Samples are taken at regular intervals, and two sampling rates 10Hz and 100Hz are available. The sampling rate allows for accurate sampling of a very low-frequency heartbeat signal. The data packet consists of a starting string, a series of queue values separated by a predefined character, and an end character. The values are coded in a way that ensures proper interpretation of the data in case of any synchronization loss.

VIII. TESTING

The sensor is affixed to the skin using proper sticky bandage. The mounting point must be decided for every patient individually, due to differences in individual's anatomy. The proper position of the sensor on patient torso was determined by performing an examination with another stethoscope. The sensor was then attached to the area marked by a physician. Other methods, like using an elastic band around the torso, resulted in decreased signal quality due to noise from breathing and muscle movement, whose spectrograms interfered with the signal's. In the study we tested a group of patients, both healthy and with diagnosed heart conditions. The device and classic ECG tests were applied simultaneously. An example ECG signal in one of the cases and sensor output are shown in fig. 8. During the testing phase, the communication protocol proved to be reliable and covered all the required functions.

IX. CONCLUSION

Correlation tests between signal parameters (period, impulse length) proved, that the sensor output and ECG signal are correlated above 0,95. This indicates that the device could be used to monitor heart rhythm with satisfactory accuracy. During future development, the protocol and the smartphone application will be improved to implement multiple sensors and allow for simultaneous sampling.

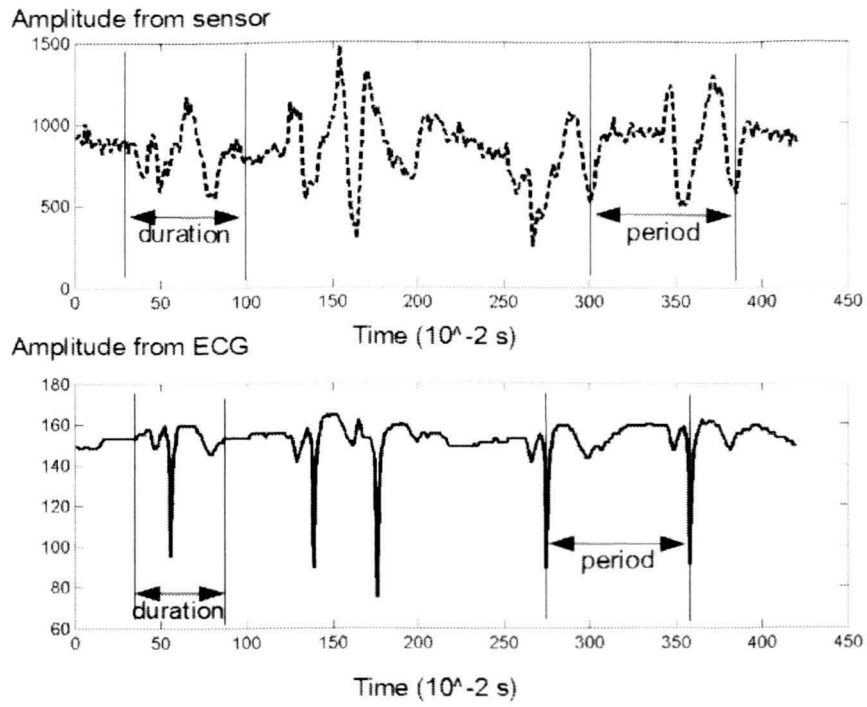


Figure 8. Example output from the the sensor (dotted line) and ECG (continuous line) in a case of multiple occurrences of supraventricular tachycardia.

REFERENCES

[1] Zhe Chen, Xiaomei Yu and David Feng, A Telemedicine System over the Internet, Biomedical and Information Technology Group, Basser Department of Computer Science, The University of Sydney, NSW 2006, Australia.

[2] California Health Care Safety Net Institute, Telemedicine Resource Guide, April 2009

[3] V. Thulasi Bai, and S. K. Srivatsa , Design and Simulation of Portable Telemedicine System for High Risk Cardiac Patients, World Academy of Science, Engineering and Technology 24 2006

[4] Scottish Intercollegiate Guidelines Network, Management of chronic heart failure, A national clinical guideline, February 2007.

[5] Tomasz Leyko, Zbigniew Ranachowski, The analysis of the heart acoustic emission – practical application in cardiologic diagnostics, Medycyna Rodzinna 1/2007, s. 10-12

[6] F. L. LEWIS Associate Director for Research Head, Advanced Controls, Sensors, and MEMS Group Automation and Robotics Research Institute The University of Texas at Arlington, Wireless Sensor Networks, Smart Environments: Technologies, Protocols, and Applications ed. D.J. Cook and S.K. Das, John Wiley, New York, 2004.

[7] Fei Hu, Member, IEEE, Yu Wang, Member, IEEE, and Hongyi Wu, Member, IEEE, Mobile Telemedicine Sensor Networks with Low-Energy Data Query and Network Lifetime Considerations IEEE TRANSACTIONS ON MOBILE COMPUTING, VOL. 5, NO. 4, APRIL 2006

[8] Measurments Specialties, Inc. "Piezo Film Sensors Technical Manual", Sensor Products Division, 950 Forge Avenue, Norristown

[9] M. Gruca, J. Grzełka, M. Pyrc, S. Szwaja and W. Tutak, „Miernictwo i systemy pomiarowe", Częstochowa University of Technology, Department of Mechanical Engineering and Information Technology, Częstochowa 2008.

Real-Time Bi-Directional People Counting Using Video Blob Analysis

Tomasz Marciniak, Adam Dąbrowski, Agata Chmielewska, Maciej Nowakowski
Poznań University of Technology, Chair of Control and Systems Engineering,
Division of Signal Processing and Electronic Systems, Poznań, Poland,
{tomasz.marciniak, adam.dabrowski, agata.chmielewska}@put.poznan.pl

ABSTRACT — **In this paper an analysis and implementation of the new algorithm for detecting and counting people in video surveillance systems is presented. First, we recapitulate methods used in the commercial systems, which can be used for people counting. Next, we concentrate on our new Matlab/Simulink model for detecting people, who are passing under a camera positioned vertically down at the entrance to a monitored area (e.g., a building). The prepared algorithm takes the direction of the transition of a person into account.**

KEYWORDS — CCTV system, video processing, people tracking and counting

I. INTRODUCTION

Modern video monitoring systems provide advanced image processing facilities including: analysis of human silhouettes, human movements, collecting information on persons entering and/or leaving a building, or testing and estimating the number of persons in a crowd [1, 2].

Commercial systems for people counting, described in Section 2, become more and more popular. They are mainly used to control movements of people in public buildings (e.g., shopping galleries), giving much useful information in real time such as e.g.: the sale success, the most popular pedestrian track, or the rush hour.

A real-time system presented in this paper monitors an entrance to a building and determines the number of people crossing the gate observed by a digital video camera. The respective video processing operations have been realized in the Matlab/Simulink environment, which was chosen because of its widespread use as well as flexible simulation and experimental testing possibilities. In addition, this environment contains a lot of libraries, which support image operations, e.g., the *Video and Image Processing* library [3] with e.g. *People Tracking* models, implemented and modified in our system. As it is shown in Fig.1, the designed and realized system detects people passing under a wireless IP camera positioned vertically down and counts the numbers of them with respect to the defined movement directions.

The paper presents important features of the prepared people counting model and an appropriate choice of the control parameters, which are necessary for the program to operate properly in the continuous mode. Problems of the proper camera placement, control of the registration area, and selection of the optimal perspective are also discussed.

II. ALGORITHMS AND SYSTEMS FOR PEOPLE COUNTING

A. *People count research and resulting algorithms*

In 2001 the authors of paper [4] proposed a real-time algorithm (over 30 fps rate) for pedestrian tracking and counting from grayscale video sequences and images. The main processing task was the blob tracking analysis and the Kalman filtering – used to estimate the pedestrian parameters. An output of the system are temporal and spatial coordinates of each pedestrian. In 2007 another approach to the blob analysis was proposed in [5]. A difference between the previous method in [4] and the new approach in [5] is that each blob in the image is trained to predict the number of the persons in the blob.

A bi-directional people counter for the pedestrian flow passing through a gate was proposed in 2006 in [6], basing on the area and the color analysis. Each person pattern can be recognized thanks to the analysis of its HSI (Hue Saturation Intensity) histogram. A result obtained with the quantized histograms of intensity (or hue) is compared with the result of the preliminary counting. Thanks to the people-touching pattern analysis the authors solved a problem of objects, which are very close to each other.

A bi-directional projection of the histogram was published in 2008 in paper [7]. The authors used the grayscale histogram of the two-frame difference image.

In 2008 the authors of article [6] proposed another method for people counting using the flow analysis [8]. A frame is divided into many blocks, then each block is classified according to its motion vector and if there is a similar movement then the respective blocks are regarded as belonging to the same moving object.

In 2011 the authors of paper [9] formulated and solved a problem of people tracking using a potentially moving and even an uncalibrated camera. A novel method, namely the racking-by-detection in a particle filtering framework, was used. This algorithm does not rely on the background modeling.

This paper was prepared within the INDECT project and partly with the DS funds for the year 2012.

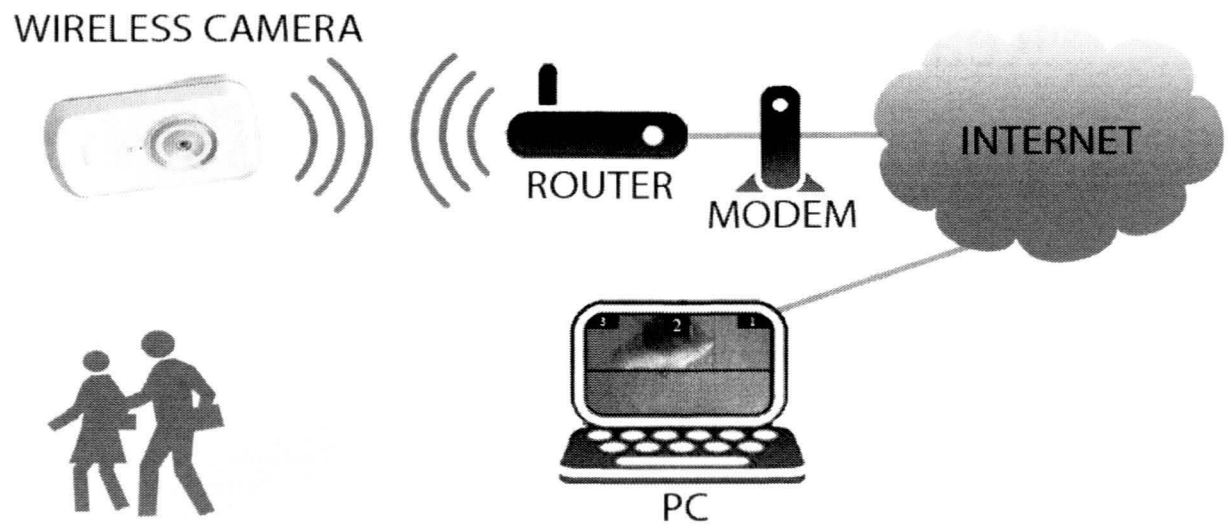


Figure 1. System for counting people with wireless transmission

In our approach we have modified the existing algorithms in order to count people entering a building, a room, etc. with merely the data from cameras placed directly above the entrance.

B. *Commercial systems*

Several companies just start to begin to offer people counting with CCTV systems. These are, e.g., the following companies: CountWise, Honeywell, Video Turnstile, and Visual Tools.

People counting system from Honeywell, is a part of the *Intelligent Video Analytics*. It is a system that can operate independently or can be integrated with other technologies such as the standard CCTV. Its output data is information about the number of people who came into and out of the room and also a check of the movement track of the humans within the field of view of the camera [10]. The Honeywell system can operate up to 40 channels, while archiving the data from all devices up to 60 days. Proper operating of the video channels is dependent on the method of recording the image. It must be at least in the QCIF (*Quarter Common Intermediate Format*), i.e., a resolution of 176 × 144.

Another DVR (*Digital Video Recorder*) with the people count function, who enter and leave a building is the VX-PECO from Visual Tools. An accuracy of counting is high, thanks to the analysis of the full-resolution image of up to 25 frames per second. The optimal placement of the camera is above an entrance to the room. The manufacturer of this system provides free software for collecting data from several PECO recorders, as well as for the later analysis, statistics, reports, and data export to the "*.csv" (*Comma Separated Values*) format. The software is available on the company website [11]. This system can effectively distinguish between silhouettes of people and ordinary objects such as e.g. suitcases, trolleys, and bags. The system works with comparing the detected object to the reference object and as one person only objects with up to 50% smaller or larger sizes are treated. Thanks to this solution, even two people who go very close to each other will be treated not erroneously as one but correctly as two persons [12].

The next quite well-known approach to count people and control their densities is based on applications offered by CountWise [13]. This is a set of products designed for various specific tasks. These include: a system to precisely examine a number of people in the crowd (I-Count), an application that monitors behavior of customers in stores and shopping centers (Z-Count); a queue management at cash registers in the stores with the use of RTC (*Real Time Control*) (Q-Count).

Another commercial system is the Video Turnstile, which is a very advanced and accurate tool for counting people. The Video Turnstile system gives a possibility of installing the counter, which displays the current number of people in the building. The counter can be connected to several CCTV cameras at all entrances to the building in order to continuously offer the current results [14].

III. AUTOMATIC COUNTING OF PEOPLE ENTERING THE BUILDING

Our system for counting people entering a room or a building is based on data from a camera placed directly above the entrance. The camera must be positioned vertically down and perpendicular to the entrance, so that the door is located at the top or the bottom of the video frame. The size of the reference object should be defined, because it is compared to the detected objects. The line placed on the floor should be pointed. Crossing this line will be the signal for counting and determining the direction of the movement of the object. We have assumed typical parameters: an average human height of approximately 175 cm and entrance to the building with a width equal to 2.5 m, the minimum height at which the camera should be mounted of 3,9 m [12]. The greater the height of the camera, the greater the accuracy of the algorithm and the less the impact resulting from the difference in the growth of people. Fig. 2 illustrates the proper placement of the camera above the entrance into the room and the field of its view.

As already mentioned, our solution of the people counting system is a modified *People tracking* model from the Matlab/Simulink environment. A block diagram explaining in detail our modified and highly efficient model is shown in Fig. 3.

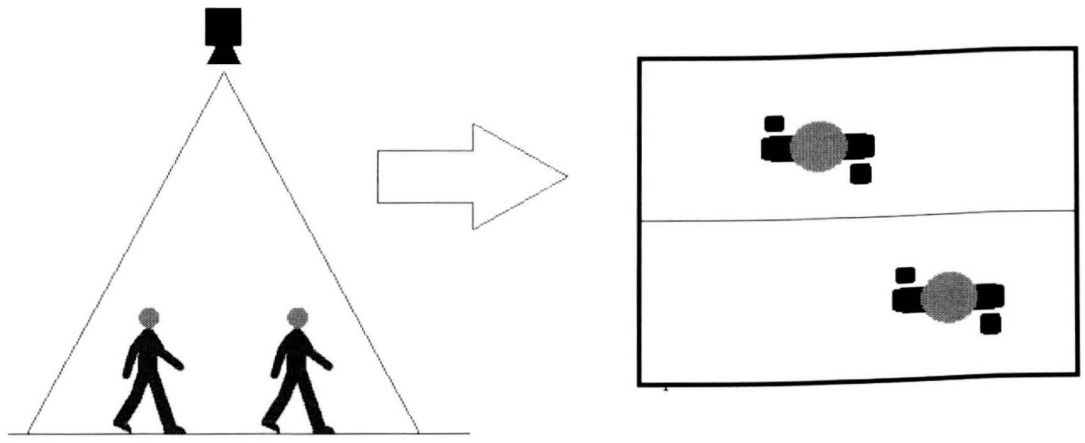


Figure 2. The optimal placement of camera above the entrance into a room and the field of view of the camera

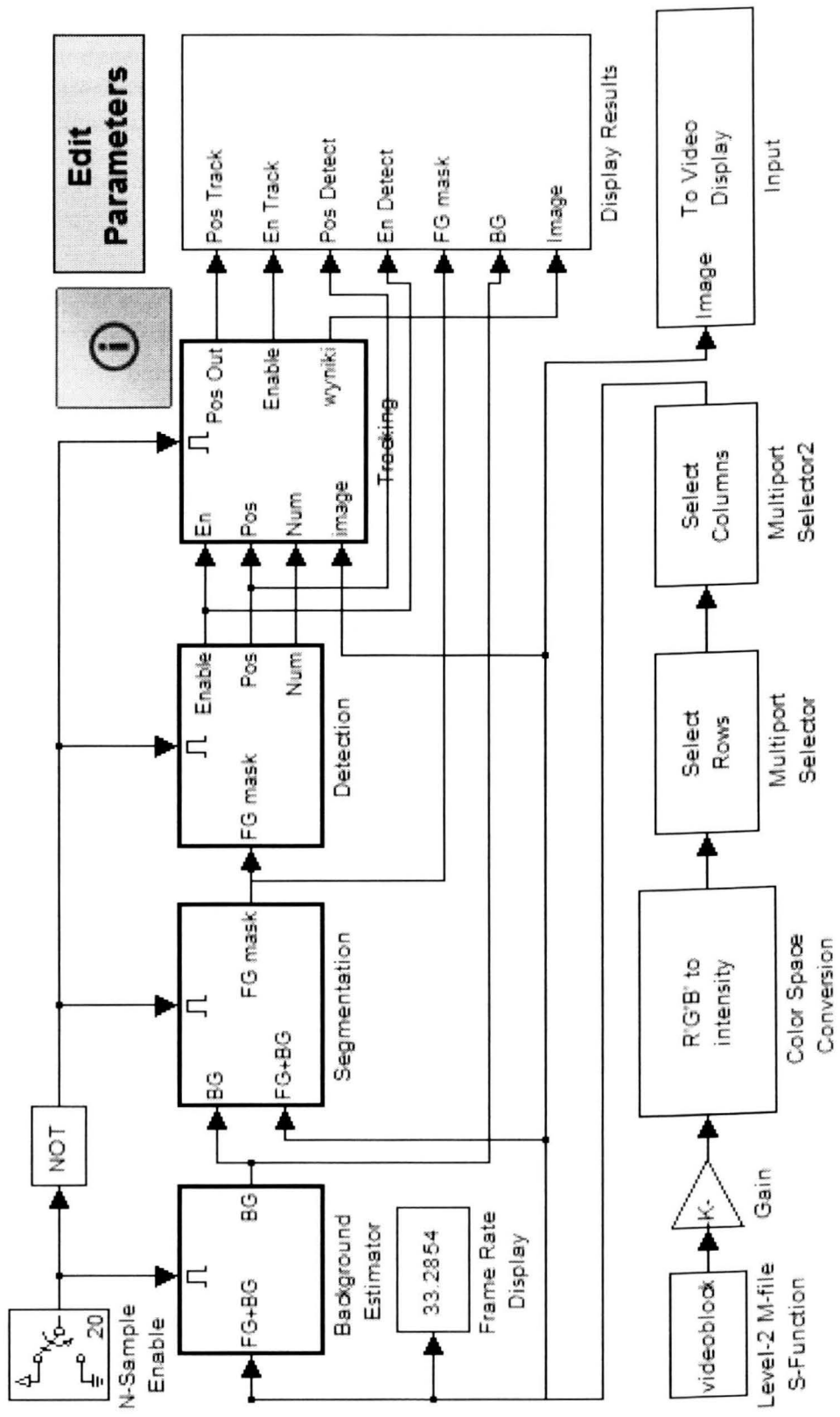


Figure 3. Block diagram of a new algorithm for counting people entering a building (in the MATLAB/Simulink environment)

Experiments of estimation of the background took place by using the continuous mode background estimation based on the first 20 samples and estimation outside the area occupied by the object (Fig. 5). The proposed method consists in accomplishing the *Compositing* block, which connects the input image, the background, and the *Detection* block. Then the result of the *Compositing* block is transported to the *Switch*, which selects and passes through the first 20 samples of the signal from the source, then passes though the output of the *Compositing* block, which is delayed by one sample. This means that a human detected during the conversion of the background values eliminates acquisition of the object to the background representation. This solution is very effective and only marginally slows down the whole system.

The model shown in Fig. 3 consists of the following steps:

- Acquisition of a live video stream with resolution of 320 × 240 pixels.
- Possible reduction of the image resolution from 320 × 240 pixels down to 160 × 120 pixels according to the computer performance.
- Conversion of color of the video frames from the RGB (*Red Green Blue*) to grayscale.
- Dividing the image into two parts: the upper and the lower, depending on the proper camera position (Fig. 2), we can tell how many people came into the room and how many of them came out.
- Object segmentation based on auto-threshold.
- Object recognition with the BLOB (*Binary Large OBject*) analysis. It is an analysis of groups of pixels defined as a large binary object, which allows for detection of a person as a whole and not separately, for example, the hands. Due to this analysis, the precision of the system has increased. The outputs of this algorithm are: the "Enable" signal that indicates whether the object was detected or not and the *Matrix of Positions* of these objects. This matrix consists of four rows. The number of columns equals the number of objects detected in one frame. Working with 160 × 120 resolution assumes that up to 8 objects can be fitted on a picture. Therefore the *Matrix of Positions* will not be larger than 4 × 8. Each of the four rows corresponds to different values assigned to the object: the first two rows are coordinates of the upper-left corner of the BBox (*Bounding Box*) on the picture, which surrounds the object. The next two rows are, respectively, the height and width of the BBox.
- Determining whether an object is located at the top or at the bottom of the image (the first row in the *Matrix of Positions).*
- Designation of the direction of moving objects. The most important in these calculations is the condition: if the object was first detected at the top, and then at the bottom of the image – it means that the object was moving from the top to the bottom – otherwise, the movement of the object was in the opposite direction.

IV. EXPERIMENTAL RESULTS

A. Cameras

Quality of the image is an important element influencing the efficiency of counting. Automatic change of camera parameters (as white balance adjustment and motion detection) can adversely affect the operation of the program. Therefore, the experimental studies used a camera that allows the disabling of that automation. Initially, we used an inexpensive webcamera Modecom VENUS [14], which additionally has four LED diodes. Convenient is the possibility of wireless transmission of the image. Thus, we finally used IP D-Link DSC-930L camera [16].

B. Efficacy analysis of distinguish objects

The input video sequence has a resolution of 320 × 240 pixels, while the processed image resolution is 160 × 120 pixels. A continuous background estimation can cause too large delays in the algorithm, thus there is a possibility of the choice between the continuous background estimation and the background estimation based on the first 20 samples.

Algorithm tests were made at a resolution of 160 × 120 pixels, on the floor surface of 320 by 240 cm. Modecom camera (type of sensor: ¼" CMOS) was placed at a height of 390 cm and directed vertically down. Using this data, we calculated that a one pixel is in reality a square with sides of 2 by 2 cm. Taking into account these consideration and the fact that the average person has a width of about 60 cm and a stride length of 80 cm, we have concluded that a single person in motion in the video frame has a surface of ca. 1200 pixels.

To reduce the possibility of an error in the acceptance of pixel groups as a single compact blob, in the Detection/ BlobAnalysis block we specify the minimum and the maximum blob area in pixels: 800 and 2000, respectively.

A real distance of two objects, at which the algorithm connects the blob, has also been found. At that distance two variables are affected. The first variable is the structuring element from the morphological closing operation on the binary image, which merges pieces into blobs. To do this the "strel function" has been used (strel('rectangle',[5,1])), which creates neighborhood of the flat, rectangle-shaped structuring element. The second variable "the box merging threshold" approves as a vector [5 0]. This means that objects separated by more than 5 pixels will be treated as different objects. In reality, with the above conditions, this means a distance of more than 10 cm.

To avoid detection of two objects as one, we have included an additional criterion. When two objects are connected within a blob, we examine the local maximum of the vertical binary image histogram within a specified range [17]. Thanks to this operation we can distinguish objects, which are very close to each other (Fig. 6).

Others parameters of the program have been set to the following values: segmentation of objects to 1.2 points, a value of the target tracking threshold to 100 points.

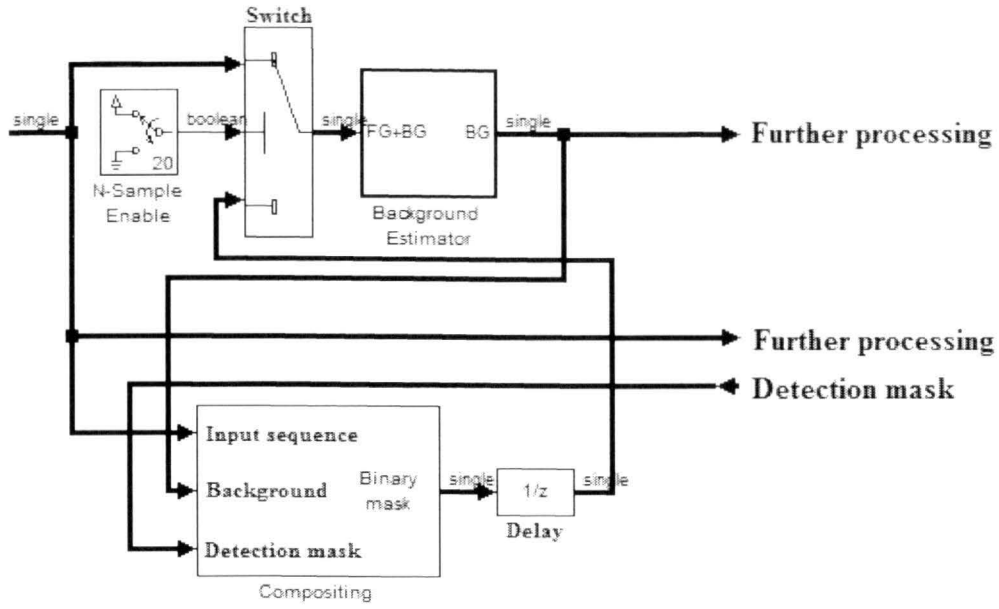


Figure 4. Estimation of the background outside the area occupied by the object

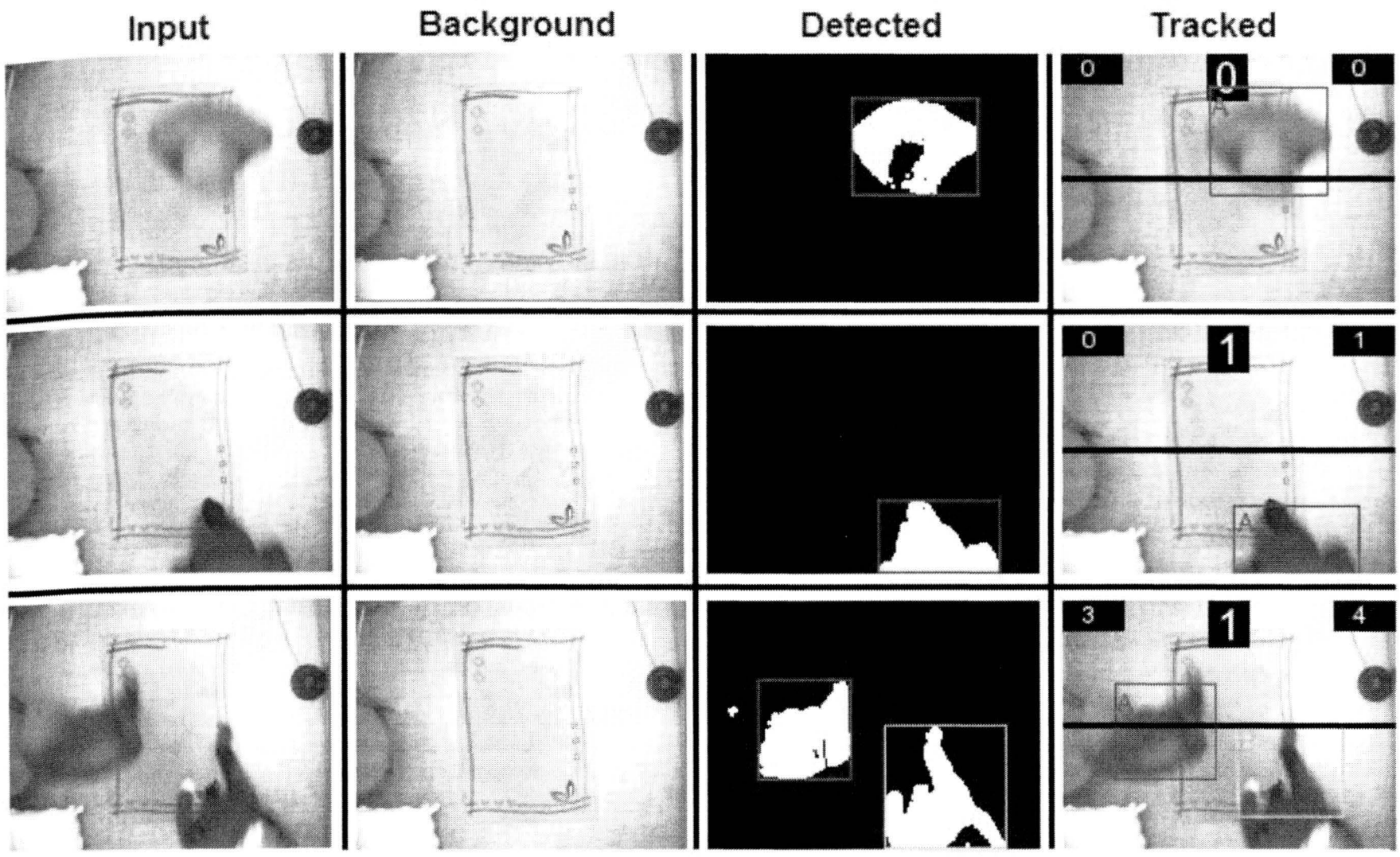


Figure 5. Results of the program for counting people entering the building

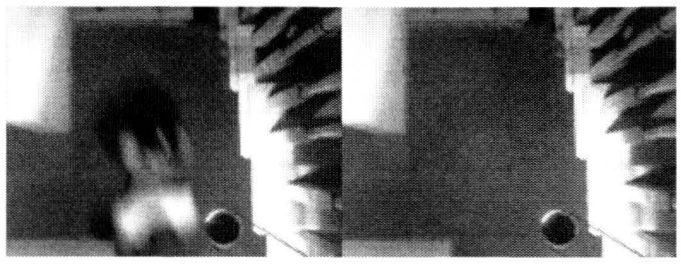
C. *Counters*

Presentation of results of the counting people model takes place through displays in the upper left corner of the image for the number of people that passed from the bottom to the top while the right corner of the image displays the number of people who passed the monitored area in the opposite direction. In the middle of the image (at the top), the algorithm displays the number of people who are currently in the room. The results of the program for counting people which entered a building are shown in Fig. 5. The first column shows the current image, the second column models the background, the third column indicates the detected and tracked objects, the fourth column shows the final results of the counters.

D. *Tests*

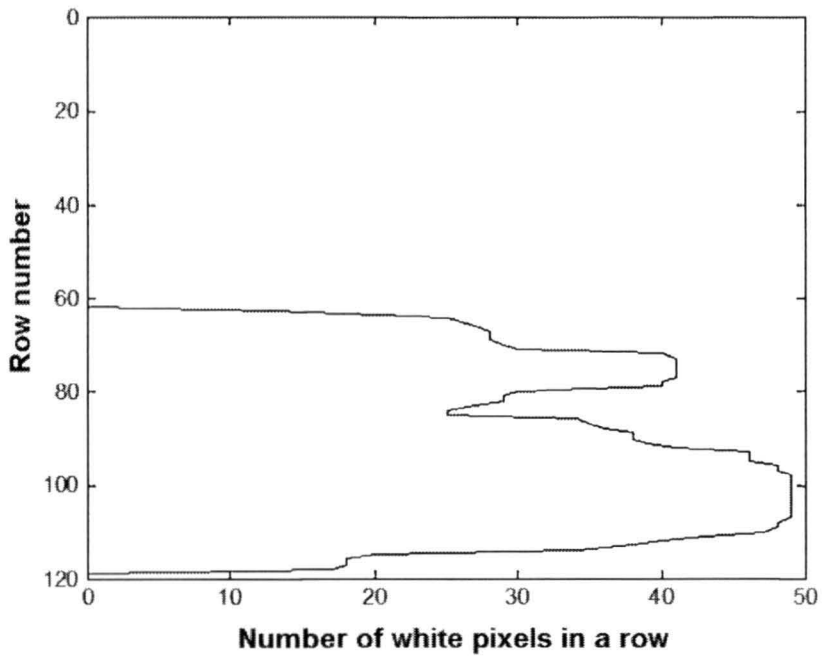
In our tests we examined a 30 minute long video sequence, which contains 175 people entering the building and the same number of leaving the building (or a room). The number of people who are in the room at the end of the algorithm should be 0, then the effectiveness of the algorithm is 100%. Our tests

have shown that the effectiveness is slightly lower equal to 94%. The results of operations on a 30 minute sequence of the algorithm indicated that the 186 people entered the room and 175 came out.



(a) (b)

(c)



(d)

Figure 6. Example of distinguish objects, which are close to each other: input image (a), background (b), binary image after morphological operations with marked maxima (c), number of white pixels in particular raws (d).

V. CONCLUSIONS

The proposed system for counting people undoubtedly has two important advantages: first, it is easy for installation, second, it is of relatively low cost compared to solutions using light curtains or commercial systems with video cameras. Our solution allows bi-directional counting of people entering and leaving a building and is quite accurate.

The solution with the location of a camera above the observed people eliminates a currently often discussed problem of collecting the personal information. Thanks that, we are not dealing with a problem of data acquisition, which is questionable as it could be used to identify people.

Using a PC computer with an i7 processor we can operate with the high processing speed (30 FPS) and a relatively high resolution of 320×240 pixels. This allows fast detection of quickly moving objects and objects in short distance to each other. In constant light conditions, counting accuracy is at the level of commercial systems i.e. ca. or even above 94%.

The proposed system requires further research. First of all, the algorithm needs to be automatically adapted to different lighting conditions. The next step should be a realization of the embedded system, which will dispense the current need for the transmission of the whole video signal from the camera to the PC.

REFERENCES

[1] C. S. Regazzoni, A. Cavallaro, Y. Wu; J. Konrad and A. Hampapur, "Video Analytics for Surveillance: Theory and Practice" [From the Guest Editors] IEEE Signal Processing Magazine, vol. 27, pp. 16-17, 2010.

[2] Proceedings of Eleventh IEEE International Workshop on Performance Evaluation of Tracking and Surveillance, PETS 2009.

[3] Description of the Matlab environment: http://www.mathworks.com/

[4] O. Masoud and N. P. Papanikolopoulos, "A Novel Method for Tracking and Counting Pedestrians in Real-Time Using a Single Camera", IEEE Transactions on Vehicular Technology, vol. 50, No. 5, pp. 1267-1278, 2001.

[5] Weizhong Ye and Zhi Zhong, "Robust People Counting in Crowded Environment", Proceedings of the IEEE International Conference on Robotics and Biomimetrics, China, pp. 1133-1137, 2007.

[6] Thou-Ho Chen, Tsong-Yi Chen and Zhi-Xian Chen, "An Intelligent People-Flow Counting Method for Passing Through a Gate", 2006 IEEE Conference on Robotics, Automation and Mechatronics, Bangkok, pp. 1-6, 2006.

[7] Honglian Ma, Hunchuan Lu and Mingxiu Zhang, "A Real-time Effective System for Tracking Passing People Using a Single Camera", Proceedings of the 7th World Congress on Intelligent Control and Automation, China, pp. 6173-6177, 2008.

[8] Thou-Ho Chen et. al. "People Counting System for Getting In/Out of a Bus Based on Video Processing", Eight International Conference on Intelligent Systems Design and Applications, Kaohsiung, pp. 565-569, 2008.

[9] M. D. Breitenstein et. al., "Online Multiperson Tracking-by-Detection from a Single, Uncalibrated Camera", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 33, No. 9, pp. 1820-1833, 2011.

[10] Honeywell International Inc. L/PEPLECTD/D, February 2010 http://www.honeywellvideo.com/products/ias/va/160978.html http://www.honeywellvideo.com/support/configurators/lens/index.html

[11] Visual Tools software to collect data from the PECO recorders: http://www.visual-tools.com/en/products

[12] Technical documentation - PeCo-POS_PL: „Wizyjne systemy zliczania osób i rejestracji transakcji kasowych (Vision systems for counting and recording of cash transactions)", 2009

[13] CountWise system: http://www.countwise.com/index.php;

Technical documentation I-Count "Accurate & Reliable People Counting", 2007; Technical documentation Q-Count "Queue Management – Improve Customer Experience", 2007; Technical documentation Z-Count "Leverage Customer Behavior Patterns", 2007.

[14] VideoTurnstile- device for counting people in the building: http://www.videoturnstile.com

[15] Modecom VENUS webcam http://www.modecom.eu/venus/webcams/peripheral/product/

[16] Description of D-Link camera: http://mydlink.dlink.com/products/DCS-930L

[17] Albiol, A.; Mora, I.; Naranjo, V.; , "Real-time high density people counter using morphological tools," Intelligent Transportation Systems, IEEE Transactions on , vol.2, no.4, pp.204-218, Dec 2001.

Determination of Blood Flow Parameters in a Cylindrical Vessel

Grzegorz Dwojakowski, Artur Klepaczko, Andrzej Materka
Institute of Electronics
Lodz University of Technology
Lodz, Poland
grzegorz.dwojakowski@dokt.p.lodz.pl

ABSTRACT — **Data from Magnetic Resonance Angiography carries information about the structure of blood vessels in the human brain. Thanks to proper segmentation methods, this information can be used in diagnosis. However, the correct assessment of these algorithms is troublesome. We have proposed a new concept of validating segmentation results. The idea is to create an MRA Simulator. The output images can then be compared with the pattern, which would not be possible using real data. This paper explains how various physical phenomena were modeled to simulate blood flow. The proper assortment of these parameters will be used to obtain images similar to those that come from the Time of Flight method.**

KEYWORDS *— Magnetic Resonance Imaging, image segmentation, blood vessel network, blood flow modeling*

I. INTRODUCTION

Magnetic Resonance Imaging (MRI) is one of radiology techniques used to visualize the internal structure of the human body [1]. Modern MRI scanners can receive high resolution 3D images with good contrast between different tissues. In comparison to traditional X-rays or Computer Tomography, MRI is noninvasive because it does not use ionizing radiation. Another advantage is the number of different sequences. Using angiographic techniques such as Time of Flight [2] and Susceptibility Weighted Imaging [3] methods combined together, results in a full map of veins and arteries [4]. Such a map carries important information about a patient's health and can be used in diagnosis and planning surgical operations.

Anomalies, such as clots at latter stage of diseases or neoplasmic diseases, are clearly visible and can be detected by radiologists without difficulty. But is it possible to spot narrowings in small vessels and clots in early stages of diseases? Because of the high complexity of the vessel networks there is a significant risk of omitting those areas.

In these situations image segmentation and visualization methods can be useful [5][6]. Separating vessels from other tissues allows to show arteries and veins as a 3D model (Fig. 1). Data in this form is much easier to be analyzed than 2D cross-sections [7]. Additionally segmentation allows to automatically search for risk regions based on vessel diameter. As shown above, benefits of segmentation are significant.

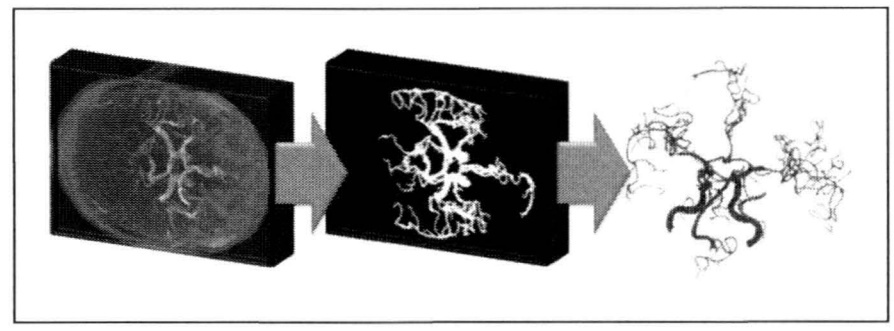


Figure 1. Results of brain vessels segmentation and visualization.

II. VALIDATION METHODS

Every image processing method, before used in a hospital, must be validated first [8][9]. It is necessary to give medical doctors a reliable tool to make correct diagnoses. How many vessels were detected? Were diameters and shapes reconstructed correctly? Were artifacts resulting from the imaging acquisition technique minimized? In order to validate the segmentation algorithm it is necessary to answer these questions.

Usually results are compared with the pattern. This time the pattern is the brain vessel network. It is not possible to properly measure such a complex structure inside a human skull. In this case, the validation process is different.

The easiest method of validation is based on a medical knowledge and subjective assessment made by doctors [10]. However, gathered information cannot be used as a reliable pattern due to uncertainty and poor reproducibility.

Second group of validation is based on physical phantoms. These artificial structures are mainly used to calibrate MRI scanners, but can also be used to test results of segmentation methods. This time, the pattern is familiar and comparison can be made. This method of validation is much more accurate, but has its drawbacks. Physical phantoms are expensive and, so far, there is no structure similar to the real blood network. This is a result of complicated topology and small vessel diameters.

The third group consists of digital phantoms [11]. Similarly to physical structures they are a familiar pattern. A complex geometrical shape is much easier to create using computer 3D graphics. What is more, they are much faster and cheaper to create. They can be duplicated with different parameters and can be used to create large set of test objects. The only disadvantage is that it is not possible to use these phantoms in a real MRI scanner.

To solve this problem, and make digital phantoms usable, we must create an MRI simulator. Using this program we should be able to obtain images similar to this from a real scanner (with noise and distortion caused by imaging sequence). A properly implemented simulator working on digital phantoms will create an effective and objective criterion of validation for image segmentation methods.

III. MRI SIMULATOR

Simulating the physical phenomena of an MRI scanner is a complex problem. Only a few attempts to this issue can be found in the literature [12][13][14][15], but they work only for invariable objects. In case of angiographic imaging, these methods cannot be used because sequences such as Time of Flight (ToF) and Susceptibility Weighted Imaging (SWI) are based on blood flow in arteries and veins. In ToF, image contrast is acquired by unsaturated molecules of blood which flows through acquisition volume in a given time moment. SWI depends on blood oxygenation and uses amplitude and phase gradient echo, with compensated blood flow effect.

As it was shown above, in order to simulate MRA sequences, there is a need to determinate fluid flow parameters in an artificial blood vessel network.

IV. MODEL SETUP:

Appropriate simulation of blood flow is a crucial issue for the whole project. Only the right model and its positive results will provide the possibility of using MRI simulator as a reliable validation tool. To determine the flow parameters we use COMSOL Multiphysics environment [16]. Results will be validated using real phantoms. After the comparing process, parameters will be ready to implement in the system.

A. Geometry

As mentioned earlier, model geometry should be identical to the physical phantom in order to perform validation. This phantom should have a familiar geometry accuracy and fluid flow must be possible in it. From among several models, we chose the Flow Phantom Set (Fig. 2) produced by Shelley Medical Imaging Technologies [17]. This Model is compatible with the high class CompuFlow 1000 MR pump [18].

This Phantom Set consists of 4 straight and 1 U-bend tubes. Diameters are between 5-8 millimeters. There is no fluid leek through the walls and no deformation caused by flow. This simplification makes the model easier to implement.

The 3D model of a simple cylinder can be constructed with COMSOL drawing Tools. To create more complex shapes including sinusoidal stenosis and bifurcations it is necessary to create geometry in the outside program. Models in this article were constructed using the Visualization Toolkit for C++ [19]. We also tested geometry created in Google Scatchup [20]. In each case, models were interpreted correctly by COMSOL.

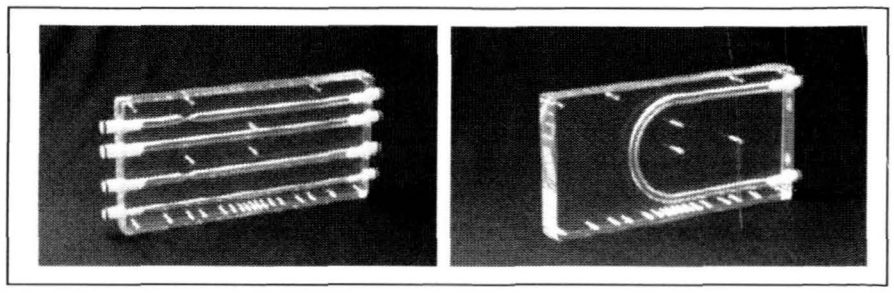


Figure 2. Phisical phantoms made of sylicon [17[. Left: Stright cylinders with stenosis. Right: U-bend tube.

B. Fluid

After creating geometry, vessels must be filled with blood. Two main parameters describing this liquid are viscosity and density. Based on literature these values were set to 1060 Kg/m^3 (density) and 0.005 Ns/m^2 (dynamic viscosity) [21]. These two parameters are sufficient to model the flow.

Blood transport in vessels is generally modeled using laminar flow equations [22]. This model assumes that fluid flows in parallel layers. Each layer has its own speed and slides past one another so there is no lateral mixing. For simulating blood flow in a pipe with ideal circular cross-section, this model is sufficient. Blood flows in one direction. Velocity is greatest in the middle of the cylinder. The value decreases as we approach the vessel wall.

Real vessels are not ideal tubes. Diameter is not a constant value. It can change gradually or rapidly in stenosis. Bifurcations are another obstacle for laminar flow; when one cylinder is divided into 2 smaller ones with different directions. In those type of situations fluid is no longer flowing in layers and turbulence appears. However, the vast majority of attempts of blood flow simulations in vessels relay on a laminar model. The main reason for this is the complexity of turbulent flow phenomenon. In this paper we modeled flow in vessel bifurcation using the laminar and turbulent model. Our goal is to compare these two results and decide if the idea of simplifying calculations is justifiable.

COMSOL Multiphysics gives possibility to simulate both types of flow. Laminar flow is generated by solving incompressible Navier-Stokes equations [23]. To simulate turbulences one of three Reynolds-averaged Navier-Stokes (RANS) model can be used [24].

A very interesting package was added to COMSOL since the 4.2a version. The name of this module is Particle Tracing [25]. Its main feature is an estimation trajectory of a chosen molecular. We have been looking forward to it since it was announced. The reason to this was the fact that information about direction and speed of blood particles is necessary to

simulate such sequences as ToF or SWI. Now the version 4.2a has been released, it can be tested and validated.

V. Results

To test the described modules, we used 3 types of digital phantoms (Fig. 3). Straight cylinder, tube with stenosis and simple bifurcation model. First two were created in the likeness of the synthetic model. After positive test results, both phantoms can be compared. The third model is a brunch which divides into two smaller tubes according to the rule of bifurcation.

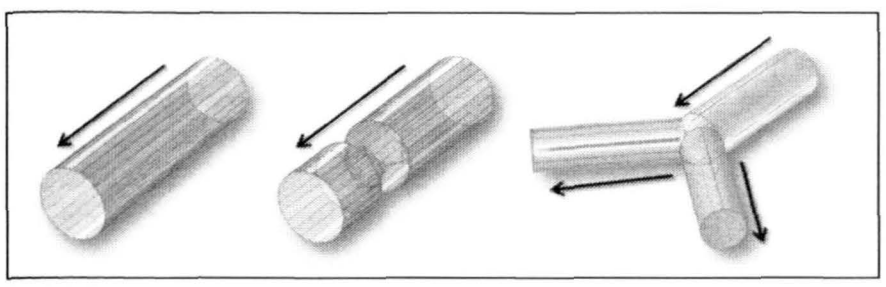


Figure 3. Three digital models. From left: Stright tube, tube with stenosis, bifurcation. Arrows indicate the direction of blood flow.

A. Straight tube

Straight tube is a perfect model to simulate Laminar flow. The geometrical shape of all orthogonal cross-sections is an ideal circle. The diameter is set to 8 mm and the length of a tube is equal to 100mm. Fluid flow is forced by setting pressure difference between input and output boundaries. There is no slip allowed through phantom walls. This assessment was made for three reasons:

- In vessels with 8mm diameter blood transfer through the wall is negligible.
- Synthetic phantom has no slip.
- It simplifies calculations.

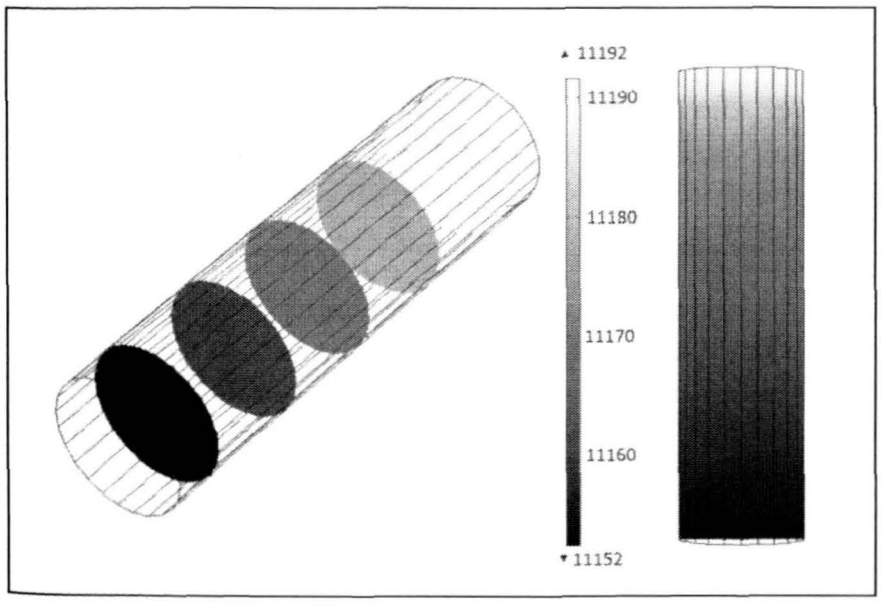


Figure 4. Preasure distribution in straight tube. Preasure range is 11.208-11.148 Pa

The results of simulation can be visible on Figure 4 and 5. According to the theory of laminar flow, highest velocity is obtained in the center of the tube (bright color). Moving towards the wall this value decreases.

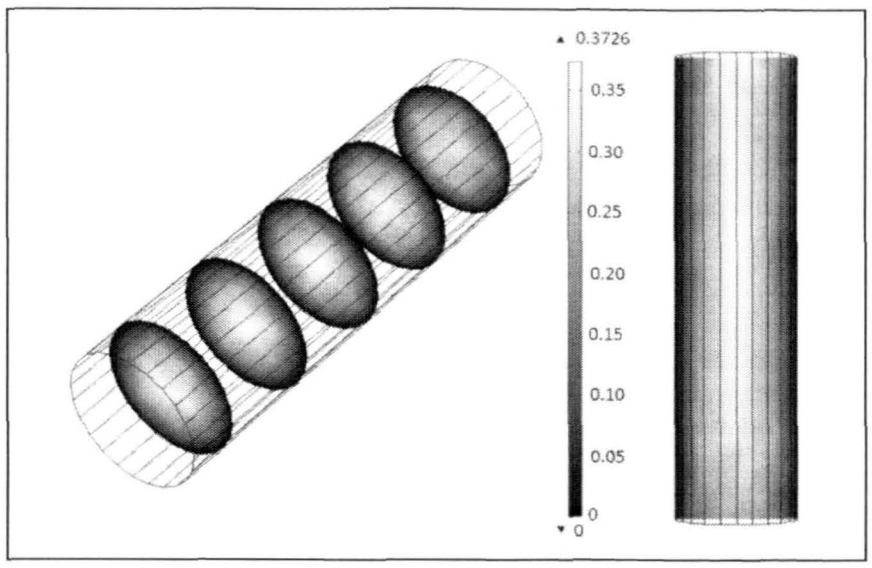


Figure 5. Velocity magnitude in straight tube. Value range is 0.0-0.37 m/s

Based on the laminar flow solution, particle tracing was performed. The Number of molecules was set to 128. All of them were defined by density and diameter. At the beginning, all particles are located at input boundary. With time, each element moves towards the exit (Fig. 7, top 3 pictures). All trajectories are straight lines parallel to the main axis of the vessel. The fastest particles reach its goal in less than one second; ones next to the wall are over five times slower. Results from this study are relevant to values obtained in laminar flow simulation. What is very important, besides getting visually attractive animations, the user is able to save 3d coordinates of all particles in chosen time steps as a spreadsheet file. There is also a possibility to analyze and process this data using output programs.

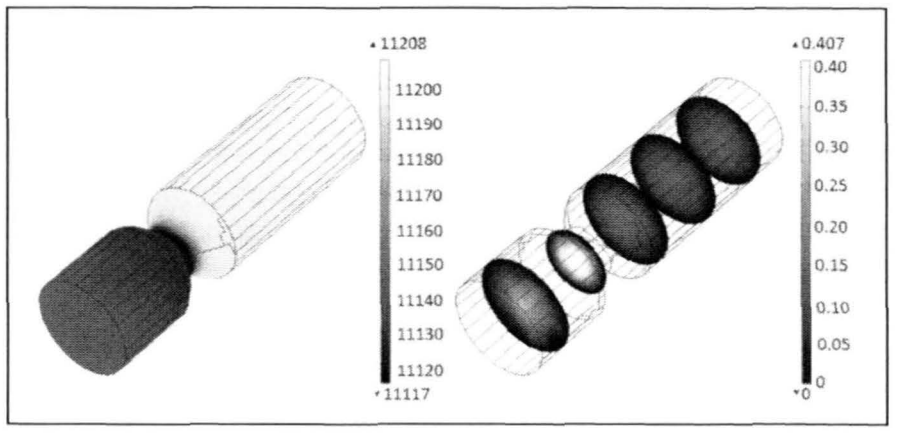


Figure 6. Tube with stenosis. Left: Preasure distribution (range is 11.208-11.148 Pa). Right: Velocity magnitude 0.0-0.43 m/s

B. Straight tube with stenosis

Another phantom created based on the synthetic model is similar 8mm diameter tube but with 50% sinusoidal stenosis by diameter. Parameters of the flow are the same as at the first test. Due to this narrowing, pressure distribution is not so linear as in the first test (Fig. 6). This time, pressure force applied on the particle before reaching stenosis is almost constant and it equals the input parameter. The pressure changes rapidly and reaches output value. In these conditions, velocity values are also different. For the first straight tube the

fastest particles have constant speed over 0.35 m/s. They reach the goal in less than 0.1 s. For cylinder with sinusoidal narrowing, the velocity value changes between the input and the output. Through the first half of the tube particles in the middle will not reach 0.20 m/s. In the region of a stenosis they double their speed for a while to slow down once again. Time of the flow is almost twice as long. After the narrowing, the density of particles is higher because they do not return to their previous positions (Fig. 7).

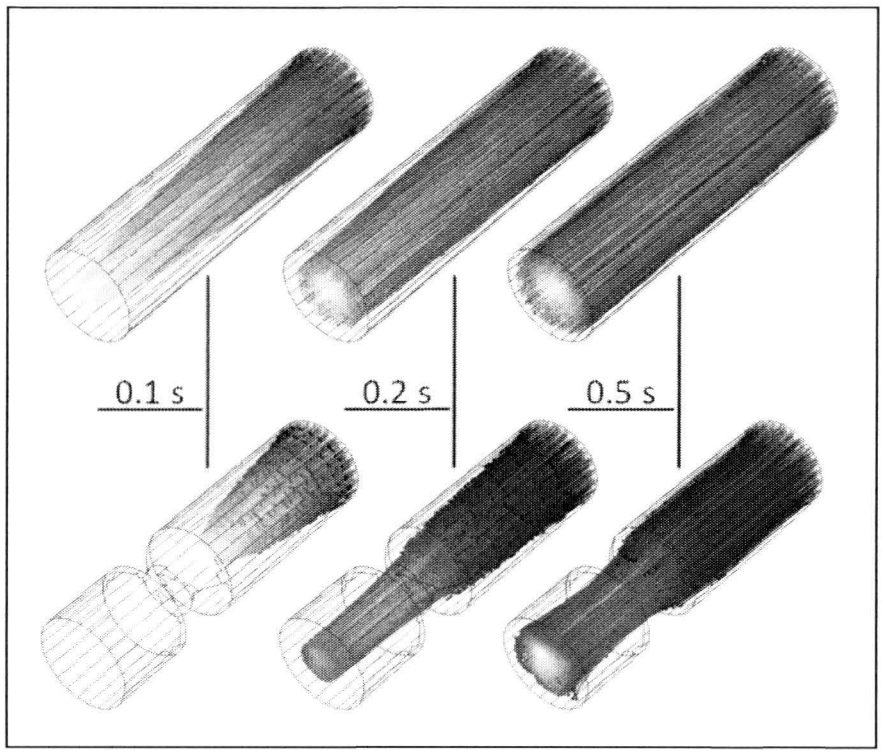


Figure 7. Particle tracing for two phanthoms with the same length and diameter.

C. *Bifurcation model*

The third phantom consists of 3 tubes: the one with the largest diameter is called the ancestor branch, the other two – descendent branches with smaller radius based on bifurcation rule. The center point of a base is common for all tubes. the deviation angle for both descendent branches is identical. In this way a simple bifurcation model was created with one input and two outputs. For this object two types of flow were implemented.

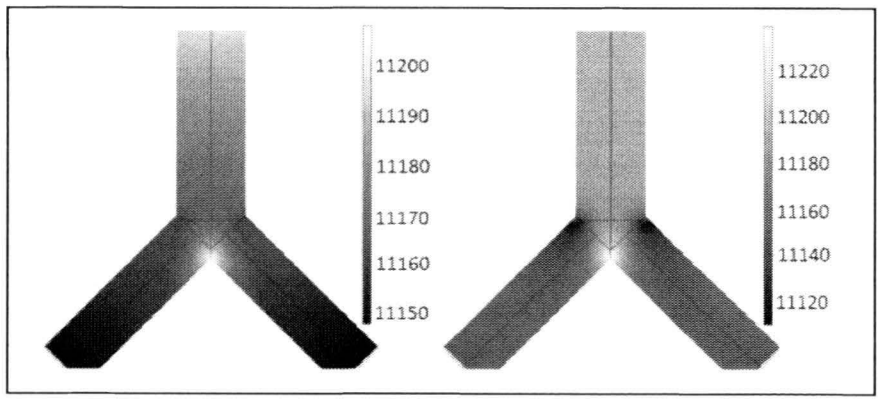


Figure 8. Preasure distribution in bifurcation model. Left: Laminar flow (range 11.208-11.148 Pa). Right: Turbulent flow (range 11.239-11.111 Pa).

For laminar flow, pressure distribution is linear (Fig. 8, left picture), similarly to the first phantom. the Velocity magnitude values are also consistent with the theory. In the ancestor branch, the direction of the velocity field is in line with the y axis. After ramification, this direction consists of two component vectors(x and y). These components have the same absolute values for both descendent brunches (Fig. 9, left picture).

Pressure and velocity values for turbulent flow are different from those obtained using the laminar model. Pressure distribution changes rapidly in the center of bifurcation. Velocity values are no longer dependant on the distance from the center of a container. In the descendent branches, blood flows according to the law of inertia, along centre walls.

For the both models particle tracking analysis was performed (Fig. 10). Trajectories from laminar flow in the ancestor branch are similar to that in straight tubes but due to different pressure at the end they move slightly towards the central axis. After bifurcation, particles travels in both tubes in the middle of the vessel. This flow is no longer laminar; velocity direction is different for every particle.

Trajectories obtained from the turbulent model of flow can be seen at figure 10. At first, all molecules have the same speed and direction. After dividing them into two groups particles move in a disorderly way.

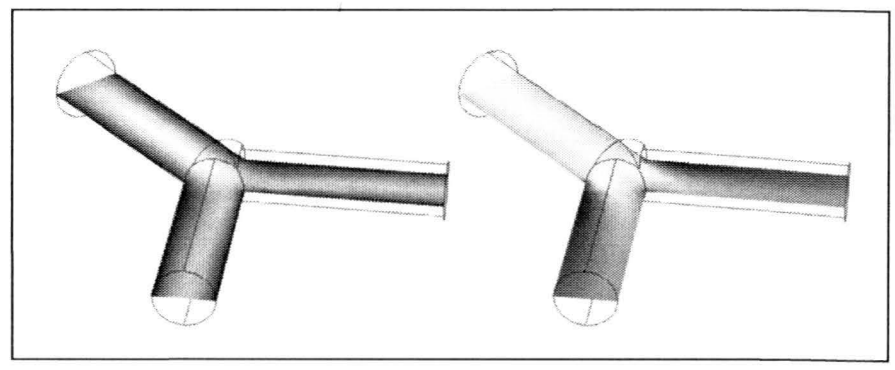


Figure 9. Velocity magnitude in bifurcation model. Left: Laminar flow (range 0.0-0.32 m/s). Right: Turbulent flow (range 0.0-0.44 m/s).

VI. PROBLEMS

Results are satisfactory and the obtained data is useful, however, there are a few problems. In our case we want to trace all particles from the beginning to the end with short timeframes (39 us). For phantoms with 10 cm length or more, it takes almost 5 seconds for all molecules to reach output. If we divide that value by timeframes, we obtain over 25.000. Each moment consists of x,y,z coordinates for all 128 particles. In result, huge amount of data is generated which is difficult to compute even by modern computers. Also the COMSOL interface was not created to handle big datasets and widgets simply stop working. The only way to deal with this situation is to divide created study into few smaller ones and connect them outside the COMSOL.

There is also a problem with the trajectory of particles near vessel walls. If a single molecule travels close enough to the edge of a cylinder, its velocity is inherited from the boundary wall. This value is 0, so the examined particle stops at one point. It happened only when input geometry was read from

the output file and cylinder orientation wasn't set along the main axis. The reason lies in limited mapping accuracy for rounded objects.

The Final problem is associated with the interpretation of results. Values obtained in the descendent branches in bifurcation model raise our concerns. There is a need to simulate these phenomena using other tools and compare results.

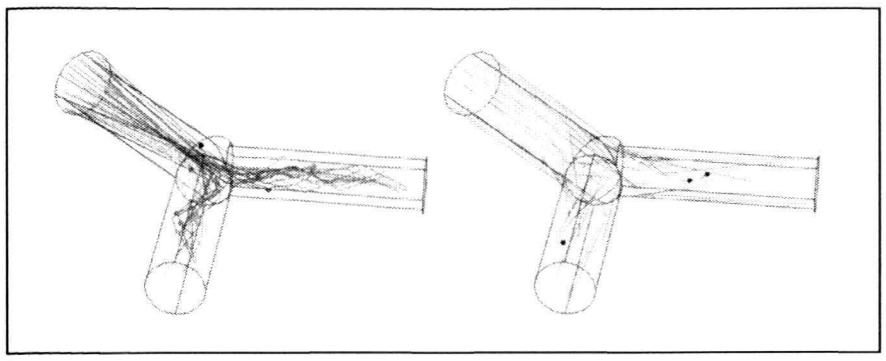


Figure 10. Particle tracing based on Laminar (left) and Turbulent model (right).

VII. Conclusion:

The goal of this paper was to carry out a simulation of blood flow in selected tubular objects. This task was successfully completed. Two types of flow were tested in different digital models. Expected values were obtained. Based on these simulations particle tracing was performed. 3D coordinates in time for each molecule was gathered. These values will be used in an MRI simulator to produce images similar to ToF. Due to the fact that Particle Tracing is a new module, several limitations and errors were found during the project. We proposed a way to deal with them.

Acknowledgment

This work was supported by the Polish National Science Centre grant no. 6509/B/T02/2011/40.

References

[1] P. Bogorodzki, "Obrazowanie czynności mózgu techniką rezonansu magnetycznego", Przegląd Elektrotechniczny, 2009, no.9, pp. 40-45.

[2] M. A. Bernstein, J. Huston , C. Lin , G. F. Gibbs, J. P.Felmlee, "High-resolution intracranial and cervical MRA at 3.0T: technical considerations and initial experience", Magn Reson Med, 46/2001, no.5, pp. 955-962.

[3] J. R. Reichenbach, E. M. Haacke, "High resolution BOLD venographic imaging: A window into brain function, NMR Biomed, 14/2001, vol. 7-8, pp. 453-467.

[4] A. Deistung, E. Dittrich, J. Sedlacik, A. Rauscher, J. Reichenbach, "ToF-SWI Simultaneous time of flight and fully flow compensated susceptibility weighted imaging", Journal of Magnetic Resonance Imaging, 29/2009, no.6, pp. 1478-1484.

[5] C. Kirbas, F. Quek, "A review of vessel extraction techniques and algorithms", ACM Computing Surveys, 36/2004, no.2, pp. 81-121.

[6] M. Nałęcz, Tom 8 – „Obrazowanie biomedyczne", Akademicka oficyna wydawnicza EXIT(2003), Biocybernetyka inżynieria biomedyczna, 2000, vol. 8, pp. 351-372.

[7] M. Strzelecki, A. Materka, M. Kociński, P. Szczypiński, Deistung A., Reichenbach J., „Ocena metody zbiorów poziomicowych w zastosowaniu do segmentacji trójwymiarowych obrazów fantomów cyfrowych oraz obrazów naczyń krwionośnych mózgu ToF-SWI rezonansu magnetycznego", InżynieriaBiomedyczna, Acta Bio-Optica et Informatica Medica,16 /2010 , pp. 167-172.

[8] P. Jonnin, E. Krupinski, S. Warfield, „Validation in Medical Image Processing", IEEE Transactions on Medical Images, vol. 25, 11/2006, pp. 1405-1409.

[9] W. R. Crum, O. Camara, L. Derk, G. Hill, "Generalized Overlap Measures for Evaluation and Validation in Medical Image Analysis", IEEE Transactions on Medical Images, vol. 25, 11/2006, pp. 1451-1461.

[10] F. Javier Sanchez Castro, "A Cross Validation Study of Deep Brain Stimulation Targeting: From Experts to Atlas-Based, Segmentation-Based and Automatic Registration Algorithms", vol. 25, 11/2006, pp. 1440-1448.

[11] B. Aubert-Broche, M.Griffin, G. Bruce Pike, A. C. Evans, D. L. Collins, "Twenty New Digital Brain Phantoms for Creation of Validation Image Data Bases" vol. 25, 11/2006, pp. 1410-1416.

[12] H. Benoit-Cattin, G. Collowet, B. Belaroussi, H. Saint-Jalmes, C. Odet. "The simri project: a versatile and interactive MRI simulator", Journal of Magnetic Resonance, no.173(1), 2005, pp. 97–115.

[13] K. Jurczuk, M. Kretowski, "Virtual magnetic resonance imaging - parallel implementation in a cluster computing environment",. Biocybernetics and Biomedical Engineering, 29(3), 2009, pp. 31–46.

[14] R.K.-S. Kwan, A.C. Evans, G.B. Pike, "MRI simulation-based evaluation of image-processing and classification methods", Medical Imaging, IEEE Transactions on, 18(11), 1999, pp. 1085–1097.

[15] J.S. Petersson, J.O. Christoffersson, K. Golman, "MRI simulation using the k-space formalism", Magnetic Resonance Imaging, 11(4), 1993, pp. 557–568.

[16] COMSOL Multiphysics, "Realease Notes", COMSOL, COMSOL 4.1 edition, 2010.

[17] Shelley Medical Imaging Technologies,"MRI Quality Assurance Flow Phantom Set", www.simutec.com

[18] Shelley Medical Imaging Technologies,"CardioFlow 1000 MR Computer Controlled Flow System", www.simutec.com

[19] W. Schroeder, K. Martin, B. Lorensen, „The Visualization Toolkit", Prentice Hall PTR, 1998.

[20] Google, "Google SketchUp 8", http://www.sketchup.google.com/

[21] K. Cieślicki "Hydrodynamiczne uwarunkowania krążenia mózgowego", Medycyna Informatyczna, Warszawa 2001, pp. 123-190.

[22] K. Jeżowiecka-Kabsch, H.Szewczyk, „Mechanika płynów", Wrocław 2001, pp. 163-245.

[23] COMSOL Multiphysics, "Sloved with COMSOL Multiphysics 4.1", COMSOL, COMSOL 4.1 edition, 2010.

[24] M Szydłowski, P. Zima 'Two-Dimensional Vertical Reynolds-Averaged Navier-Stokes Equations Versus One-Dimensional Saint-Venant Model for Rapidly Varied Open Channel Water Flow Modeling", Archives of Hydro-Engineering and Environmental Mechanics, 53(4), 2006, pp. 295-309

[25] COMSOL Multiphysics, "Chemical Engineering Module – User's Guide", COMSOL, COMSOL 4.1 edition, 2010.

Visualization and Volume Calculation of the Hand's Small Vessels

Michał Strzelecki, Sławomir Skonieczka
Institute of Electronics
Technical University of Lodz
Lodz, Poland

Ludomir Stefańczyk, Marek Olszycki,
Katarzyna Skrobisz, Agata Majos, Piotr Grzelak
Department of Radiology and Diagnostic Imaging
Medical University of Lodz
Lodz, Poland

ABSTRACT — **The aim of this study was to evaluate magnetic resonance angiography (MRA) images of the small vessels of the hand's fingers, with the use of the new MRA blood pool contrast and the custom software. The software not only visualizes the hand vascularity but also estimates the blood vessel volume in each finger along with the volume of remaining tissues. Analysis results for sample image sets were presented and discussed.**

I. INTRODUCTION

The evaluation of the state of human organs, which is necessary for medical diagnosis and/or disease treatment, requires objective information about physical properties of various tissues and organs. Since usually there is no direct access to a living tissue, the noninvasive imaging techniques of acquiring quantitative information about parts of the human body are of great value for the task [1]. Thus one of the main areas of researchers activity is aimed at developing fast, objective and precise means aiding medical doctors in their profession [2]. An excellent example of a success in this field is a revolutionary technique of magnetic resonance imaging (MRI) - developed in the seventies of the past century and now widely applied to hospitals and clinics. The recent advent of new MRI blood pool agents offers a new approach for monitoring the vessel in MR angiography (MRA). They produce a higher signal and have a longer half-life, in comparison to standard agents, so a specific region of interest can be evaluated with a longer time of acquisition at a higher spatial resolution [3], [4], MR angiography techniques are characterized by higher temporal and spatial resolution when compared to digital subtraction angiography, moreover they are much less invasive. Thus MRA is preferable for diagnosing upper extremity vascular disorders and characterization of complex arteriovenous anatomy such as in vascular malformations and for evaluation of dialysis fistulas and grafts [4]. Also, a wide variety of diseases like atherosclerosis or embolism, which are the most common in affecting the upper extremity can be diagnosed by MRA methods. There are not many references related to the quantitative analysis of vessels in extremities images. In [9], the vascular abnormalities in the upper extremities echo images were described by an image-based flow model to detect major bleeding. Maximum intensity projection maps were applied in [3], [4] to identify vessel abnormalities.

II. MATERIALS AND METHODS

The survey was conducted in a group of 5 healthy volunteers, who declared no diseases or past injuries of their hands. The examination consists of contrast agent application (i.v. 0.1 ml/kg c.c. Vasovist) and consecutive MRA data acquisition in the so-called high-resolution sequence, using parallel imaging and echo-sharing technique. Patients have one hand (left) cooled down and used to lie in prone and headfirst position (called "Superman" position) with the hands placed above the head and under the coil, with fingers speeded in order to avoid wraparound artifact. The examination was repeated at regular intervals to determine the degree of extend of vessels. As a reference, warm (right) hand was used. Under these conditions visualization of peripheral circulation of upper extremity was much more pronounced than in the MRA obtained with the standard contrast agent and the same acquisition parameters as shown in 1.

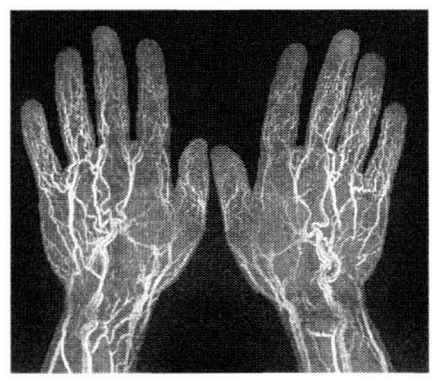


Fig. 1. The maximum intensity projection

Magnetom Avanto 1.5T (Siemens Medical Systems, Erlangen, Germany), with a dedicated send-receive multi-channel phased-array surface coil which enables quite high signal-to-noise ratio, small pixel size and high spatial resolution have been used. High resolution is desirable to properly depict the palmar arches and the digital vessels, which are often of sub-millimeter diameter. A fast three-dimensional (3D) sequence T1-weighted "vibe": (T1_vibe_we_cor_320_iso: TE=5.5, TR=13.1, FoV=30x30 cm, matrix 704x704, Slice Thickness 0.40, Pixel Spacing 0.426x0.426) was applied. Slice thickness was 1.6 mm, consideration of 24 partition images results in 38.4 mm thick imaging slab, that was adequate to cover the both hands. Acquisition time was approximately 3

minutes, including oversampling by doubling the number of phase-encoding steps. This fast interpolated 3D acquisition permit near isotropic depiction of hand anatomy. The native resolution of the images was increased with use the "zero-filling k space interpolation" which allows to display them with a higher-resolution matrix [10].

For blood vessel differentiation the following algorithm was proposed, as presented in 2

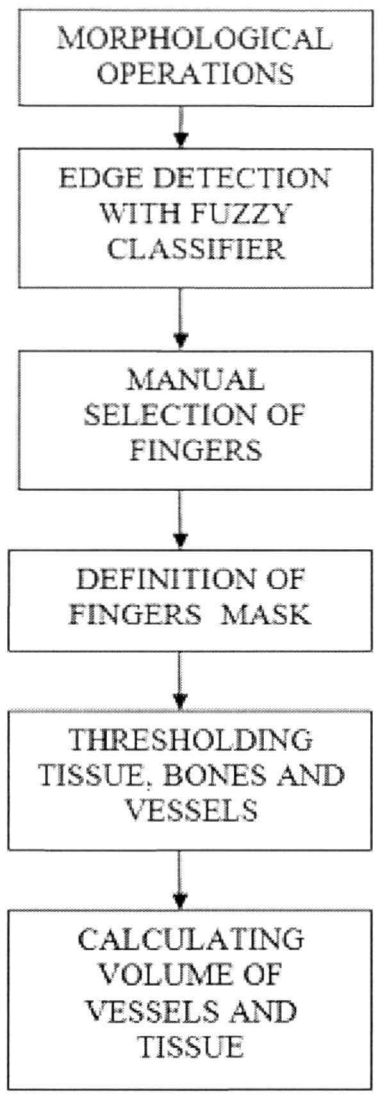


Fig. 2. Proposed algorithm for vessel image analysis

Step l: To remove the acquisition noise from the image, the morphological operations were performed (opening and closing) [8]. Step 2: In order to obtain the hand boundary to separate it from the background fuzzy edge classifier detection algorithm [7] was applied. 3 x 3 x 3 mask was used, with the central pixel situated at coordinates (i, j, k). An edge may appear in many directions and it is necessary to calculate the sum of the differences of the bidirectional amplitudes between the given pixel and its neighbors. For each pixel in the input image, a vector that contains previously calculated distances must be formed. The next step is to divide the input image pixels into classes. For this purpose the following classes were defined: edges (it corresponds to a single vector of amplitudes), background and noisy edges. The amplitudes will be related only to the minimum and maximum values that they may assume. The appropriate class for the image background will correspond to any pixel in whose neighborhood the amplitude difference in all directions is small. The last class, in which an edge is regarded as containing noise, the amplitude change in the vicinity of a pixel in all directions is considered to be high. Thanks to this background signals are not confused with the signal coming from the bones of the hand. 3 shows the result of hand detection algorithm

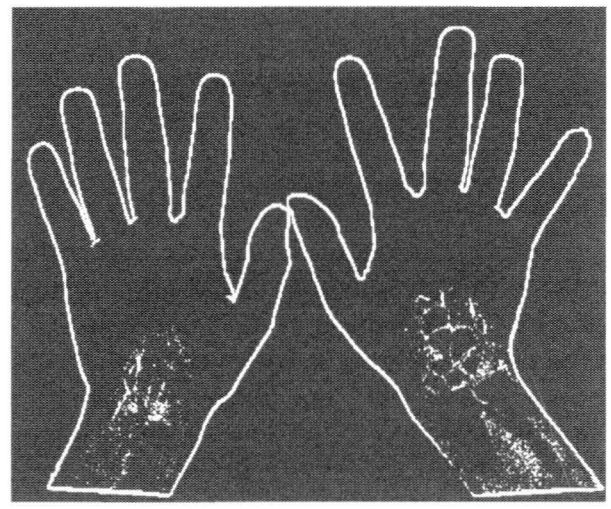


Fig. 3. Outline of detected hands

Step 3: Manual fingers selection. 4 shows a window of the developed software with manually defined masks to differentiate all fingers.

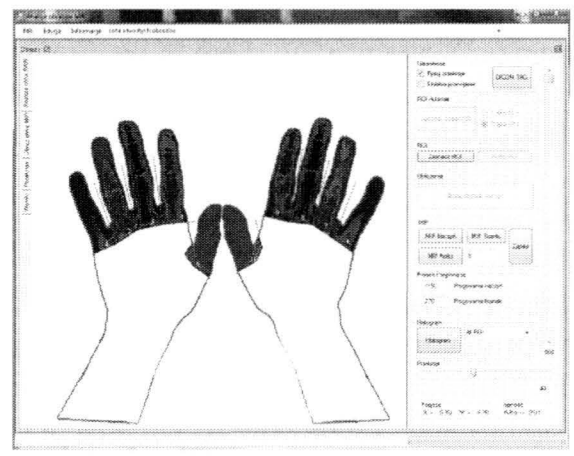


Fig. 4. Fingers separation

Step 4. Calculation a number of voxels belonging to the soft tissue, bone and blood vessels based on thresholding method. The threshold value can be set manually; default one is equal to 1150. 5 illustrate separated blood vessels by algorithm from DICOM 3D sequence.

III. RESULTS

Symmetrical and qualitatively satisfactory MRA raw and MIP anatomical images of the small vessels of both hands were obtained, including the level of fingers in all of our healthy subjects. The results of volume calculation of the hand's small vessels summarized Table 1. I

IV. CONCLUSION

MRA with the use of the blood pool contrast agent demonstrates the ability of presenting the hand's peripheral circulation, including the level of the fingers. Obtained results demonstrated that developed software provides quantitative

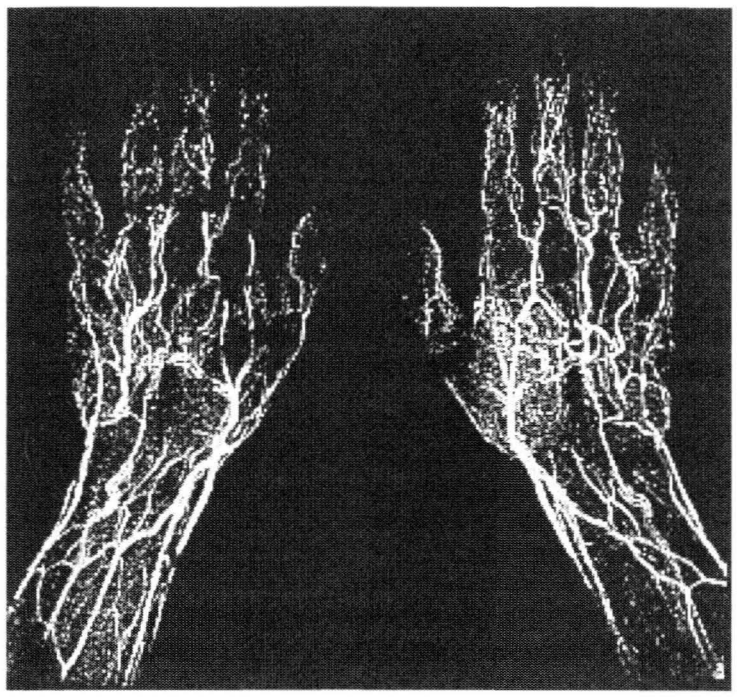


Fig. 5. Thresholded blood vessels

TABLE I
EXAMINATION RESULTS

EXAMINATION	Hand	[Voxels]	Volunteer I	Volunteer II	Volunteer III	Volunteer IV
I	LEFT	VESSEL	1936	8775	5878	5161
		TISSUE	784869	963652	1268383	891082
		VESSELS/mm3	0,0114	0,0367	0,0196	0,0280
	RIGHT	VESSEL	6284	17179	31177	21189
		TISSUE	801477	1434014	1325614	809752
		VESSELS/mm3	0,0374	0,0609	0,1175	0,1247
II	LEFT	VESSEL	1670	9969	6242	5973
		TISSUE	807976	1178752	1229973	908915
		VESSELS/mm3	0,0098	0,0398	0,0239	0,0325
	RIGHT	VESSEL	4518	13412	30897	21526
		TISSUE	743977	1108758	1276601	848393
		VESSELS/mm3	0,0290	0,0586	0,1181	0,1169
III	LEFT	VESSEL	2790	10704	2677	8190
		TISSUE	907372	1170637	1340701	944497
		VESSELS/mm3	0,0148	0,0435	0,0098	0,0425
	RIGHT	VESSEL	6845	10208	21709	21145
		TISSUE	867458	1131299	1293184	885243
		VESSELS/mm3	0,0383	0,0445	0,0822	0,1145
IV	LEFT	VESSEL	5809	12212	8092	13425
		TISSUE	949723	1170877	1301358	993012
		VESSELS/mm3	0,0299	0,0496	0,0234	0,0712
	RIGHT	VESSEL	6780	16107	29745	21327
		TISSUE	878040	1146875	1288803	845223
		VESSELS/mm3	0,0374	0,0689	0,1124	0,1233

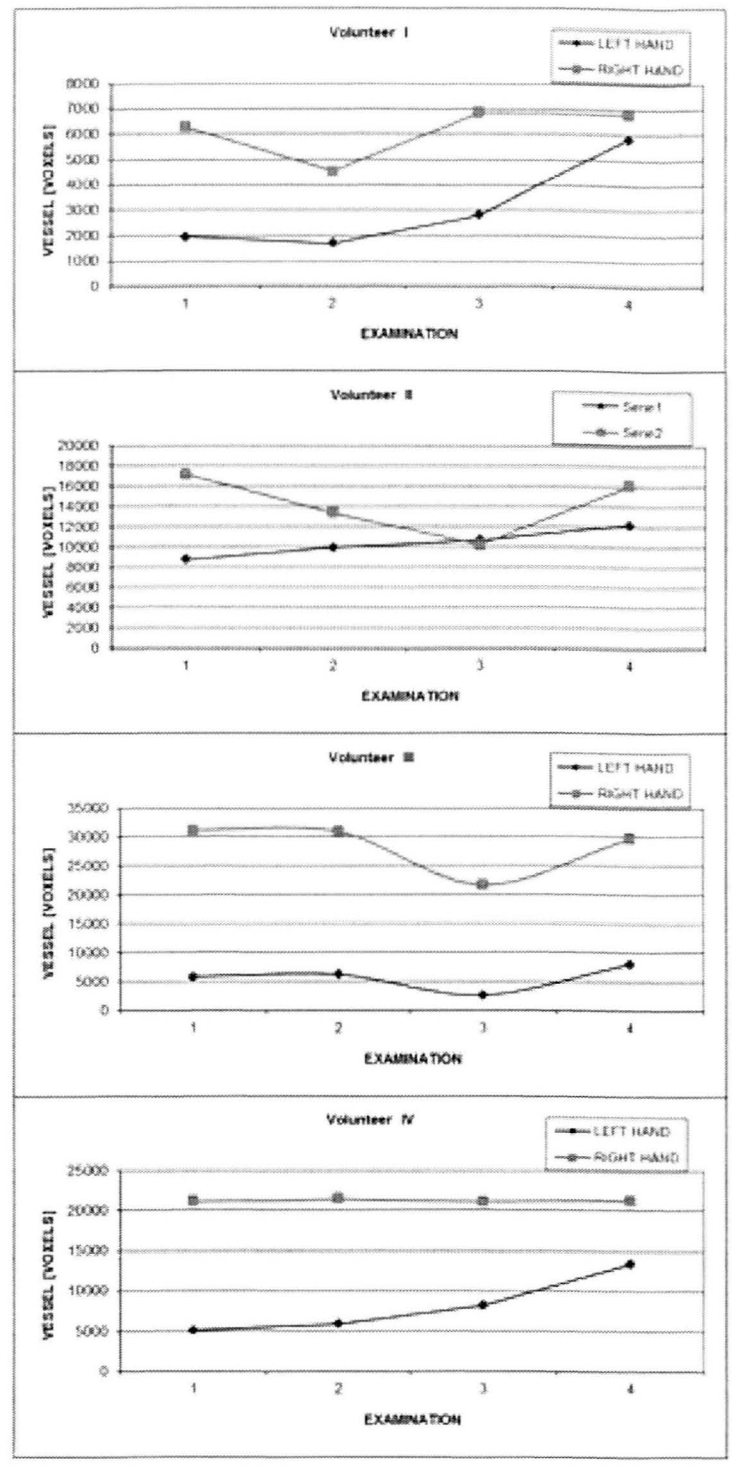


Fig. 6. Diagrams of vessel volume change

and objective evaluation of peripheral circulation. 6 illustrate process of increasing the vessels volume as a result of hand warming. This process takes app. 20 minutes. Such plots are useful for physicians in the assessment of circulation disorders. Our algorithm allowed quantifying the research - both in terms of vascular and soft tissue. This is a great advantage in comparison with the previous only subjective - qualitative assessment, typical for routine analysis of standard MRI. In clinical MR equipment is the lack of tools for quantitative measurements of the image. An analysis of the data obtained enables detection of vascular anatomy differences between patient's hands, and even between each finger. So far, such differences were perceptible subjective and dependent on the observer - his experience, knowledge of anatomy, as well as the time spent on assessment tests. The applied algorithm helped demonstrate statistically significant differences in the evolution of vasospasm in each successive time of measurement in cooled down hands. These observations are consistent with data from other diagnostic techniques used to assess these pathologies - thermography and ultrasound-Doppler. Also, it was possible to reveal another feature of vascular reactivity in response to the cool down test - vascular spasm in the hands of the control (not cooled down), which was previously suggested by other studies (ultrasound, thermography), but not clearly observed in our study diagnostic angio-MR. Data obtained from the analysis of histogram hand controls allow investigation the vascular response of the hand to thermal stimulation. In the future, the proposed algorithm can be applied not only for routine diagnosis of peripheral vascular limb - as anatomy and function, but also to monitor clinical trials of drugs changing vascular game and monitor treatment process in subsequent controlled examinations. Also, new approaches to vessel detection will be considered, based on structural pattern recognition methods including e.g. syntactic image analysis[11].

REFERENCES

[1] M. Kociski, A.Klepaczko, A. Materka, A. Lundervold, M. Chekenya, "3D image texture analysis of simulated and real- world vascular trees", Computer Methods and Programs in Biomedicine (2011)

[2] P. Szczypinski, M. Strzelecki, A. Materka, MaZda - a Software for Texture Analysis, Proc. of ISITC 2007, November 23-23, 2007, Jeonju, Korea, pp. 245-249

[3] Corot C, Violas X, Robert P, Gagneur G, Port M Comparison of different types of blood pool agents (P792, MS325, USPIO) in a rabbit MR angiography-like protocol. Invest Radiol 38(6):2003; 311-319

[4] Nikolaou K, Kramer H, Grosse C, Clevert D, Dietrich O, Hartmann M, High-spatial-resolution multistation MR angiography with parallel imaging and blood pool contrast agent: initial experience. Radiology 2006; Dec; 241(3):861-72.

[5] Stepansky F., Hecht E., Rivera R,, Hirsh L., Taouli B., Kaur M. Dynamic MR Angiography of Upper Extremity Vascular Disease: Pictorial Review Radiographics January 2008 ; Published online October 29, 2007, doi:10.1148radiol.e28.

[6] Connell D., Koulouris G., Thorn D., Potter H. Contrast-enhanced MR angiography of the Hand Radiographics, May 2002, 22:583-599

[7] C. Suliman, C. Boldior, R.Bazavan F. Moldoveanu, A fuzzy logic based method for edge detection. Bulletin of the Transilvania University of Brasov, 4 (53), 2011:156-164.

[8] Graham AD, Turner MJ, Hewitt T, Multi-scale image de-noising and de-cluttering using cross-scale mathematical morphology, Proceedings of the Ninth IASTED International Conference on Computer Graphics and Imaging, 2007

[9] Bech, F. Lee, J.; Taylor, C.A. ; Liang, D.H. Wang, A.S. , Developing an arterial bleed detection algorithm for diagnostic ultrasound Ultrasonics Symposium, 2008. IUS 2008. IEEE, pp: 1627 - 1630

[10] Korosec R F., Frayne R., Grist M. T., Mistretta A. C. Time-resolved contrast-enhanced 3D MR angiography, AJNR Am Neurocardiol, 2002, 22:1615-1619

[11] Ogiela, M.; Tadeusiewicz, Syntactic reasoning and pattern recognition for analysis of coronary artery images. Artificial Intelligence in Medicine 2002, 26(1-2): 145-159

Influence of Pose Angle on Face Recognition from Very Low Resolution Images

Tomasz Marciniak, Radosław Weychan, Agata Chmielewska, Adam Dąbrowski
Poznań University of Technology, Chair of Control and Systems Engineering,
Division of Signal Processing and Electronic Systems, Poznań, Poland,
{tomasz.marciniak, radoslaw.weychan, agata.chmielewska, adam.dabrowski}@put.poznan.pl

ABSTRACT — **This paper presents analysis of impact of the image resolution on the efficacy of the automatic face recognition. During experimental studies three databases were tested, where the head in the photos is set at different angles. The effectiveness of face location detection was examined with the use of the skin color and geometric models. Next, we tested the influence of the head position and the image resolution on the people identification effectiveness. Effect of lowering the image resolution is presented in a FAR / FRR graph form.**

KEYWORDS — ***face detection, face recognition, low resolution, face databases***

I. INTRODUCTION

Automatic face recognition based on images taken by surveillance cameras is a task that requires several problems to be taken into account in comparison to the standard access verification systems. Majority of the recognition techniques assume a full-frontal view of the face. During the image acquisition in CCTV (*closed-circuit television*) systems, we are dealing with changes of illumination, changes of scale, and non-frontal views of the subject [1]. This article focuses on exploring the impact of the image resolution and issues associated with various angles of the face settings.

Typically, the video monitoring systems are built according to the EN standard 50132-7 [2]. In section 7.6 of this standard there is a recommendation concerning the size of an object on the screen. The subject should be related to the tasks of the operator, such as identification, recognition, detection, or control. If the object is a person, and the resolution limit of the CCTV installed exceeds 400 television lines, then for purposes of identification, the object (person) should occupy at least 120 % of the screen. In the case of currently used identification algorithms suggested facial image resolutions are 32×32 or 64×64 pixels [3]. Such resolutions are marked as VLF (*very low resolution*).

This paper is a continuation of our previous research works presented in [4, 5]. All experiments were carried out in Matlab with the use of the public databases briefly described in the next section.

II. FACE DATABASES USED IN RESEARCH

In order to verify influence of the head position on the effectiveness of face recognition three databases were tested. Yale [6], FullFaces [7] and MUCT [8] are free databases shared by university research teams. Basic statistics of the databases are presented in Fig. 1.

Yale database contains 5760 images of 10 people, each individual was photographed at 9 positions and under 64 light conditions – the total number of images of every individual is 576. Grayscale images were saved in PGM (*portable graymap format*) at 640×480 resolution.

Achermann database – called also FullFaces from University of Bern in Switzerland includes files saved in Sun Image Raster Bitmap (RAS) format at 512×342 resolution. This database contains 10 grayscale images of every 30 individuals (i.e., only 300 images in total). The photos were taken in constant light conditions and in various positions of the head – frontal, profile view and face directed up and down.

The third tested database is the MUCT face database from the University of Cape Town. It consists of 3755 faces of 624 individuals. As in the case of FullFaces database, the MUCT database is prepared at constant light conditions and at different angles of the head positions, with the difference, that the MUCT database files are saved as color images at 480×640 resolution in the JPEG format.

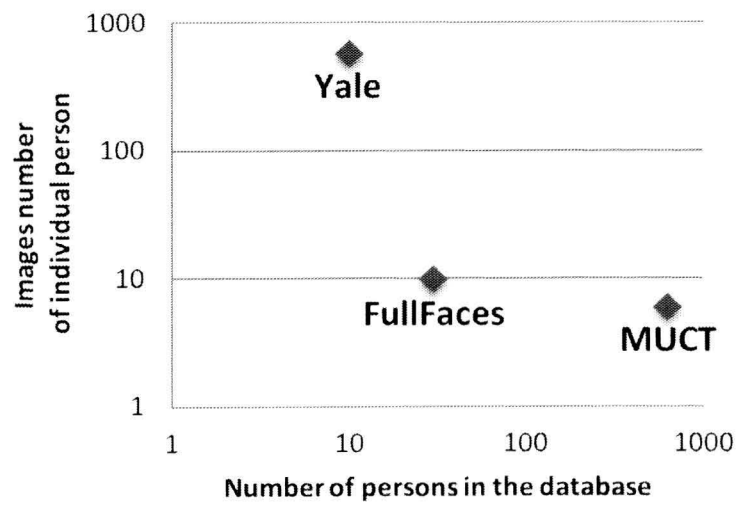


Figure 1. Statistics of databases used in scientific research

III. FACE DETECTION STAGE

The first stage of face recognition is the face localization. During this process a position and a scale of face (or faces) is estimated and these values are used as the input data in the identification/verification algorithm. The face detection subsystem should correctly identify objects similar to faces, like a hand (in the case of the skin color algorithm) or a ball (in the case of the geometric modeling) and reject them. Location of a human face in the image is one of the most important steps in the process of face recognition. This stage reduces the calculation time by reducing the analyzed area in the identification/verification stage.

The face detection is typically implemented in three steps. The first step is reduction of a distortion with the use of image processing techniques such as noise reduction and equalization of the histogram. The next step consists in finding areas, where faces are most likely to appear. The final step verifies the previously selected areas.

Most face recognition techniques and databases assume that the image is normalized by ISO/IEC 19794-5:2005 [9] (Information technology – Biometric data interchange formats – Part 5: Face image data) standard or ANSI/INCITS 385-2004 [10] (Information technology – Face Recognition Format for Data Interchange) norms. These standards indicate, for example, the positions, size, and rotation of face in the image or the width-to-height ratio of the image.

Below we briefly describe how distortion of images in relation to standards such as: lack of regular source of light, various image sizes, head or image rotation, hidden part of the face affect abilities of the face detection. During our research two public software packages for face detection are used. First of them is based on the skin detection algorithm [11] and the second uses geometric models [12] exploiting the Hausdorff distance [13].

A. Face localization using skin color detection

In the case of color images we can detect the face with the use of the skin color filter – this is the first factor that can be taken into account in detection of the human face. The human skin color differs in intensity (luminance), not the color (chroma) – thanks to which, skin color differences can be effectively removed.

In our research we used "Face Detection in Color Images" software [11]. This environment is very comfortable due to embedded basic functions for image processing. Input image in RGB space format is converted to the IRgBy color space, then Rg and By matrices are filtered by the median filter. Hue and saturation values are used to select those areas, where the color matches the color of the skin – the result is a binary skin map (Fig. 2).

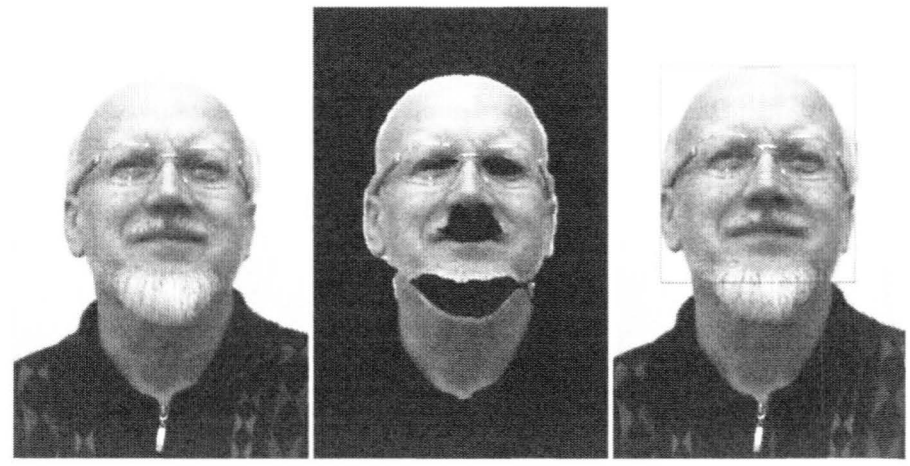


Figure 2. (From left) Original image, skin map multiplied with original image in the grayscale, image with detected face

The binary skin map and the original image are used to detect faces in the picture. This technique is based on selecting the designated regions that have holes – eyebrows, eyes, mouth, or nose. Theoretically, all regions, where the skin is detected without holes, are not faces. The above operations – finding holes in the binary image – are performed using morphological operations. An additional advantage of this algorithm is compensation of lighting effects in the image, by the histogram extension. This operation helps to eliminate the darker and lighter regions of the face image.

Figure 3 shows examples of face detection. A problem with this method is small resistance to low lighting or intensive side illumination. In this case the algorithm incorrectly selects the face area in the image. Improper lighting conditions can be compensated with the use of the histogram equalization.

It can be observed that the algorithm correctly detects faces, which have dark skin tone, are partially hidden, e.g., by a scarf, and even when the head is rotated.

B. Face localization using geometric models

If we deal with gray-scale images, we cannot use an algorithm that searches for skin areas. In this case the solution can be the use of geometric models. These methods are based on the knowledge of geometry of a typical human face, e.g., assumes its natural symmetry. An appropriate algorithm finds a rule to describe the shape, size, and other face characteristic points such as eyes, nose, or chin. Relationships between them (positions and distances) are also important.

For testing, an experimental software "FDMver1.0" was used [12]. As previously mentioned, an algorithm applied in this software uses the Hausdorff distance and is based on gray-scale images.

Figure 3. Examples of face detection based on skin detection

Our experiments have confirmed that this technique is not resistant to changes of orientation or face position and to low light conditions as it is shown, for example, in Fig. 4. In full light conditions the program works correctly as well as in the case of side lighting. Problems can occur when the light comes from above, because of the shadows falling on the face. Detection during rotations of the head, as long as the facial image is symmetrical (in the horizontal orientation), works properly. Additional accessories like a hat, glasses, and a scarf generally do not influence the proper detection. Wrong detection can occur in the case of a complex background or patterned clothes.

IV. INFLUENCE OF IMAGE RESOLUTION AND HEAD ROTATION

Two mentioned databases (Yale and FullFaces) have been downsampled 2, 4, 8, and 16 times for recognition issues. In the reported experiments we have used the *eigenfaces* approach implemented in the Matlab software [14] modified to batch processing [4]. This software works in two modes:

- Model creating – creates model of every individual. Each subsequent folder, which includes face pictures, becomes a new class. The more data (i.e. images) in the folder, the better the model is created. The generated database includes coefficients representing each class in separated rows.
- Recognition – creates table [*N M*] (*N* number of pictures in the test folder, *M* class number generated in the training phase), which includes distances between every tested picture to the class obtained in the previous mode.

Figure 4. Examples of proper and improper face detection with use of Hausdorff distance

In the case of the FullFaces database, 5 pictures of every individual have been used for training mode, and 5 for the recognition mode. A database for the recognition phase

includes 150 rotated face pictures (pitch and roll). The same types of frames were used to create a model of each person. Experimental results are presented in Fig. 5. As it can be observed, similar results are given for the original picture resolution, downsampled 2 and 4 times. In this case the EER (*equal error rate)* is about 5 %. For pictures downsampled 8 and 16 times the results are only by 2 % lower (the EER is then about 7 %). The obtained values are very similar to the results presented in [4].

Yale database includes 585 pictures of each person. It gives a possibility to distinguish the amount of frames used to create the model and those for the recognition phase. 117 face pictures of each person, rotated (pitch and roll) and in various (13 types) light conditions were used for the training stage. The rest of frames, including faces photographed in 52 other light conditions, were used in the recognition phase. In this stage 4700 files were processed. Due to a large number of comparisons, the recognition task was a very time-consuming process. For original resolution it took about 48 hours. In case of the parallel processing, every downsampling by 2 lasts about 8 hours less.

Results of the experiments for the Yale database are presented in Fig. 6. FAR/FRR curves are in this case much smoother in comparison to those for the FullFaces database (Fig. 5), because now there are 5760 images compared to only 300 for the FullFaces database. It can be noticed, that similarly to the previous results, the recognition quality with the use of the original data and the data downsampled 2 and 4 times is almost the same. However, the EER is higher in the case of the larger database and various light conditions (Yale database) by about 16 %. Subsequent downsamplings 8 times and 16 times give larger results, i.e., 19 % and 20 %, respectively.

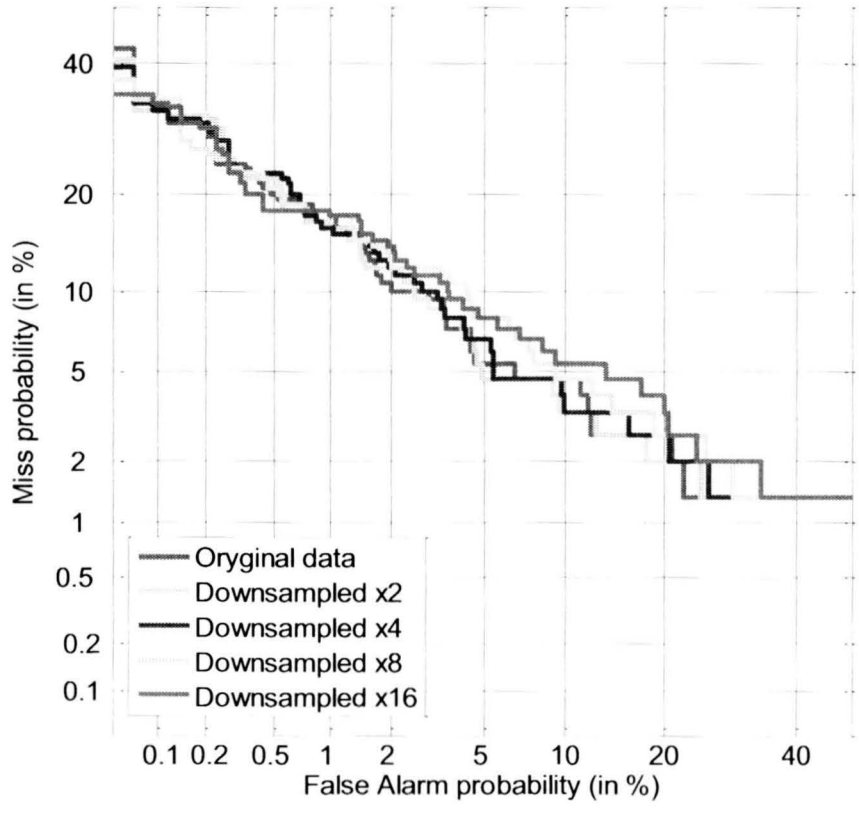


Figure 5. FAR/FRR plot for FullFaces database downsampled up to 16 times

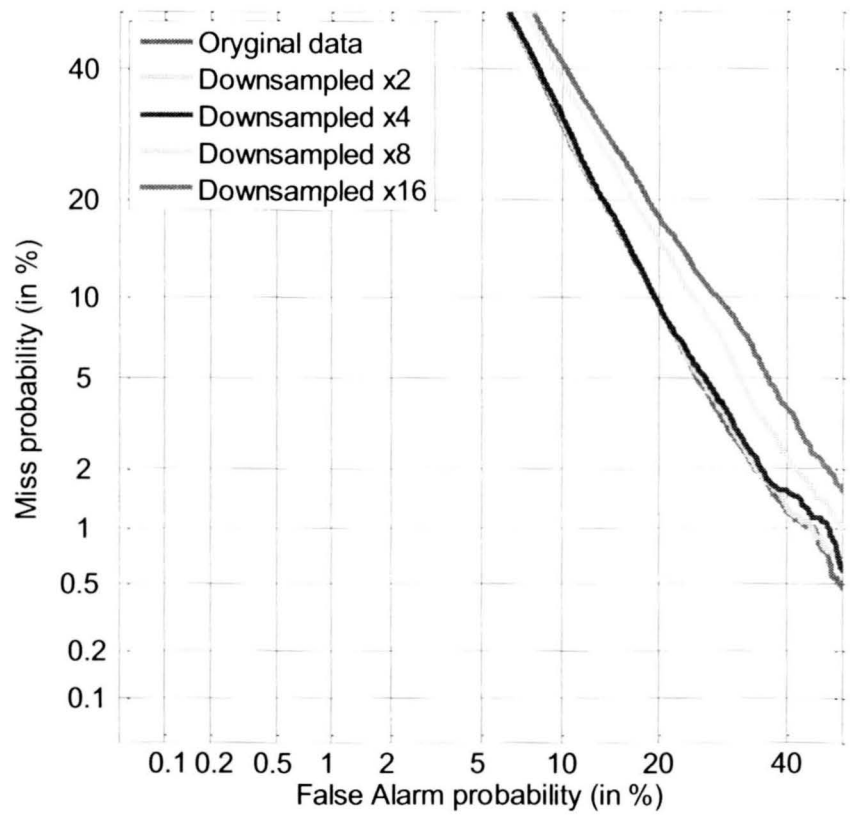


Figure 6. FAR/FRR plot for Yale database downsampled up to 16 times

V. CONCLUSIONS

Rotation of the head in the picture and variable lighting conditions can significantly degrade effectiveness of the identification based on the face image. Experimental results showed that the face identification efficiency decreases due to face rotation (pitch and roll) and downsampling by about 10 %. Moreover, a combination of unfavorable lighting conditions together with facial rotation causes degradation of the results up to 12 % (according to experiments with the Yale database).

Another problem that can occur during the image acquisition in video surveillance systems is a low resolution of those image portions where faces can appear. The tests show that even 4 times downsampling (output image resolution of 160×120 pixels) does not yet influence the face recognition accuracy. However, 8 times and 16 times downsamplings make the EER worse by about 2–3 %.

Generally, face detection based on the skin color recognition better detects faces in a picture in comparison to the geometric method. The considered version of the geometric method based on the Hausdorff distance has problems with poor lighting, lighting from above, and rotation of the head but it more accurately determines location of the face.

On the other hand, the analyzed algorithm for the skin color detection supplemented with the holes search in the recognized object in order to consider it as a face adds the neck to the face in most cases, which is not required for the face identification process.

REFERENCES

[1] M. Davis, S. Popov and C. Surlea, "Real-Time Face Recognition from Surveillance Video", studies in: Computational Intelligence, vol. 332/2010, pp.155-194, 2010.

[2] EN standard 50132-7: 1998 Alarm Systems CCTV surveillance systems for use in security applications Part 7.

[3] Zou, W.W.W.; Yuen, P.C.; , "Very Low Resolution Face Recognition Problem," *Image Processing, IEEE Transactions on* , vol.21, no.1, pp.327-340, Jan. 2012.

[4] Marciniak, T., Dąbrowski, A., Chmielewska, A., Weychan, R., Face Recognition from Low Resolution Images to be published in: A. Dziech and A. Czyżewski (Eds.): MCSS 2012, *Communications in Computer and Information Science* 287, pp. 220-229. Springer, Heidelberg, 2012.

[5] Marciniak, T., Drgas, Sz., Cetnarowicz, D., Fast Face Localisation Using AdaBoost Algorithm and Identification with Matrix Decomposition Methods, Multimedia Communications, Services and Security 2011, *Communications in Computer and Information Science*, 1, Volume 149, , pp. 242-250

[6] A. S. Georghiades, P. N. Belhumeur, D. J. Kriegman, "From Few to Many: Illumination Cone Models for Face Recognition under Variable Lighting and Pose", IEEE Trans. Pattern Anal. Mach. Intelligence Vol. 23, No. 6, pp. 643-660, 2001.

[7] Achermann database, The face database of The University of Bern, 1995 http://www.cs.columbia.edu/~jebara/htmlpapers/UTHESIS/node81.html

[8] S. Milborrow, J. N. Morkel, "The MUCT Landmarked Face Database", Pattern Recognition Association of South Africa 2010, http://www.milbo.org/muct/, 2010.

[9] ISO/IEC 19794-5:2005, Information technology Biometric data interchange formats Part 5: Face image data, 2005.

[10] ANSI/INCITS 385-2004, Information technology Face Recognition Format for Data Interchange, 2004.

[11] J. P. Kapur, "Face Detection in Color Images", EE499 Capstone Design Project, University of Washington, Department of Electrical Engineering, 1997; http://www.oocities.org/jaykapur/face.html

[12] Algorithm for face detection using Hausdorff distance: FDMver1.0, http://en.freesion.com/3865/4898908/30215154/

[13] O. Jesorsky, K. J. Kirchberg, and R. W. Frischholz: "Robust Face Detection Using the Hausdorff Distance", BioID AG, Berlin, Germany 2001

[14] L. Rosa, Face Recognition System 2.1, http://www.advancedsourcecode.com//face.asp , 2006.

This work was prepared within the INDECT and DS projects.

Optimization of the FPGA Parallel Image Processor

Przemysław Brylski, Michał Strzelecki
Institute of Electronics, Technical University of Lodz
Wolczanska 211/215, 90-924 Lodz, Poland
e-mail: mstrzel@p.lodz.pl

ABSTRACT — **This paper describes a hardware implementation of parallel digital image processor in FPGA technology. The architecture and algorithm modifications presented in this paper are aimed to enable fully parallel processing and reduction of FPGA resources. The circuit core is 64x64 array for image processing and analysis. The proposed processor was tested on labeling of binary images; obtained analysis results are presented and discussed.**

I. INTRODUCTION

The image processing system presented in this paper is a continuation of work on the parallel, digital image processor proposed in [1] and modified in [2,3,4].

It implements the matrix of active nodes, which correspond to the image pixels. These nodes are connected by weights that depend on gray level of neighboring pixels. Weight values are processed into the form containing information about the local neighborhood of node and stored in nodes before segmentation. These data are used in the image region growing process.

Version of the image processor described in this article was implemented in VHDL [5,6] language. The processor has been synthesized for 64x64 active elements matrix and neighborhood size N = 8. It has been demonstrated that proposed processor could be successfully applied for segmentation of binary images.

Presented in [4] an IP Core of the image processor with 16x16 elements matrix size was synthesized for XCS500E (Spartan-3E family device) with Xilinx ISE 13.1. Operation of the image processor was successfully verified during segmentation of binary images. Both high speed of labels propagation (t_{LP}=60ns for 6-bit label) and segmentation time of single areas in the range of microseconds were achieved. However, this solution has several drawbacks. Images are processed in a sequential manner, object by objects. Additionally, system performance is limited by the need to re-send commands to nodes matrix for every processed image area.

This paper presents algorithm and architecture changes, introduced to increase the system efficiency and reduce a utilization of FPGA logic by a single image processing unit (node). Section 2 outlines the new processor architecture and algorithm. Sections 3, 4 and 5 contain implementation details and achieved results. Finally, section 6 presents discussion and conclusion.

II. IMAGE PROCESSOR ARCHITECTURE

Fig. 1. presents a block diagram of the parallel image processor. It contains microcontroller, control unit and a matrix of NxN nodes, which represent image pixel processing units.

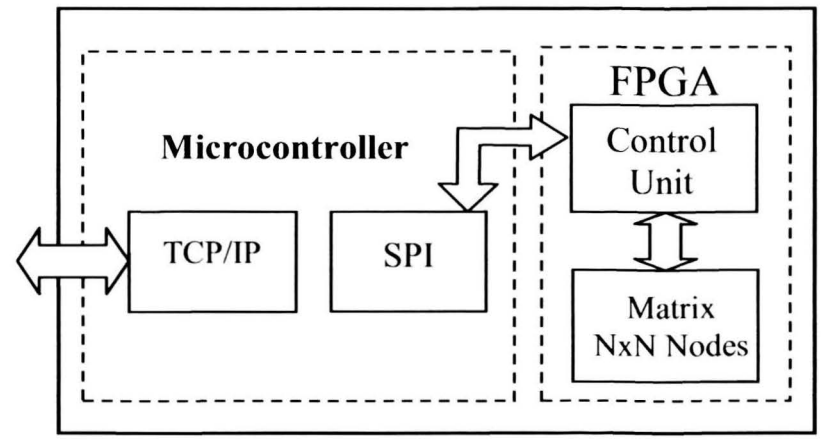


Fig. 1. Block diagram of image processor.

Microcontroller carries out the following tasks:

- loading the image from the host computer
- evaluation of node weights and neighborhood mask values
- controlling the segmentation process
- delivering the processed image to the host computer

Central Unit performs the following functions:

- downloading the preprocessed image data from the microcontroller into processing matrix
- providing the clock an control signal for nodes
- detection of the end of segmentation process
- delivering the analyzed image back to the microcontroller

Detailed information about algorithm performed by the image processor is presented in [2].

To overcome disadvantages of previous solution [4] and achieve fully parallel image processing, several algorithmic and structural modifications have been introduced:

- storage of neighborhood data in form of mask register

- acquired label by the node (in a single propagation cycle) is equal to maximum value of previously stored label and label received from a neighbor
- removal of FSM logic froma node
- introduction of the label comparator unit to node

Modifications of the algorithm

In previous implementation [4], during the simultaneous activation of all leaders, a problem with label propagation in homogeneous image regions may appear. It is possible to generate such label propagation cycles, which does not allow detecting the end of the segmentation. To solve this problem we have proposed that the acquired label by the node (in a single propagation cycle) is equal to maximum value of previously stored label and label received from a neighbor.

Next modification is related to method of storing in a single node information about its local neighborhood. Currently, weights are transformed and stored in each particular node in form of neighborhood mask shown in Fig. 2.

Neighborhood mask replaces the idea of activation table presented in [2] and introduces the N_ADDR signal (common for all nodes), used for addressing of local node's neighborhood.

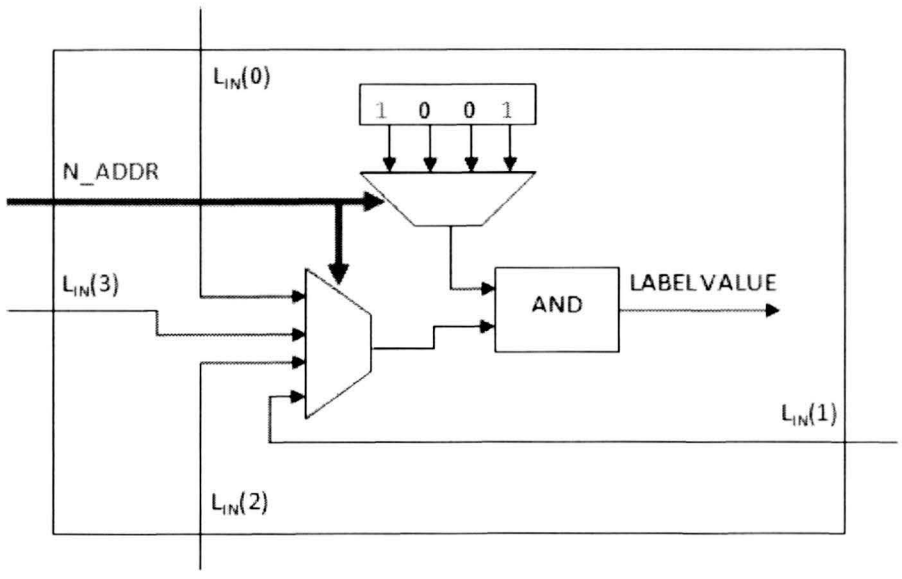


Fig. 2. The idea of the mask register for neighborhood size N=4.

The N_ADDR signal selects the direction of label propagation for the each particular propagation cycle. It is presented in Fig. 3. Signal is provided by the Sequencer unit and has a value in range from 0 to N-1, where N is a neighborhood size.

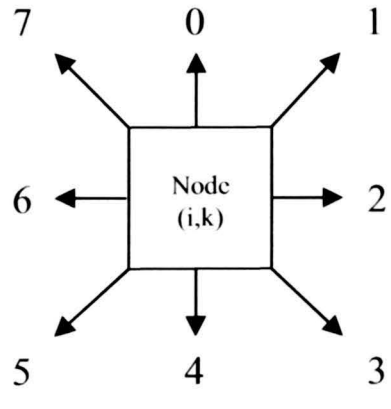


Fig. 3.Addressing scheme for the neighborhood size N=8.

Neighborhood mask is evaluated for each node by the microcontroller based on weights value. It selects possible neighbors to receive its label. Logic '0' value on the position corresponding to local address of node's neighbor excludes it from the process of label propagation. It is due to performing a logical AND operation on the acquired label with mask value for a selected neighbor.

Neighborhood mask allows to store node's local neighborhood data more efficiently. It reduces (to one for each neighbor) the number of necessary registers (bits) for the information storage related to node neighborhood and saves usage of FPGA resources. This solution enables for a very effective implementation in distributed RAM shift registers (an alternative functionality of LUTs–SRLs [7]). In comparison to idea of activation table, increased speed of neighborhood addressing.

Another modification concerns the method of leaders selection. In the presented solution an unlimited number of leaders can be selected. The leader is a node with initial label value greater than zero.

Selection of leaders is performed by the Control Unit, during the loading of neighborhood masks from the microcontroller. It is based on value of neighborhood mask assigned to a node. If its mask value is greater than zero, the node becomes a leader.

Nodes chosen as leaders, take the initial values of the labels, which are consecutive natural numbers. The remaining nodes are initialized with label value equal to zero.

An effective storage of neighborhood data in a node, allows analysis of larger size of local neighborhood, using the similar number of hardware resources.

This enabled increasing of size of the analyzed neighborhood from 4 to 8 nodes.

III. Hardware implementation

Microcontroller

Previously used ATMega128 [9] microcontroller has by replaced by the ARM family processor, because of performance problems. We have chosen the EM2440-III [10] development platform. It is an ARM-9 based board, populated with the S3C2440A (up to 400MHz, 64MB SDRAM) microcontroller. Rich peripherals such as Ethernet and SPI make this evaluation board suitable for this application. We chose the Linux (2.6.30.4) because it is an efficient, flexible and reliable operating system with very little memory needs.

FPGA

As the target platform the XUPV505-LX110T [8] evaluation board has been used. It features the XC5VLX110T chip, which is a medium size FPGA device from the Xilinx Virtex-5 family.

The introduction of the global addressing scheme simplified the internal structure of the node. The address generation and command processing unit (implemented as a Finite State Machine logic) has been removed and node is now stateless. This reduces utilization of FPGA logic by a node and simplifies operations performed by the Control Unit.

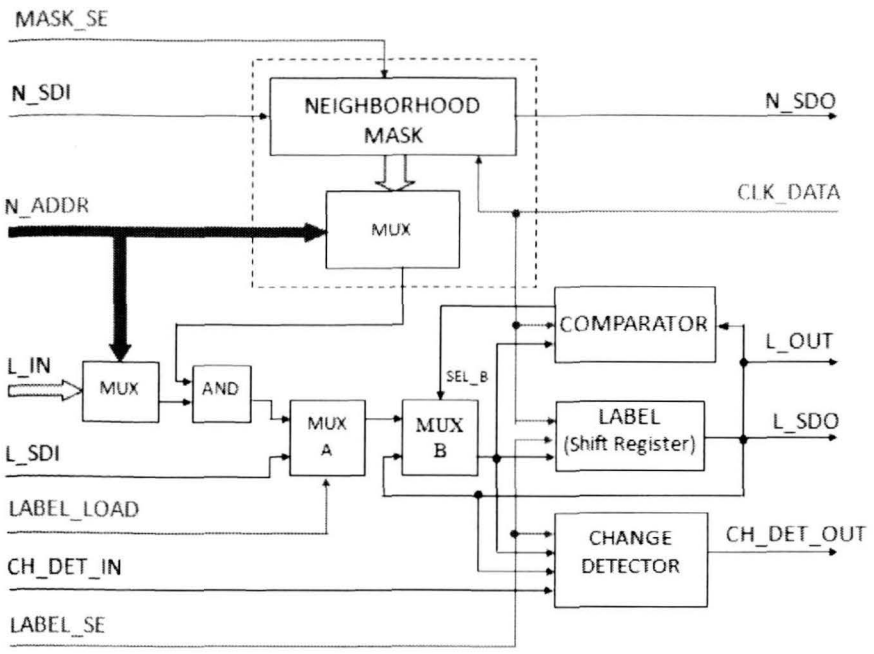


Fig. 4. Single node block diagram.

The node block diagram is shown in Fig.4. It consists of two parts. The first part (at the top of the figure), is responsible for addressing the local neighborhood and contains the mask register along with integrated multiplexer. The second part (at the bottom), is responsible for label exchange. It consists of a series of multiplexers that determine the label propagation path. The role of label shift register and change detector unit has been no changed.

Introduction of comparator is a novelty. It ensures the new label value (if acquired) will be greater than the previously keep in node. This requires that propagation of label starts from the least significant bit.

IV. Implementation results

The image processor presented in [4] was synthesized fora Spartan-3E family device, while the implementation described in this paper is targeted to a Virtex-5 family device.

The direct comparison of the synthesis results for both FPGA families is difficult because of architecture differences. Spartan-3E family devices can implement 4-input LUTs, while Viertex-5 feature 6-input LUTs.

For the above reasons, the IP Core of the presented image processor with 16x16 elements matrix size has been synthesized for Spartan-3E and Virtex-5 platforms.

Resource allocation for previous and current implementation of the image processor is summarized in Tab.1. We can observe fourfold reduction of occupied slice registers, while the number of used 4-input LUTs and slices has been twice reduced.

Storing information about local neighborhood of a node in the form of neighborhood mask has very strong influence to the FPGA optimization results. All these factors allow to process larger images in the same FPGA device.

Tab.1. Comparison of resource utilization between previous and current implementation targeted on the XC3S500E – 16x16 matrix size.

Logic utilization	Implementation [4]	Present implementation
Number of Slice Flip Flops	5,433	1,229
Number of 4-input LUTs	7,221	3,117
Number of occupied Slices	4,527	2,146
Number of bonded IOBs	8	6
Number of BUFGMUXs	3	3

The image processor with 64x64 matrix size was synthesized then placed and routed for XC5VLX110T with Xilinx ISE 13.1.

The implemented design utilizes approximately 93% slices of the target platform device. Detailed statistic of the FPGA logic utilization is presented in Tab.2.

Tab.2. Statistics of FPGA resource (XC5VLX110T) used for the implementation of the image processor - 64x64 matrix size.

Logic utilization	Used	Utilization
Number of Slice Flip Flops	20,755	30%
Number of 6-input LUTs	37,150	53%
Number of occupied Slices	16,155	93%
Number of bonded IOBs	28	4%
Number of BUFGMUXs	5	15%

The Xilinx XPower tool estimated the power dissipation to be less than 0.160W for a toggled rate of 50%.

High-speed operation of the FPGA logic enables rotation of labels with frequency of 100MHz. Thus label propagation time (for 12-bit label) between directly connected nodes is 120 ns.

V. Experimental results

As the source of binary images the Microsoft Windows application called "IPU Tester" was used.

Although the FPGA part of system can process images up to 64x64 pixels, the size of the test image is currently limited to 32x32 pixels. This is due to GUI limitations of the IPU Tester.

On the left side of application front view (shown in Fig.5) the image composer is placed and the result of segmentation is displayed on the right side.

Sample binary image (shown on the left side in Fig.5) with 9 objects (black pixels) on the background (white pixels) was used for system testing.

As can be seen, the image processor correctly recognized nine homogeneous areas on a sample image within 66.12 μs. Pixels belonging to the first object (the arrow) were marked by label "1" within 10 μs. Background of the image was not labeled and highlighted (the initial labels values equal to 0 were maintained).

The previous version of the processor [4] has marked the same object within 2.5μs, while the total processing time for a sample image was tens of milliseconds.

Increasing the size of the analyzed neighborhood from 4 to 8 nodes, allows propagation of labels along diagonals and proper recognition of object marked by label "4".

We can see a large performance boost of the image processor and reduction of the label growth speed in a single image area. This is due to increased length of labels from 6 to 12 bits and thus twice longer propagation time (t_{LP}).

Fig.6 presents the processing results for the test image with a spiral curve (segmented within 281.58μs). The next Fig. 7 shows the case, where single object fills the entire image (processing time was 26.85 μs).

The last example, Fig. 8 shows the sample image, consisting of multiple identical objects (segmented within 6.28 μs), while results are presented in Fig. 9.

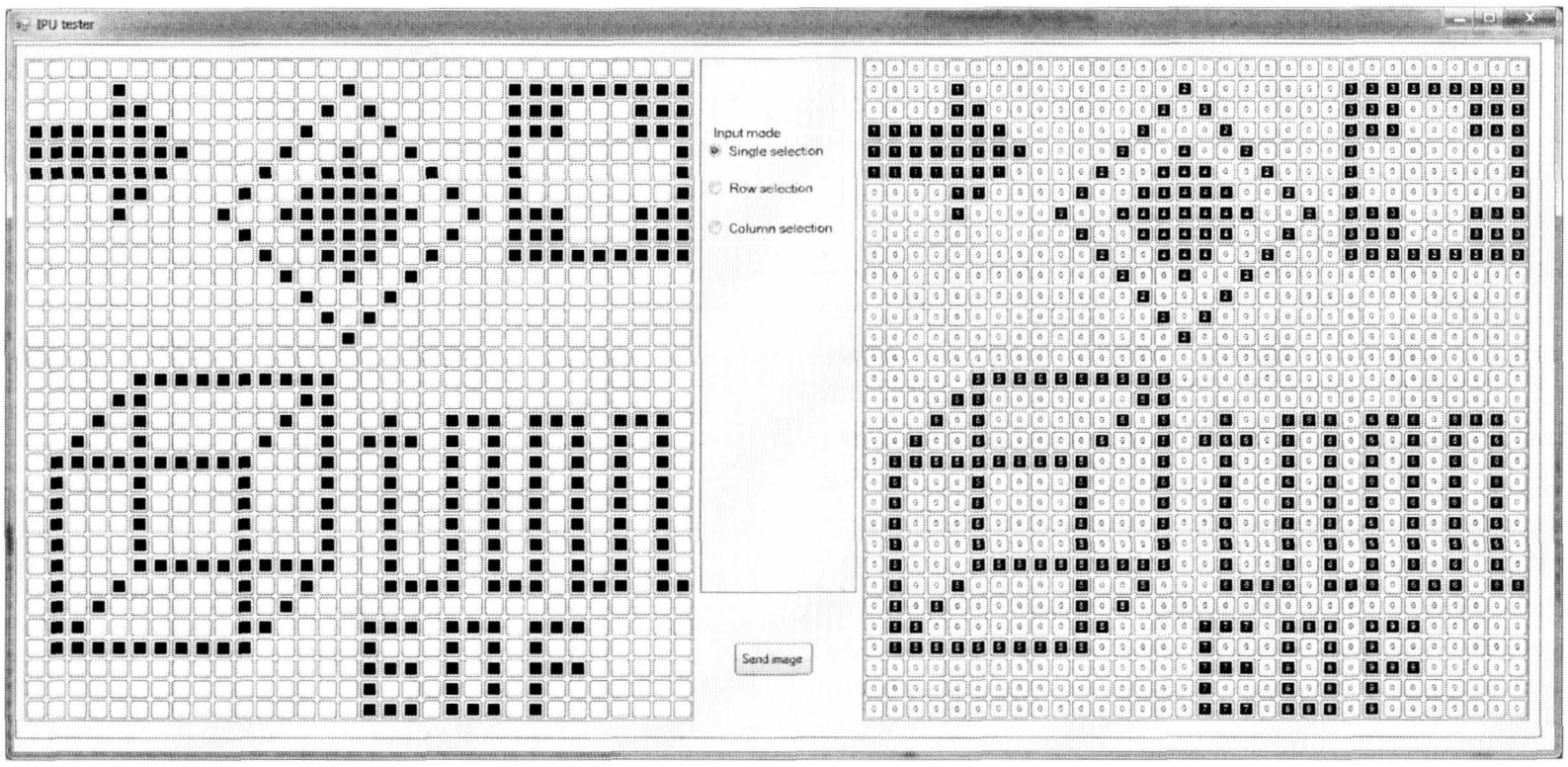


Fig. 5. Segmentation result of sample binary image – processed within 66.12 μs.

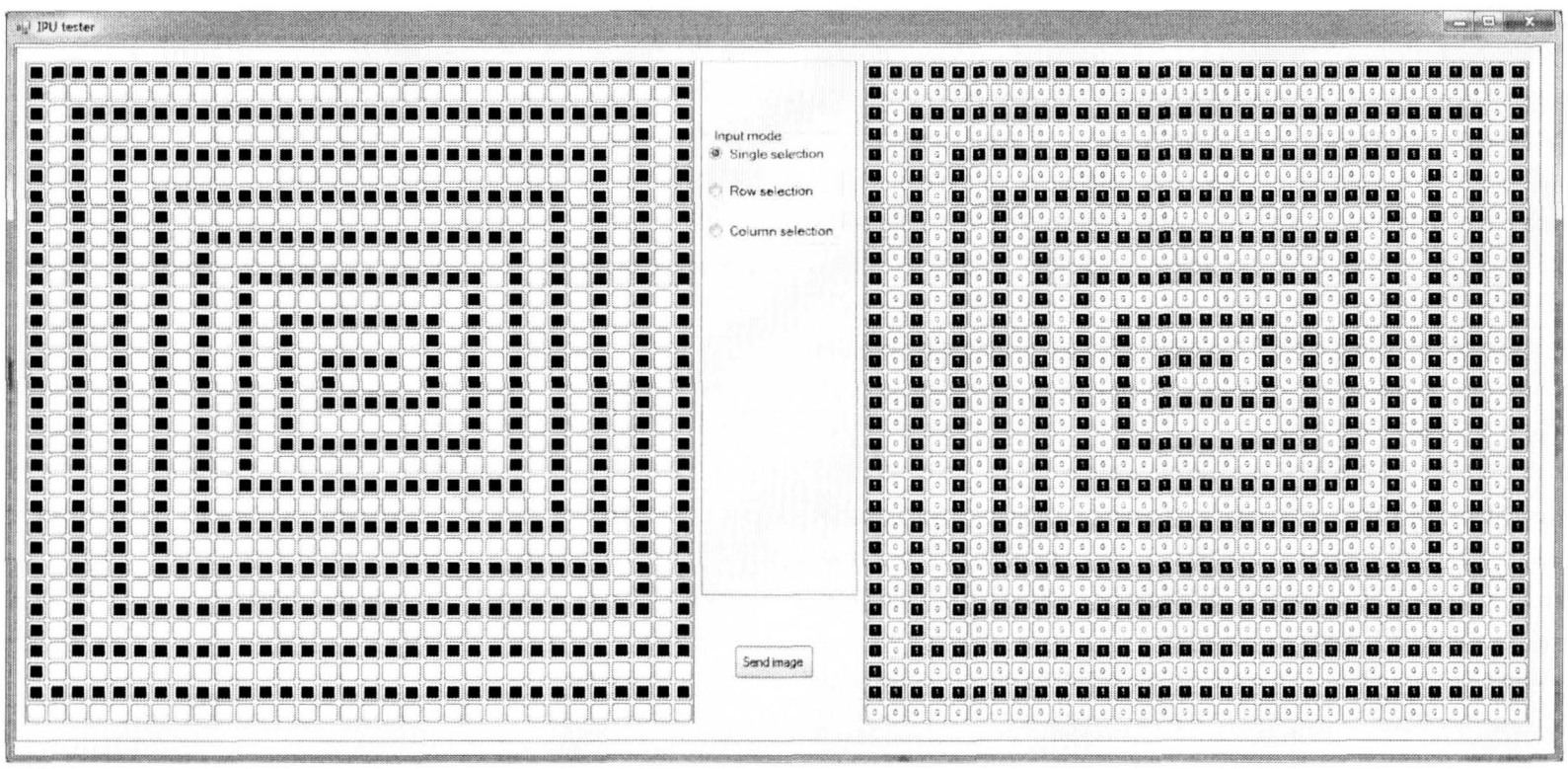


Fig. 6. A spiral curve test image – processed within 281.58 μs.

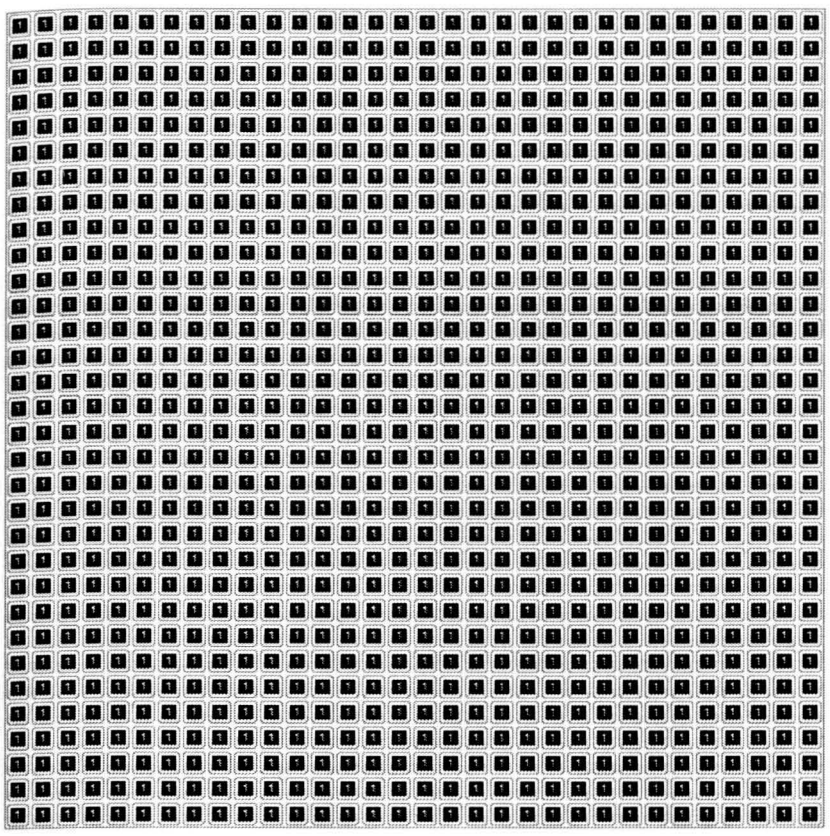


Fig. 7. Results of segmentation for an image consisting of a single large object – processed within 26.85 μs.

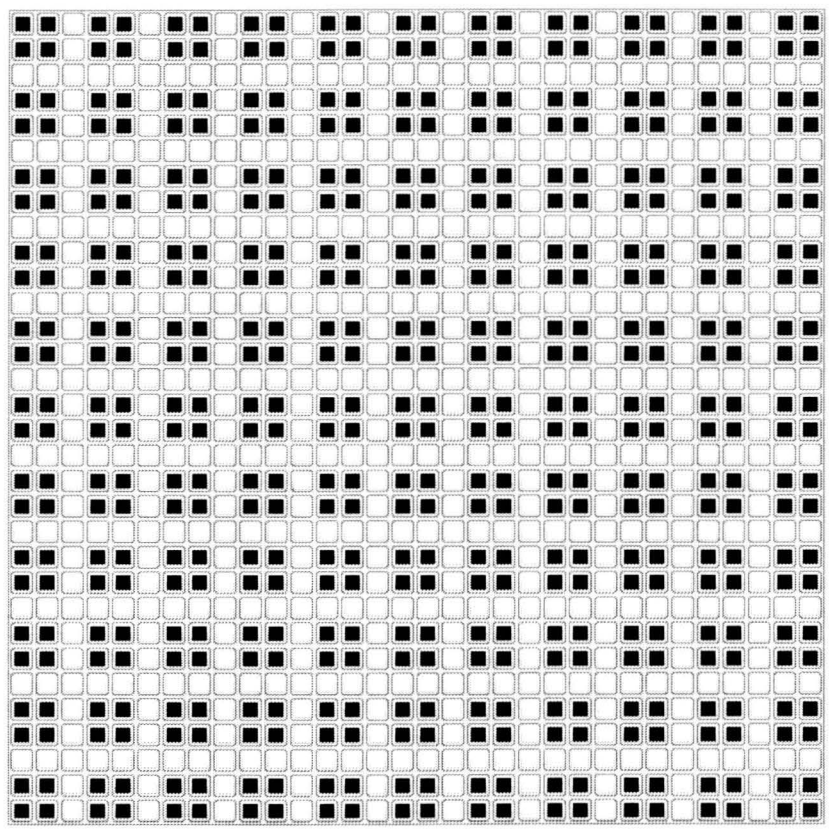


Fig. 8. Sample image consisting of multiple identical objects – processed within 6.28 μs.

```
001,001,000,002,002,000,003,003,000,004,004,000,005,005,000,006,006,000,007,007,000,008,008,000,009,009,000,010,010,000,011,011,
001,001,000,002,002,000,003,003,000,004,004,000,005,005,000,006,006,000,007,007,000,008,008,000,009,009,000,010,010,000,011,011,
000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,
012,012,000,013,013,000,014,014,000,015,015,000,016,016,000,017,017,000,018,018,000,019,019,000,020,020,000,021,021,000,022,022,
012,012,000,013,013,000,014,014,000,015,015,000,016,016,000,017,017,000,018,018,000,019,019,000,020,020,000,021,021,000,022,022,
000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,
023,023,000,024,024,000,025,025,000,026,026,000,027,027,000,028,028,000,029,029,000,030,030,000,031,031,000,032,032,000,033,033,
023,023,000,024,024,000,025,025,000,026,026,000,027,027,000,028,028,000,029,029,000,030,030,000,031,031,000,032,032,000,033,033,
000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,
034,034,000,035,035,000,036,036,000,037,037,000,038,038,000,039,039,000,040,040,000,041,041,000,042,042,000,043,043,000,044,044,
034,034,000,035,035,000,036,036,000,037,037,000,038,038,000,039,039,000,040,040,000,041,041,000,042,042,000,043,043,000,044,044,
000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,
045,045,000,046,046,000,047,047,000,048,048,000,049,049,000,050,050,000,051,051,000,052,052,000,053,053,000,054,054,000,055,055,
045,045,000,046,046,000,047,047,000,048,048,000,049,049,000,050,050,000,051,051,000,052,052,000,053,053,000,054,054,000,055,055,
000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,
056,056,000,057,057,000,058,058,000,059,059,000,060,060,000,061,061,000,062,062,000,063,063,000,064,064,000,065,065,000,066,066,
056,056,000,057,057,000,058,058,000,059,059,000,060,060,000,061,061,000,062,062,000,063,063,000,064,064,000,065,065,000,066,066,
000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,
067,067,000,068,068,000,069,069,000,070,070,000,071,071,000,072,072,000,073,073,000,074,074,000,075,075,000,076,076,000,077,077,
067,067,000,068,068,000,069,069,000,070,070,000,071,071,000,072,072,000,073,073,000,074,074,000,075,075,000,076,076,000,077,077,
000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,
078,078,000,079,079,000,080,080,000,081,081,000,082,082,000,083,083,000,084,084,000,085,085,000,086,086,000,087,087,000,088,088,
078,078,000,079,079,000,080,080,000,081,081,000,082,082,000,083,083,000,084,084,000,085,085,000,086,086,000,087,087,000,088,088,
000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,
089,089,000,090,090,000,091,091,000,092,092,000,093,093,000,094,094,000,095,095,000,096,096,000,097,097,000,098,098,000,099,099,
089,089,000,090,090,000,091,091,000,092,092,000,093,093,000,094,094,000,095,095,000,096,096,000,097,097,000,098,098,000,099,099,
000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,
100,100,000,101,101,000,102,102,000,103,103,000,104,104,000,105,105,000,106,106,000,107,107,000,108,108,000,109,109,000,110,110,
100,100,000,101,101,000,102,102,000,103,103,000,104,104,000,105,105,000,106,106,000,107,107,000,108,108,000,109,109,000,110,110,
000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,
111,111,000,112,112,000,113,113,000,114,114,000,115,115,000,116,116,000,117,117,000,118,118,000,119,119,000,120,120,000,121,121,
111,111,000,112,112,000,113,113,000,114,114,000,115,115,000,116,116,000,117,117,000,118,118,000,119,119,000,120,120,000,121,121,
```

Fig. 9. Segmentation results for the image from Fig. 8.

We can observe that the image processing time strongly depends on the size and shape of the objects, while their number has no influence on it.

In the current implementation the length of label (fixed to 12 bits) depends on the number of image pixels. It must be sufficient that every node being a leader, took a unique value of the initial label. In extreme cases, the number of segmentation leaders is an order of magnitude the number of pixels. So the required label length L can be determined according the formula:

$$L = \log_2(x \cdot y) \tag{1}$$

where, x – image width, y – image height.

Twice longer label only partially explains the reason of slowing down growth of the image area. The second factor is the way of analysis of its local neighborhood by the node. It turns out that the concept of the table activation is more efficient, in terms of acquisition of new label value by the node. This is because the activation table orders only neighbor nodes, for which the determined weight value is sufficiently high. This demonstrates the performance advantage of the activation table over the neighborhood mask.

VI. Conclusion

Verification of the hardware implementation of the parallel digital image processor was described in this paper. Performed tests confirm correct operation of the system.

Introduced algorithm and structural processor modifications solve the propagation cycle issue. It enables the parallel analysis of whole image regions and increases the system performance.

A significant reduction of FPGA logic usage allows the processing of larger images in the same FPGA device.

Further improvement of the system performance can be achieved by increasing the frequency of labels rotation.

Finally, different modes of node weight estimation will be implemented to enable segmentation of gray level images and other image processing operations e.g. edge detection of morphological filtering.

References

[1] P. Brylski, M. Strzelecki, Network of Synchronized Oscillators – Digital Approach, SPA 2008, Poznan, Poland, pp. 161–164.

[2] P. Brylski, M. Strzelecki, The architecture of a digital network for image analysis, SPA 2009 Poznan, Poland, pp. 25–29.

[3] P. Brylski, M. Strzelecki, FPGA Implementation of Parallel Digital Image Processor, SPA 2010 Poznan, Poland, pp. 25–28.

[4] P. Brylski, M. Strzelecki, Parallel digital image processor implemented in FPGA technology, SPA 2011 Poznan, Poland, pp. 52–55.

[5] K. Skahill, *VHDL for Programmable Logic*, Addison-Wesley 1996.

[6] Volnei A. Pedroni, *Circuit Design with VHDL*, MIT Press 2004.

[7] http://www.xilinx.com/support/documentation/user_guides/ug190.pdf, last visited in June 2012.

[8] http://www.xilinx.com/univ/xupv5-lx110t.htm, last visited in June 2012.

[9] http://www.atmel.com/dyn/products/product_card.asp?part_id=2018&category_id=163&family_id=607&subfamily_id=760, last visited in March 2012.

[10] http://www.armdesigner.com/EM2440-III.html, last visited in June 2012.

SESSION 6:
AUDIO PROCESSING III

Design of DSP Supported Systems for Real-Time Voice Watermarking

Paweł Pawłowski, Radosław Weychan, Adam Dąbrowski, Szymon Drgas
Poznan University of Technology, Department of Computing Division of Signal Processing and Electronic Systems
Poznań, Poland
{pawel.pawlowski, radoslaw.weychan, adam.dabrowski, szymon.drgas}@put.poznan.pl

ABSTRACT — **A problem of design of a system for real-time voice watermarking using a digital signal processor (DSP) is studied and presented in this paper. The authors prepared and compared three versions of the considered system using different types of data formats (fixed-point and floating-point) and various ways of programming (Matlab/Simulik compiler for Code Composer Studio and plain C/C++ programming).**

KEYWORDS — DSP, watermark, voice, DWT, real-time, design

1. Introduction

Real-time signal processing realized with digital signal processors (DSP's) becomes more and more important and popular. In fact, nowadays DSP's support many devices exploited even in a broad sense everyday life.

Shrinking time to market (TTM) requirements together with rising equipment demands make the DSP based design process more and more difficult and error prone. In order to discuss the respective problems, the authors consider and present in this paper a design example of the DSP based platform, namely the system for real-time watermarking of telephone conversations. It is thought to be a part of a larger system for reliable segmentation of speakers during telephone calls (e.g., to numbers of emergency services). A DSP adds, in real-time, an almost not hearable and not annoying watermark to the voice of one side of the telephone conversation. During inserting the watermark the signal distortion must be low and the voice delay should be not disturbing the conversation. The target system should be standalone with no need to program it after the power-on cycle. Proper algorithms for the considered purpose should be selected and realized using modern hardware/software co-design techniques.

The authors already proposed an experimental version of the mixed hardware/software watermark inserting system in [1]. The present paper presents three other versions of this system, designed with the use of various types of data formats (i.e. with the use of fixed-point and floating-point arithmetics), different ways of programming (i.e. with Matlab/Simulik compiler for Code Composer Studio and with plain C/C++ programming) and also using different hardware. Finally, the presented design variants are evaluated and compared.

The paper was supported by INDECT and DS 2012 projects.

2. Voice watermarking system

For real-time generation of the watermark during telephone conversations the authors prepared a DSP based hardware/software platform [1]. Its block diagram is presented in Fig. 1. The speaker voice processing is performed with the TMS320C6713 DSK (DSP starter kit).

The input signal comes from the microphone placed in the telephone handset. The output signal is a properly watermarked input signal, which then goes through the separation module to the telephone and further to the telephone line. Because the DSP board is powered by the external power adaptor and additionally is connected to the host computer for programming, it should be galvanically separated from the telephone (as well as the telephone line). It is realized with the separation module (see Fig. 1 and Fig. 2) [1].

Due to a variety of the telephone handset wiring methods, an additional wiring selector acts between the handset and the system.

The just presented system idea was used to realize three versions (called platform I, II, and III). They are described in this paper in the following sections.

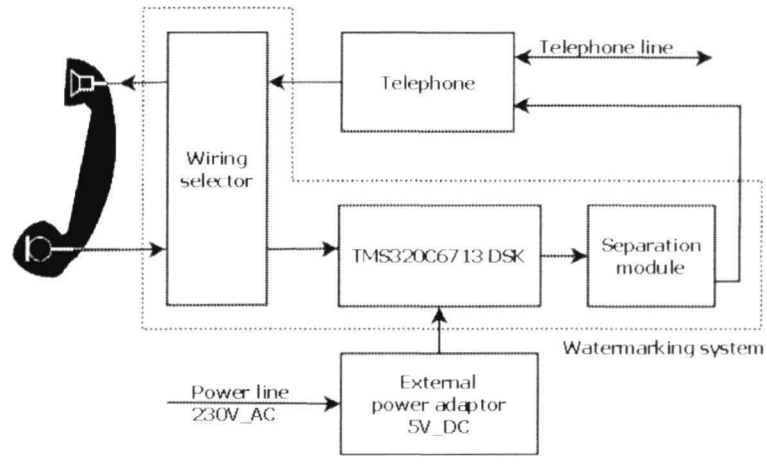


Fig. 1. A block diagram of the watermarking system (platforms I and II)

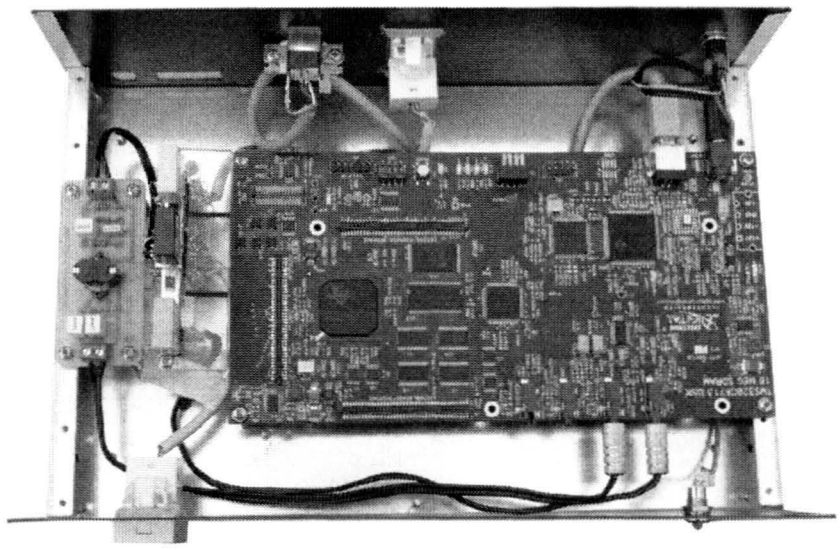


Fig.2. Hardware of watermarking system (platforms I and II)

3. Voice watermarking algorithm

A problem of segmentation of speakers, e.g. during the telephone conversation, has already been considered in the literature [3, 4].

Typical approaches that can be found in [3, 4] are based on assumption of some knowledge of the speaker voice and the results are often erroneous. Our method does not need any knowledge of this kind. However, we assume that we have access to one side of the telephone conversation (which is just the case, e.g., in the emergency telephone service centers) and we can specially prepare this side speaker voice by adding a watermark precisely in the speaker activity instants. Thus our segmentation method is free from errors.

We have already presented an audio watermarking method for the emergency telephone speaker segmentation [1]. In this specific application there is no need to transmit any data via the watermark. The segments are determined only by the existence of the watermark in the given time instances. The method proposed in [1] was designed to be fast and simple.

The proposed audio watermarking method is based on the discrete wavelet packet transform (DWPT). In the discrete wavelet transform (DWT) domain, the analyzed signal is represented as a weighted sum of the so called wavelet functions $2^{j/2}\psi(2^j t - k)$, with scale 2^{-j} (j being resolution) and shift k. Thus, DWT can be interpreted as a fixed octave-band filter bank. Unlike DWT, DWPT provides a wide range of bases and analyzes in the less than the octave bands.

In our case the DWPT is used to analyze the telephone band (0–4000 Hz) in 256 uniform bands. The watermark is added into one of these bands of the DWPT domain.

This method was first modeled and positively verified in the Matlab/Simulink environment. This was the initial step for the design of the DSP based Matlab/Simulink platform I (Section 4).

For the C/C++ programmed platforms II and III a simplified method of the watermark inserting has been proposed. Instead of calculation of the whole wavelet packet analysis and synthesis filter coefficients, the following simplified and computationally efficient procedure has been proposed.

The signal is processed independently frame by frame, which are all 256 samples long. Both, the i'th frame vector $\boldsymbol{x}_i$ and the watermark vector $\boldsymbol{w}$ contain 256 samples. The watermark signal $\boldsymbol{w}$ is just the symlet wavelet [2] with the center of the bandwidth in 2000 Hz. Next, for each frame a part of the signal correlated with the watermark is subtracted

$$\boldsymbol{x}_i \leftarrow \boldsymbol{x}_i - (\boldsymbol{w}^T \boldsymbol{x}_i)\boldsymbol{w}\,. \tag{1}$$

The obtained signal corresponds to the no watermark case in the current instant.

If one needs to watermark the current instant that corresponds to i'th frame, the watermark is added using the following expression

$$\boldsymbol{x}_i \leftarrow \boldsymbol{x}_i + \alpha \boldsymbol{w}. \tag{2}$$

A value α needs to be chosen according to the tradeoff between watermark audibility and its robustness.

Detection of the watermark is performed off-line by the PC, which analyses files registered with the digital telephone line recorder. The telephone line may, however, be (and often is) analog. In such a case the watermark can only be detected after the digital-to-analog and the analog-to-digital conversions. Thus no direct connection or synchronization with the described DSP based watermark insertion system is possible. However, because both the watermark inserting system (transmitter) and the receiver of the watermark operate in the digital domains some post synchronization of the respective signals, i.e., selection of the optimum positions of frames of samples is necessary. This is realized by means of the correlation maximization as explained below.

The signal is again divided into frames and each frame $\boldsymbol{x}_i$ again contains 256 samples numbered with indices $i = 1, \ldots, 256$. Let us define

$$\boldsymbol{w}_j = \begin{bmatrix} (\boldsymbol{w})_{(256-j+1):256} \\ (\boldsymbol{w})_{1:(256-j)} \end{bmatrix}, \tag{3}$$

where $(\boldsymbol{w})_{a:b}$ denotes a vector, which contains elements of the watermark $\boldsymbol{w}$ from indices a to b if $a \leq b$. If $a \geq b$, vector $(\boldsymbol{w})_{a:b}$ disappears from (3). For each frame the inner product

$$y_j = \boldsymbol{w}_j{}^T \boldsymbol{x}_i \tag{4}$$

is evaluated for $j = 0, \ldots, 255$. Finally, we check if $\max y_j > \theta$, where θ is an experimentally selected threshold. If this inequality is true, it means that the watermark has been detected in the i'th frame.

4. Platform I realized with Matlab/Simulink

Design of the DSP solution using Matlab/Simulink [6] (platform I) is possible for digital signal processors from Texas Instruments. The authors used Matlab/Simulink

(ver. 7.1) model-based DSP design, Code Composer Studio (ver. 3.1) and a floating-point DSP embedded in a typical starter kit, i.e., the TMS320C6713 DSK [8].

A fact that the used DSP is floating-point simplifies translation of data from Matlab/Simulink (which is typically 64-bit floating-point data) to C/C++ and finally to the processor assembler.

A diagram of the Matlab/Simulink model-based DSP design is presented in Fig. 2. It allows for a rapid implementation of algorithms on variety of DSP platforms. A high-level programming model in Matlab/Simulink must first be prepared. Very often such a model is prepared for testing of the ideas and algorithms through the simulation, independently to the further implementation. Due to this, the project must, in the next step, be adapted to work on a given DSP platform. Currently, Mathworks together with Texas Instruments support many (but not all) processors from families: TMS320C2000, TMS320C5000, and TMS320C6000. Unfortunately, this set is not constantly changing. In new versions of Matab/Simulink some of processors are added but some of them are removed (e.g. the last Matlab version that correctly supports TMS320C6713 DSK is ver. 7.1 from 2006). This situation narrows down the set of the target platforms and makes the design quite cumbersome.

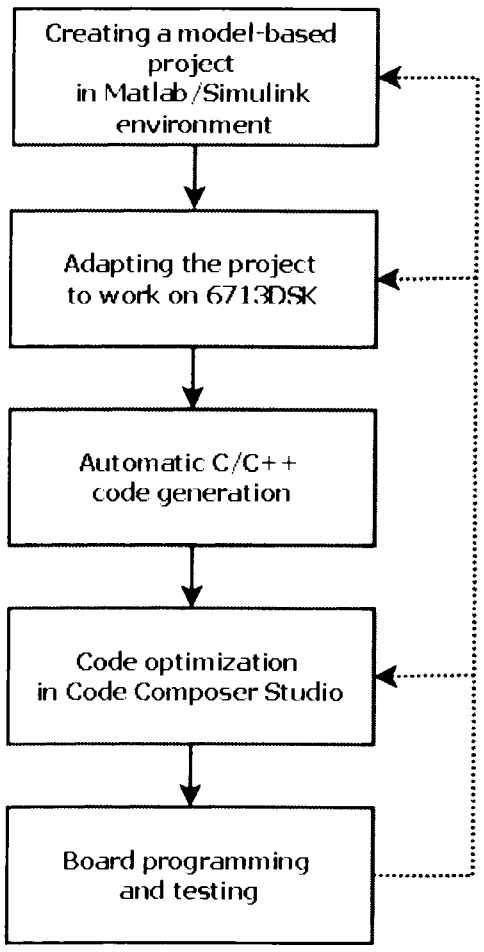


Fig. 3. Matlab/Simulink model-based DSP design

Next step, i.e. automatic C/C++ code generation (see Fig. 3) is performed during compilation of the project. During this stage, Matlab/Simulink connects to the Code Composer Studio [7] and generates the C/C++ code for each graphical model. This kind of programming is very simple and can be the fastest way to check the algorithm accuracy. However, it has some disadvantages. The automatically generated source code is hard to understand and optimize. Optimization is a problem in itself but it is additionally augmented by another problem: automatic translation from the high level Matab programming to the low level C/C++ and finally to the assembler language. This translation is very hard to be done optimal. In fact it tries to translate one type, model-based, to the quite different types of programming with the process control like e.g. procedures, interrupts, processes, threads.

Finally, an unnecessary large amount of code lines is generated and, in many cases, also an unnecessary large delay in the signal processing (i.e. between the input and the output signals) occurs,. Automatically generated binary file, to be loaded into the memory, is also much larger than that prepared directly in the C/C++ language under the Code Composer Studio from Texas Instruments.

It has major influence in stand-alone platforms, i.e. those without the host computer. The stand-alone system must store the entire program data in a non-volatile memory, typically it is a FLASH memory. Despite reading from the FLASH memory is much faster than writing, the FLASH memory is too slow to be the program memory during the run-time of the DSP. Typically, during the booting process the DSP uses the booting kernel and uploads the FLASH memory into the internal RAM (IRAM). After this it executes the program from the internal RAM. If the program does not fit into the internal memory, it should be divided into sectors. Some of the sectors are placed in the internal RAM, some of them are stored into the external RAM (SDRAM). This case is very hard to manage, especially under the Matlab/Simulink environment.

All mentioned above problems took place in the initial realization of the platform for inserting of the watermark.

First problem was related to the size of the output file. In the most cases, the entire program can be loaded into the IRAM. On the TMS320C6713 DSK it has size of 196 kB. The memory map in this case is shown in Fig. 4.

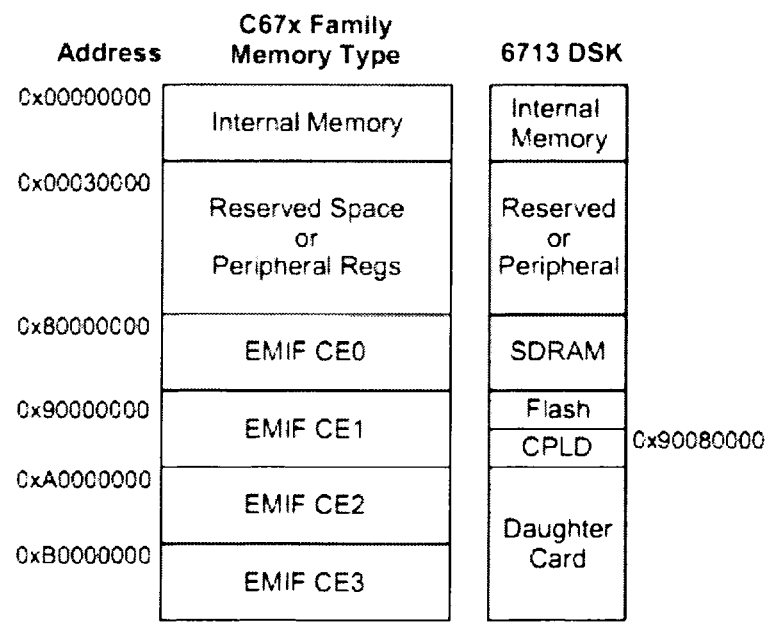


Fig. 4. Memory map of TMS320C6713 DSK [8]

The size of the file generated by Matlab was equal to 185 kB, but it does not mean that the internal memory size was enough. The internal memory must store additional objects like the kernel or the stack. Other address spaces are reserved or fixed to given components (with no possibility to reconfigure). Finally, it may be necessary to use external 128 Mb of SDRAM. While the program is running under control of the Code Composer Studio (e.g. during programming, debugging, or testing), the problem is not the case, but the DSP must be programmed by the host computer after each power-up. To make the system the standalone type, it is necessary to use a nonvolatile memory. On the board a 4 Mb parallel FLASH memory is available. To use it, at first an additional software converter (hex6x.exe [9]) must prepare the required output file and must expand the program code with the booting kernel and divide the program using the table of sections linked to the proper sectors in the memory map (see the vector table in [10]). In practice, the mentioned software loses some sections, making the program impossible to operate correctly after loading it from the FLASH memory. In fact, even the documentation from Texas Instruments [10] and the related internet forums do not solve this problem

The project developed in platform I contains 42 files. The watermark inserting algorithm is realized in about 700 lines of C/C++ code but it includes also linking functions to other files, making the whole algorithm complex.

In case of the not optimized source code the output signal is delayed in relation to the input signal by 1.756 s. Such a delay is very annoying and makes the work of the phone operator very difficult or even impossible during the conversation he or she can hear him- or herself after an unacceptably long delay.

5. Platform II with C/C++ programming

The second realization of the watermarking system (called platform II) was realized with the same hardware as platform I. In contrast, in the previous solution the authors decided not to use the Matlab/Simulink environment, but the Code Composer Studio ver. 3.3 and the C/C++ programming only. The C/C++ DSP software design is presented in Fig. 5.

During the design, all algorithms must be manually programmed in the C/C++ language but even in this case some steps like the configuration of the hardware modules can be semi-automatic or library supported. The optimization loop (dotted lines in Fig. 5) makes ity possible to reach a stable and efficient solution. This optimization loop is much easier to manage.

The genesis of the decision to prepare platform II was related to problems with making the system standalone under the Matlab/Simulink environment with a link to the Code Composer Studio library.

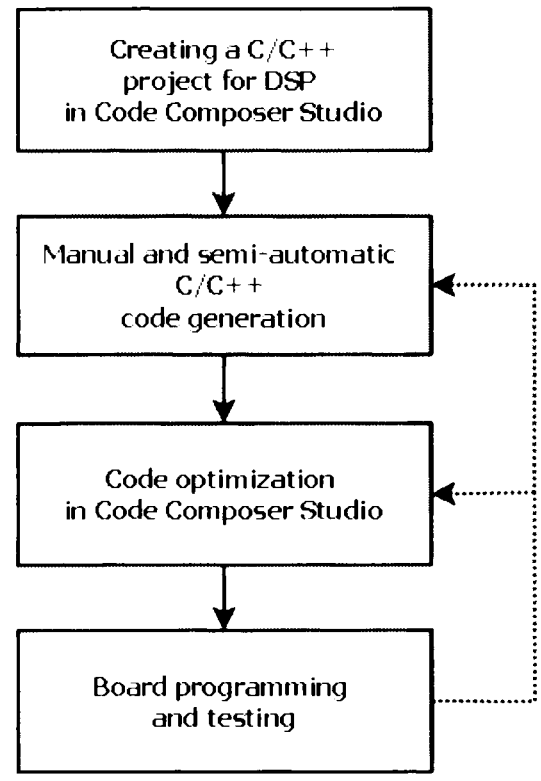


Fig. 5. Typical C/C++ DSP software design

The project realized under the Code Composer Studio includes the following set of files:

- DSP and peripheries configuration, e.g. PLL (phase-locked loop, for CPU clocking), audio codec, EMIF (external memory interface), interrupt vectors, memory map, etc.
- Source code main function calling DSP initialization functions and then performing the signal processing algorithms

This implementation includes 21 files and about 120 lines of the C/C++ code in the main function only. It is enough to realize the same application as in previous solution, i.e. the watermark inserting algorithm.

This implementation includes 21 files and about 120 lines of the C/C++ code in the main function only. It is enough to realize the same application as in previous solution, i.e. the watermark inserting algorithm.

First advantage of this is a size of the output file. It has 76 kB and can be fully loaded into the IRAM of the DSP core. Therefore, in this case, the standalone system can be realized with no problems. Nowadays, the C/C++ language programming is a standard in the processor and microcontroller programming. Thus for an experienced programmer it should be not a problem. The C/C++ offers the best compromise between the high level programming and the direct low level control of the hardware by the assembler.

In the presented case, due to relative low software complexity, putting the changes to the source code is very easy. The method used in the implementation of the algorithm reduced the delay between the input and the output to 32 ms only. It is equivalent to one frame of 256 samples with the sampling rate of 8000 samples per second.

Thus, platform II allows very fast processing with no hearable delay. Tthe main disadvantage is still the same size as in platform I (for details see Tab. 1).

6. Platform III with fixed-point processor and C/C++ programming

The third realization has been made using the fixed-point DSP TMS320C5505 embedded in the eZDSP USB STICK starter-kit [11] and Code Composer Studio ver. 4.0. The C55x DSP architecture offers a high performance and low power through the increased parallelism and focus on power savings [12]. A block diagram of platform III is presented in Fig. 6 and its realization in Fig. 7. The DSP board is very small (93×40 mm), low power, dedicated to mobile equipment. It allows to miniaturize the system.

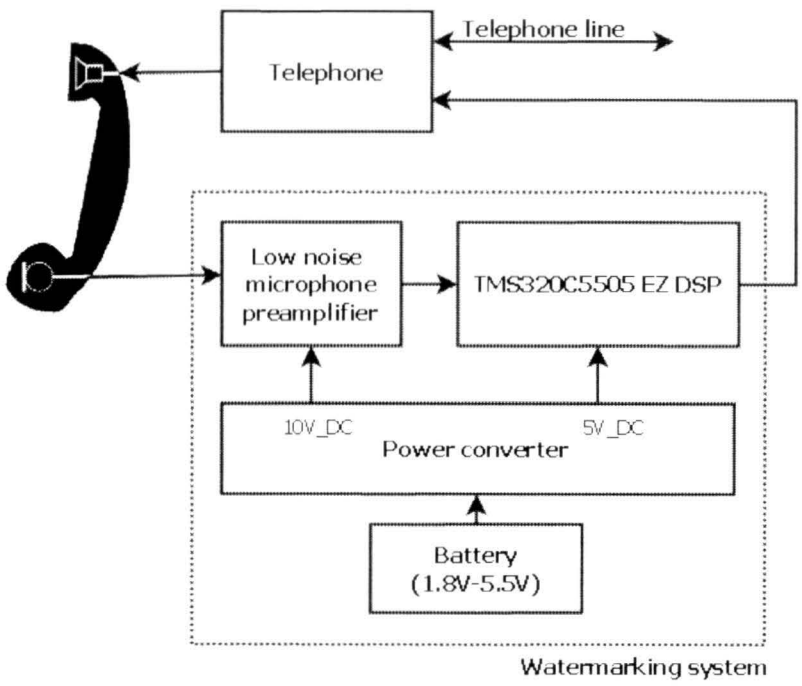


Fig. 6. A block diagram of the watermarking system (platform III)

Beside the DSP board, the platform has the second PCB (printed circuit board), designed and realized by the authors, which consists of the low voltage power converter and low noise microphone preamplifier (c.f. Fig. 5). The system is battery powered and therefore the separation module, that was required in platform I and II can be omitted. The third platform does not require an external power converter, as well as is much smaller than the previous solutions. This means that it can be applied in mobile equipment.

Internal, high efficiency low power converter produces 5V to supply the DSP and 10V to supply the analog preamplifier. The input power comes from batteries and can vary from 1.8 V to 5.5 V.

However, the codec in this platform has an electret microphone interface and offers variable amplifications of the input signal. The authors designed an additional low-noise microphone preamplifier, to separate the low voltage analog signal coming from the digital system part. It brings better signal-to-noise ratio.

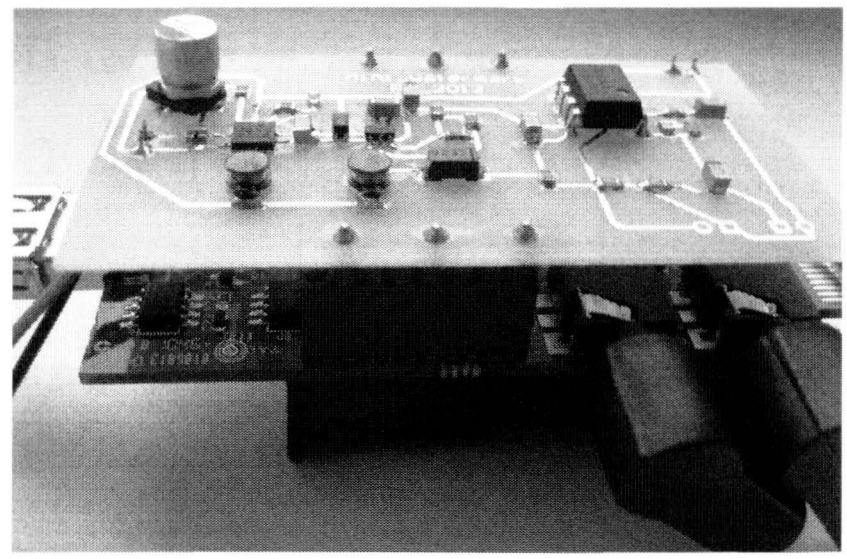


Fig. 7. Hardware of watermarking system (platform III)

The DSP software is similar to this in platform II but is adapted to the fixed-point arithmetic and different kind of architecture. As the TMS320C6713 allows to acquire and process 16-bit samples from input, the TMS320C5505 acquires 16-bit samples as well but should convert them to 32-bit variables in case of processing. Any multiplying of 16-bit values gives a 32-bit output, that has to be right shifted to shrink the format to 16-bit. An architecture of the floating-point unit allows to do this operation on 2-byte variables, however, the fixed point TMS320C5505 unit loses data when using variables of the same size (the C55x CPU provides two multiply-accumulate (MAC) units, each capable of 17-bit x 17-bit multiplication in a single cycle). The input samples are acquired in the interrupt routine, evoked by the audio codec, also as in the TMS320C6713. The line-in input does not prevent to connect the signal from the microphone in the case of ability to control the software gain by setting the proper bits in the audio codec registers. A size of the generated binary file is only 30 kB. It fits into the external non-volatile EEPROM (connected to EMIF by I2C interface) and during the booting process is uploaded into the internal memory DARAM in full, what allows proper work in the standalone mode [8]. In fact in the TMS320C5505 eZDSP USB STICK board there are no other possibilities than to fit all the code into the internal RAM, because, in opposite to platforms I and II, this board has no external RAM chip. Fig. 8. shows the memory map of the TMS320C5505.

7. Comparison of platforms

To summarize, Table 1 presents the main technical data and performance of the watermarking platforms I, II, and III. Beside the hardware specifications, some raw performance measurements are added. It can be noticed, that platforms II and III, programmed in plain C/C++ language are faster (produce lower delay), are more stable, easier to manage, and what is not obvious, they require less supply power than platform I, programmed in Matlab/Simulink. The most economical is platform III and it should be dedicated to the mobile applications.

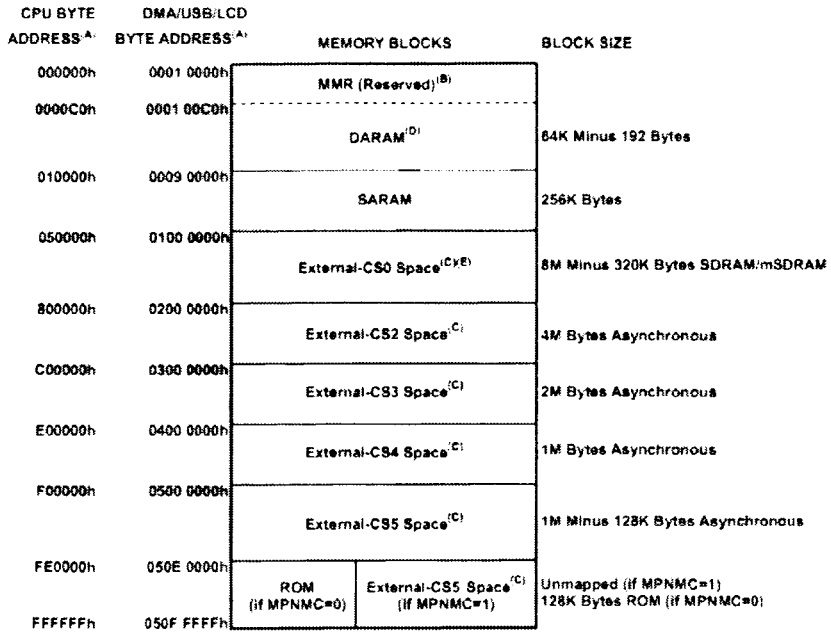


Fig. 8. Memory map of TMS320C5505 unit [11]

Table 1. Technical data and performance of watermarking platforms I, II, and III

Platform /Parameter	I	II	III
DSP platform and processor	TMS320 C6713 DSK	TMS320 C6713 DSK	TMS320 C5505 eZDSP USB STICK
Program memory	4Mb FLASH with parallel interface	4Mb FLASH with parallel interface	512Kb EEPROM SPI
External RAM	128Mb SDRAM	128Mb SDRAM	N.A.
Programming	Matlab /Simulink	C/C++	C/C++
Code size	185 kB	76 kB	30 kB
Audio codec	16-bit 8kS/s TLV320 AIC23B	16-bit 8kS/s TLV320 AIC23B	16-bit 8kS/s TLV320 AIC3204
Programming interface	USB	USB	USB (FT2232)
Signal delay	1.756 s	32 ms	32 ms
Power dissipation (run)	1.75 W(*)	1.4 W	0.4 W
DSP board dimensions	222x112 [mm]	222x112 [mm]	93x40 [mm]
External dimensions	85x196x310 [mm]	85x196x310 [mm]	50x55x110 [mm]
Power supply	5V DC (external power adaptor)	5V DC (external power adaptor)	1.8-5.5V DC e.g. battery 2x1.2V or 3.7V
Telephone connection	Galvanically separated	Galvanically separated	Direct
Additional functionalities	Hardware selector for various wiring of handset	Hardware selector for various wiring of handset	Low noise microphone preamplifier

(*) run under software debugger

8. Conclusions

In this paper the design of the DSP supported platforms for the real-time voice watermarking has been shown. Three presented solutions explained advantages, differences and problems during the fast prototyping such Matlab/Simulink way in comparison to manually designed C/C++ programming. The first one is much easier during the design but is also slower, requires more resources, and is less tractable in the meaning of advanced configuration e.g. FLASH memory booting for the mobile solutions.

In future the authors plan to prepare a more detailed comparison between the Matlab/Simulink based design and the manually tuned C/C++ programming. First rather raw measurements, presented in this paper, show that quickly designed platforms do not offer the optimal performance and require more hardware resources and more supply power.

References

[1] A. Dąbrowski, P. Pawłowski, R. Weychan, A. Meyer, M. Portalski, A. Chmielewska, T. Janiak, "Real-time watermarking of one side of telephone conversation for speaker recognition", Przegląd Elektrotechniczny (Electrical Review), Nr 6/2012, pp. 36 – 41.

[2] H. G. Stark, "Wavelets and Signal Processing", Springer, Germany, 2005.

[3] W.N. Chan, at al., "Use of vocal source features in speaker segmentation", 2006 IEEE Conference on Acoustics, Speech and Signal Processing, 2006, vol. 1, pp. 657 – 660.

[4] B. Ziółko, S. Manandhar, R.C. Wilson, "Phoneme segmentation of speech", 18th International Conference on Pattern Recognition, 2006, vol. 4, pp. 282 – 285.

[5] P. Meena, K. U. Rao, D. Ravishankar, "Real-time detection and analysis of PQ disturbances with DSP using Matlab embedded link to code composer studio", International Conference on Power Systems, ICPS '09, 2009, pp.1 5.

[6] Description of the environment Matlab/Simulink, http://www.mathworks.com/, 2012.

[7] TMS320C6000 Code Composer Studio Tutorial, Texas Instruments, spru301c.pdf, 2000.

[8] TMS320C6713 DSK technical reference, Texas Instruments, dsk6713_TechRef.pdf, 2003.

[9] TMS320C6000 Assembly Language Tools v 6.1 Texas Instruments, spru186q.pdf, 2008.

[10] TMS320C6000 Tools: Vector Table and Boot ROM Creation, Texas Instruments, spra544d.pdf, 1999.

[11] Technical Reference Manual of the TMS320VC5505 eZDSP USB Stick, Texas Instruments, usbstk5505_TechRef_revb.pdf, 2009.

[12] TMS320C5505 Fixed-Point Digital Signal Processor, Texas Instruments, sprs660e.pdf, 2012.

Influence of GSM Coding on Speaker Recognition Using Short Polish Sequences

Agnieszka Krzykowska, Tomasz Marciniak, Radosław Weychan, Adam Dąbrowski
Division of Signal Processing and Electronic Systems, Chair of Control and Systems Engineering
Poznań University of Technology
Poznań, Poland
tomasz.marciniak@put.poznan.pl

ABSTRACT — **The paper presents a comparison of speaker models used for fast speaker identification in short recordings of telephone conversations. The knowledge of the encoder type used during the transmission of speech allows to apply a model that takes specific characteristics of the encoder into account. This improves efficiency of the speaker recognition process. The influence of the following GSM encoders was tested: FR, HR, EFR and AMR. During the experimental research we used our database of short voice phrases that usually occur during emergency calls. This paper is based on our studies related to techniques for the GSM encoding detection and to algorithms for removing silence in the voice recordings.**

KEYWORDS — GSM, GMM, speaker identification

I. INTRODUCTION

Speaker recognition a is a new and attractive functionality, which often occurs e.g. in various types of call-center systems. An important factor influencing effectiveness of the speaker recognition (verification / identification) is e.g. the quality of transmission / recording of the speech signal. In case of the public switched telephone network (PSTN) and the typical PCM bitstream of 64 kbit/s (8-bit quantization with sampling rate equal to 8000 samples per second) the speaker verification performance is about 95 %. An intrinsic use of the speech codecs applied in mobile networks decreases efficiency of the speaker identification [1, 2]. It can currently be observed that most of the calls are performed using the mobile network. For example, in 2010, the conventional telephone network density (a number of lines per 100 inhabitants) in Poland was 21.6 only and is declining since 2004. In the same time the mobile network penetration rate (a number of SIM cards per 100 people) increased up to 123.4 [3].

The problem of building a correct model of a person to be verified by a biometric system is an issue that requires consideration of the specific acquisition and transformation of the signal. Knowing properties of the signal under test, we can compare it with patterns held in the dedicated database. A general idea of our speaker verification system is shown in Fig. 1. For each of the speakers we have developed models incorporating different types of speech coders.

Discussion of encoders used in the experimental studies is given in Chapter 2 followed by description of the detection of the encoder type, based, according to our previous studies [4], on the mel-frequency cepstral coefficient (MFCC) parameters and the mean square error (MSE).

A selection of the speaker during the verification stage is realized by means of the Gaussian mixture model (GMM) approach. The speech preprocessing includes an algorithm for automatic removal of silence in the speech signal sequence as discussed in [5].

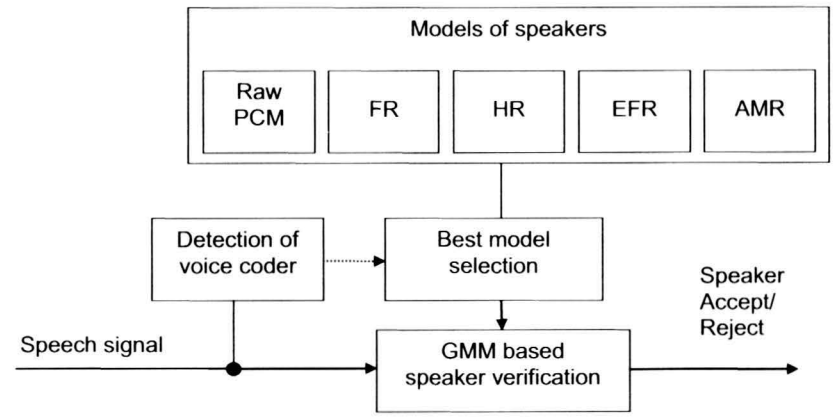


Figure 1. Speaker verification with automatic selection of speaker models

We assume that the analysis is based on the resynthesized speech. Direct application of the encoder parameters seems to be impractical according to [2], since such approach can be realized only in the systems implemented by the operators of the cellular network. Another assumption of our experiments is the use of the database of short speech utterances. Thus we assume that the detection of people should be very fast [6].

II. SPEECH CODERS

A. GSM speech encoders overview

As mentioned in the introduction, an important element of the speaker verification systems is selection of the proper comparison model, depending on the speech coding technique. The following encoders used in the mobile telephony were tested during our experimental studies:

- full rate (FR) encoder [7],
- enhanced full rate (EFR) encoder [8],
- half rate (HR) encoder [9],
- adaptive multi-rate (AMR) encoder [10].

These encoders can be divided into two groups, which use:

- full rate linear-prediction based on the analysis-synthesis with the RPE-LTP algorithm (*regular pulse excitation - long term prediction*) that generates a 13 kbps bitstream with 8000 samples per second sampling rate; every 40 samples are computed with the use of the previous 120 and split into frame, which consists of 160 samples, being equivalent to 20 ms of speech in the time domain
- CELP (*code excited linear prediction*) algorithm – enhanced full rate / adaptive multi-rate or half rate; these encoders use ACELP (*algebraic-code-excited linear predictive*) algorithm (both EFR and AMR) or VSELP (*vector-sum excited linear prediction*), respectively, generating bitstream of 4,75-12,2 kbps.

Implementation of AMR [11], EFR [12], and HR [13] encoders have been based on the official ANSI C code, adapted to the authors' for batch processing. An interesting fact is that the EFR implementation [12] cuts about 1024 final samples from every speech file – in this case all sequences have to be extended with additional 1024 zero values before the next transcoding. Implementation of the FR encoder has been realized with Matlab with the use of equations described in [7].

B. *SNR after GSM speech coding*

It is obvious that speech coders degrade signal quality. Errors between original and processed speech files can be estimated calculating SNR (signal to noise ratio). This value also gives us possibility to check how subsequent transcodings influence unprocessed speech. In [14] the authors proposed the following equation

$$SNR = 10\log\frac{\sum(S_{ORG})^2}{\sum(S_{ORG} - S_{CODED})^2} \quad (1)$$

This equation assumes that we can directly compare samples of the original and encoded sequence. However, it is impossible to calculate SNR in this way in case of a delay between the unprocessed and processed speech. In order to obtain proper SNR results, we have used the correlation function to find the mentioned delay and shift of the coded sequence. In [14] the authors used only linear-prediction based GSM algorithms, making SNR calculation much easier.

SNR values have been computed for our database of short sequences. It consists of 7200 speech files recorded with 22050 samples per second and downsampled to 8000 samples per second with the use of high order polyphase filters. The database has been transcoded four times by four main GSM encoders: full rate (FR), enhanced full rate (EFR), half rate (HR), and adaptive multi-rate (AMR). The last three encoders use adaptive and fixed codebooks. In this case, it is impossible to calculate SNR value even with the use of correlation function.

Fig. 2 shows the corresponding set of samples (after fitting with the correlation function) of the original speech and the transcoded once by the AMR coder. As it can be noticed, calculation of the SNR values sample by sample may give incorrect results if large differences occur between the corresponding samples. The EFR and HR encoders give very similar results. In the case of the FR encoder, which uses predictive algorithms, the corresponding transcoded frames are more similar to the original (using the correlation function gives better results than those presented in [14]).

Fig. 3 shows the transcoding effects that occur during the FR coding. The SNR values for subsequent tandems (multiple subsequent transcodings) are presented in Table I. Subsequent tandems affect the signal less and less but the SNR values decrease nonlinearly, as it was expected.

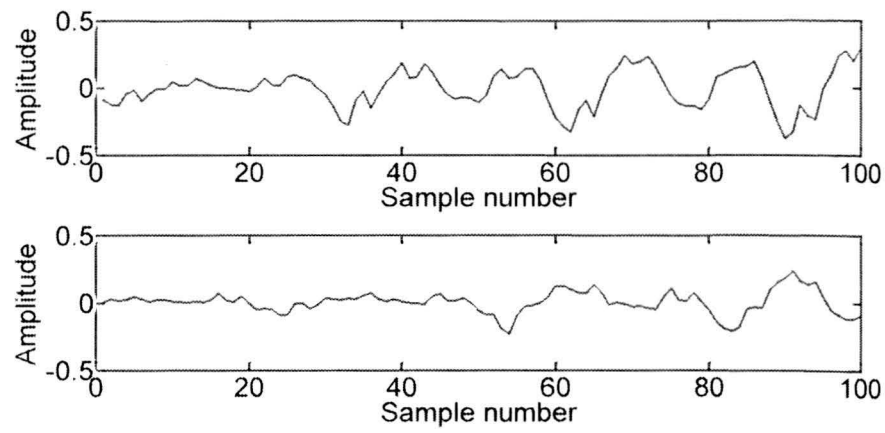


Figure 2. Corresponding subset of frames of original speech and transcoded once by AMR ecnoder

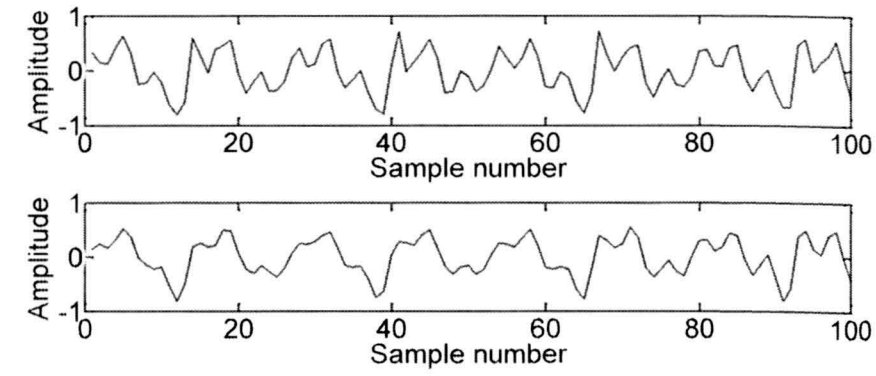


Figure 3. Corresponding subset of frames of original speech and transcoded once by FR ecnoder

TABLE I. SNR VALUES COMPUTED FOR TANDEMS IN CASE OF FULL RATE ENCODER

Encoder type	**SNR [dB]**			
	1 coding	2 codings	3 codings	4 codings
Full rate	31.7701	29.9596	28.4174	27.0173

III. ALGORITHMS AND DATABASE

A. *Speaker modeling method*

An extraction of the speaker features from a particular speech sequence starts with the division of the sampled signal into blocks of the length equivalent to 16 ms. Next, each block is multiplied by the Hamming window function, and the DFT (*discrete Fourier transform*) is calculated to finally receive 12 MFCC (*mel frequency cepstral coefficients*) by mel-scaling. Obtained feature vectors are used to train the model with the GMM. The expectation-maximization (EM) algorithm is used during this step.

B. *End-point detection – removing silence from speech*

Experiments described in [5, 15] proved that removing silence from speech significantly improves speaker recognition even if the modeled speech contains only several words. From four of the presented EPD (*end-point detection*) algorithms (both in their basic and extended forms) we chose two for further tests: the energy analysis and the Jang HOD (*high-order differences*) method (presented by Roger Jang in [16]). Table II summarizes the main ideas of the mentioned methods.

TABLE II. EPD METHODS USED FOR SREAKER RECOGNITION

Method	Short Description
Energy analysis	Calculation of energy
Jang HOD	Use of volume and high-order differences

In our application the described algorithms were enhanced by detection and removal of silence in the beginning, middle, and at the end of each sentence (extended version), therefore we are calling them "middle energy" and "middle Jang HOD" algorithms.

1) *Middle energy algorithm*

This EPD algorithm is based on the analysis of the signal energy, which can be computed from equation

$$E_i = \sum_{n=k_{ip}}^{k_{ik}} x^2(n) \qquad (2)$$

where i stands for the number of the window of the signal x, k_{ip} is the first sample of the i-th window, and k_{ik} is the last one. Number of samples in one window used to count the energy is 80, and the length of each window is 0.01 s. The offset of the window is equal to 0.001 s.

Selection of threshold parameters were tested on wave sequence: "Chciałbym zgłosić wypadek" (in English: "I would like to report an accident") which has 16 383 samples and lasts for 2.05 seconds. A mean value of the energy in this wave is 1.7578 with the minimal value very close to zero (0.00006) and the maximum value of 15.384. The energy threshold was chosen experimentally and is equal to 0.1. After removing the silence from the tested sentence the length of raw speech decreased to 1.04 second i.e. to 8335 samples. Mean energies of the voiced and silence parts are 3.6 and 0.016, respectively.

2) *Middle Jang HOD algorithm*

This end-point detection algorithm uses high-order differences of the speech signal to detect speech. Individual steps of the algorithm are as follows:

a) Computing volume (V) with the use of equation (3) and the absolute value of the sum of the j-order difference (H) with equation (4):

$$V_i = \sum_{n=k_{ip}}^{k_{ik}} |x(n)| \qquad (3)$$

$$H_i = \sum_{n=k_{ip}}^{k_{ik}} \frac{\Delta^j x(n)}{\Delta n^j}, \qquad (4)$$

where x describes the signal, n – the sample number, i stands for the number of the time window, k_{ip} and k_{ik} are the first and the last sample of the i-th window, respectively, Δ represents the differential. Next, the values of V and H are normalized.

b) Computing the VH curve from volume (V) and sum of obtained differentials (H):

$$VH = (V + H)/2 \qquad (5)$$

c) Computing threshold t for VH in order to determine the end-points. The threshold is defined as

$$t = VH_{\min} + (VH_{\max} - VH_{\min}) \times r \ . \qquad (6)$$

A default value of $r = 0.125$. A length of the time window to compute the volume is 0,016 s. A number of samples in a window is 128. Adjacent windows are not overlapping.

A number of differences (in our case $j = 4$) was chosen experimentally for the sentence mentioned in the description of the previous method. In this example the mean value of VH was 0.1757, while the range of values fluctuated between 0.0015 to 0.9611. The threshold value computed from equation (6) equals 0.0922. In this case the length of the voiced signal was 0.75 second i.e. 6022 samples. The mean VH values of the speech samples and the removed silence are 0.3762 and 0.0274, respectively.

C. *Database*

The database (described in [5]) used in our experiment consists of 6 sentences repeated 30 times by each of 40 speakers (males and females between 20 and 55 years of age). An average length of the recorded sentences is 1.6 second.

Recordings were realized during three sessions. Every speaker repeated each sentence 10 times at once. Time break between sessions was 1 to 6 weeks.

Each of 7200 samples was recorded in an anechoic chamber with the use of an omnidirectional condenser microphone. The sampling rate of the recorded samples was set to 22050 samples/second and to 16 bit resolution. Next, the recorded phrases were downsampled to 8000 samples/second to compare with the telephone quality speech. An average duration of speech sentences is about 1 s, thus during the experiment the features obtained from 5 random sequences were combined to create a model of each speaker.

IV. INFLUENCE OF GSM CODERS ON SPEAKER RECOGNITION

To determine how each coder influences the speaker recognition accuracy, we tested four GSM coders (AMR, EFR, FR, and HR) and the unprocessed (raw) speech in matched conditions. This means that the speaker model and the speaker sentence used for tests were taken from the same database, thus coded with the same coder.

As it can be seen in Fig. 4, coding of speech decreases accuracy of the speaker recognition, which was expected. However the results show that the FR coder apparently extracts some crucial speaker features because effectiveness of the speaker recognition system for speech encoded with this coder is increased in comparison to raw (unprocessed) speech. Other coders in the order from the best to the worst result are: EFR, AMR, and HR. EER (*equal error rate*) values for this experiment are presented in Table III in bold.

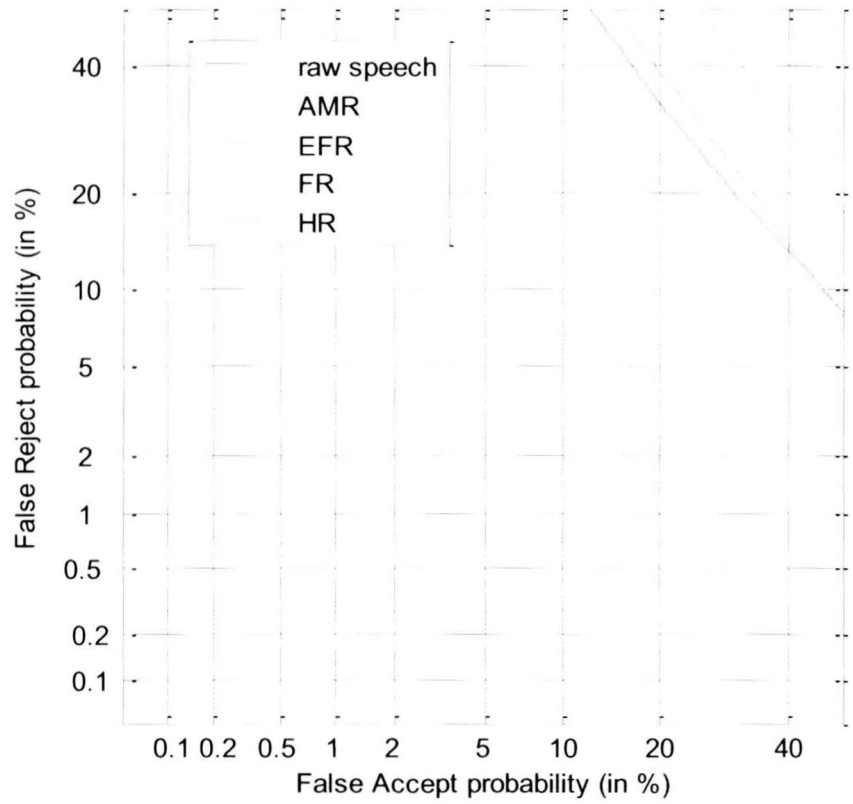


Figure 4. Influence of GSM coding on speaker recognition for matched conditions

V. INFLUENCE OF TANDEMING AND SILENCE REMOVAL ON SPEAKER RECOGNITION

In the next step of our experiment we tested the effect of multiple coding and decoding of speech on speaker recognition. For that, we used speaker models coded once with one of the coders, and test sequences coded one and four times with each coder.

Furthermore, we tested the influence of silence removal algorithms on already transcoded speech. Two methods described earlier were used and the effects are described below. Tables III to V show the results of the performed experiments both in matched and in mismatched conditions. Gray highlight shows the lowest value of EER for a given test sequence.

A. *Tandeming*

Every cellular phone call comes through the base station and the base station controller. There can be several switches between the sender and the receiver. Every operation of coding and decoding has impact on the signal quality. An idea of tandeming (transcoding the speech several times) is to check how it influences the speaker recognition accuracy.

Table III shows EER values for matched and mismatched conditions for four GSM coders. Speech samples used for the test and the training part were transcoded 1 and 4 times. It can be observed that in cases of AMR, EFR, and HR encoders EER value increases with the number of tandeming. An interesting fact is that for the FR encoder, recognition accuracy is higher than for the raw data even when transcoded 4 times. It can be observed in the highlighted fields in Table III, whose structure tends to be diagonal.

TABLE III. EER OF SPEAKER RECOGNITION FOR TRANSCODED SPEECH (IN %)

Test / Model	Raw speech	Speech transcoded 1 time				Speech transcoded 4 times			
		AMR	EFR	FR	HR	AMR	EFR	FR	HR
Raw speech	**25.7**	30.6	26.3	24.7	35.2	38.9	37.5	24.3	43.6
AMR	33.9	**33.0**	29.8	31.7	35.5	39.4	39.2	30.9	42.2
EFR	31.2	31.0	**27.3**	28.9	33.7	39.0	38.2	28.3	42.2
FR	27.2	30.4	26.5	**24.4**	34.0	37.9	35.5	23.2	43.4
HR	38.2	36.4	34.3	35.4	**35.3**	40.1	40.7	34.7	40.1

Fig. 5 presents plotted lines of values included in Table III. The abscissa axis is described with the tested sequence transcoded by a specific encoder, while the ordinate axis shows the EER value. Colors of the plots correspond to the used models. As it can be observed, the full rate encoder gives better results even than the raw data.

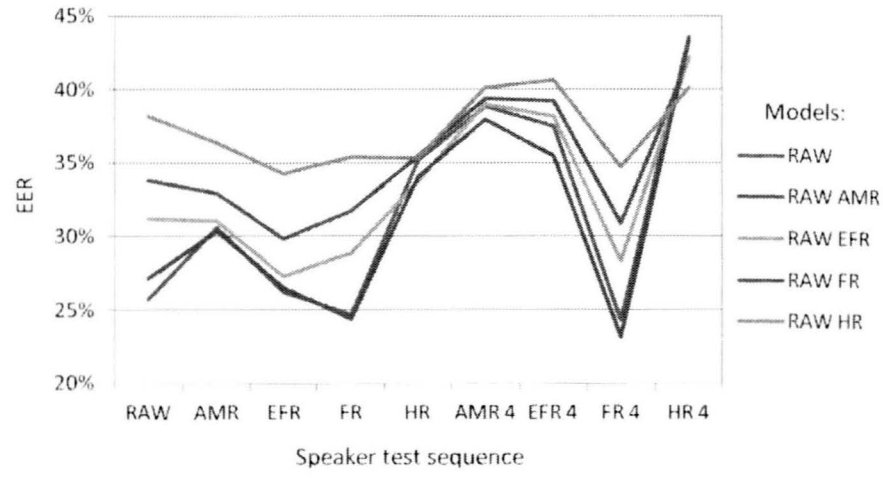


Figure 5. EER of speaker recognition system for transcoded speech

B. *End-point detection*

As it can be inferred from comparing Tables III, IV, and V removing silence has a significant influence on the speaker recognition system. Even for raw speech the results are better (an improvement from 26% to 17% and 16% for the middle energy and the middle Jang HOD algorithms, respectively). Though, both algorithms give quite similar accuracy.

1) *Middle energy algorithm*

Table IV shows results for matched and mismatched conditions with the use of four GSM coders.

Removing silence, both from the model and the tested speech, improved the recognition efficiency. The best results were obtained for the test speech coded only once with any coder and the model, which was adequate to the coder.

In case of tandeming for the tested speech coded with the EFR coder, the best results were acquired for the speaker models coded with the AMR and EFR coder. When voice of the verified speaker were coded 4 times with the FR coder the best accuracy can be observed for the speaker modeled after coding with the FR encoder. However, good results are also received for the unprocessed speech and the EFR model.

TABLE IV. EER OF SPEAKER RECOGNITION FOR TRANSCODED SPEECH WITH REMOVED SILENCE BY MIDDLE ENERGY METHOD (IN %)

Test / Model	Raw speech	Speech transcoded 1 time				Speech transcoded 4 times			
		AMR	EFR	FR	HR	AMR	EFR	FR	HR
Raw speech	**17.1**	18.9	17.8	18.7	27.5	29.3	22.3	19.3	44.2
AMR	19.7	18.1	17.7	19.7	23.3	**24.5**	**18.7**	20.9	43.3
EFR	18.3	**17.7**	**16.6**	18.3	23.4	26.7	19.3	19.3	43.1
FR	19.0	19.9	18.6	**16.9**	26.0	31.2	23.9	**17.3**	44.6
HR	25.5	23.0	22.8	24.6	**20.7**	**24.6**	20.9	25.1	**34.2**

Fig. 6 illustrates the obtained EER values. Designation ME stands for the model processed with the middle energy algorithm, and each line represents speaker model tested against the speaker test sequence listed on the horizontal axis.

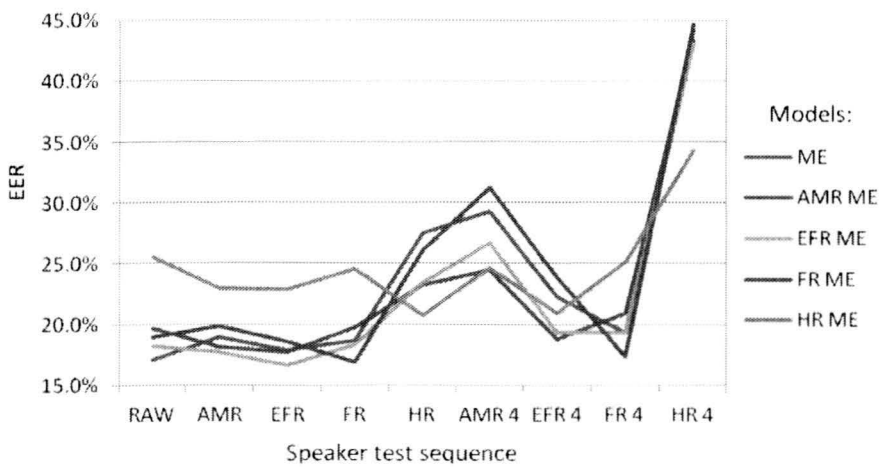


Figure 6. EER of speaker recognition system for transcoded speech with removed silence with use of middle energy method

Every model (apart from the one coded with the HR coder) gives similar results for testing speech coded only once. HR coded test speech stands out here (the only exception is when the model is coded also with the HR encoder). For testing the speech transcoded 4 times with the HR coder all results are worse, though the HR coded model is better in this case.

2) Middle Jang HOD algorithm

As it can be inferred from Table V and Fig. 7, the middle Jang HOD algorithm, used for removing unvoiced parts of speech, shows significant improvement of the speaker recognition accuracy. It brings a very similar effect to that of the previously studied middle energy method. The differences between them are almost unnoticeable.

It can be also observed that in case when the tested speech is transcoded 4 times with the HR coder no EPD algorithm can improve recognition accuracy (except when the model is also coded with the HR encoder).

As in the previous method, the best results can be observed when the model and the speech of the verified speaker are coded with the same coder (this conclusion is certainly valid in both cases of 1 and of 4 transcodings).

In Fig. 7 the abbreviated designation MJH stands for the speaker model processed with the middle Jang HOD algorithm. The corresponding graph shows recognition accuracy (EER) in reference to the transcoded test speech sequence.

TABLE V. EER OF SPEAKER RECOGNITION FOR TRANSCODED SPEECH WITH REMOVED SILENCE BY MIDDLE JANG HOD METHOD (IN %)

Test / Model	Raw speech	Speech transcoded 1 time				Speech transcoded 4 times			
		AMR	EFR	FR	HR	AMR	EFR	FR	HR
Raw speech	**16.4**	18.7	17.5	18.6	27.5	29.9	22.5	19.7	44.7
AMR	19.1	**17.7**	17.5	19.4	24.1	**25.5**	**19.4**	20.8	42.4
EFR	17.9	**17.6**	**16.5**	18.2	24.0	27.2	**19.8**	19.7	44.0
FR	18.8	20.0	18.7	**16.2**	26.2	31.6	24.0	**16.9**	44.2
HR	25.5	22.7	22.5	23.9	**20.6**	**25.7**	21.4	25.2	**34.1**

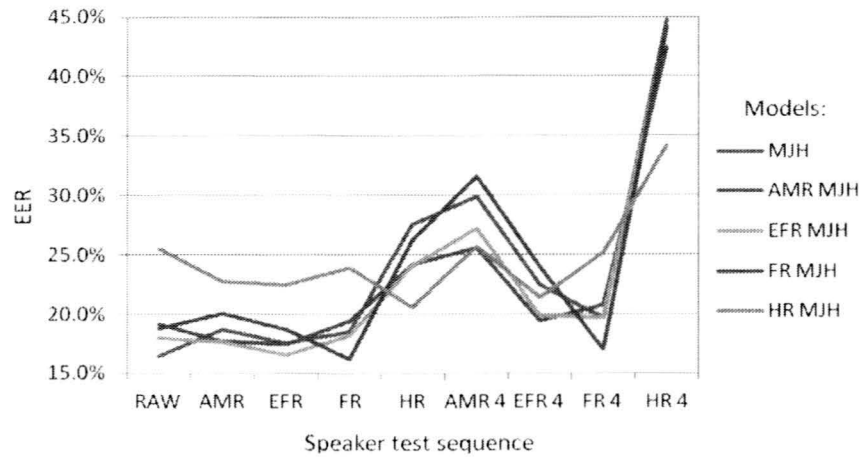


Figure 7. EER of speaker recognition system for transcoded speech with removed silence with use of middle Jang HOD method

VI. TIME OF PROCESSING

Presented experiments were performed in the Matlab environment on the Linux operating system. The hardware used for tests involved Xenon Quad Core E5405 2.0 GHz CPU and 2 GB RAM.

Table VI and Fig. 8 present average time of computations for each part of the speaker recognition system, i.e.: silence removal, feature extraction, creation of model, and verification of one particular speaker with one model. The time unit is milliseconds.

TABLE VI. PROCESSING TIME (IN [MS])

	Raw speech	Speech with removed silence	
		Middle energy	Middle Jang HOD
Silence removal	---	54.0	236.2
Feature extraction	2.7	1.7	1.6
Model creation	181.7	65.6	61.4
Verification of speaker	1.8	1.3	1.2
Overall	**186.2**	**122.6**	**300.4**

The tested algorithms are of various complexities, thus the time consumption of their performance is adequate – removing silence with middle Jang HOD algorithm takes more time than computation of the signal energy.

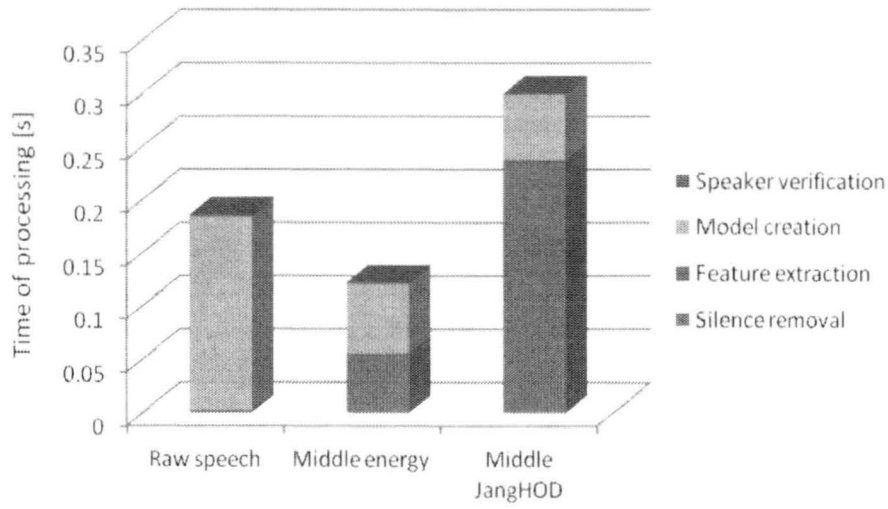


Figure 8. Time of processing raw speech and speech with removed silence

Even though detection of silence is time consuming, speech containing only voiced parts enables to create the speaker model faster (because of the much smaller amount of the samples left). Therefore, in case of the middle energy algorithm the sum of the time spent on detecting silence and creating the speaker model is shorter than for the unprocessed speech. Only when the middle Jang HOD method is used the computational time is longer.

VII. CONCLUSIONS

The article shows that selection of the appropriate model significantly improves verification of the speaker. In general, the studied verification is the most plausible when the base models of the speakers include the FR encoder parameters. Such a situation occurs if the EPD algorithms are not used.

Removal of silence improves verification, but it is closely linked with the dedicated model. Removing of silence from the transcoded speech leaves the voiced speech parts only, which is significant for coding features of the particular speaker. The computation time for cutting silence by the middle energy method is shorter than for the unprocessed speech. Thus, applying this algorithm greatly enhances the speaker recognition correctness and can be used for creating better speaker identification algorithms.

REFERENCES

[1] Vuppala, A.K.; Sreenivasa Rao, K.; Chakrabarti, S.; , "Effect of speech coding on speaker identification," *India Conference (INDICON), 2010 Annual IEEE* , pp.1-4, 17-19 Dec. 2010

[2] Grassi, S.; Besacier, L.; Dufaux, A.; Ansorge, M.; Pellandini, F., "Influence of GSM Speech Coding on the Performance of Text-Independent Speaker Recognition", Proc. Of European Signal Processing Conference (EUSIPCO) 2000, pp. 437-440, Tampere, Finland,September 4-8, 2000

[3] COMMUNICATION - ACTIVITY RESULTS IN 2010, Central Statistical Office, Warsaw, Poland, 2011

[4] Dabrowski, A.; Drgas, S.; Marciniak, T.; , "Detection of GSM speech coding for telephone call classification and automatic speaker recognition," *Signals and Electronic Systems, 2008. ICSES '08. International Conference on* , pp.415-418, 14-17 Sept. 2008

[5] Dąbrowski, A.; Marciniak, T.; Krzykowska, A.; Weychan, R., "Influence of silence removal on speaker recognition based on short Polish sequences", *Proc. Of SIGNAL PROCESSING SPA'2011*, Poland Section, Chapter Circuits and Systems IEEE, pp. 159-163, September, 29-30th 2011

[6] Marciniak, T.; Weychan, R.; Drgas, S.; Dabrowski, A.; Krzykowska, A.; "Speaker recognition based on short polish sequences," *Signal Processing Algorithms, Architectures, Arrangements, and Applications Conference Proceedings (SPA), 2010* , pp. 95-98, 23-25 Sept. 2010

[7] FR description http://www.3gpp.org/ftp/Specs/archive/06_series/06.10/0610-820.zip

[8] EFR description http://www.3gpp.org/ftp/Specs/archive/06_series/06.60/0660-801.zip

[9] HR description http://www.3gpp.org/ftp/Specs/archive/06_series/06.20/0620-801.zip

[10] AMR description http://www.3gpp.org/ftp/Specs/archive/06_series/06.90/0690-721.zip

[11] AMR ANSI C source http://www.3gpp.org/ftp/Specs/html-info/26073.htm

[12] HR ANSI C source http://www.3gpp.org/ftp/Specs/html-info/46006.htm

[13] EFR ANSI C source http://www.3gpp.org/ftp/Specs/html-info/46053.htm

[14] Lilly, B.T.; Paliwal, K.K.;, "Effect of speech coders on speech recognition performance," *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on* , vol.4, pp. 2344-2347 vol.4, 3-6 Oct 1996

[15] Marciniak, T.; Krzykowska, A.; Weychan, R., "Speaker recognition based on telephone quality short Polish sequences with removed silence", *Przegląd Elektrotechniczny*, vol. 06/2012, pp. 42-46, June 2012

[16] Jyh-Shing Roger Jang, "Audio and Signal Processing and Recognition", available at the links for on-line courses at the author's homepage at http://www.cs.nthu.edu.tw/~jang

This work was prepared within the INDECT and DS projects.

Feature Generator for Speaker Recognition Using the Fusion of Cepstral and Melcepstral Parameters

Ewelina Majda, Andrzej P. Dobrowolski
Military University of Technology, Faculty of Electronics
Warsaw, Poland
EMajda@wat.edu.pl, ADobrowolski@wat.edu.pl

ABSTRACT — **The paper examines issues related to the determination of features distinctive to sound generators using a fusion of mel-cepstral and cepstral information for an Automatic Speaker Recognition (ASR) system. Parameterization of the speech signal is crucial to these systems, as the chosen parameterization dictates the effectiveness of the diagnosis and the speed of the system. The authors focus on the use of speech signal processing methods that consider the phenomena connected with the speech generation process while searching for features related to speech characteristics. A well-designed system should be able to extract speech characteristics independent of the linguistic content of the speech. The research presented in this paper focuses primarily on multicriteria optimization of the generator parameters based on a set of descriptors derived from the fusion of the melcepstra and cepstra and also considers the use of additional features. Finally, the evaluation of the results was based on the analysis of the Fisher coefficients and main components of a set of descriptors.**

KEYWORDS — ***automatic speaker recognition, feature generator, features selection, PCA***

I. INTRODUCTION

Speech is a natural and one of the most effective means of communication between humans. Automatic speaker recognition has a variety of technical solutions. The common feature of these solutions is the processing of the speech signal using a digital device to extract the required information for specific applications. In general, the procedure for identification of persons can be divided into three phases. The pre-processing block is responsible for receiving the signal from the microphone and its initial processing, involving also quality enhancement of the signal. The second stage involves analysis of the speech signal, in order to obtain parameters carrying information about the individual characteristics of the voice of the speaker, regardless on the speech content. The final stage of classification is based on similarity of obtained parameters of the signal sample to their previously acquired references (in the so-called teaching process) for particular persons [1].

II. PARAMETRIZATION OF THE SPEECH SIGNAL

The most important step performed by a speaker recognition system is to generate an appropriate set of numerical descriptors that best characterize the recognized speakers. The purpose of the parameterization of the speech signal for ASR is transformation of the temporary input process to obtain the smallest possible number of descriptors containing information relevant to the speaker, while minimizing their sensitivity to variation in the signal that is irrelevant for ASR. The selection of these descriptors was guided by the analysis presented above, the process of speech generation and the searching for elements related to the individual characteristics..

A. Research method

The recordings of the acoustic signal were made at the *Institute of Electronic Systems Faculty of Electronics WAT* using a *Monacor DM-500* dynamic microphone, a computer sound card and *Matlab* software. During the test, the distance from speaker's mouth to the microphone was approximately 10 cm. Additionally, the microphone was equipped with a shield, which prevents distortion associated with whistling sounds (s-, s-, cz-, ć-) and explosive sounds (p-, b-, t).

The phonetic material included a variety of phonetic text produced from fragments of a typical dialogue, that may occur for example at the airport or at the border and a sublime and fun expression In this way the authors have received series of speeches significantly different in terms of intonation of a similar nature to the expected expression, people, who will be identified in the future. The total length of all recordings was about 4 minutes. The study group consisted of 38 men and 10 women. The signals were sampled at a frequency of 22.050 Hz with 16-bit amplitude resolution and recording of a single channel (mono). A database containing the speaker identifier has been created from recorded expressions and the corresponding samples of the acoustic signal.

B. Pre-processing

Pre-processing of the speech signal is a very important step in data processing because it precedes the introduction of the signal to the features generator and has a fundamental effect on the quality of the speaker identification process. The main purpose of pre-processing the speech signal is to ensure the greatest independence of the acoustic signals from the settings of the recording equipment. In the pre-processing stage, the filtration - lowpass type II Chebyshew filter: (4.6 kHz, -3dB),

This work is supported by the Polish Ministry of Science and Higher Education in the years 2010-2012 as a development project.

(5kHz, -6 dB), and normalization is performed to eliminate differences between different frequency characteristics and the measurement circuits. The return loss, noise and disturbance were bypassed by assuming no distortion and signal noise issues. However, these issues will be taken into account in further research.

C. *Feature generator*

Speech signals have a variable frequency structure in time. Thus, the parameterization is subject to successive signal fragments and not the signal as a whole. Sections of the divided signal are called frames. Generally, the frame length Δt is related to the shift (leap) τ, as follows

$$\tau = \frac{1}{3}\Delta t \qquad (1)$$

One of the first tasks for the authors was to establish the basic parameters of the features generator, which is the frame length. Durations of the individual phonetic units are different and depend on the speaker. Units consisting of voiced sounds are characterized by a duration ranging from 10 ms to even 200 ms. The range of variation is substantial, so the decision concerning the choice of the frame length is extremely important in the ASR. Studies on the optimization of the individual parameters features generator are presented in the chapter 4. Framing of a signal causes discontinuities in the processed signal, which is associated with frequency leakage. To minimize this effect, the signal of each frame must be windowed by multiplication with an appropriate window function. Windowing results in the smoothing of the discontinuity and the removal of false spectral components. The Hamming window with good properties has been applied by the authors.

$$w(n) = 0{,}54 - 0{,}46\cos\left(\frac{2\pi n}{N}\right); \quad 0 \le n \le N \qquad (2)$$

Because important information related to the speaker is contained only in the voiced parts of speech, only the "voiced frame" should be considered during the analysis. Voiced fragments are characterized by the occurrence of regular peaks (with the period of a basic tone). The voiceless parts are similar to an aperiodic signal. In the system, the classification of the speech signal into voiced or unvoiced parts is performed using the autocorrelation function. To verify if a sound is voiced, the second global maximum is determined and checks one level (the first maximum is in zero). If the level is higher than a reference value p_v, then this part is considered to be voiced; otherwise, it is deemed voiceless. Determining the optimal level of p_v is another step in the optimization described in Chapter 4.

By choosing representative frames for each speaker, an additional constraint was applied by the authors - the detection of speaker activity. During the registration, parts of the signal in which the speaker is not active occurred. Use of another parameter responsible for the rejection of frames without speech is aimed at eliminating the silence of the recording and the rejection of frames that are potential noise, which can cause erroneous feature extraction. The authors decided to base on the power of the variable component (the variance of the signal by threshold. The establishment of an additional parameter, the power level, is the next task to optimize.

III. GENERATION OF NUMERICAL DESCRIPTORS

ASR systems must numerically express speech signals using well- defined descriptors that characterize the speaker. The features utilized by these systems must be describe the human voice as a means of distinguishing between different speakers. After appropriate feature selection, a feature vector will be created and used as the basis for classification (identification and verification). The authors decided to search for distinctive features by considering phenomena related to the internal structure of the source of the speech signal. In practice, the feature generation is based on three cepstral analysis techniques. In each method, a set of preliminary pre-selection characteristics is created, and then all generated descriptors are fused.

A. *Cepstral features*

The primary and basic form in which the speech signal is registered is its temporal form. The time domain is not the most appropriate to perform further operations because the speech signal is characterized by significant redundancy. Further analysis of the speech signal is more efficiently performed in the frequency domain. A significant number of computer methods are based on spectral analysis, which replaces the convolution in the time domain with the product of the spectrum of stimulation (laryngeal) and the spectrum of transmittance of the voice track (variable in step of articulation) in the frequency domain [2]. However, as the amplitude of the speech signal is modulated by the passing through the voice path, it is preferable to calculate in the first phase the logarithm of the spectrum. This way, the multiplicative relationship between the stimulation and the voice path is re-placed by an additive relationship, which greatly simplifies the subsequent separation of the two components. The reasoning presented above leads directly to homomorphic processing methods, in particular to the concept of cepstrum [4, 5].

The amplitude spectrum of the speech is usually determined using a Fast Fourier Transform. The signal is composed of a rapidly changing factor (arising from the stimulation) and a slowly changing one (arising from the current construction of the organ of articulation) that modulates the amplitude of successive pulses resulting from the stimulation. Interpretation of the spectrum amplitude logarithm is similar, but the slowly changing component is not multiplied by the amplitudes of individual pulses from stimulation. In-stead, the slowly changing component is added to the amplitude of the individual pulses. The calculation of the spectrum of such signals shows that the low-frequency waveforms associated with the transmittance of the voice path are close to zero on the pseudo-time axis, and pulses associated with laryngeal sound begin approximately at the laryngeal signal period and repeat periodically. Information related to the voice path transmittance is focused near zero time, and therefore, one

should look for concise information on what is being said in this area. On the other hand, for the time period above the laryngeal sound. Because the laryngeal sound is closely connected to anatomy of the larynx and glottis, it is a good carrier of individual information A thorough analysis led us to conclude that the characteristic descriptors include the fundamental frequency F_{av} (Descriptor 1), corresponding to the inverse of the first maximum of the cepstrum, and the values of the 4 successive maxima of the cepstrum normalized by the value of the first maximum [4].

B. *Mel cepstral features*

The most popular method of parameterizing speech signals is to use the MFCCs (*Mel-Frequency Cepstrum Coefficients*). This method is based on sub-band analysis of the signal using bandpass filters equally spaced in the mel scale of frequency. A major feature of this transformation is the conversion of the spectrum to a linear scale, which accounts for the nonlinear perception of sound frequency by humans and significantly reduces the size of the data. A general diagram of this method is shown in Fig. 1 [3].

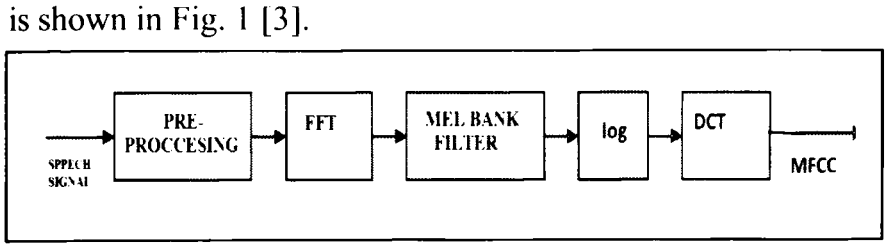


Figure 1. Diagram of the calculation of coefficients MFCC

The first step in determining the MFCCs is to calculate the discrete spectrum of the signal's representative frames using the *Fast Fourier Transform* (*FFT*). The spectrum is then filtered using characteristics inspired by the physiology of the human ear. The response of the human ear depends nonlinearly on the frequency of the received sound. The details of speech are more easily detectable in the low frequency range (up to approximately 1 kHz) than in the higher frequencies of the audible spectrum. It follows that the accuracy is reduced at higher frequencies, and a wider bandwidth is required to compensate for this loss of accuracy. The mel scale was determined empirically by the following process[3]:

$$f[\text{mel}] = 1127 \ln\left(1 + \frac{f[\text{Hz}]}{700}\right) \tag{3}$$

To perform the non-linear transformation of the spectrum, a set of filters spaced evenly in the nonlinear mel scale has been created. Triangular-shaped filters are defined in the frequency domain, which allows their responses to be computed using the sum of the products of the module spectra and the triangular function. The logarithm of the outputs of all filters is then taken, and the result is then subjected to the discrete cosine transform. The resulting vector of MFCCs has a length equal to the number of bands.

Thirty filters were applied during the MFCC generation, providing 30 distinct coefficients, ie 30 filters used in the band from zero to half the sampling frequency.Determining which of the MFCC features are representative of the pronounced sound and which are representative of the speaker is a difficult task. Features that are related to the linguistic content of the speech should not be considered and, as described above, the cepstral reconvolution technique should only consider features above a certain threshold. The authors applied an initial pre-selection of relevant features and reduced the length of the MFCC vector to 7 while minimizing any loss in the vector's representativeness. The results were checked using Principal Component Analysis (PCA). This method was used because of the large initial dimension of the preliminary vector of MFCC features. Display 30 - dimensional vector of MFCC features on plane, enabled the efficient initial pre-selection of features relevant to the modeled of feature generator.

C. *Weighted cepstral features*

The authors, inspired by the idea of the MFCC method, attempted to extend the feature vector to include other features defined in the cepstrum by using sub-band bleeder filters. The proposed algorithm does not produce the same peaks at their expected positions; rather, it sums the amplitudes of all of the relevant bands with certain weights. To optimize the system, the optimal characteristics of the filter (weighting function) and the optimal widths of the bands must be selected. The rectangular function was found to be optimal. The second through the fifth cepstral maxima represent the 4 relevant weighted cepstral features and are normalized to the sum of the amplitudes received in the first band, which corresponds to the fundamental frequency.

At the feature generation step, 16 numerical descriptors are defined to differentiate speakers C_1-C_{16}. These descriptors include the fundamental frequency F_{av} (C_1), corresponding to the inverse of the first maximum of the cepstrum; four weighted cepstral features (C_2-C_5); the 4 successive normed maxima of the ordered cepstrum (C_6-C_9) and seven mel-cepstral features (C_{10}-C_{16}). Each set of features for each speaker was averaged over a set of representative frames.

IV. MULTICRITERIA SYSTEM OPTIMIZATION

The previous section showed a general diagram of the designed system. Depending on what function the system is to fulfill (recognizing the content of speech or the identity of the speaker), the optimal parameters of the system must be chosen with a consideration of the procedure of features extraction and the registration mode. The authors had the task to optimize the system based on four basic parameters: the length of the frame (Δt) and its shift (τ), the threshold of voiced frame (p_v) and the level of power (p_p).

Due to the wide ranges of changes of all the optimized parameters, the authors decided arbitrarily to make an initial choice of the value of the parameters based on the coefficient of significance that Fisher defined in the following function

$$F_{ij}(f) = \frac{c_i - c_j}{\sigma_i + \sigma_j} \tag{4}$$

The quantities c_i, c_j and σ_i, σ_j denote the sample mean values and the sample standard deviations of features for classes i and j, respectively [8].

The Fisher coefficients of significance were determined for nine descriptors based on the eight classes consisting of four women and four men. The even partition of men and women was not accidental. Note that the value of the descriptor may have high discriminative power between women but much less for men. Thus, the Fisher coefficient of significance was categorized into three subclasses: *Women*, *Men* and the subclass of *All*. Because the number of classes is more than two, the Fisher coefficient of significance was calculated for all pairs and was subsequently summed (the total Fisher coefficient of significance). In the first stage, the parameter to optimize was the frame length (Δt). The results are illustrated in Fig. 2.

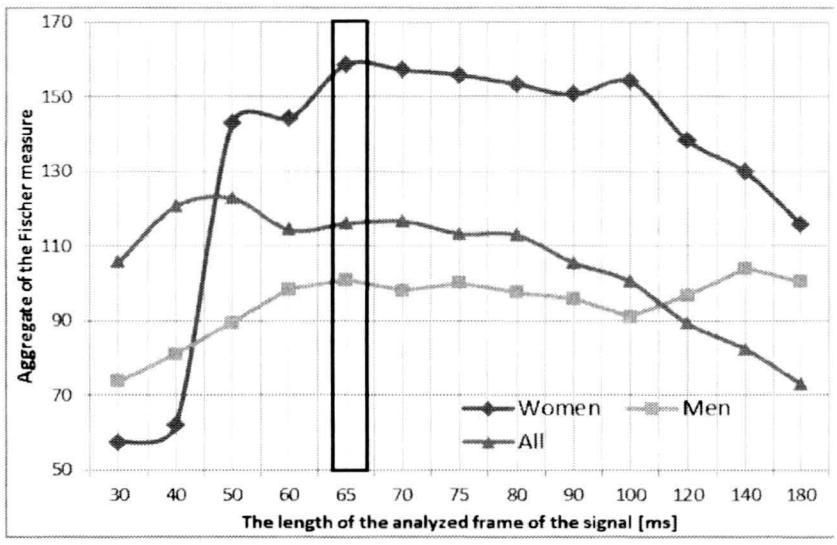


Figure 2. Aggregate of Fisher measure for each subclass depending on the length of the analyzed frame of the signal

It is clear from the graph that for a small frame length (30-40 ms) Fisher coefficients are small. The strong growth begins approximately 50 ms, and for the frame lengths exceeding 90 ms, the value of coefficients in subclasses Women and All are significantly decreased. Thus, a frame length ranging from 60 to 80 ms was chosen. Note that there is no such frame length for which the Fisher coefficient of significance reaches a maximum in all three subclasses. Thus, we attempted to compromise. Finally, a frame length of 65 ms was chosen. To validate the choice, several series of detailed studies have been performed. These studies confirmed that the optimal frame length was $\Delta t = 65$ ms.

Another parameter to optimize was the shift with which the frame will move along the analyzed speech signal. To solve the problem, note that a smaller step value yields a larger number of frames, which translates into a longer calculation time. We attempted to seek the shift value of the frame run in parallel with the optimization of the two other parameters (p_v, and p_p). Due to the large amount of information contained in the input data (11-dimensional vectors of features), we decided to optimize based on the analysis of the main components (*Principal Component Analysis – PCA*). The essence of the PCA method is the transformation of a large amount of information contained in the mutually correlated input data into a set of statistically independent components arranged according to their validity. The PCA step was one of the most laborious research stages. The work relied on the observation of the change of position of the feature vectors for a speaker on the PCA_1/PCA_2 plane. The research was based on three sets of 8 speakers.

According to the literature [4], calculating the fundamental frequency by the cepstral method is less accurate but more robust than the autocorrelation method, especially for an extremely noisy speech signal. To achieve greater stability of the descriptors, we decided to introduce an additional constraint used in selecting the correct frames. The fundamental frequency will be compared based on the autocorrelation function and the cepstrum. Finally, we decided that if differences between the values of the fundamental frequency frame by these two methods differ by more than 15%, the frame will be rejected automatically and will not be involved in the generation of descriptors. The set of optimized parameters for the features generator of 15-second segments of voice are shown in Tab. 1.

TABLE I. OPTIMIZED PARAMETERS OF THE FEATURE GENERATOR

Parameter		***Value***
Frame length	Δt	65 ms
Shift frame	τ	16 ms
Voiced level	p_v	10%
Power level	p_p	20%
Level of differences in the fundamental frequency	p_f	20%

V. FEATURE SELECTION

The set of descriptors defined at the stage of features generation are the maximum set of distinctive features. These descriptors can be used in automatic pattern recognition systems that represent the tested object. The maximum set of features has been shown to often not lead to the best results because they may have different impact on the pattern recognition. Some of the features resemble noise, thereby reducing the recognition efficacy. Some of the features are strongly correlated with the others, thereby adversely impacting the quality of classification by dominating over others and dampening their beneficial effects[6]. The important element is thus the assessment of the quality of each feature and selection of the best set of features on which the classification will be performed

Two strategies can be used to study the quality of these features. The first strategy is to test each feature regardless of the method of classification (the so-called filtering features) and assess their ability to differentiate the speakers without considering the specific classifier. The second strategy is to select the features based on the characteristics of the classifier [6]. The authors decided to filter the characteristics because

a final decision regarding the specific classifier has not yet been made.

The authors applied a pre-selection based on the Fisher method to extract the numerical properties of the various parameters characterizing the speakers. According to the assumptions of the Fisher method, a large aggregate value coefficient of significance indicates a good potential separation between classes. In contrast, a small value means that the feature values belonging to both classes are scattered and potentially intermingled with each other, thereby disqualifying one as a diagnostic feature. The total Fisher coefficients of significance of each descriptor are shown in Fig. 3.

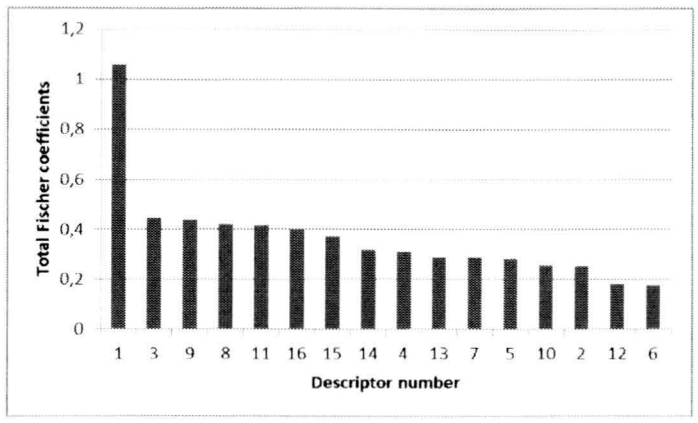


Figure 3. Total Fisher coefficients of significance of each descriptor.

It is difficult to clearly determine the optimal set of features carrying the important differentiating information on the basis of these results. Another solution may be offered by analyzing the distribution of the two principal components found using the *PCA* analysis when including different numbers of significant descriptors. It is important to maintain proper normalization when using this method so that the values of all features are expressed at similar numerical magnitudes. This standardization can be performed in many different ways. In this study, the features were normalized using the maximum value of each descriptor. Regardless of the total discriminant value of each feature, when building the automatic classification system, it is worth checking the discriminative power of the descriptors employed. However, it is known that the feature discriminative ability may change when used in co-operation with the others. Some features (even the worst ones) can be mutually enhanced, thereby raising their discriminative ability. We performed that analysis by following changes in the positions of the each vectors defining the speaker on the PCA_1/PCA_2 plane. Fig. 4 shows two examples of this distribution. Fig. 4.a applies to the full set of features, and Fig 4.b shows the results using only the 11 best features.

Based on the Fisher coefficients of significance of each descriptor and of the observed changes in the feature vectors based on the PCA transformation, the optimal 16-dimensional feature vector *VP*, called the *Voice Print*, has been determined. For each speaker were made averaging of selected set of features based on the 15-second excerpts recorded speech, taking into account only the correct frame selected at the pre-processing (among others discarded voiceless fragments and fragments of silence). As a result of applied the feature generator were obtained for each speaker eight voiceprint of the 15-second fragments expression, distributed evenly over the range from zero to 4 minutes. It is described using the following relations:

$$
\begin{aligned}
&F_{av} = \frac{1}{N}\sum_{j=1}^{N} F_j, \\
&S_i = \frac{1}{N}\sum_{j=1}^{N} s_{i,j}, \quad M_i = \frac{1}{N}\sum_{j=1}^{N} m_{i,j}, \quad MC_i = \frac{1}{N}\sum_{j=1}^{N} mc_{i,j}, \\
&\mathbf{VP} = [vp_1,\ldots,vp_{11}] = \\
&= \left[F_{av}, \frac{S_2}{S_1}, \frac{S_5}{S_1}, \frac{M_2}{M_1}, \frac{M_3}{M_1}, \frac{M_5}{M_1}, MC_{11}, MC_{12}, MC_{13}, MC_{15}, MC_{17}\right]
\end{aligned}
\tag{5}
$$

where: N – number of correct frames,

F_j – fundamental frequency of the j-th frame,

$s_{i,j}$ – sum of value of the real cepstrum surrounded by the i-th maximum for the j-th frames (equivalent to the mean value in subband),

$m_{i,j}$ – value of the i-th maximum of the real cepstrum for the j-th frames,

$mc_{i,j}$ – i-th coefficients of mel-frequency cepstrum (*MFCC*) for j-th frames.

Detection of individual peaks was carried out on a searched around the maximum values predicted peaks set on the the fundamental frequency.

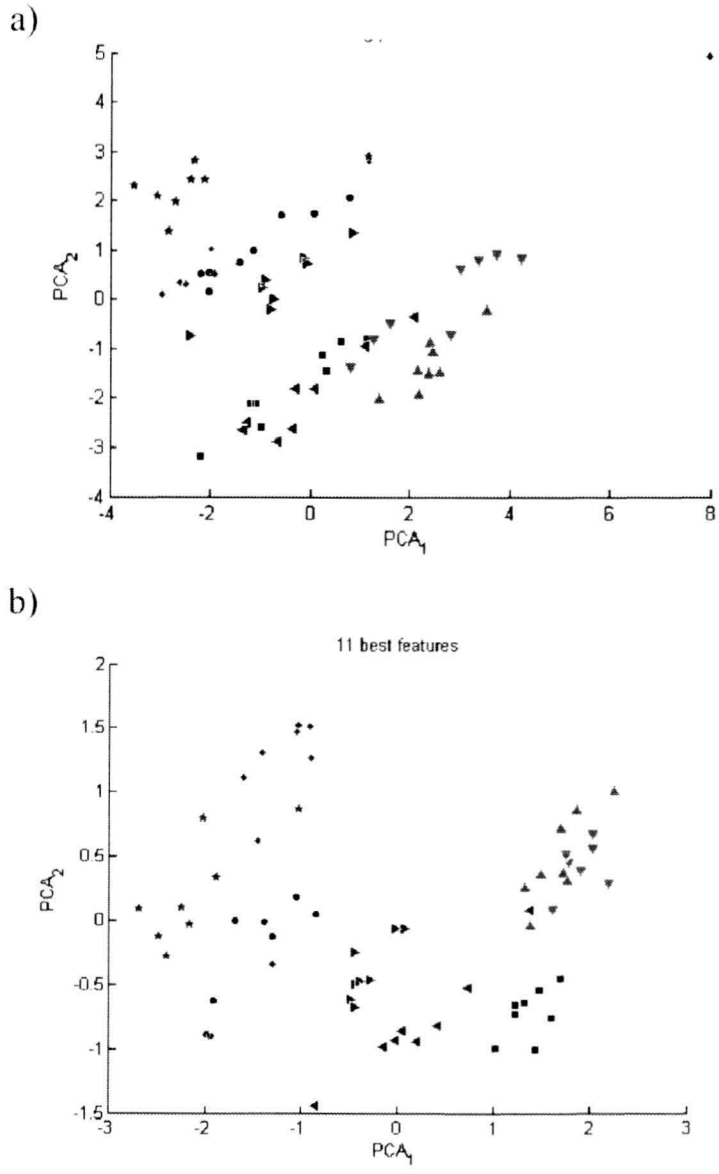


Figure 4. The distribution of the datafocused on the two principal components in terms of; a) all features, b) the 11 best features

VI. RESULTS OF STUDY

As a result of the multi-criteria optimization of parameters and selection of descriptors, we have obtained the final model of the features generator for the proposed speaker recognition system. Examples of the results of the PCA transformation are shown in Fig. 5. Individual results relate to the three separable sets of speakers (each set includes 8 speakers), treated as a representative group of 48 of personal database of speakers. PCA had to be split into separate set, due to the necessity of correct the visual evaluation of the results. Each speaker is represented by eight separate 11-dimensional Voice Prints. Note the analysis presented in Figure 6c. It shows the set consist of equal numbers of women and men (the men are grouped by the ellipse). Other results of the analysis presented in Figure 6a and 6b, have been developed mainly based on male voices.

The main advantage of the PCA transformation is an opportunity to observe the distribution of individual feature vectors on the plane despite the fact that the original feature vector is 11-dimensional. The vectors were obtained for each speaker using the optimized features generator. This generator enables an initial classification of the different speakers. Note that for each speaker we obtained reproducible results, despite the large diversity of recorded speech (dialogue, voice serious and humorous speech).

VII. CONCLUSION

The stage of parameterization of the signal is very important because the incorrect results of this stage cannot be corrected in further stages. The conducted experiments have allowed for the optimization of the model of features generator in the proposed speaker recognition system. A multi-criteria optimization of the selected parameters and the selection of the descriptors were performed. It is worth noting that the selection of the features affects the optimization, so the two processes must be repeated to obtain the optimal solution.

In the authors' assessment, the obtained results are very promising, in the context of having to go to the last stage the designed system, which is the identification. The statements of each speaker are concentrated in separate regions, and additional tests that considered more of the principal components confirm the conclusions presented. The analyses demonstrate that the Voice Print exhibits resistance to the linguistic content of the speech. The final stage of the speaker recognition process is the classification step, and the authors are currently studying this stage. The present analysis demonstrates that a classification according to the method of nearest mean would be sufficient, but the high complexity of the speech signal requires a set of several classifiers applied simultaneously; possible classifiers include k-means classifiers and linear and nonlinear SVM networks. The main objective of the paper will be to select the optimal set of classifiers that provide the minimum classification error.

The authors are also working on methods to reduce the time required by the system to recognize the speaker. The presented results utilize speech segments with a length of 15 seconds. Preliminary results show that there is a real possibility to reduce the speech segment needed to classify the speaker by two or three times. Undoubtedly, reducing this length to five seconds would represent a great advantage of this system, as it would allow speaker classification virtually in real-time.

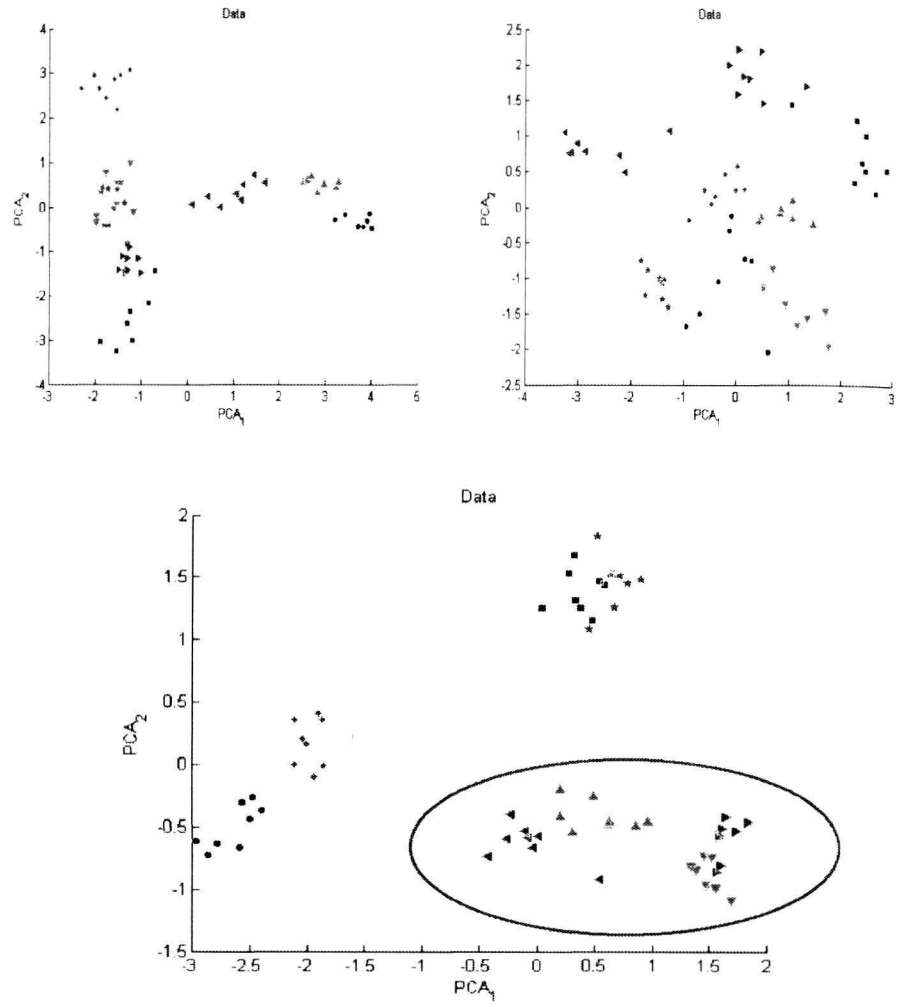


Figure 5. PCA for sets 1, 2 and 3 of the speakers. On the last figure men awarded the ellipse.

REFERENCES

[1] S. Furui, *Recent advantages in speaker recognition*, Pattern Recognition Letters 18, 1997, pp. 859-1872

[2] A. V. Oppenheim, R. W. Schafer, *From Frequency to Queferency: A history of the cesptrum,* IEEE Signal Processing Magazine, 2004 pp.95-106

[3] S.K. Kopparapu, M. Laxminarayana, *Choice of Mel Filter Bank in Computing MFCC a resamples Speech,* 10 th International Conference on Information Science, Signal processing and their Applications, Malaysia 2010, pp. 121-124

[4] A. Dobrowolski, E. Majda, *Cepstral analysis in the speakers recognition systems*, 15th IEEE SPA Conference, pp. 85-90, Poznań 2011

[5] A. Dobrowolski, E. Majda, *Application of homomorphic methods of speech signal processing in speakers recognition system*, Electrical Review, R. 88 NR 6/2012

[6] M. Kruk, S. Osowski, R. Koktysz, *Recognition of Colon Cells Using Ensemble of Classifiers*, International Conference on Neural networks, Orlando, Floria, USA, 2007.

[7] Z. Zhao, J. Zhang, J. Tian, Y. Lou, *An effective identification method for speaker recognition based on PCA and double VQ,* Proceedings of the Eight Conference on Machine Learning and Cybernetics, Baoding, 2009, pp. 1686-1689

[8] T. W. Anderson, *R. A. Fisher and multivariate analysis*, Statistical Science, vol. 11(1), 1996, pp. 20-34

Sample Rate Conversion with Fluctuating Resampling Ratio

Marek Blok
Faculty of Electronics, Telecommunications and Informatics
Gdańsk University of Technology
11/12 G. Narutowicza Street, 80-233 Gdańsk, Poland
e-mail: mblok@eti.pg.gda.pl

Piotr Drózda
ADVA Optical Networking Sp. z o.o
35/37 Śląska Street, 81-310 Gdynia, Poland
e-mail: PDrozda@advaoptical.com

ABSTRACT — **In this paper a sample rate conversion with continuously changing resampling ratio has been presented. The proposed implementation is based on variable fractional delay filter implemented using a Farrow structure. It have been demonstrated that using the proposed approach instantaneous resampling ratio can be freely changed. This allows for simulation of audio recored on magnetic tape with nonuniform velocity as well as removal of these distortions using the same algorithm.**

I. INTRODUCTION

A huge number of sample rate standards [1] available today create a demand on development of sample rate conversion (SRC) algorithms [2]–[4]. The digital resampling algorithm replaces digital-analog conversion followed by analog signal sampling allowing for more flexible implementations. A common example used to demonstrate usefulness of such algorithms is the conversion between compact disc (CD) with $F_s = 44.1$ kHz (multimedia standard) and digital tape (DAT) with $F_s = 48$ kHz (communications standard) [3], [5]. In this paper we present SRC implementation based on variable fractional delay (VFD) filter. Using this algorithm we not only can implement an arbitrary constant resampling ratio but fluctuating changes in resampling ratio can be readily implemented as well. Based on this tool we propose a novel application for resampling algorithm which is correction of signals with unintentional nonuniform sampling. For example correction of old recordings with distortions resulting from nonuniform velocity of the media [6]–[8] which is typically addressed with interpolation techniques [9]. On the other hand, with correctly sampled signal we might simulate such distortions using the same VFD SRC algorithm. In both cases, we make use of the fact that nonuniformly sampled signal when reconstructed with uniform sampling changes its pitch inversely proportionally to the sample rate changes.

II. VFD FILTER

The SRC algorithm investigated in this paper is based on FD filters and its performance depends on the design method used to calculate coefficients of FD filter. In this paper to the approximate ideal FD filter the FIR FD filter with the frequency response

$$H_N(f) = \sum_{n=0}^{N-1} h[n] \exp(-j2\pi f n) \tag{1}$$

is used where $h[n]$ is the impulse response of the length N. The designer tries to find the coefficients of this impulse response which offer the best performance of FD filter, which is usually evaluated using frequency domain error function [10]

$$E(f) = H_N(f) - H_{id}(f) \tag{2}$$

where the ideal frequency response $H_{id}(f)$ of the FD filter with total delay τ_d is defined by the following formula [10]

$$H_{id}(f) = \exp(-j2\pi f \tau_d), \quad f \in [-0.5, 0.5) \tag{3}$$

which corresponds to the ideal impulse response

$$h_{id}[n] = \mathrm{sinc}(n - \tau_d) \tag{4}$$

Because of the causality requirement, high performance FD filters are characterized with nonzero integer delay $D = \mathrm{round}(\tau_d)$, which for FIR filters is usually selected close to the bulk delay $\tau_N = (N-1)/2$. With those two delays defined, we receive the following formula for the total delay

$$\tau_d = D + d = \tau_N + \varepsilon \tag{5}$$

where $d \in [-0.5, 0.5)$ is the fractional delay and ε is the net delay.

There are several design methods offering optimal FD filters with maximally flat (MF), least squares (LS) and minimax being the most popular methods. The optimality criteria for each of these methods is based on complex approximation error (2). For MF filters approximation error and its $N-1$ derivatives must be equal to zero. The MF filter offers excellent performance but only around zero frequency. Differently, the LS and minimax filters allow the designer to specify the approximation band $f \in [0, fa]$ in which the error is minimized. The LS FD filter [10] has the energy of error minimized

$$E_{\mathrm{LS}}(f_a) = 2 \int_0^{f_a} |E(f)|^2 \, df \tag{6}$$

while the minimax FD filter [10], [11] has minimized peak error (PE)

$$E_{\mathrm{PE}}(f_a) = \max_{f \in [0, f_a]} |E(f)| \tag{7}$$

in the approximation band.

The coefficients of the impulse response, vector $\mathbf{h}$, of all these optimal filters, MF, LS and minimax, with fractional delay d can be found solving the following matrix equation [10]

$$\mathbf{Ph} = \mathbf{p} \tag{8}$$

where the coefficients of matrix $\mathbf{P}$ and vector $\mathbf{p}$ depend on the optimization criteria. For MF filter matrix $\mathbf{P}$ is a Vandermonde matrix

$$\mathbf{P}_{k+1,n+1} = n^k \tag{9}$$

and vector $\mathbf{p}$ has elements

$$\mathbf{p}_{1,k+1} = \tau_d{}^k \tag{10}$$

where $k, n = 0, 1, ..., N-1$. To find the LS filter we need only to change coefficients of matrix $\mathbf{P}$

$$\mathbf{P}_{k+1,n+1} = f_a \operatorname{sinc} f_a(n-k) \tag{11}$$

and column vector $\mathbf{p}$

$$\mathbf{p}_{1,k+1} = f_a \operatorname{sinc} f_a(k-\tau_d) \tag{12}$$

A minor modification is needed for minimax filters. First a set of $N+1$ frequency points f_k, called extremal points, must be found using recursive complex Remez algorithm [12]. Then coefficients of matrix $\mathbf{P}$ and vector $\mathbf{p}$ can be computed using the following formulas

$$\mathbf{P}_{k+1,n+1} = \cos(2\pi f_k n) - \sin(2\pi f_k n) \tag{13a}$$
$$\mathbf{P}_{k+1,N+1} = (-1)^k \tag{13b}$$

and

$$\mathbf{p}_{1,k+1} = \cos(2\pi f_k \tau_d) - \sin(2\pi f_k \tau_d) \tag{14}$$

where $k = 1, 2, ..., N+1$ and $n = 1, 2, ..., N$. In this case vector $\mathbf{h}$ has one additional element with magnitude equal to peak approximation error (7).

Since VFD filter needs to be able to change its delay for each output sample the high numerical costs related to solving matrix equation (8) become a significant problem. The most popular solution is the Farrow structure [11], [13]–[15]. The idea of this structure is to approximate each sample of the impulse response with a separate polynomial of the order q dependent on fractional delay d

$$h[n] = \sum_{m=0}^{q} c_m[n] d^m \tag{15}$$

Now, the output samples of the FD filter can be expressed with the following formula

$$y[n] = \sum_{k=0}^{N-1} h[n]x[n-k] = \sum_{m=0}^{q} y_m[n] d^m \tag{16}$$

where

$$y_m[n] = \sum_{k=0}^{N-1} c_m[k]x[n-k] \tag{17}$$

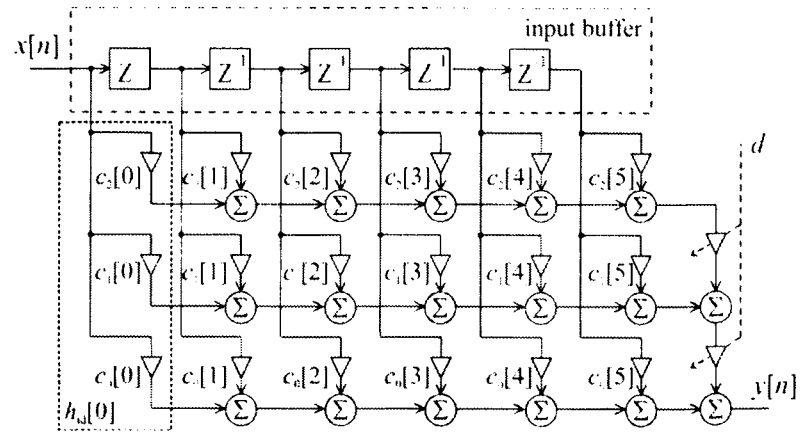


Fig. 1. Farrow structure of the order $q = 2$ implementing VFD filter of the length $N = 5$.

Formulas (16) and (17) define the Farrow structure presented in Fig. 1 where each row coefficients implement separate filter with impulse response $c_m[n]$.

To find coefficients $c_m[n]$ we need to compute only a few impulse responses of FD filters for fractional delays d uniformly spread in range $[-0.5, 0.5]$. In practice, to compute approximation polynomials it is enough to use just $q+1$ impulse responses. The polynomials of order q equal to 6 or 7 offer performance adequate for high quality FD filters with approximation error about -100dB [11].

III. Sample Rate Conversion Using VFD Filter

Classic three rate sample rate conversion algorithm is presented in Fig. 2. Input signal samples with sample rate F_{s1} is up-sampled by means of insertion of $L-1$ zeros between each pair of consecutive input samples. Next, at the intermediate sample rate, the lowpass interpolation filter with upper frequency $f_u = \min(0.5/L, 0.5/M)$ prevents the aliasing and removes spectral images resulting from zeroinserting. At last stage, the sample rate is reduced to the desired one by means of decimation, only every M-th sample remains in the output signal.

This approach, because of its simplicity, is well suited for simple cases requiring constant sample ratio with small factors L and M. In other cases the intermediate sampling ratio is very high and the passband of the interpolation filter becomes extremely narrow. In the result numerical costs increase drastically and, what's more important, the interpolation filter becomes very difficult to design. Therefore for arbitrary resampling ratios the VFD filter is used (Fig. 3).

As we can see in Fig. 4, the output sample $y[m]$ can be interpreted as the closest input sample delayed by a fraction of sampling period $d[m]$. This fractional delay can be computed

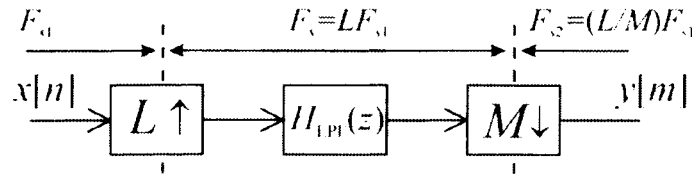


Fig. 2. Classic three-rate sample rate conversion algorithm by rational factor L/M.

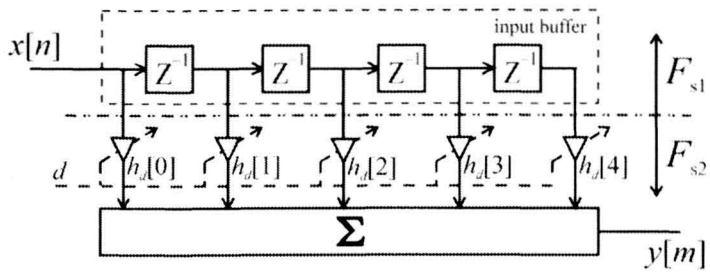


Fig. 3. General VFD filter structure applied to SRC.

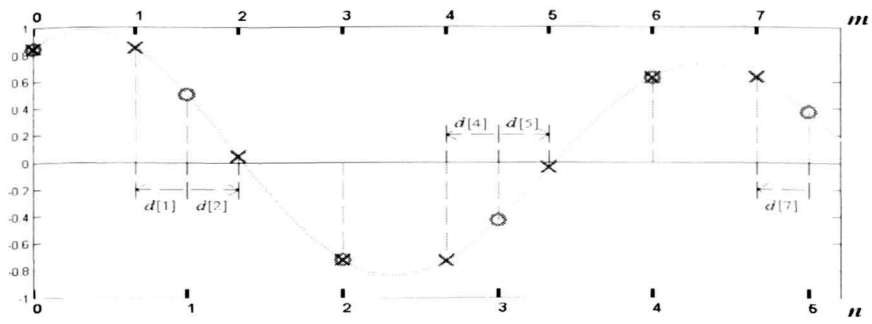


Fig. 4. Illustration of sample rate conversion by $L/M = 2/3$. Input samples $x[n]$ – o, output samples $y[m]$ – x.

using the following recursive formula [16]

$$d[m] = d[m-1] - r[m] + \Delta n[m] \tag{18}$$

where the reversal of instantaneous resampling ratio

$$r[m] = F_{s1}[m]/F_{s2}[m] = T_{s2}[m]/T_{s1}[m] \tag{19}$$

and $\Delta n[m]$ is a number of new samples required in input buffer to compute the next output sample

$$\Delta n[m] = \text{round}(r[m] - d[m-1]) \tag{20}$$

For rational resampling ratio (Fig. 2)

$$r[m] = M/L \tag{21}$$

sequences $d[m]$ and $\Delta n[m]$ are periodic with period L, but in general with VFD filter the resampling ratio can be an arbitrary positive number and can change in time. Nevertheless, if we want to avoid resampled signal distortions then $r[m]$ must be limited by the instantaneous signal oversampling ratio.

With the above two parameters defined ((18) and (20)) the resampling algorithm is following (Fig. 3):

1) start with $d[0] = 0$ and $\Delta n[0] = 0$,
2) wait for $\Delta n[m]$ new samples in input buffer,
3) find output sample $y[m]$ delayed by $d[m]$ using FD filter,
4) calculate $\Delta n[m]$ and $d[m]$ for next m and go back to step 2.

For every output sample the resampling algorithm requires different fractional delay (18). This means that for each output sample we need to compute a new impulse response of the FD filter. For rational resampling ratio L/M we actually need only L impulse responses which can be stored in look-up-table (LUT) [3] but when ratio is arbitrary and additionally changing in time, the filters needed in resampling cannot be specified beforehand and must be computed during runtime which can be done readily using Farrow structure presented in the previous section.

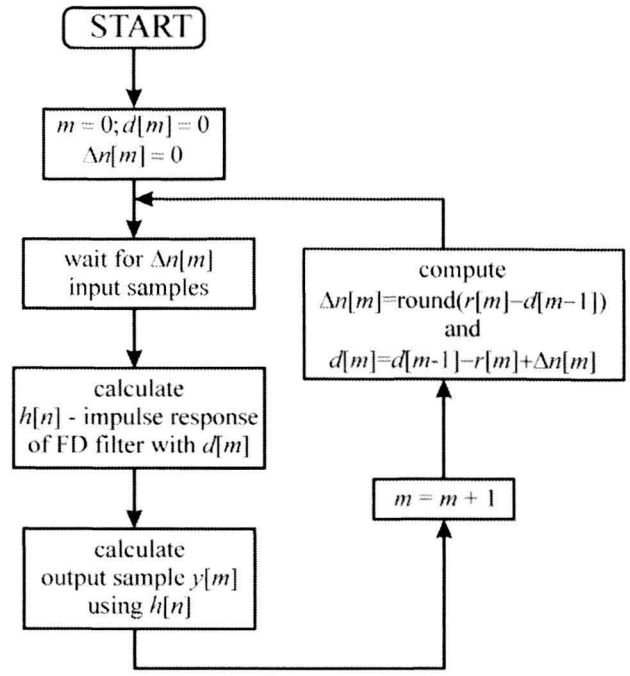


Fig. 5. Diagram of SRC algorithm based on VFD filter.

IV. FD FILTER DESIGN FOR SRC

If the best performance of SRC algorithm must be achieved then the optimal FD filters like minimax or LS [10], [12], [16] with approximation band specified by the designer seem to be the best option. Especially that with the Farrow structure the complexity of the FD filter design algorithm does not affect the runtime VFD filter implementation since all structure coefficients can computed beforehand.

Nevertheless, we must notice that it is not sufficient to know just the errors of FD filters to assess the performance of the SRC algorithm based on these filters. It can observed that the SRC algorithm based on FD filters (Fig. 3 and 5) is equivalent to the classic approach (Fig. 2) [16]. We only need to replace the interpolation filter in the classic approach with the overall filter which is composed of FD filters used in resampling [16]. This can be done only for rational resampling ratios but the conclusions can be readily adapted to arbitrary resampling ratios.

To obtain the impulse response of the overall filter [16] we need to interleave impulse responses $h_{d[m]}[n]$ of FD filters with fractional delays $d[m]$

$$h_o[m + nL] = h_{d[m]}[n]; \quad m = 0, 1, \ldots, L-1 \tag{22}$$

with delays $d[m]$ arranged in decreasing order

$$d[m-1] = d[m] + 1/L; \quad m = 1, \ldots, L-1 \tag{23}$$

Using the overall filter (22) we can readily analyze distortions introduced by the SRC algorithm based on FD filters since this filter must fulfill the same requirements as the interpolation filter in the classic approach (Fig. 2).

In Fig. 6 we can see overall filter obtained for minimax FD filters with upper frequency of approximation band $f_a = 0.4$. The problem with the SRC based on optimal filters is that the overall filter demonstrates large lobes in stopband which may result in aliasing when input signal has components above f_a. Moreover, the transition band location of the overall filter

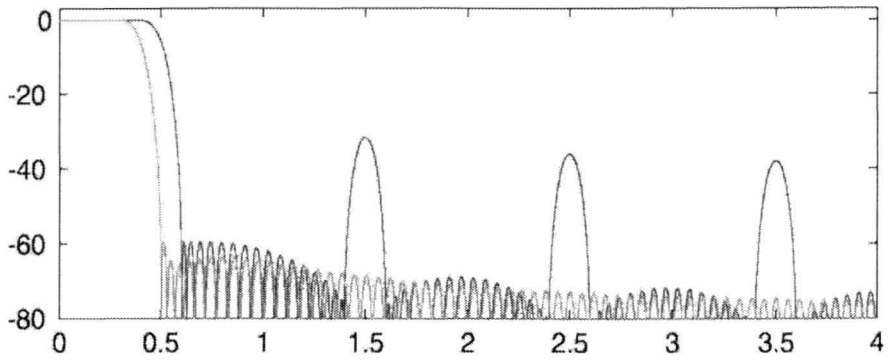


Fig. 6. Comparison of overall filters composed of minimax FD filters (blue) and composed of FD filters designed using offset window method (red) of the length $N = 17$ and with $f_a = 0.4$ and offsetting MF filter of the length $N_{\text{off}} = 5$ for $L = 9$.

cannot be manipulated. These disadvantages of optimal FD filter can be overcome with the use of offset window method [17]–[20]. Using filters designed with offset window method for window extracted from minimax filter, the large lobes in stopband are eliminated and the transition band can be readily shifted [20] (Fig. 6). The numerical cost of the VFD implementation do not change since, as we have observed, the Farrow structure of the same order can be used also in this case.

Although the overall filters in Fig. 6 are composed of just $L = 9$ filters their properties are maintained for any resampling ratio. An excellent testing signal for the SRC algorithm is the constant amplitude LFM chirp, which has been used in Fig. 7. We can see there that the signal converted using minimax FD filters demonstrates high distortions when input signal frequency exceeds f_a. On the other hand, the signal converted using FD filters designed with offset window has no components caused by nonlinear distortions larger than -60 dB which is directly related to the peak error (7) of the worst FD filter used in the SRC algorithm [16].

V. Processing with fluctuating resampling ratio

In this section we will demonstrate SRC with changing ratio. The first simple yet spectacular example is the conversion of sinusoidal signal with constant frequency F_{in} into chirp signal with linear instantaneous frequency

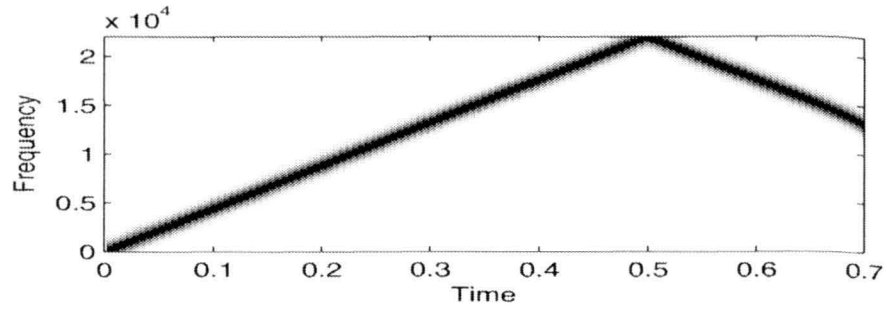
$$F_{out}[m] = F_0 + m\Delta F \tag{24}$$

where F_0 is the initial output sample rate and ΔF is the chirp rate. In this example the input sample rate is constant and the output sample rate must change in such a way that the oversampling ratio will change linearly according to output time index m.

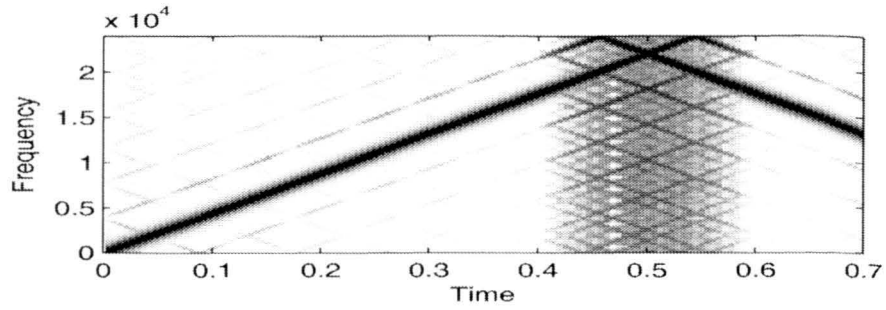
$$F_{s2}[m] = F_{in}\overline{F_{s2}}/F_{out}[m] \tag{25}$$

where $\overline{F_{s2}}$ is the constant output sample ratio required for resampled signal to observe chirp with assumed instantaneous frequency (24). Since in this example input sampling frequency F_{s1} is constant the formula for the reversal of the instantaneous resampling ratio can be readily obtained

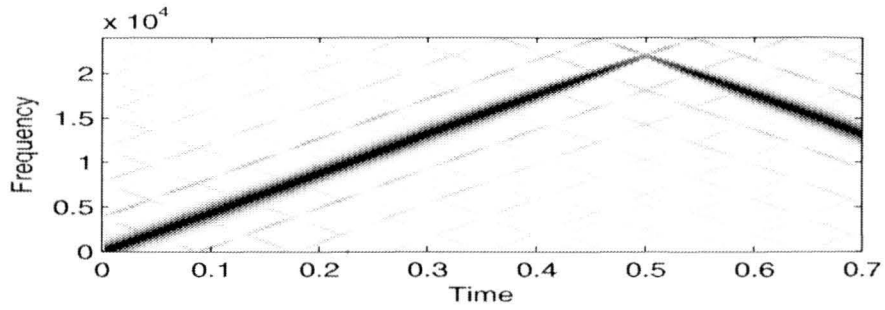
$$r_{\text{chirp}}[m] = F_{s1}/F_{s2}[m] = \frac{F_{in}(F_0 + m\Delta F)}{F_{s1}\overline{F_{s2}}} \tag{26}$$



(a) Input chirp signal.



(b) Chirp signal resampled using minimax FD filters.



(c) Chirp signal resampled using FD filters designed using window offset with MF filter of the length $N_{\text{off}} = 5$.

Fig. 7. Spectrograms illustrating resampling with constant resampling ratio $160/147$. VFD filter of the length $N = 17$ and Farrow structure order $q = 5$.

The effects of resampling are presented in Fig. 9. Let us notice that when the output signal is reconstructed using variable output sample rate (25) the input (Fig. 9a) and output (Fig. 9c) signals represent the same analog signal. However, if we assume constant output rate $\overline{F_{s2}}$, the output samples of SRC algorithm represent the chirp signal (Fig. 9b) which we wanted to obtain.

The process described above can be reversed. The chirp signal obtained in the previous step can be converted back to sinusoid but selection of the $r[m]$ is now more difficult. In the first scenario the sample rate and the frequency of the input signal are constant which simplifies the derivations. In this problem we need to assume that either the frequency of the input signal or its sample rate is changing. Since we want to demonstrate how to reverse the resampling process we will assume that the signal frequency is constant with variable distance between input samples.

From (25) we know the sample ratio and input sampling instants but need to find the ratio $r[m]$ (19) specified in equidistant output instants m. Let us assume that we know the instantaneous input sample rate $F_{s1}[n] = 1/T_{s1}[n]$ sampled in the same instants as the input signal $x[n]$ and instantaneous output sample rate $F_{s2}[m] = 1/T_{s2}[m]$ sampled in the same instants as the output signal $y[m]$. We are looking for ratio $r[m] = T_{s2}[m]/\hat{T}_{s1}[m]$ where $\hat{T}_{s1}[m]$ is the input sampling

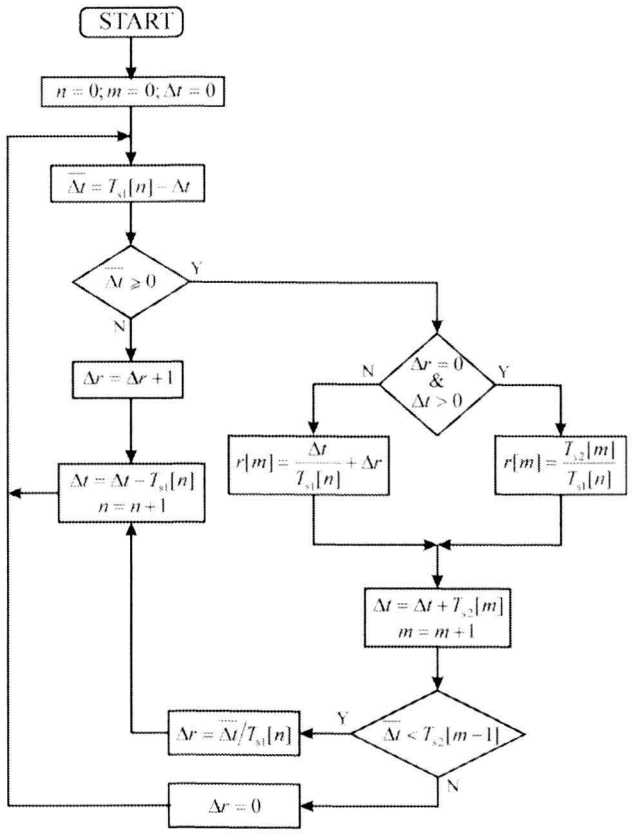


Fig. 8. Diagram for computation of $r[m]$ (19) based on instantaneous input and output sample rates.

period corresponding to the output time instant m.

Assuming that we know the positions of input sampling instants, which are our output instants from the previous problem

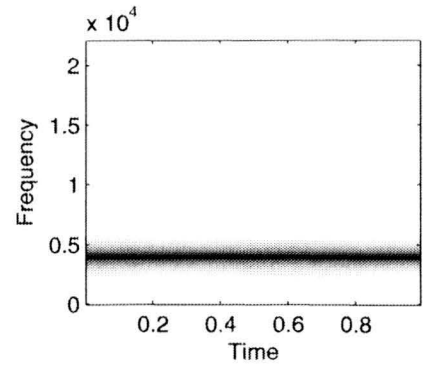
$$t_{in}[n] = \sum_{i=1}^{n} T_{s1}[n] = t_{in}[n-1] + T_{s1}[n] \qquad (27)$$

and output sampling instants, which are our input instants from the previous problem

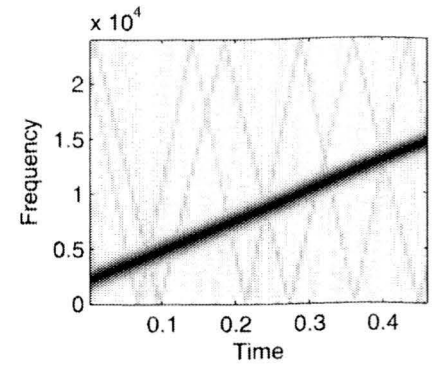
$$t_{out}[m] = \sum_{i=1}^{m} T_{s2}[m] = t_{out}[m-1] + T_{s2}[m] \qquad (28)$$

We propose the following algorithm for computation of the reversal of the instantaneous resampling ratio $r[m]$.

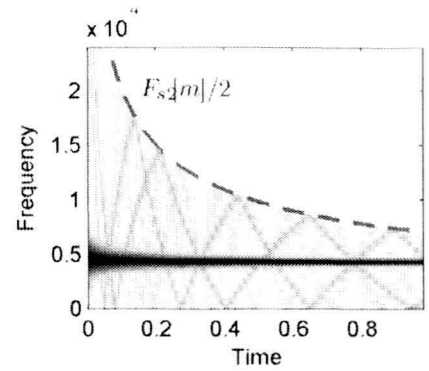
1) Start with input and output discrete time indexes $n := 0$ and $m := 0$ with corresponding continuous time instants $t_{in} := 0$ and $t_{out} := 0$.
2) Compute distance from the current output time instant to the current and the next input time instant:
$$\Delta t := t_{out} - t_{in},$$
$$\overline{\Delta t} := T_{s1}[n] - \Delta t.$$
3) If $\overline{\Delta t} >= 0$ then
   a) if the previous output sample is located in the same input sampling interval ($\Delta r = 0$ and $\Delta t > 0$) then
   $$r[m] := T_{s2}[m]/T_{s1}[n],$$
   otherwise
   $$r[m] := \Delta t/T_{s1}[n] + \Delta r,$$
   b) If $\overline{\Delta t} < T_{s2}[m]$ then
   $$\Delta r := \overline{\Delta t}/T_{s1}[n]$$
   and move to the next input instant
   $$t_{in} := t_{in} + T_{s1}[n],$$
   $$n := n + 1,$$
   otherwise
   $$\Delta r := 0,$$
   c) move to the next output instant
   $$t_{out} := t_{out} + T_{s2}[m],$$
   $$m := m + 1,$$
4) otherwise ($\overline{\Delta t} < 0$)
   a) $\Delta r := \Delta r + 1$,
   b) move to the next input instant
   $$t_{in} := t_{in} + T_{s1}[n],$$
   $$n := n + 1,$$
5) Go to point 2.



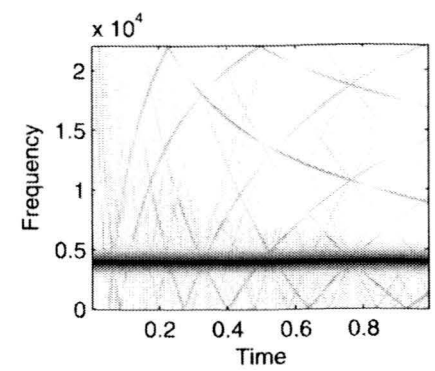
(a) Spectrogram of input sinusoidal signal.



(b) Spectrogram of chirp signal obtained from sinusoidal signal. Assumed constant sample rate $\overline{F_{s2}} = F_{s1}$.



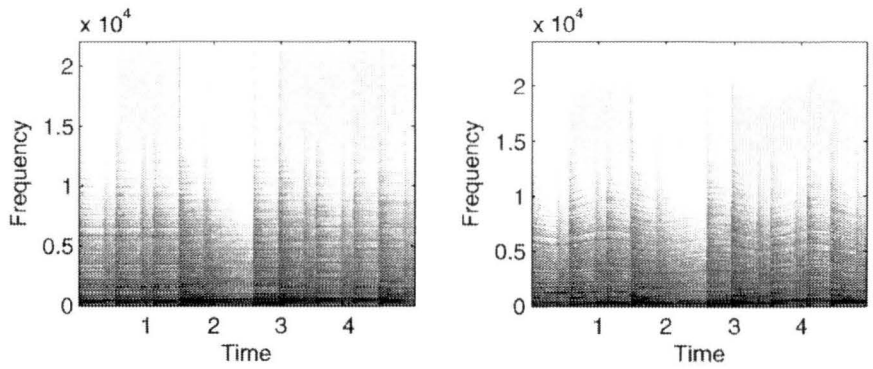
(c) Spectrogram from Fig. 9b reshaped with accordance to variable sample rate. Dashed line indicates the folding frequency.



(d) Spectrogram of restored sinusoidal signal.

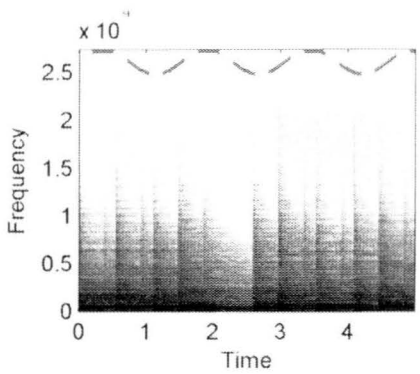
Fig. 9. Conversion from sinusoidal signal into chirp and back using SRC based on VFD filters. VFD filter with $f_a = 0.4$ of the length $N = 17$ and Farrow order $q = 5$ designed using window offset using MF FD filter of length $N_{\text{off}} = 5$.

The proposed algorithm is universal and can be used for computation of instantaneous ratio $r[m]$ for any variable input and output sample rates. Fig. 8 presents the modified version of the proposed algorithm. The introduced modifications allow to eliminate from algorithm the continuous accumulation of input and output times which eventually would lead to roundoff errors. In Fig. 9d spectrogram of sinusoidal signal recovered from chirp signal generated in previous example is presented. We can observe the nonlinear distortions but their level can be controlled with the selection of the overall filter (Fig. 6)

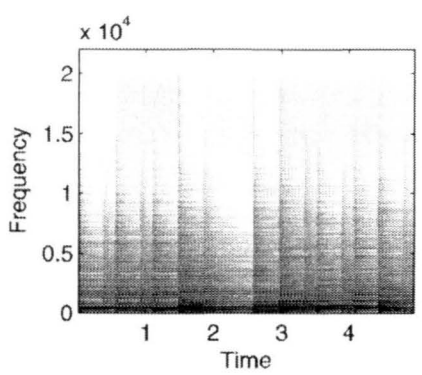


(a) Spectrogram of the input piano signal.

(b) Spectrogram of piano signal with sinusoidal fluctuations in sample rate introduced.



(c) Spectrogram from Fig. 10b reshaped with accordance to variable sample rate.



(d) Spectrogram of the restored signal.

Fig. 10. Manipulations on piano music. Introduction and removal of sinusoidal fluctuations into instantaneous sample ratio. VFD filter with $f_a = 0.45$ of the length $N = 47$ and Farrow order $q = 6$ designed using window offset method using MF FD filter of length $N_{\text{off}} = 6$.

attenuation in the stopband.

Fig. 10 presents the example in which a sinusoidal sample rate changes are introduced and later removed from the processed signal. The signal contains the piano music sensitive to such distortions. Presented example simulates the case in which a velocity of magnetic tape changes because of mechanical problems. As we can see, the proposed solution can be used efficiently to remove such distortion if only we are able to find out how the velocity/sampling ratio changes [6]–[8].

VI. Conclusion

We have demonstrated that using Farrow structure implementing VFD filter audio signal can be resampled with continuously changing sample rate ratio. The proposed approach can be used to simulate signal distortions, for example change or remove speech intonation, as well as to correct old recordings distorted because of non-constant media velocity, e.g. magnetic tape.

In this paper we have only demonstrated that the proposed solution can be used in aforementioned applications. Further research should focus on comparison with other resampling methods and automatic selection on sample rate ratio based for example on changes of pitch period for speech processing or on the properties of recorded distortions for reconstruction of old recordings [6]–[8].

Acknowledgment

This work was in part supported by the Polish Ministry of Science and Higher Education under the research project financed from the state budget designated for science in the years 2010-2012.

References

[1] AES5-2008, "AES recommended practice for professional digital audio – preferred sampling frequencies for applications employing pulse-code modulation," Audio Engineering Society, Tech. Rep., 2008, revision of AES5-1997.

[2] A. Tarczynski, W. Kozinski, and G. D. Cain, "Sampling rate conversion using fractional-sample delay," in *Proc. IEEE ICASSP'94*, Adelaide, Australia, May 1994, pp. 285–288.

[3] E. Hermanowicz, M. Rojewski, and M. Blok, "A sample rate converter based on a fractional delay filter bank," in *Proc. ICSPAT 2000*, Dallas, Tx, USA, Oct. 16–19, 2000.

[4] G. Evangelista, "Design of digital systems for arbitrary sampling rate conversion," *Signal Process.*, vol. 83, no. 2, pp. 377–387, Feb. 2003.

[5] K. Rajamani, L. Yhean-Sen, and C. W. Farrow, "An efficient algorithm for sample rate conversion from CD to DAT," *IEEE Signal Process. Lett.*, vol. 7, no. 10, pp. 288–290, 2000.

[6] A. Ciarkowski, A. Czyzewski, M. Dziubinski, A. Kaczmarek, B. Kostek, M. Kulesza, and P. Maziewski, "Methods for detection and removal of parasitic frequency modulation in audio recordings," in *AES Conference: 26th International Conference: Audio Forensics in the Digital Age*, Jul. 2005.

[7] A. Czyzewski, A. Ciarkowski, A. Kaczmarek, J. Kotus, M. Kulesza, and P. Maziewski, "Dsp techniques for determining ŞwowŤ distortion," *J. Audio Eng. Soc*, vol. 55, no. 4, pp. 266–284, 2007.

[8] A. Czyzewski, P. Maziewski, and A. Kupryjanow, "Reduction of parasitic pitch variations in archival musical recordings," *Signal Processing*, vol. 90, no. 4, pp. 981 – 990, 2010.

[9] P. Maziewski, "Wow defect reduction based on interpolation techniques," *Bulletin of the Polish Academy of Science: Technical Sciences*, vol. 54, pp. 469–477, 2006.

[10] T. I. Laakso, V. Välimäki, M. Karjalainen, and U. K. Laine, "Splitting the unit delay — tools for fractional delay filter design," *IEEE Signal Process. Mag.*, vol. 13, no. 1, pp. 30–60, 1996.

[11] M. Blok, "Farrow structure implementation of fractional delay filter optimal in Chebyshev sense," in *Proc. SPIE*, vol. 6159, Wilga, Poland, May 30 – Jun. 2, 2005, p. 61594K.

[12] ——, "Optimal fractional sample delay filter with variable delay," in *OSEE 2002*, TechOnLine, Bedford, Massachusetts, USA, Mar. 18, 2002. [Online]. Available: http://www.eetimes.com/design/analog-design/4018005

[13] C. W. Farrow, "A continuously variable digital delay element," in *Proc. IEEE ISCAS'88*, Espoo, Finland, 1988, pp. 2641–2645.

[14] F. J. Harris, "Performance and design of Farrow filter used for arbitrary resampling," in *Proc. DSP'97*, vol. 2, Santorini, Greece, 1997, pp. 595–599.

[15] E. Hermanowicz, "On designing a wideband fractional delay filter using the Farrow approach," in *Proc. EUSIPCO'2004*, Austria, Sep. 6–10, 2004, pp. 961–964.

[16] M. Blok, "Collective filter evaluation of an FSD filter-based resampling algorithm," in *OSEE 2002*, TechOnLine, Bedford, Massachusetts, USA, Jan. 15, 2002. [Online]. Available: www.eetimes.com/design/signal-processing-dsp/4017905

[17] A. Yardim, G. D. Cain, and P. Henry, "Optimal two-term offset windowing for fractional delay," *Electron. Lett.*, vol. 32, no. 6, pp. 526–527, Mar. 1996.

[18] A. Yardim, G. D. Cain, and A. Lavergne, "Performance of fractional-delay filters using optimal offset windows," in *Proc. IEEE ICASSP'97*, vol. 3, Apr. 21–24, 1997, pp. 2233–2236.

[19] E. Hermanowicz, "A nearly optimal variable fractional delay filter with extracted Chebyshev window," in *Proc. IEEE ICECS'98*, vol. 2, Lisboa, Portugal, Sep. 7–10, 1998, pp. 401–404.

[20] M. Blok, "Fractional delay filter design with extracted window offsetting," in *Proc. MixDes'2012*, Warsaw, Poland, May 24–26, 2012.

Influence of Bitrate on the Loudness Measured in the Audio Broadcast Stream

Przemysław Plaskota
Wroclaw University of Technology
Wroclaw, Poland
przemyslaw.plaskota@pwr.wroc.pl

ABSTRACT — **Radio and television broadcasters increasingly encounter the problem of loudness in broadcast programmes. The problem occurs at all stages beginning with the production of a programme, ending with the transmission. In order to meet the expectations of broadcasters and receivers of radio and television content the International Telecommunication Union (ITU) has developed an algorithm for measuring the loudness of programmes simultaneously introducing a measurement unit for loudness, LU. In the paper the influence of bitrate on the loudness measured in the audio stream of TV programmes has been presented.**

KEYWORDS — ***loudness, lossy audio compression, broadcasting***

I. INTRODUCTION

Measurement of the loudness of the television programmes has recently become a necessity. Problem connected to loudness of the programmes has been growing since a long time. Radio and television broadcasters increase the loudness of the transmitted programmes competing for the audience's attention. The idea is that the loudness attracts the attention of the viewer/auditor, and consequently makes a programme more attractive. It particularly concerns the commercials – loud commercial stands out in comparison to other adverts and makes the viewer or auditor pay greater attention to a particular content. Advertising agencies assume that even negative emotions, in this case annoyance of the viewer with excessive loudness, still benefit the products advertised. No matter what means are being employed, the point is to attract the attention of a potential client. The result of this way of thinking is a great loudness disproportion between the emitted programmes and the commercial blocks.

The issue of loudness does not pertain to the advertisements alone, although undoubtedly these are the most annoying for the receivers of the content. Television and radio producers on everyday basis encounter the problem of varied loudness of content within one programme. It mainly concerns programmes containing portions of the material coming from the remote TV sites. Good example is a news programme. In case of television, a news programme comprises of the information presented by the announcer as well as materials prepared in various parts of the country during all kinds of local events. Although the broadcast from a single place is rather uniform in loudness, combining few presentations of different producers in one block makes it hard to preserve a stable loudness impression for the auditor. It is especially true in cases of a single or first broadcast of a particular material. At television stations broadcasting only news programmes, during a first screening of the material a sound producer often does not know what loudness is to be expected. For the next broadcasting of the same material the producer, based on the knowledge gained with the previous transmission, can set the loudness in a way that would allow achieving unity with the accompanying materials. Thus it is important to standardise the production methods in such a way that the materials gained from different sources have the same loudness and the broadcaster airing these materials knows before the first emission what is the loudness of the programme to be broadcasted in a moment.

The problem with achieving uniform loudness occurs also in cases of continuous production, when the materials come entirely from the same source. Yet it is especially challenging in a situation when speaking intervenes with the fragments in which the music is used as a background for presented images. For illustration purposes usually popular music is used, and on the production level it is set at the maximum volume. A producer most of the time assesses the loudness based on the meter indication, instead relying on his/her personal loudness impression. This method of assessment is usually employed due to a lack of time, or often also because of unfavourable audition conditions, as well as insufficient experience. This procedure brings a combination of variegated contents of diverse loudness. Such situation can be also observed in live radio programmes when a loud musical piece presentation forcing a listener to turn down the radio is followed by presenters words which can be hardly understood by the listener – until the radio receiver's volume is brought back to its original setting.

Another problem resulting from different volume levels of broadcast programmes is particularly noticeable with television. It pertains to a situation wherein a viewer looking for a proper programme to watch switches the channels one after another. Practically every station follows its own loudness standards and uses different audio processors. In this situation a viewer is forced to adjust the volume every time he or she changes a channel. This situation is mostly seen with the cable television networks, yet it is also clearly perceptible while viewing contents presented by analogue terrestrial television providers. Because of different transmission standards this problem to a smaller degree concerns television programmes transmitted digitally. Nevertheless previously presented problems do also affect the digital television.

The reason behind introducing broadcast loudness regulations is many complains having been lodged by the recipients of the transmitted contents. Therefore adhering to loudness measurement standards would improve the comfort of the audience. By looking at this problem from broader perspective one can notice that introducing such standards can be also beneficial for the producers and broadcasters.

II. LOUDNESS MEASUREMENT

In order to make the assessment of the broadcast loudness possible, the ITU (International Telecommunication Union) has developed a measurement algorithm, described in recommendation ITU-R BS.1770 [1, 2]. The subsequent versions of the document introduce modifications improving programme loudness assessment. The algorithm proposed by ITU involves double filtering of a signal (K-filter), then calculating root mean square value. The last stage is adding up values acquired for all the channels. In recommendation [2], in comparison to the previous version, gating of low level signals was introduced. This procedure prevents decreasing the loudness of the whole measured sound material. Recommendation ITU-R BS.1771 [4] contains requirements for the loudness meter. The accuracy of a meter reading can be assessed by applying recommendation ITU-R BT.2217 [5].

Based on ITU recommendations, the authority controlling the broadcasters in Poland, the National Broadcasting Council [KRRiT], drew up a regulation [3] obligating the broadcasters to control the loudness of the advertisement blocks. Since KRRiT is a regulatory body, it is interested only in those advertisement blocks whose loudness has been complained about by the viewers. But the European Broadcast Union (EBU), in order to meet the expectations of those broadcasters for who the loudness problem is much more complex, has drawn up recommendation EBU R 128 [6]. As far as the programme loudness measurement is concerned, this recommendation entirely refers to the ITU recommendation. The document contains definitions of some parameters of audio signals, e.g. Target Level, which is the loudness level value to be achieved in an entire programme; Loudness Range etc. In this document the additional audio signal gating was also added to the measurement algorithm in order to prevent the advertisement producers from manipulating the level of loudness. Moreover the EBU R128 recommendation was supported by additional documents instructing the producers of the content how to control the loudness during the entire production process [7, 8, 9, 10].

In a situation, when costs of the radio and television programmes transmission in Digital Video Broadcasting (DVB) and Digital Audio Broadcasting (DAB) standards depend on the bandwidth used, it is obvious that the producers will try to reduce the expenses by decreasing the bitrate of data stream. In case of television transmission an audio stream is definitely smaller in comparison to a video data stream. The situation differs in case of radio transmissions, as here we deal only with the audio stream.

The material recorded by the regulatory body for a control loudness measurement is usually lossy coded. Here also arises a problem of storing the recorded material. The stronger is the compression of such material the lesser are its storage costs. Since the aim of the control is detecting any loudness limit exceeded, it is important that there are no doubts as to whether such limit has been crossed. Therefore it is important to answer the question whether decreasing a bitrate of an audio stream influences the result of loudness measurement.

III. MEASUREMENT SETUP

The aim of the research is the loudness measurement of the sound recorded in the lossy coded files. The audio files contain fragments of television programme with some broadcast and advertisement block pieces. The files had been lossy coded, then again transformed into non-lossy format and in such form their loudness was measured. In DVB as well as in DAB the Advanced Audio Coding (AAC) is used. The research material has been coded in the AAC format using bitrate from 256 to 32 kbps. For the files with the bitrate ranging from 256 to 80 kbps the sampling frequency equals 44.1 kHz, whereas smaller bitrates required lower value sampling frequencies, which unquestionably entails limiting the audio signal frequency band.

The loudness of radio and television programmes should be measured with a meter compatible with the requirements of ITU-R BS. 1771 [4]. The measurement setup used for the experiment is very simple, comprising of a player and a loudness meter. (Fig. 1)

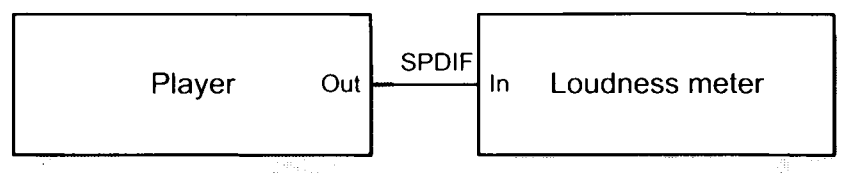


Figure 1. Meaurement setup

The measurements were conducted with the use of two measurement methods, one complying with the older version of the ITU recommendation and the other with the newer [1, 2]. This procedure has been adopted because even if the current KRRiT recommendation [3] refers to the first version of the recommendation, it should be expected that in the nearest future this recommendation might be changed and adjusted to comply with the newer one. Additionally, it will allow comparing both versions of recommendation and determining whether the introduced changes affect the result of the measurement.

IV. RESULTS

The results of the measurement with the method complying with [1] have been presented in Table I, whereas the results achieved in accordance with [2] are listed in Table II. On the basis of the measurement results it can be ascertained that the loudness level value for the files with the bitrate ranging from 256 kbps to 80 kbps does not differ from the loudness level of the original file. The loudness of the files with bitrates of 64 and 56 kbps is smaller by 0.2 LU than that of the original file, whereas the loudness of the files with the bitrate of 48, 40 and 32 kbps has decreased by 0.4 LU. The results are the same for both measurement methods, except for the file with 112 kbps

bitrate for the measurement conducted in accordance with [2]. But this inconsistency still falls within the scope of measurement uncertainty.

In Figure 2 the comparison of the results for both measurement methods is presented. The difference between the measurement results does not exceed 0.2 LU. On this basis it can be assumed that the change introduced to the algorithm influences the measurement results. It should be noted though that the comparison was made only for one type of sound material which included only short – in relation to the span of the whole sample – sections of the low level signal. Nevertheless the result achieved with the method complying with [2] shows higher level of loudness, which means that the aim of eliminating very silent fragments from a general loudness assessment has been achieved.

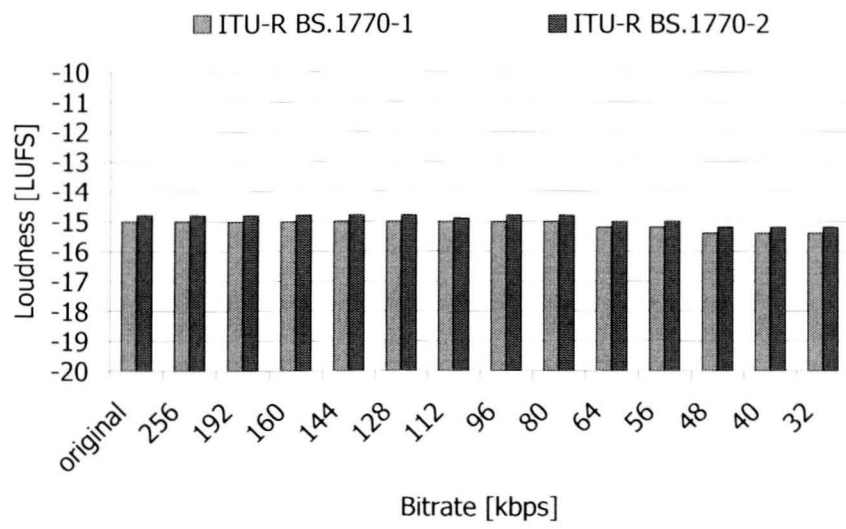


Figure 2. Comparision of the results for the loudness measurements conducted in accordance with both recommendations [1, 2]

TABLE I. RESULT OF THE PROGRAMME LOUDNESS MEASUREMENTS ACHIEVED WITH THE METHOD ADHERING TO ITU-R BS.1770-1 [1] AND THE LOUDNESS DIFFERENCE IN COMPARISON TO THE ORIGINAL FILE

Bitrate [kbps]	**Loudness [LUFS]**	**Loudness difference [LU]**
without compression	-15.0	n/a
256	-15.0	0.0
192	-15.0	0.0
160	-15.0	0.0
144	-15.0	0.0
128	-15.0	0.0
112	-15.0	0.0
96	-15.0	0.0
80	-15.0	0.0
64	-15.2	-0.2
56	-15.2	-0.2
48	-15.4	-0.4
40	-15.4	-0.4
32	-15.4	-0.4

Positive value in the third column indicates loudness of a sample higher than the loudness of the sample without compression

TABLE II. RESULT OF THE PROGRAMME LOUDNESS MEASUREMENTS ACHIEVED WITH THE METHOD ADHERING TO ITU-R BS.1770-2 [2] AND THE LOUDNESS DIFFERENCE IN COMPARISON TO THE ORIGINAL FILE

Bitrate [kbps]	**Loudness [LUFS]**	**Loudness difference [LU]**
without compression	-14.8	n/a
256	-14.8	0.0
192	-14.8	0.0
160	-14.8	0.0
144	-14.8	0.0
128	-14.8	0.0
112	-14.9	-0.1
96	-14.8	0.0
80	-14.8	0.0
64	-15.0	-0.2
56	-15.0	-0.2
48	-15.2	-0.4
40	-15.2	-0.4
32	-15.2	-0.4

Positive value in the third column indicates loudness of a sample higher than the loudness of the sample without compression

V. SUMMARY

The paper presents the result of loudness measurement for the audio files of television programmes. The files were coded with the AAC algorithm for different bitrate values and then the loudness level measurement was conducted for the files coded in this way. On the basis of the measurement results it has been determined that for the bitrate values usually used in DVB and DAB transmissions the lossy coding does not influence the loudness measurement result. For very low bitrate values small loudness level drop has been observed. Comparing the results achieved with the use of two algorithms allowed determining that very low level sounds influence the loudness measurement result.

On the basis of the conducted measurements it can be concluded that during the signals recording, which is performed by the regulatory body supervising the broadcasters for the purpose of loudness inspection measurement, should be used bitrates for which there have not been observed any level differences in comparison to the original files. The most appropriate seems to be using the bitrates identical or higher than those used by the broadcasters during the transmission of the programmes.

REFERENCES

[1] ITU-R BS.1770-1. Algorithms to measure audio programme loudness and true-peak audio level, 2009.

[2] ITU-R BS.1770-2. Algorithms to measure audio programme loudness and true-peak audio level, 2011.

[3] Regulation of the National Broadcasting Councis of 30 June 2011 concerning principles of advertising and teleshopping in radio and television services (Official Journal no. 150, item 895).

[4] ITU-R BS.1771-1. Requirements for loudness and true-peak indicating meters, 2012.

[5] ITU-R BT.2217. Compliance material for Recommendation ITU-R BS.1770, 2011.

[6] EBU Technical Recommendation R 128 Loudness normalisation and permitted maximum level of audio signals, 2010.

[7] EBU Tech Doc 3341 Loudness Metering: 'EBU Mode' metering to supplement loudness normalisation in accordance with EBU R 128, 2010.

[8] EBU Tech Doc 3342 Loudness Range: A descriptor to supplement loudness normalisation in accordance with EBU R 128, 2010.

[9] EBU Tech Doc 3343 Practical Guidelines for Production and Implementation in accordance with EBU R 128, 2010.

[10] EBU Tech Doc 3344 Practical Guidelines for Distribution of Programmes in accordance with EBU R 12, 2010.

SESSION 7:
STEREOVISION AND VIDEO CODING

NTAV/SPA 2012
27-29TH SEPTEMBER, 2012, ŁÓDŹ, POLAND
NEW TRENDS IN AUDIO AND VIDEO / SIGNAL PROCESSING ALGORITHMS, ARCHITECTURES, ARRANGEMENTS AND APPLICATIONS

Using Harris Corner Points to Reduce the Complexity of a Local Stereo Image Matching Algorithm

Mikołaj Roszkowski
Institute of Radioelectronics
Warsaw University of Technology

Grzegorz Pastuszak
Institute of Radioelectronics
Warsaw University of Technology

ABSTRACT — **The purpose of a short-baseline stereo image matching algorithm is a calculation of the disparity map of a given rectified stereo pair. Most stereo algorithms presented in the literature achieve this goal by searching all possible disparities in the assumed search range for all pixels. The proposed algorithm consists of two stages. In the first stage detected feature points in the stereo pair images are matched. In the second, a local stereo matching algorithm is used but with a disparity search range limited to a subset of the possible disparities determined by the matched feature points. It is shown that the usage of feature points is an efficient way of limiting the number of disparities searched at the cost of only minor increase in the result disparity map's error. The selected feature points are Harris corners, which can be quite easily computed. Furthermore, the algorithm works on image blocks and each block is processed almost independently. Because of that the proposed framework seems suitable for real-time parallel implementation on GPU or in the FPGA technology.**

I. INTRODUCTION

Stereo matching problem is one of the most widely studied early vision problems, and many algorithms has been created in the recent years. The input of the stereo algorithm is a pair of rectified images representing the same scene, but captured by two cameras positioned close to each other. As both images represents the same scene, the pixels from the first image can be matched with the pixels from the second image, and the displacement (disparity d) between the matched pixels in the two images can be calculated. Thus, the stereo matching problem can be defined as the problem of assigning each pixel in the image a correct disparity value. It is also a well known fact that the disparity is inversely proportional to the depth of an object in the scene.

The disparity assignment problem is usually considered only in a discrete space rather than a continuous one, and two important groups of algorithms for solving this problem can be distinguished: local and global optimisation based algorithms [1]. The local algorithms assign disparity to a pixel taking into regard only the pixel and its closest neighbourhood, while the global algorithms optimise the estimated matches for the whole image at the same time. However, both types of the algorithms have one thing in common: the need to define and compute the cost of matching the pixel (x, y) in the first image with the pixel $(x, y \pm d)$ in the second image. The sign of d depends whether the matching is performed from the left image to the right or conversely. This matching costs are represented by cost volume function $C(x, y, d)$. The matching cost function can be e.g. a simple AD (absolute difference) measure [1] or census measure [2].

In the typical local based algorithm the cost-volume function $C(x, y, d)$ is filtered like a two-dimensional image for each disparity d separately. The disparity to be assigned to the pixel is then selected as:

$$d(x, y) = \text{argmin}_d C(x, y, d) \tag{1}$$

The algorithms usually only differ in the filter used. The simplest possible filter is the box filter [1]. More advanced approach [3] uses filter which weights are determined for each pixel separately and depend on spatial distances and colour similarities between the pixels within the window. These weights are computed like the ones in the bilateral filter. Other proposed filtering methods include e.g. geodesic distance based weights [4] or guided filter based weights [5]. In [6] it has been shown that, instead of using bilateral-like filter weights such as in [3], the use of just two weights: $0, 1$ also gives reasonable results. The use of such weights hugely decreases the number of necessary calculations. Similar approach is adopted in [7], where authors presented a local algorithm producing disparity maps of quality comparable to those obtained with global algorithms.

Most commonly used global optimisation algorithms in stereo matching are graph-cuts [8], [9] and belief propagation [10], [11]. Both are energy minimisation frameworks that require the definition of the pixel matching cost and energy smoothness term. The energy optimisation is done in iterative fashion and is usually very computationally expensive. The simplest global optimisation algorithms [11] work using just a cost volume $C(x, y, d)$. The most advanced ones start with an image segmentation and simple stereo matching in order to formulate the hypothesis about the possible 3D planes appearing in the image. The final optimisation is done in the estimated planes' space [12], [13], [14].

The algorithms presented above require the cost-volume function $C(x, y, d)$ to be known for all possible (x, y, d) values. In the case of relatively small images, having resolutions lower than 512x512 and number of disparity levels smaller than 80, the near-real time or real time implementations are feasible on GPU ([5], [15]). Real-time processing systems for larger disparity ranges and higher resolutions has been proposed for FPGA platforms [16], [17]. However, it seems that the need to calculate full cost volume $C(x, y, d)$ is one of

the major obstacles to a calculation of high resolution disparity maps in real time.

In this paper a solution capable of alleviating the problem of calculating $C(x, y, d)$ for all possible (x, y, d) is proposed. This goal is reached by limiting the search range on d for each pixel to the most probable values only. The paper clearly shows that reduction in the number of the disparities searched can be significant, which means huge decrease of the algorithm computational requirements. This should allow for computation of disparity maps for images with higher resolutions and larger maximal disparities in shorter time. The rest of the paper is organised as follows: in section II some previous approaches to the problem are discussed, in section III the algorithm new algorithm for solving the problem is proposed and in section IV the obtained results are presented.

II. Related work

There seem to be two established ways of reducing the number of disparities searched established in the literature. One is the use of hierarchical matching. In this case blocks are typically first matched on one, coarse level of the hierarchy. The found matches are then passed to the finer level of the hierarchy, where they are refined [18]. Hierarchical approach can also be used to optimise the belief propagation algorithm [19].

The other way to limit the number of searched points is widely used in the wide-baseline stereo matching frameworks. The reduction of search space is achieved by first finding feature points in both images using feature points detectors like e.g. SIFT [20], SURF [21] or Harris corners [22] and matching them. The found matches are later propagated in their neighbourhood using some set of defined rules [23], [24], [25]. The problem with this kind of algorithm is that during the propagation step of the algorithm a priority queue of best possible matches is used. Therefore the algorithm usability on highly parallel platforms like GPU or FPGA is limited. Similar algorithm has also been proposed for rectified stereo pair matching [26].

Other method of reducing stereo matching algorithm complexity is proposed in [27]. It concentrates on limiting the number of disparities for which the non-linear cost volume filtering step is performed. In this proposal a full cost volume is calculated first, but only a specified number of the most probable disparities is kept for each pixel and used in the second step. The latter consists in non-linear filtering of the disparity volume. Moreover, in order to reduce the algorithm complexity even further, the authors suggest the use of spatially sampled filtering windows.

III. Proposed method of reducing the matching complexity

The analysis of the existing stereo matching algorithms indicate that the major obstacle in obtaining stereo algorithms running in real time for large images is the need to calculate a full disparity cost volume $C(x, y, d)$. In this section a simple method greatly reducing the number of searched disparities is proposed. The method makes use of the feature points detection and matching like in the match propagation algorithms [23], [24], [25]. However, unlike these algorithms, in the next algorithm step a normal, stereo algorithm is used, with the number of disparities searched limited to these found by matching the feature points. The algorithm consists of the following steps:

1) Divide the image into tiles.
2) Find feature points in the images.
3) Match feature points.
4) Cross-check feature points.
5) Set the searched disparities for the tile to the disparities of the feature points belonging to the tile and its neighbours.
6) Run the local stereo matching algorithm with the disparities established for each tile.

The detailed description of each step is given in the following sections.

A. Division into tiles

It is fairly obvious that disparities assigned to pixels belonging to one object should be equal or have very close values, as the objects are smooth in 3D. Therefore, in the first step, an algorithm could try to find correct disparities for some pixels of the objects. In the second step it would be only necessary to search the disparities close to these already found for the remaining pixels of the object. Unfortunately, a segmentation of an image is usually a time-consuming and computationally demanding process. Because of that, it is easier to divide the image into a grid of square tiles, which size is a power of two (see Fig. 1). Next, feature points in the image are found using

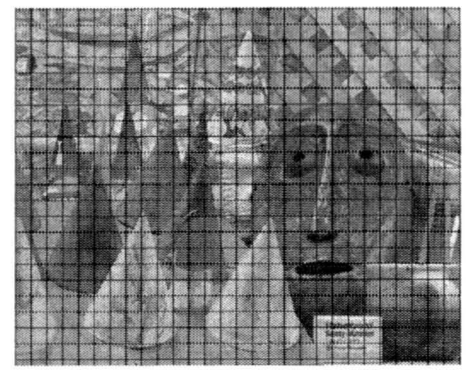


Fig. 1. Left image of the Cones stereo pair divided into tiles;

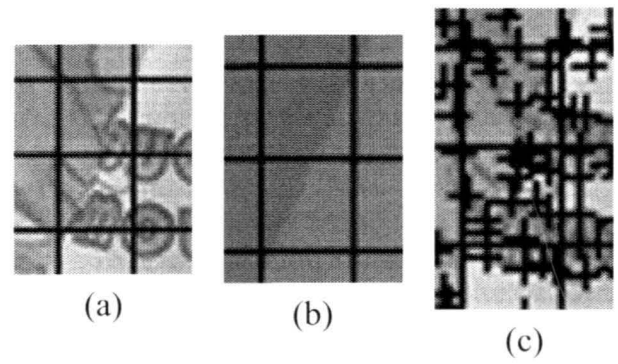


Fig. 2. An example of an image fragment containing two different objects. (a) a fragment of one cone with the background; (b) disparity map for the fragment, two different shades of grey indicate that there are two objects; (c) the fragment with feature points marked with crosses.

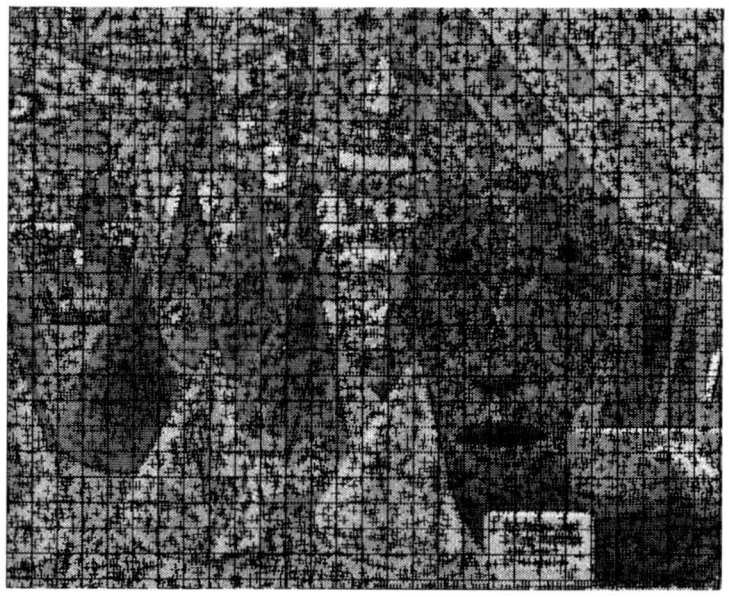


Fig. 3. Left image of the Cones stereo pair divided into tiles and with the detected Harris corners.

the selected feature detector, and matched giving a disparity value for each feature point. Fig. 2 shows an example of a part of an image, which contains two objects, and is divided into tiles. Some of the tiles contain pixels belonging to two objects, some only pixels belonging to one object. Fig. 2 (c) shows the feature points detected for the tiles marked with crosses. Some of them are found on one object, others on the second one. If each tile contained correctly matched feature points belonging to all the objects in the tile, it would be possible to search in the tile only for the disparities close to those of the matched feature points. In the case of the example in Fig. 2, this would mean a search around either one or two disparity values, depending on the number of objects in the tile. Unfortunately, in the real world situation finding good match for at least one pixel of each object in each tile might prove difficult. Thus, the disparity search range is propagated from each tile to the tile's neighbours.

B. Feature points detection and matching

There are many existing feature detectors, for instance SIFT [20], SURF [21] or Harris corners [22]. SIFT and SURF are powerful feature points detectors and descriptors that has been designed to be invariant to rotations and illumination differences in the images. The Harris corner detector is quite different — it is simpler and robust to neither rotation nor illumination changes. Nevertheless, it is just as useful in the case of the short-baseline rectified stereo pairs. For such images, the illumination differences are rarely significant. Moreover, as the images are rectified, the orientation of the points to be matched do not differ in the images.

The Harris corner detector has been chosen for proposed algorithm, for the sake of its simplicity. It has also another big advantage: the number of detected feature points can be changed by modifying only two parameters: window size, and a threshold used for determining whether point is a corner or not. As pointed out in section III-A it is very important to find every possible disparity in each tile, which is easier when the number of feature points is higher. The modifications of the two Harris detector algorithm's parameters allow for easy change of the number of feature points discovered, and ensuring that enough features are found. Obviously the number of features should be enough for detecting all existing disparities. At the same time it shouldn't be too high, as every feature points has to be matched, which increases the complexity of the algorithm. Fig. 3 shows an example of an image divided into tiles with and with marked Harris corners in a number we consider sufficient — about 2%–3% of total number of image pixels.

The feature points are matched using simple SAD window based approach. However, instead of using original colour images for matching, it is performed on gradients' magnitude images, obtained with Sobel filter. The calculated gradients' magnitudes are also clipped to the maximal value equal to 255. Each feature point is matched only with the points in the other image being closer than the allowed maximal disparity. In the case of rectified stereo pairs, the corresponding points in the two images must lie on the same line. Nevertheless, the corresponding feature points often do not meet this requirement due to the fact that the Harris detector does not guarantee perfect localisation of the feature points. Thus, the algorithm has to match each feature point with feature points lying not only on the same line but also on neighbouring ones. Matching process is performed for each image separately and subsequently the matches are cross-checked. This means that each point in the left images is assigned a best matching feature point in the right image, and each point in the right image is assigned a best matching point in the left image. It is then verified that the matches agree. Otherwise they are rejected.

C. Dense stereo matching

The previous steps of the algorithm provide the information which disparities should be searched in each tile of the image. This way the sparser disparity cost volume $C(x, y, d)$ can be constructed, with the values corresponding to the not searched disparities set to "unknown". Such a cost volume can be filtered using any method, for example one of the proposed in [3], [5], [6]. The only restriction is that the $C(x, y, d)$ values marked as "unknown" should be excluded from filtering. In the tests presented in the paper the colour-cross algorithm described in [6] is used. Such a decision is motivated by a simplicity and fairly good results of this algorithm. The pixel dissimilarity measure used to build the disparity cost volume is Census measure [2] with neighbourhood size equals to 5. The cross is restricted to a 16x16 window. The following colour similarity function is used for determining pixel similarity (in RGB colour space):

$$S(p, p') = \max_{i \in \{R,G,B\}} |p_i - p'_i| \tag{2}$$

Such a function has been chosen as it has been already effectively used in stereo matching [7]. The pixels for which the value of the similarity function is lower than threshold (set to 30) are considered as similar in cross-construction. It should only be noted, that because "unknown" disparities

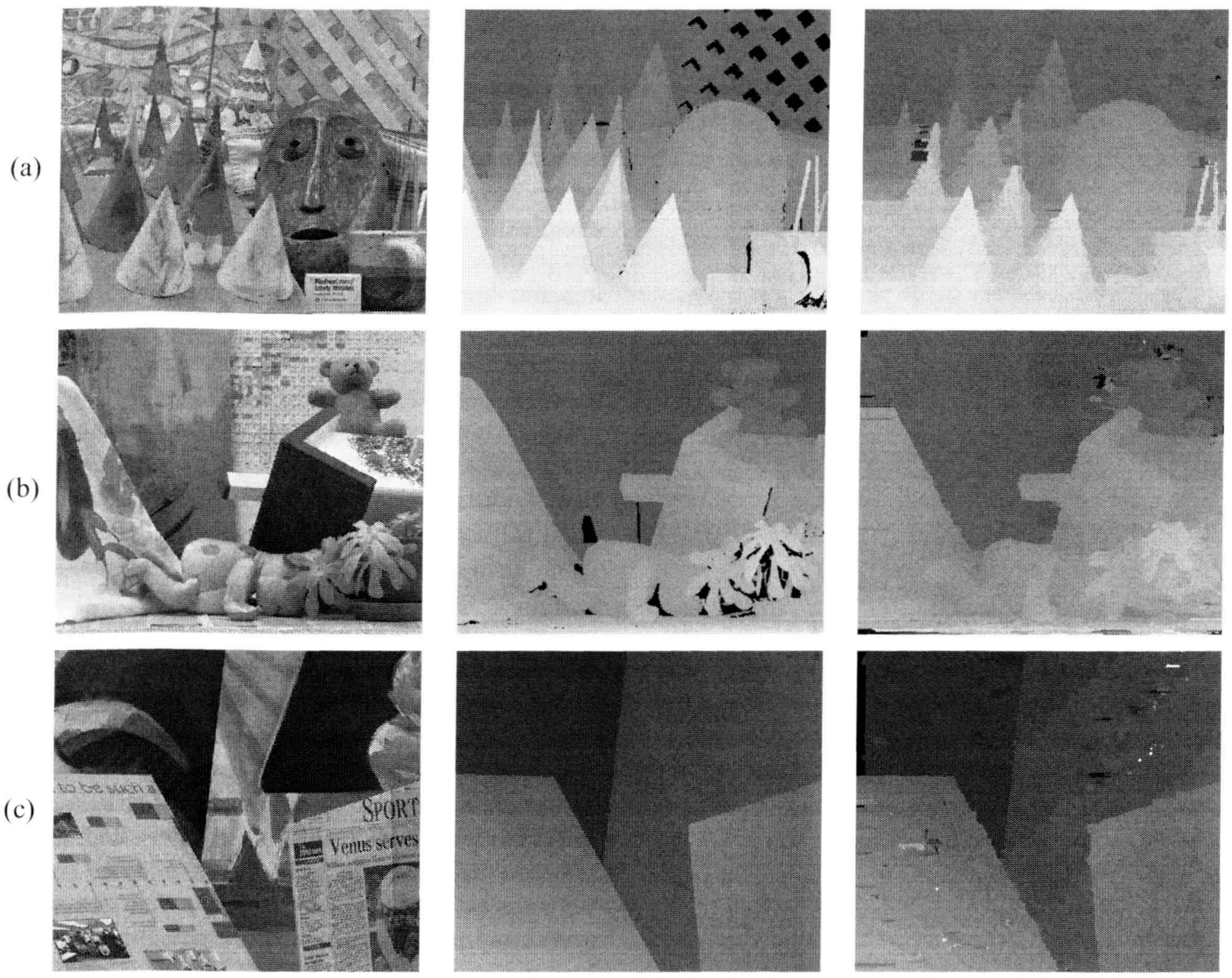


Fig. 4. Results obtained for Cones (a), Teddy (b) and Venus (c) datasets. The first image is a left image of the stereo pair, the second is a reference disparity map and the third is a disparity map produced by our algorithm.

are excluded from filtering, filter normalisation factor for each pixel should be changed accordingly.

After the matching step, a simple occlusion detection and filling is performed. The calculated disparities are cross-checked and the pixels that do not pass cross-check are filled with the closest lowest disparity value.

IV. Results

A. Conducted tests

In the experiments standard stereo pairs from the Middlebury stereo evaluation suite are used: Cones, Teddy and Venus [28] and selected stereo pairs from the extended Middlebury set: Art, Baby1, Dolls, Wood1 [29], [30]. We decided not to use popular Tsukuba scene, as it has very limited disparity search range equal to 16. The results of the algorithm are evaluated by specifying the percentage of the wrongly assigned disparities for the image. A disparity is considered to be wrong if it differs from the ground truth disparity by more than one.

The proposed algorithm, presented in the previous section, is compared with an algorithm performing full disparity search. The full search algorithm used is the colour-cross algorithm with parameters described in III-C. This means that it differs from the proposed algorithm in two ways: all disparities are searched and the disparity cost volume $C(x, y, d)$ built by this algorithm does not contain "unknown values",

The proposed algorithm depends on fairly large number of parameters, which include:

- the parameters of the Harris detector algorithm,
- the size of the tiles into which the image is divided (only 16x16 and 32x32 sizes were considered),
- the size of the window used for matching feature points,
- the range of the disparities that should be searched around each disparity value found by matching feature points (disparity refinement range), values varying from 2 to 8 are tested

The selection of the best parameter values is not a straightforward task, because there are two conflicting goals: obtaining as good disparity map as possible while checking as little disparity values as possible. Because of that, the parameter optimisation process has been done in the following way:

1) The results for the algorithm with full cost volume search are obtained.
2) The threshold t is set. It specifies how much worse the results are allowed to be when compared to the original ones, after limiting the disparity search range.
3) The parameters that give the lowest average number of searched disparities per pixel are selected.

TABLE I
RESULTS OBTAINED FOR THE LIMITED DISPARITY SEARCH ALGORITHM FOR $t = 3\%$ AND THE SMALLEST NUMBER OF DISPARITIES SEARCHED.

Pair name	Number of feature points in image[%]	Average feature points comparisons	Average disparities searched	Disparities searched / max disparity [%]
Baby1	1.80	12.83	10.98	13.73
Books	2.85	20.14	16.57	20.71
Dolls	1.94	13.06	13.76	17.19
Wood1	3.24	21.88	18.63	23.29
Cones	2.19	14.88	13.11	16.39
Teddy	2.40	16.27	15.02	18.78
Venus	2.51	18.38	13.57	16.96

TABLE II
ERROR DIFFERENCES BETWEEN THE FULL AND LIMITED DISPARITY SEARCH ALGORITHMS FOR $t = 3\%$.

Pair name	Error		
	full search	limited search	difference
Baby1	6.89	9.88	2.99
Books	23.46	26.43	2.97
Dolls	18.18	20.16	1.98
Wood1	9.5	12.48	2.98
Cones	9.64	11.88	2.24
Teddy	13.71	15.34	1.63
Venus	1.78	3.41	1.63

The best results, obtained for $t = 3\%$ for the left stereo pair image, are presented in Table I and Table II. The used parameter values are: tile size equal to 16, disparity refinement range equal to 2, and feature points matching window size equal to 13. The first column in the Table I specifies how many feature points has been detected in the image, as a percentage of the number of pixels in the original image. The second column shows how many comparisons had to be done an average for each feature point in order to match it. The last column contains the average number of disparities searched for each pixel divided by the maximal disparity value md for the image. This value presents what percentage of disparities that would be searched by the full-search algorithm is checked in the proposed one. The maximal disparity for the selected stereo image pairs varies from 64 to 80, but 80 was used as a search value for all pairs.

The results for $t = 2.8\%$ are presented in Table III and Table IV. The conducted tests have proven that for the selected image set it is impossible to further lower the allowed maximal error difference t. The error difference for at least one image pair (Books) always remains around 2.8%. In this case the disparity refinement range had to be increased to 8, and tile size to 32.

B. Discussion

The results in Table I and Table II indicate that the number of average disparity hypothesis tested for each pixel can be decreased by about 75 % at the cost of increasing the number of mismatched disparities by not more than 3%. However, the results presented in Table III and Table IV show that the increase in the number of mismatched disparities, compared to the original algorithm results, cannot be easily lowered to zero. Even as the number of average disparities searched, approaches 80, which is a maximal disparity value, the results are not much better than those presented in Table I and Table II. This is most likely caused by the fact that some $C(x, y, d)$ values are marked as "unknown", which means that there is less information available than in the full-search case. The worst results are obtained for the Books stereo pair, which contains fairly numerous low-texture areas. In such areas it is usually better to use larger filter windows in order to aggregate costs from larger neighbourhood. In the case of the used local stereo matching algorithm this means construction of a larger colour-cross. However, such a cross has to span multiple tiles and the fact that disparity search range is determined for each tile almost independently severely limits the benefits of this approach. This is because a lot of larger crosses include many "unknown" $C(x, y, d)$ values. As a result an effective size of the cross is smaller, which may result in observed worse results.

TABLE III
RESULTS OBTAINED FOR THE LIMITED DISPARITY SEARCH ALGORITHM FOR $t = 2.8\%$ AND THE SMALLEST NUMBER OF DISPARITIES SEARCHED.

Pair name	Number of feature points in image[%]	Average feature points comparisons	Average disparities searched	Disparities searched / max disparity [%]
Baby1	3.31	21.68	66.70	83.38
Books	4.03	27.04	69.89	87.36
Dolls	3.42	22.39	60.46	75.58
Wood1	4.68	30.96	73.86	92.33
Cones	3.57	23.61	62.32	77.90
Teddy	3.91	25.23	68.70	85.88
Venus	3.92	26.55	63.57	79.46

TABLE IV
ERROR DIFFERENCES BETWEEN THE FULL AND LIMITED DISPARITY SEARCH ALGORITHMS FOR $t = 2.8\%$.

Pair name	Error		
	full search	limited search	difference
Baby1	6.89	9.17	2.28
Books	23.46	26.20	2.74
Dolls	18.18	20.09	1.91
Wood1	9.5	11.36	1.86
Cones	9.64	11.35	1.89
Teddy	13.71	15.11	1.40
Venus	1.78	2.80	1.02

The other problem is that the Harris corner feature points are found near the edges in the image, which are also often real objects' edges. In such a case, these feature points lie on disparity discontinuities. This means that when they are matched, the matching window covers pixels that should be assigned different disparities, which may result in a mismatch. However, the cross-checking of the Harris corner matches eliminates most bad matches, but sometimes also results in an exclusion of a disparity of a narrow object from the second stage search.

V. CONCLUSION

The incorporation of a Harris corner detector into a stereo matching algorithm can bring many benefits. As the Harris corners are features that are fairly stable across the images of the stereo pair, they can serve as a useful guide for selecting

disparities to be searched. The novel scheme of the disparity space search limiting that we propose yields disparity maps of the quality at most 3% worse than those obtained by the selected full search algorithm. The huge benefit is the reduction in the number of searched disparities by even 75%. The only additional cost is the need to calculate and match detected Harris corners, but their number is very low compared to the image size. Thus it does not add much complexity to the algorithm. Moreover, the way the proposed algorithm limits searched disparities is block based and each block is treated almost independently. This means that the algorithm can be very easily parallelised and implemented on the GPU or FPGA platforms.

REFERENCES

[1] D. Scharstein, R. Szeliski, and R. Zabih, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," in *Stereo and Multi-Baseline Vision, 2001. (SMBV 2001). Proceedings. IEEE Workshop on*, 2001, pp. 131 –140.

[2] R. Zabih and J. Woodfill, "Non-parametric local transforms for computing visual correspondence," in *European Conference on Computer Vision*, May 1994, pp. 151–158.

[3] K.-J. Yoon and I. S. Kweon, "Adaptive support-weight approach for correspondence search," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 28, no. 4, pp. 650 –656, 2006.

[4] A. Hosni, M. Bleyer, M. Gelautz, and C. Rhemann, "Local stereo matching using geodesic support weights," in *Image Processing (ICIP), 2009 16th IEEE International Conference on*, 2009, pp. 2093 –2096.

[5] C. Rhemann, A. Hosni, M. Bleyer, C. Rother, and M. Gelautz, "Fast cost-volume filtering for visual correspondence and beyond," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, june 2011, pp. 3017 –3024.

[6] K. Zhang, J. Lu, and G. Lafruit, "Cross-based local stereo matching using orthogonal integral images," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 19, no. 7, pp. 1073 –1079, 2009.

[7] X. Sun, X. Mei, S. Jiao, M. Zhou, and H. Wang, "Stereo matching with reliable disparity propagation," in *3D Imaging, Modeling, Processing, Visualization and Transmission (3DIMPVT), 2011 International Conference on*, may 2011, pp. 132 –139.

[8] V. Kolmogorov and R. Zabih, "Computing visual correspondence with occlusions using graph cuts," in *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, 2001.

[9] ——, "Multi-camera scene reconstruction via graph cuts," in *Computer Vision — ECCV 2002*, ser. Lecture Notes in Computer Science, A. Heyden, G. Sparr, M. Nielsen, and P. Johansen, Eds. Springer Berlin / Heidelberg, 2002, vol. 2352, pp. 8–40, 10.1007/3-540-47977-5_6. [Online]. Available: http://dx.doi.org/10.1007/3-540-47977-5_6

[10] P. Felzenszwalb and D. Huttenlocher, "Efficient belief propagation for early vision," *International Journal of Computer Vision*, vol. 70, pp. 41–54, 2006, 10.1007/s11263-006-7899-4. [Online]. Available: http://dx.doi.org/10.1007/s11263-006-7899-4

[11] J. Sun, N.-N. Zheng, and H.-Y. Shum, "Stereo matching using belief propagation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 25, no. 7, pp. 787 – 800, 2003.

[12] M. Bleyer and M. Gelautz, "A layered stereo algorithm using image segmentation and global visibility constraints," in *Image Processing, 2004. ICIP '04. 2004 International Conference on*, vol. 5, 2004, pp. 2997 – 3000 Vol. 5.

[13] M. Bleyer and M. Gelautz, "Graph-based surface reconstruction from stereo pairs using image segmentation," in *SPIE Symposium on Electronic Imaging 2005 (Videometrics VIII)*, vol. 5665, January 2005, pp. 288 – 299.

[14] H. Tao, H. Sawhney, and R. Kumar, "A global matching framework for stereo computation," in *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, vol. 1, 2001, pp. 532 –539 vol.1.

[15] M. Humenberger, C. Zinner, M. Weber, W. Kubinger, and M. Vincze, "A fast stereo matching algorithm suitable for embedded real-time systems," *Computer Vision and Image Understanding*, vol. 114, no. 11, pp. 1180 – 1202, 2010, special issue on Embedded Vision. [Online]. Available: http://www.sciencedirect.com/science/article/B6WCX-500SK2G-1/2/3ba40f1b09b54eeadde9df20105a819e

[16] J. Morris, K. Jawed, G. Gimel'farb, and T. Khan, "Breaking the ton achieving 1% depth accuracy from stereo in real time," in *Image and Vision Computing New Zealand, 2009. IVCNZ '09. 24th International Conference*, 2009, pp. 142 –147.

[17] S. Sabihuddin, J. Islam, and W. MacLean, "Dynamic programming approach to high frame-rate stereo correspondence: A pipelined architecture implemented on a field programmable gate array," may. 2008, pp. 001 461 –001 466.

[18] L. Zhang, "Fast stereo matching algorithm for intermediate view reconstruction of stereoscopic television images," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 16, no. 10, pp. 1259 –1270, 2006.

[19] Q. Yang, L. Wang, and N. Ahuja, "A constant-space belief propagation algorithm for stereo matching," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, june 2010, pp. 1458 –1465.

[20] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, pp. 91–110, 2004, 10.1023/B:VISI.0000029664.99615.94. [Online]. Available: http://dx.doi.org/10.1023/B:VISI.0000029664.99615.94

[21] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," in *Computer Vision – ECCV 2006*, ser. Lecture Notes in Computer Science, A. Leonardis, H. Bischof, and A. Pinz, Eds. Springer Berlin / Heidelberg, 2006, vol. 3951, pp. 404–417, 10.1007/11744023_32. [Online]. Available: http://dx.doi.org/10.1007/11744023_32

[22] C. Harris and M. Stephens, "A combined corner and edge detector," in *Fourth Alvey Vision Conference*, 1988, pp. 147–151.

[23] J. Kannala and S. Brandt, "Quasi-dense wide baseline matching using match propagation," in *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, 2007, pp. 1 –8.

[24] M. Lhuillier and L. Quan, "Match propagation for image-based modeling and rendering," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 8, pp. 1140 – 1146, Aug. 2002.

[25] J. Yao and W.-K. Cham, "3d modeling and rendering from multiple wide-baseline images by match propagation," *Signal Processing: Image Communication*, vol. 21, no. 6, pp. 506 – 518, 2006, special issue on multi-view image processing and its application in image-based rendering. [Online]. Available: http://www.sciencedirect.com/science/article/B6V08-4JKJSDF-3/2/81c4cd55f400ba7efe0d4dffceefcf66

[26] J. Cech and R. Sara, "Efficient sampling of disparity space for fast and accurate matching," in *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, june 2007, pp. 1 –8.

[27] D. Min, J. Lu, and M. Do, "A revisit to cost aggregation in stereo matching: How far can we reduce its computational redundancy?" in *Computer Vision (ICCV), 2011 IEEE International Conference on*, nov. 2011, pp. 1567 –1574.

[28] D. Scharstein and R. Szeliski, "High-accuracy stereo depth maps using structured light," vol. 1, jun. 2003, pp. I–195 – I–202 vol.1.

[29] H. Hirschmuller and D. Scharstein, "Evaluation of cost functions for stereo matching," jun. 2007, pp. 1 –8.

[30] D. Scharstein and C. Pal, "Learning conditional random fields for stereo," in *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, june 2007, pp. 1 –8.

An Approach to Adjustment and Reduction of the Number of Controlling Parameters for Simple 2D to 3D Image Conversion Schemes

Adam Dąbrowski, Julian Balcerek, Adam Konieczka
Poznań University of Technology, Chair of Control and System Engineering, Division of Signal Processing and Electronic Systems
ul. Piotrowo 3a, 60-965 Poznań, Poland
e-mail: {Adam.Dabrowski, Julian.Balcerek, Adam.Konieczka}@put.poznan.pl

ABSTRACT — **In this paper an approach to adjustment and reduction of the number of parameters controlling simple 2D to 3D image conversion schemes is presented. With the reported experiments five 2D to 3D conversion methods were examined and compared. The controlling parameters were experimentally adjusted by the viewers. The results of experiments indicate a linear dependence between the parameters. Thus it is possible to reduce them to only one parameter.**

KEYWORDS — stereovision impressions, 2D to 3D conversion, anaglyph technique

I. INTRODUCTION

Nowadays the image and video content is still to a large extent two-dimensional (2D), i.e., geometrically flat. Thus there is an urgent need for preparation of simple and effective tools (although possibly only approximate) for the 2D to 3D image conversion. Taking simplicity of visualization into account, which is adequate to our approximate approach, we have decided to consider the anaglyph technique [11–16] but our ideas may also be implemented with other visualization techniques like, e.g., the autostereoscopic screens [17].

In this paper we analyze and verify five variants of efficient real time 2D to 3D conversion schemes, proposed in our previous publications [1, 2], based on the following processing of the red color component:

- direct shift
- direct shift with interpolation
- mirroring with differential filtering
- segment shifting
- segment scaling.

All considered methods are based on shifts of the red color component, separately for objects and for the background, according to the depth (disparity) map [3–9].

The first method in the above list consists merely of the shifts. Using the second method, i.e., the direct shift with interpolation, the information holes occurring after shifting the red color component (i.e., regions without this component) are filled in by means of linear interpolation. In the third approach the red color component information holes are filled in with the respective regions of the mirror image followed by differential filtering. In case of the segment shifting the red component information holes are filled with the rescaled neighboring segments of the image. In the last considered method the existing image segments are scaled up to cover also the information holes.

In our approach we use binary depth maps. Thus, we have two red color component shift parameters: the first one is referred to the object and the second – to the background. Both parameters should be adjusted experimentally, separately for the object and for the background for all considered methods in order to obtain the best 3D effect. A problem of the proper choice of these two parameters for the described five 2D to 3D conversion methods, together with the report about our related research works in this field, is described in next paragraphs.

II. STEREOVISION EXPERIMENTS

A. Conditions of experiments

In the experiments three test images were used: *Lena_color* (the standard test image) with resolution 512×512 pixels, *bugatti* with resolution 518×389 pixels, and *slup* with resolution 509×382 pixels (Fig. 1).

Fig. 1. Images used for investigations during experiments: a) *lena_color*, b) *bugatti*, c) *slup*

56 students of the Poznan University of Technology participated in the experiments. The distance from the viewer to the 24 inch diagonal LCD screen was about 0.8 m. Viewers wore anaglyph glasses with the red and cyan filters.

In the experiments our test application implemented in the MATLAB environment, version 7.11.0 (R2010b) was used. The graphical user interface (GUI) is screened on Fig. 2.

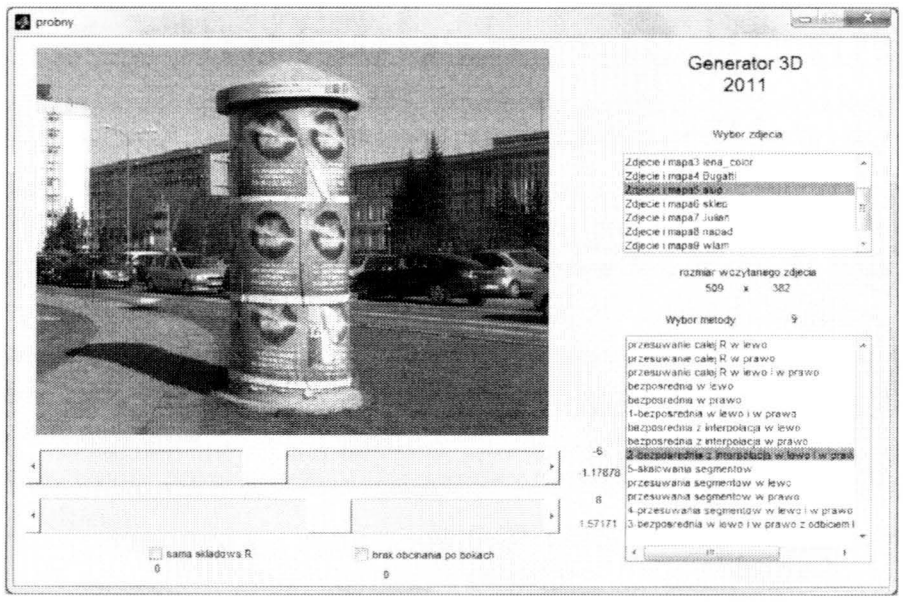


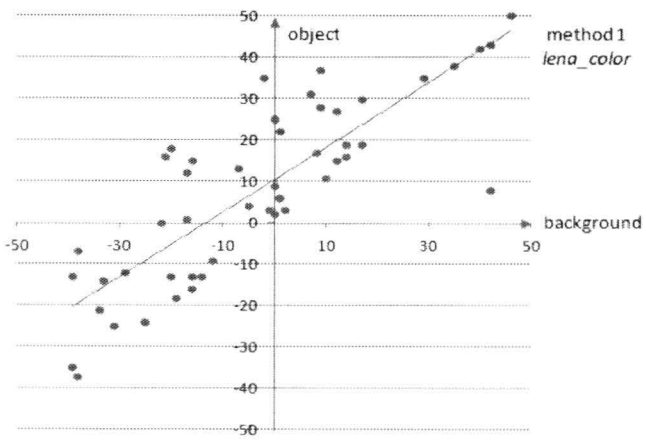
Fig. 2. GUI of MATLAB application used in experiments

Each examined person saw three test images for five methods. As the first method we chose the direct shift method, as the second – the direct shift with interpolation, as the third – the mirroring with differential filtering, as the fourth – the segment shift method, and as the fifth – the segment scaling method.

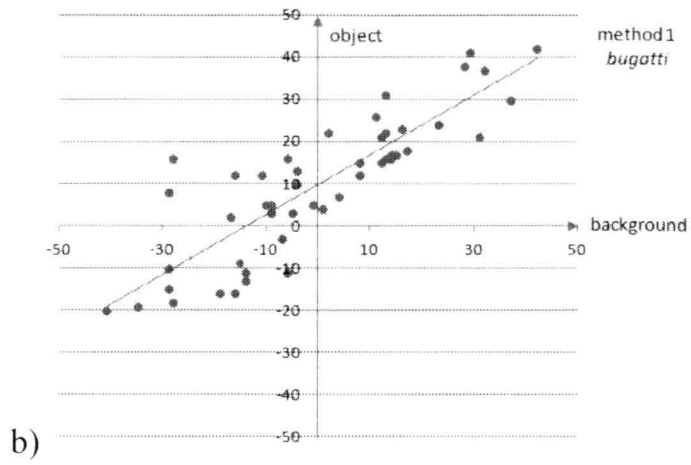
For each image the viewer had to adjust two best parameter values referred to each method for the background and for the object or affirm that the adjustment of the optimal parameters is impossible. These parameters are referred to the red color component shift or the level of rescaling [1, 2]. Viewers using two sliders arbitrary operated in the range from –50 to 50 pixels for parameters of the methods 1 - 4 and in the range from 1 to 6 for parameters of the method 5. The results of each parameter change were visualized on the application window in real time or in almost real time.

B. Results of experiments

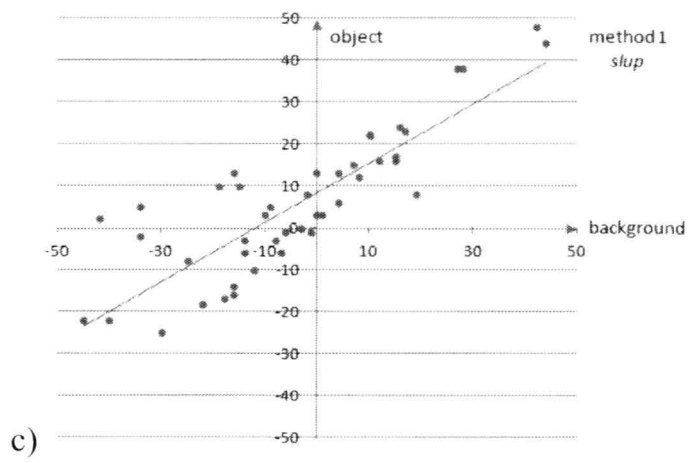
Figure 3 presents distribution of the chosen best values of parameters referred to the background versus the object chosen by the viewers separately for each image for all methods. Five most distinct points were removed for each experiment for the calculation of the linear regression and the square of the Pearson product moment correlation coefficient.



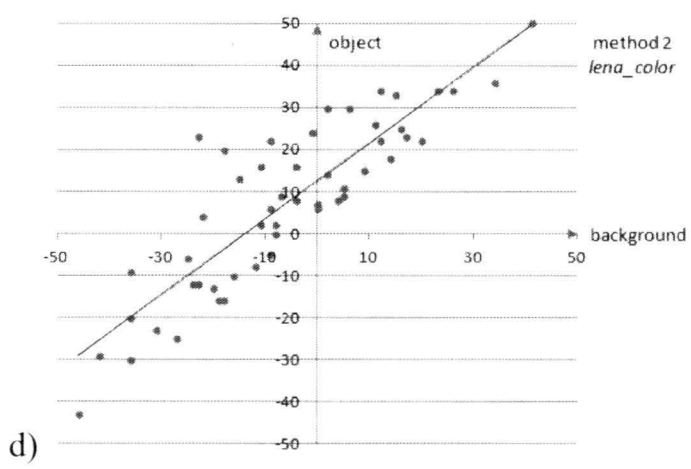
a)



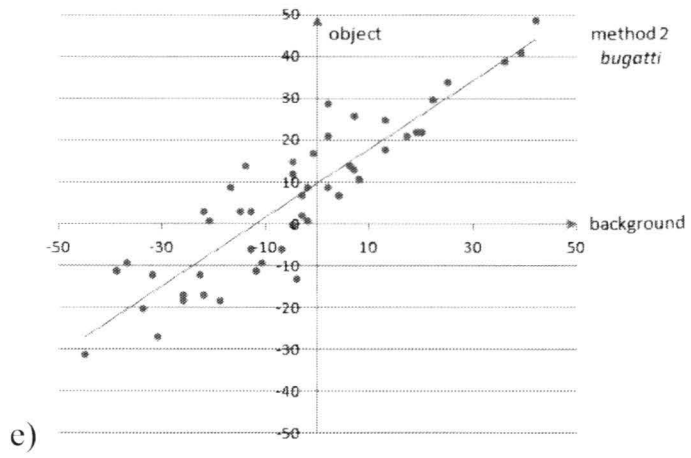
b)



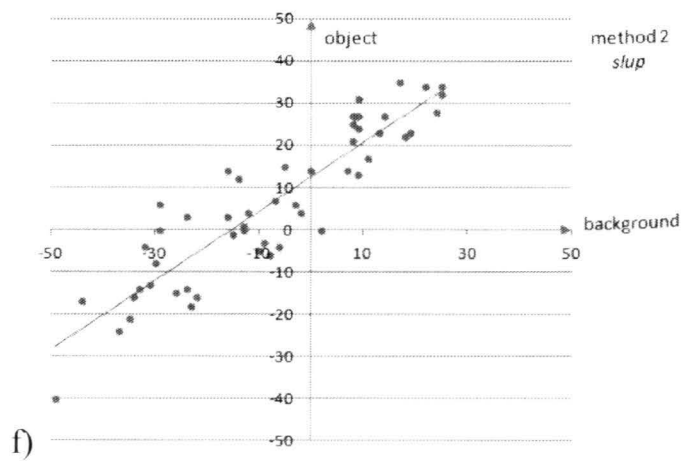
c)



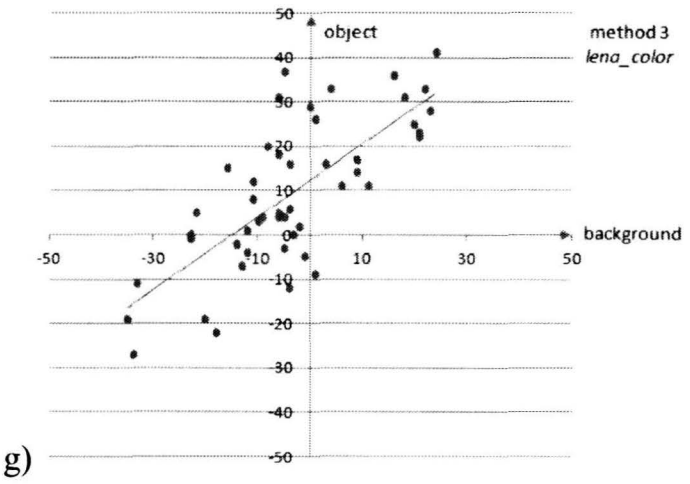
d)



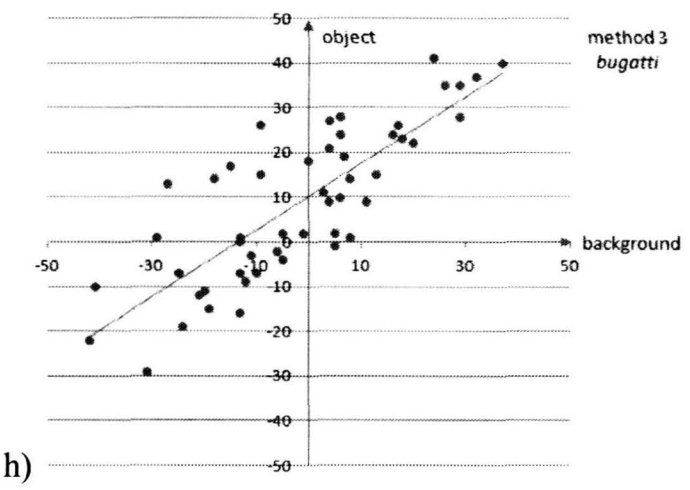
e)



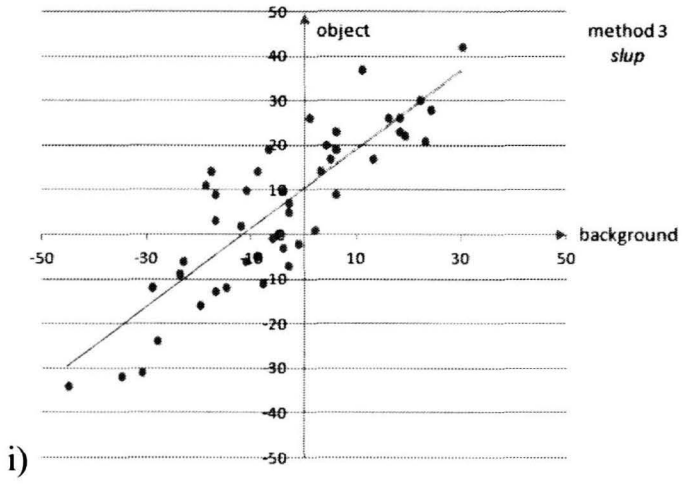
f)


object
method 3
lena_color
background

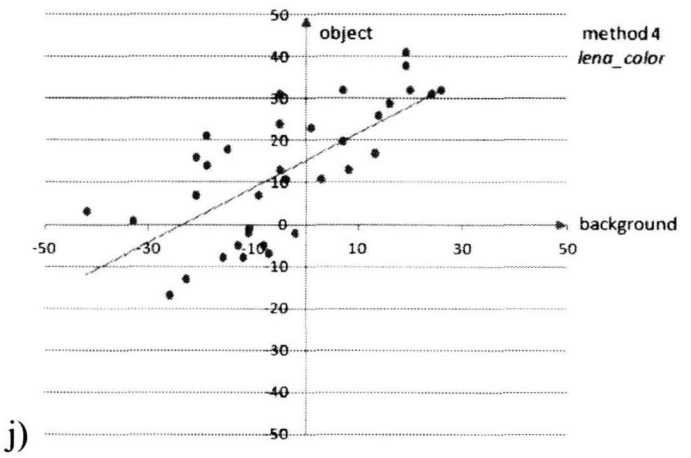
g)


object
method 3
bugatti
background

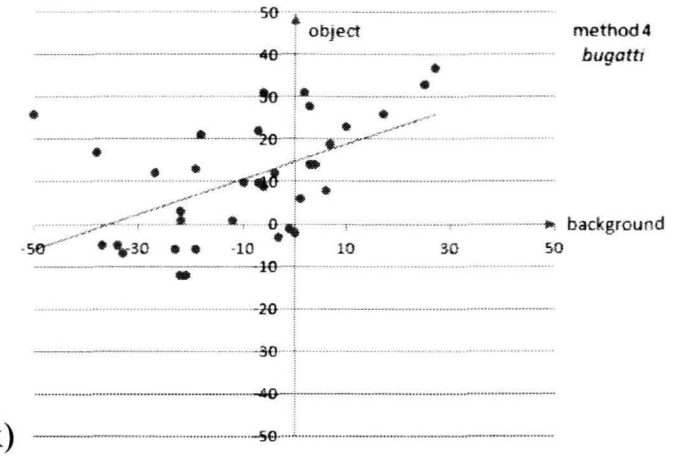
h)


object
method 3
slup
background

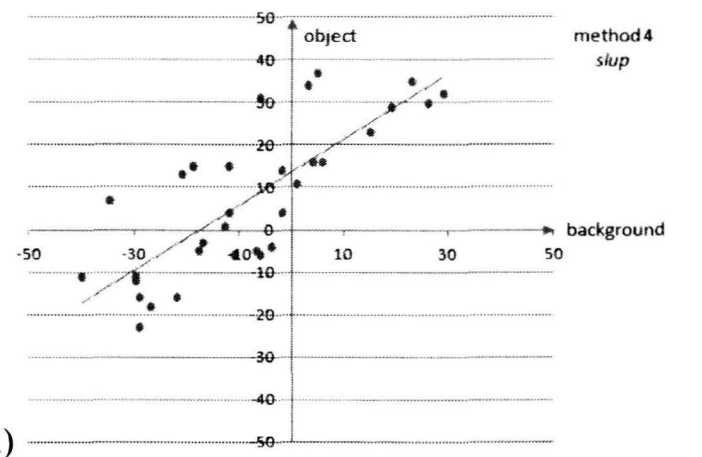
i)


object
method 4
lena_color
background

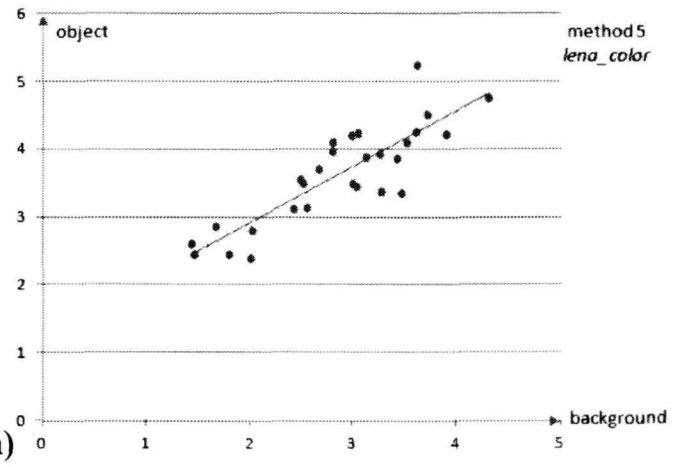
j)


object
method 4
bugatti
background

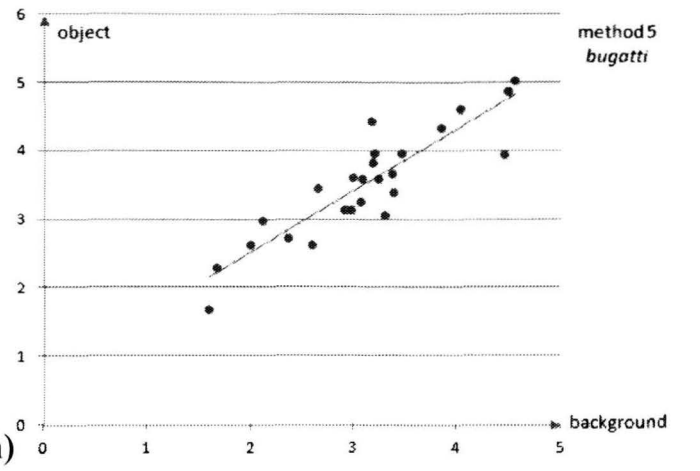
k)


object
method 4
slup
background

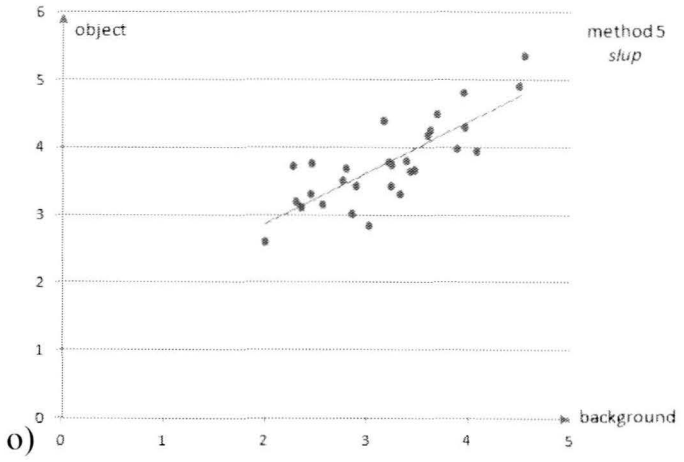
l)


object
method 5
lena_color
background

m)


object
method 5
bugatti
background

n)



o)

Fig. 3. a) - o) Distribution of best values of parameters referred to background versus object chosen by viewers for each method for all images

In Table 1 a comparison of the best parameters for the proposed conversion methods is presented. In cases of the direct shift, direct shift with interpolation as well as for the mirroring with differential filtering more than 91 % of viewers perceived a 3D effect and could adjust the optimal values of the available control parameters. A smaller number of people (less than 74 %) perceived the 3D effect and could adjust the optimal parameter values of for the segment shifting and the segment scaling methods. Thus these two methods have to be improved for future experiments.

A clearly linear relationship between parameters for the object and the background can be observed. Equations for the linear regression and the square of the Pearson product moment correlation coefficient are also presented in Table 1.

TABLE 1. COMPARISON OF 2D TO 3D CONVERSION METHODS

Methods	Image	Percentage of people who saw 3D effect and could adjust optimal parameters [in %]		Linear regression	Square of Pearson product moment correlation coefficient
1. direct shift	*lena_color*	100	mean: 95.24	$y = 0.78x+10.48$	$R^2 = 0.68$
	bugatti	94.64		$y = 0.71x+9.77$	$R^2 = 0.73$
	slup	91.07		$y = 0.71x+8.46$	$R^2 = 0.71$
2. direct shift with interpolation	*lena_color*	100	mean: 98.81	$y = 0.90x+12.59$	$R^2 = 0.76$
	bugatti	96.42		$y = 0.82x+9.75$	$R^2 = 0.81$
	slup	100		$y = 0.82x+12.36$	$R^2 = 0.82$
3. mirroring with differential filtering	*lena_color*	96.43	mean: 97.02	$y = 0.82x+12.18$	$R^2 = 0.55$
	bugatti	98.21		$y = 0.75x+10.13$	$R^2 = 0.66$
	slup	96.43		$y = 0.88x+10.36$	$R^2 = 0.72$
4. segment shift	*lena_color*	73.21	mean: 70.83	$y = 0.65x+15.26$	$R^2 = 0.49$
	bugatti	73.21		$y = 0.41x+14.72$	$R^2 = 0.30$
	slup	66.07		$y = 0.77x+13.49$	$R^2 = 0.62$
5. segment scaling	*lena_color*	58.93	mean: 57.74	$y = 0.83x+1.26$	$R^2 = 0.73$
	bugatti	53.57		$y = 0.90x+0.70$	$R^2 = 0.80$
	slup	60.71		$y = 0.76x+1.33$	$R^2 = 0.65$

III. CONCLUDING REMARKS

The results of our experiments demonstrate that the best 3D illusion visibility is obtained for the direct shift, direct shift with interpolation, and mirroring with differential filtering. These methods are simple, low computational cost (thus applicable in real time), and offer a relatively good 3D quality anaglyph images.

Experimentally determined linear correlations (the obtained values of the square of the Pearson product moment correlation coefficient) show strong linear dependences between the adjustable parameters for all examined 2D to 3D conversion methods. In result, it is possible to reduce the number of the adjustable parameters to only one with the second being controlled according to the predetermined linear function. The user may conveniently operate on this one parameter only to obtain the best perceptible and the most pleasant 3D effect (Fig. 4).

The presented results may be adapted to other visualization methods than the anaglyphs and may be used not only for the 2D to 3D conversion but also in the 3D imaging.

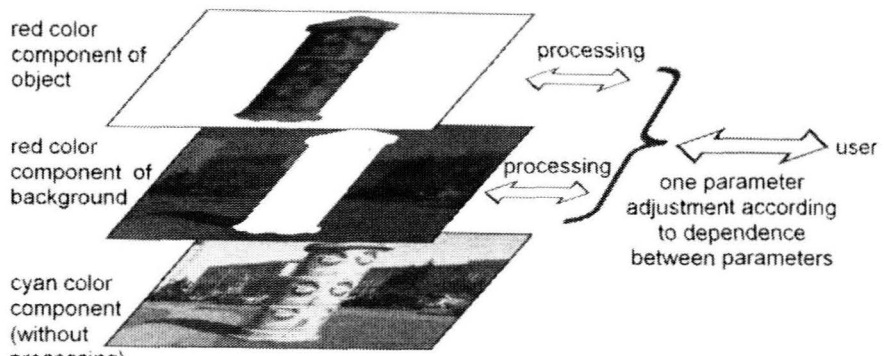


Fig. 4. Concept of interface for only one parameter manually adjustable by user for 3D effect visualization

In future, we plan to perform new experiments using more advanced depth maps (i.e. those with higher numbers of distances), using other (more advanced) visualization methods like e.g. the autostereoscopy, and improving the proposed conversion methods. We will also try to qualify and quantify influences of all important (e.g., environmental conditions) in order to automatically determine not only the best conversion method but also its best controlling parameters, thus making a fully automatic 2D to 3D conversion possible.

REFERENCES

[1] J. Balcerek, A. Dąbrowski, A. Konieczka, Simple efficient techniques for creating effective 3D impressions from 2D original images, Proc. of New Trends in Audio and Video / Signal Processing NTAV / SPA'2008, IEEE Poland Section Chapters Signal Processing, Circuits and Systems, pp. 219-224, Poznań, Poland, 25-27th September 2008.

[2] J. Balcerek, A. Konieczka, A. Dąbrowski, T. Marciniak, Binary depth map generation and color component hole filling for 3D effects in monitoring systems, Proc. of Signal Processing SPA'2011, IEEE, Poland Section, Chapters Signal Processing, Circuits and Systems, pp. 138-143, Poznań, Poland, 29–30 September 2011.

[3] W.J. Tam, L. Zhang, 3D-TV Content Generation: 2D-to-3D Conversion. Multimedia and Expo, 2006 IEEE International Conference on July 2006, Pages: 1869–1872.

[4] Y.L. Chang, C.Y. Fang, L.F. Ding, S.Y. Chen, L.G. Chen, Depth Map Generation for 2D-to-3D Conversion by Short-Term Motion Assisted

Color Segmentation. Multimedia and Expo, 2007 IEEE International Conference on 2–5 July 2007, Pages: 1958 1961.

[5] I. Ideses, L.P. Yaroslavsky, B. Fishbain, Real-time 2D to 3D video conversion. Real-Time Image Proc (2007) 2:3 9.

[6] A. Redert, R.P. Berretty, C. Varekamp, O. Willemsen, J. Swillens, H. Driessen, Philips 3D Solutions: From Content Creation to Visualization. 3D Data Processing, Visualization, and Transmission, Third International Symposium on June 2006, Pages: 429–431.

[7] D. Kim, D. Min, K. Sohn, Stereoscopic Video Generation Method Using Motion Analysis. 3DTV Conference, 2007, 7 9 May 2007, Pages: 1 4.

[8] I. Ideses, L. Yaroslavsky, I. Amit, B. Fishbain, Depth Map Quantization How Much is Sufficient? 3DTV Conference, 2007, 7–9 May 2007, Pages: 1 4.

[9] L. Zhang, W.J. Tam, Stereoscopic image generation based on depth images for 3D TV. Broadcasting, IEEE Transactions on Volume 51, Issue 2, June 2005, Pages: 191 199.

[10] J. Balcerek, A. Konieczka, A. Dąbrowski, M. Stankiewicz, A. Krzykowska. Brightness Correction and Stereovision Impression Based Methods of Perceived Quality Improvement of CCTV Video Sequences. Proc. of 4th International Conference MCSS 2011, 64 72, June 2 3, Krakow, Poland. Springer. Communications in Computer and Information Science, 2011.

[11] L.Onural, An Overview of Research In 3DTV. Systems, Signals and Image Processing, 2007 and 6th EURASIP Conference focused on Speech and Image Processing, Multimedia Communications and Services. 14th International Workshop on 27 30 June 2007, Page: 3.

[12] A.A. Dumbreck, C.W. Smith, 3-D TV displays for industrial applications. Stereoscopic Television, IEE Colloquium on, 15 Oct 1992, Pages: 7/1 7/4.

[13] E. Dubois, A projection method to generate anaglyph stereo images. Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP'01). 2001 IEEE International Conference on, Volume 3, 7 11 May 2001 Pages: 1661 1664 vol. 3.

[14] W.A. IJsselsteijn, H. de Ridder, J. Vliegen, Subjective evaluation of stereoscopic images: effects of camera parameters and display duration. Circuits and Systems for Video Technology, IEEE Transactions on, Volume 10, Issue 2, March 2000, Pages: 225 233.

[15] A.C. Gallagher, Detecting anaglyph images with channel alignment features. Proc. of 2010 17th IEEE International Conference on Image Processing (ICIP), 26 29 September 2010, Hong Kong, Pages: 2985 2988.

[16] L. Stelmach, W.J. Tam, D. Meegan, A. Vincent, Stereo image quality: effects of mixed spatio-temporal resolution. Circuits and Systems for Video Technology, IEEE Transactions on, Volume 10, Issue 2, March 2000, Pages: 188 193.

[17] H. Urey, K.V. Chellappan, E. Erden, P. Surman, State of the Art in Stereoscopic and Autostereoscopic Displays. Proceedings of the IEEE, Volume: 99 , Issue: 4, April 2011, Page(s): 540 555.

The paper was prepared within the INDECT and DS projects.

A VLSI Architecture for Intra Prediction for a HEVC Decoder

Andrzej Abramowski
Institute of Radioelectronics
Warsaw University of Technology
Warsaw, Poland 00–665
e-mail: A.Abramowski@ire.pw.edu.pl

Grzegorz Pastuszak
Institute of Radioelectronics
Warsaw University of Technology
Warsaw, Poland 00–665
e-mail: G.Pastuszak@ire.pw.edu.pl

ABSTRACT — **One of the key changes introduced in the High Efficiency Video Coding (HEVC) in comparison to the H.264/AVC is a revised algorithm of Intra prediction. Unfortunately, along with a significant improvement in the performance, its complexity has also been increased. This growth is particularly noticeable in the design of potential hardware architectures, due to a substantial augmentation in the number of supported modes and block sizes. This article presents an Intra prediction module architecture for the decoder, compliant with the Main profile of the sixth version of the HEVC draft. It allows the throughput of at least one sample per clock cycle at a moderate consumption of hardware resources, what should be enough to provide support for HDTV sequences in real-time for a 100 MHz clock.**

I. INTRODUCTION

The rapid development of multimedia technologies provides users with a better quality of experience almost every day. This particularly affects the video coding, where in recent years increasing demands have been put on an improvement of the quality of the compressed video and a support for sequences with increasingly higher resolutions (HDTV, UHDTV). As a result, even the latest solutions as the current state-of-the-art H.264/AVC, often are insufficient. To address these needs, ITU-T and ISO/IEC came together and formed a group of experts, named Joint Collaborative Team on Video Coding (JCT-VC), in order to prepare a new standard. The call for proposals on video compression technology was announced in January 2010 and 27 full propositions were submitted from over 20 companies in response. After more than two years of work and eight meetings of the group a new solution is on the verge of ratification under the name of High Efficiency Video Coding (HEVC). Preliminary tests indicate that it allows almost 40% coding efficiency improvement over the H.264/AVC High profile [1], while maintaining flexibility necessary for a wide variety of applications.

One of the main areas of work while creating a new standard was the development of Intra prediction. The result is a significant increase in the complexity of the algorithm, described in the sixth version of the HEVC draft [2] in comparison to the H.264/AVC. The number of available prediction modes has been increased from 9 to 35 for the Main profile and 36 overall. Moreover Intra prediction can operate on larger blocks, since in addition to standard 4x4, 8x8 and 16x16 blocks, 32x32 and 64x64 blocks are also available. This should meaningfully improve the compression quality, especially for large homogeneous areas of an image. Unfortunately, these changes may hinder hardware implementation. Firstly, they impose substantial requirements on the amount of resources needed to store reference samples. Secondly, it is a complex task to design a uniform architecture for all possible Intra prediction modes.

In this paper an efficient FPGA architecture for Intra prediction for a HEVC decoder, compliant with the sixth version of the HEVC draft, is presented. It supports all the prediction modes defined for the Main profile with moderate resource consumption thanks to a reduced number of registers in favor of memory cells. The module is able to process sequences with different chroma sub-sampling: 4:2:0, 4:2:2 and 4:4:4 and may be easily configured to work with sample bit depths higher than eight bits.

The rest of the paper is organized as follows. Section II provides a short overview of the Intra prediction in the sixth version of the HEVC draft. Section III contains the description of proposed FPGA architecture. The simulation and synthesis results are summarized in Section IV. Finally, Section V concludes the paper.

II. INTRA PREDICTION ALGORITHM

Abrupt changes in pixel values do not appear frequently in natural images, the transition between neighboring samples is smooth and continuous. This property makes it possible to forecast the pixel value on the basis of its surroundings. The technique is widely known as Intra prediction. The HEVC draft [2] contains several different approaches to the presented problem, described in the following subsections.

A. Prediction Unit

Prediction Unit (PU) is a square block with sizes ranging from 4x4 to 64x64. The process of calculation of Intra prediction for pixels inside the PU is based on reference samples located at the upper and left border. Their number can be up to twice the length of the edge of the PU for each direction (Fig. 1). If any of required reference samples is unavailable, it should be replaced with the closest attainable sample. When none of the reference samples is available, they

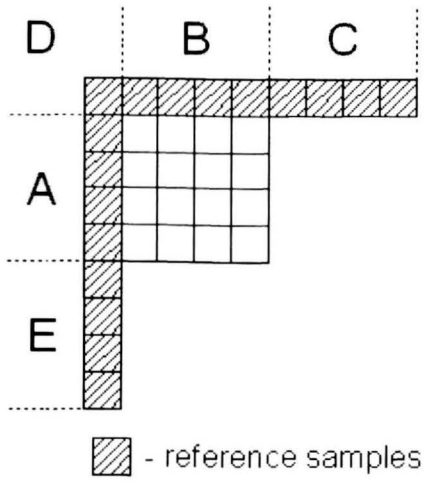


Fig. 1. Visualization of the potential reference samples for the 4x4 PU, divided into five regions labeled A, B, C, D and E.

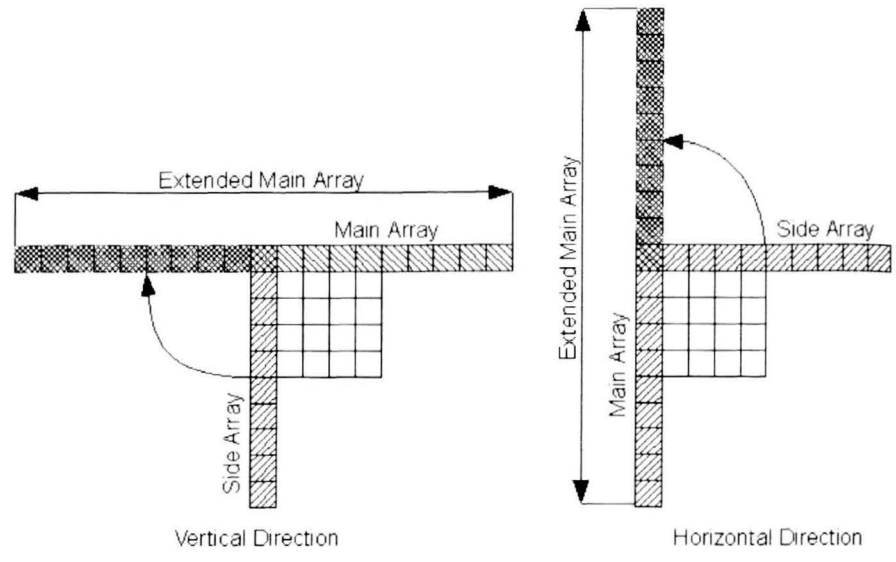


Fig. 2. The process of creation Extended Main Array in 4x4 PU.

are assigned the default value defined by a given bit depth. Some combinations of the prediction mode and block size also requires a pre-filtration of reference samples, calculated as a normalized weighted sum of three adjacent samples:

$$pF[x] = (p[x-1] + 2 \cdot p[x] + p[x+1]) >> 2 \qquad (1)$$

B. Planar prediction

The planar prediction type is designed to ease the reconstruction of smooth image segments, through removing the negative impact of the division into blocks. It ensures the continuity of the image plane at the PUs' borders, through a creation of a gradual change in the value of pixels. The values of the prediction are calculated using the following equation:

$$\begin{aligned} p[x,y] = ((nS-1-x) \cdot p[-1,y] + \\ (x+1) \cdot p[nS,-1] + (nS-1-y) \cdot p[x,-1] + \\ (y+1) \cdot p[-1,nS] + nS) >> (k+1) \end{aligned} \qquad (2)$$

where nS denotes the PU size, $x, y = 0 \ldots nS-1$ and determines position within the PU and $k = log_2(nS)$.

C. DC prediction

The DC mode assigns the mean of the reference samples from the upper and left neighbor (A and B in Fig. 1) to predicted values. To improve the subjective quality of the algorithm an additional filtration process is present for the luminance component. For each sample lying on the lupper and left edge of the PU, the average is modified by taking into account the value of the reference sample (or samples for the corner position) adjacent to the position of prediction, leading to a smooth transition between blocks.

D. Angular 10 and 26 prediction

In the angular 10 mode the prediction is equal to the reference sample from the left neighbor in the same row, whereas in the angular 26 mode, the prediction is equal to the reference sample from the upper neighbor in the same column. An additional filtration is required for luminance samples lying on the edge of the PU adjacent to the neighbor, which normally does not participate in the computation of the prediction.

E. Angular (2..9, 11..25, 27..34) prediction

To calculate the prediction in the angular mode, it is necessary to create two separate arrays, corresponding accordingly to reference samples located above the current block for the first one and reference samples on the left side for the second one. Based on the main direction of the prediction (vertical or horizontal) one of these arrays is designated as Main Array, while the other is named Side Array. If it is the vertical direction mode, Main Array is located above the current block. Otherwise, Main Array consists of samples from the left side of the PU.

The angular prediction mode determines a value of an angle. When the angle value is positive, only reference samples from Main Array are used. Otherwise reference samples from Main and Side arrays may be utilized, creating Extended Main Array (Fig. 2) [3], [4]. The prediction value is determined by linear interpolation of neighboring reference samples, stored in the final array:

$$iIdx = ((y+1) \cdot intraPredAngle) >> 5 \qquad (3)$$

$$iFact = ((y+1) \cdot intraPredAngle) \& 31 \qquad (4)$$

$$\begin{aligned} p[x,y] = ((32 - iFact) \cdot refArray[x + iIdx + 1] + \\ iFact \cdot refArray[x + iIdx + 2] + 16) >> 5 \end{aligned} \qquad (5)$$

where nS denotes the PU size, $x, y = 0 \ldots nS-1$ and determines position within the PU.

F. Intra From Luma

Intra From Luma is a prediction mode not available in the Main profile. It defines a method to predict the chrominance values for a given PU on the basis of the reconstructed luminance samples. It implies the largest computational burden from all the Intra prediction modes in the HEVC standard.

III. Proposed architecture

When designing this Intra prediction module the primary goal was to ensure a throughput of at least one sample per clock cycle, which should allow the decoder to process HDTV sequences with 4:2:0 chroma subsampling in real time

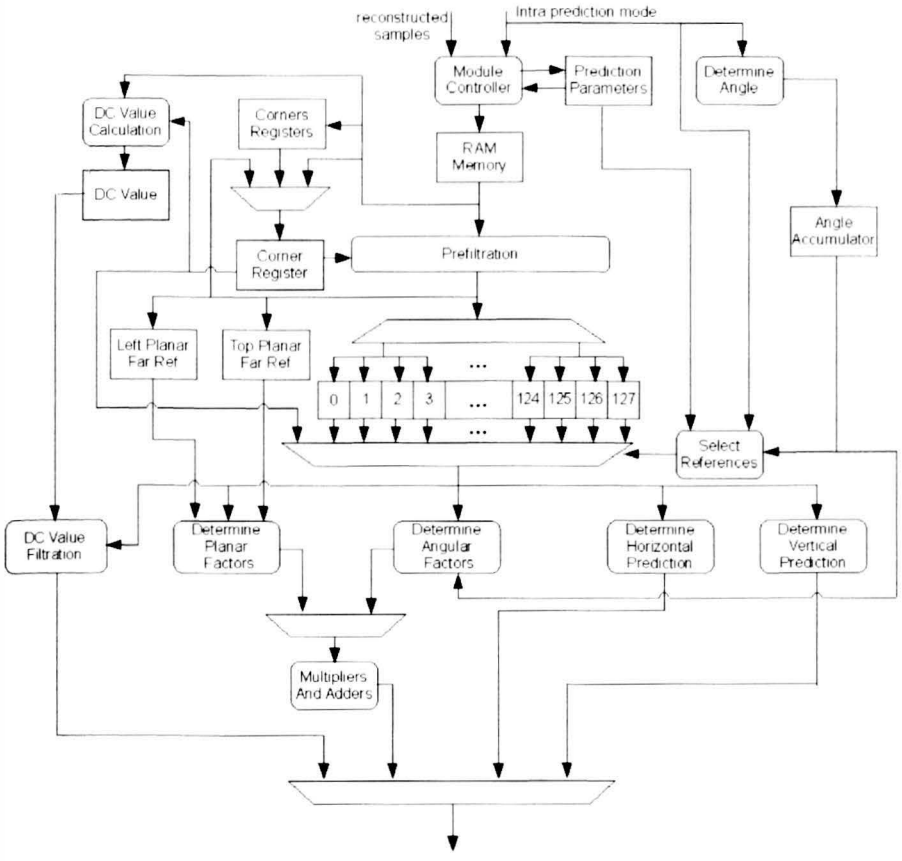


Fig. 3. A proposed architecture for intra prediction in the HEVC decoder.

using 100 MHz clock. To achieve this objective, regardless of the prediction mode, the module calculates a prediction for the next four pixels in the raster scan order. The approach leads to a regular structure, shown in general in Fig. 3.

In the proposed solution reference samples for the upper and left neighborhood of each image component are stored in a RAM memory in independent address subspaces. The subspace for the top neighborhood is significantly larger, as it is necessary to keep bottom line of reconstructed samples from the PUs for the whole image. Each memory cell stores four horizontally adjacent samples. Samples, which may become corners for further blocks, are stored separate in registers (*Corners Registers*). The first step to prepare prediction for each mode is to read necessary reference samples from the memory and save them to a registers array (*0 - 127*). A relatively complicated algorithm determining the necessary reference samples, especially for the angular prediction, may be significantly simplified. The four locations can be specified in dependence on the prediction mode and using the designation of areas in Fig. 1:

- reference samples situated in areas marked as A, D and E for Intra prediction modes from 27 to 34 inclusive
- reference samples situated in areas marked as B, D and C for Intra prediction modes from 2 to 9 inclusive
- reference samples situated in areas marked as A, B and D for Intra prediction modes 1 and from 10 to 26 inclusive, if the pre-filtration of reference samples isn't required
- reference samples situated in areas marked as A, B and D plus the first two reference samples from areas C and E for Intra prediction modes 0 and from 10 to 26 inclusive, if the pre-filtration of reference samples is required

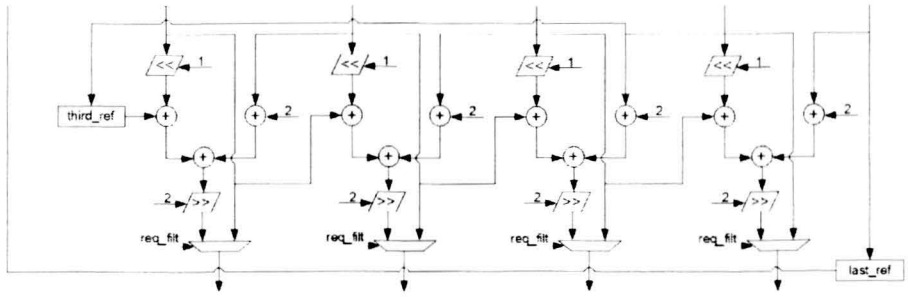


Fig. 4. A pre-filtration of a single reference samples set.

The most time consuming is the last possibility, as it takes half of the PU side in pixels plus four clock cycles to obtain all the necessary samples. Taking into account that the maximum block size is 64x64 in the worst case 131 reference samples have to be read. Most of them is written to the registers array. The exception is the corner sample (*Corner Register*), as it serves as a replacement for unavailable pixels, and two values of samples for the planar prediction, labeled as $p[nS, -1]$ (*Top Planar Far Ref*) and $p[-1, nS]$ (*Left Planar Far Ref*) in Eq. 2.

When the pre-filtration of the reference samples is required, for the first, second, and third sample, it is carried out immediately after the reading from the memory, while the last sample is stored in the register unchanged. At the same time, third sample is saved in the collateral register. It is used to calculate the value of the filtered fourth sample at the time of the next reading. This is shown in detail in Fig. 4.

The set of reference samples is subsequently obtained from the registers based on the prediction mode and the position within the PU, then it goes in parallel to one of five possible modules responsible for the calculation of the individual prediction type. Finally, the correct values for the four samples in the raster scan order are chosen by the mode.

A. Planar prediction

It is possible to optimize the Eq. 2 by changing the order of arithmetic operations, similarly as it is performed in the HM 6.1 reference software:

$$\begin{aligned} p[x,y] = (&nS + nS \cdot p[-1,y] + nS \cdot p[x,-1] + \\ &(x+1) \cdot (p[nS,-1] - p[-1,y]) + \\ &(y+1) \cdot (p[-1,nS] - p[x,-1])) >> (k+1) \end{aligned} \tag{6}$$

The size of the PU can be only a power of two, so the $nS \cdot p[-1,y]$ and $nS \cdot p[x,-1]$ parts of the equation can be calculated merely by shifting bits, eliminating two multiplications. This means, it is possible to significantly reduce the logic responsible for the calculations, as it is shown in Fig. 5. Moreover, derived form is so similar to the Eq. 5 it allows sharing multipliers with the angular prediction, which is presented symbolically in Fig. 3.

B. Angular (2..9, 11..25, 27..34) prediction

The basis of the prediction is the table, which allows a mapping of the index value calculated from Eq. 3 to the position of a reference sample in the registers array. This table

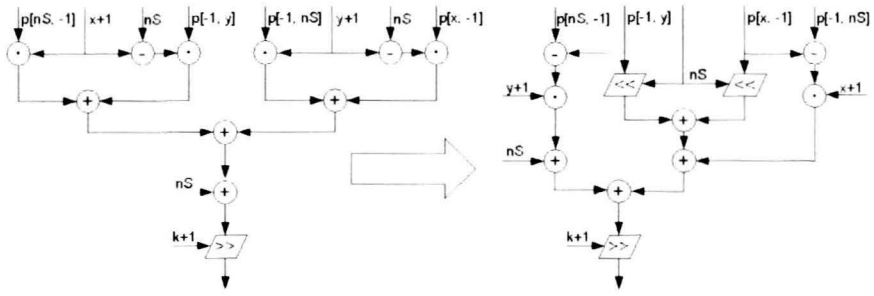


Fig. 5. An optimization of the planar prediction mode equation.

is constructed according to the size of the PU and the prediction mode and implemented as a LUT. Further optimizations include replacing multiplication from Eq. 3 and Eq. 4 with an angle accumulator, which value increases at each transition to the next row of the PU, as in [3]. Thanks to this $iIdx$ and $iFact$ values may be determined as the six most significant and the five least significant bits of the accumulator respectively. The result of Eq. 5 is calculated using combinational logic shared with the planar prediction path.

C. Other prediction modes

The DC value is calculated in parallel to the transfer of reference samples from the memory to the registers array. The angular 10 and 26 modes come down to the assignment of the adequate reference sample to the prediction value. Any filtration, if needed, is performed on an ongoing basis for each of these modes.

D. Reconstruction

Reference samples from the upper and left PU surroundings, which can be used as corner samples for the subsequent PUs, are transferred to dedicated registers when Intra prediction is prepared. Thus, immediately after the end of calculations, the module waits for the reconstructed sample. The interface is designed to accept four reconstructed samples in the raster scan order. Received samples, located at the right and bottom edges of the block, are stored at the appropriate addresses in the memory. Estimated number of clock cycles allocated for this process corresponds to the number of samples within the PU divided by four, and increased by one. The additional clock cycle is required due to the fact that the last sample must be stored as the potential reference sample from the upper and from the left neighbor, which means that it must be written to two different locations in the memory.

IV. Implemmentation results

The correctness of the proposed architecture was verified on the basis of the HM 6.1 reference software, modified in order to provide the necessary information in the text form. This software is released in successive versions along with the emergence of updates of the HEVC draft and reflects the state of development of the standard at a time.

The Intra prediction module was synthesized using Synopsys tools using the TSMC 0.13μm technology. Results are presented in the Tab. I and show more than an quintuple increase in the consumption of hardware resources in relation to the synthesis results presented in [3]. The main factor causing such an adverse outcome for the proposed architecture is probably the ability to handle the full range of PU sizes, instead of just 4x4 as in [3]. This entails the need to use considerably larger number of registers and more combinational logic. The minor differences, which may also influence the comparison, are the support of the planar prediction mode and significantly higher throughput. Unfortunately, the novelty of the standard limits the number of available solutions, which one could refer to.

TABLE I
SUMMARY OF SYNTHESIS RESULTS AND COMPARISON WITH OTHER WORKS.

Architecture	Proposed	[3]
Technology	TSMC 0.13μm CMOS	
Logic gate count	54745	9020
Maximum operation frequency	150 MHz	150 MHz
Average cycles to generate a pixel	< 1	1.5

TABLE II
TIME DISTRIBUTION FOR THE INDIVIDUAL STEPS OF THE ALGORITHM FOR THE PROPOSED ARCHITECTURE, AND AN INDICATION OF THE AVERAGE NUMBER OF CLOCK CYCLES SPENT ON THE CALCULATION OF THE PREDICTION FOR A SINGLE PIXEL, DEPENDING ON THE SIZE OF THE PU.

PU Size	Gathering reference samples	Preparing Intra prediction	Reconstruction	Sum	Cycles per pixel
4x4	6	4	5	15	0.94
8x8	8	16	17	41	0.64
16x16	12	64	65	141	0.55
32x32	20	256	257	533	0.52
64x64	36	1024	1025	2085	0.51

Tab. II shows the number of clock cycles needed to determine the prediction depending on the size of the block in the worst case. The presented results are based on the assumption that the transfer of reconstructed samples to the module will start immediately after the end of the calculation of Intra prediction. The most important parameter is the average number of clock cycles needed to calculate the prediction for a single sample, as it allows an easy estimation of the speed of the module. Its value drops with the increase in the size of the PU, due to the relative minimization of the time spent on obtaining reference samples from the memory. This feature appears to be particularly useful for high resolution video sequences. It is also worth noting that the maximum throughput of the module comes to almost two samples per clock cycle.

V. Conclusion

In this paper, an efficient FPGA architecture of intra prediction for the HEVC decoder is proposed. It is fully compliant with the Main profile defined in the sixth version of the HEVC draft. Thanks to using an uniform structure for angular and planar prediction modes, the throughput of at least one sample per clock cycle is assured.

Additional work associated with the development of the architecture should focus on providing support for the Intra From Luma mode. This is certainly the most complicated Intra prediction algorithm, available in the standard,

and as such is not included at the present time in the Main profile.

REFERENCES

[1] B. Li, G. J. Sullivan, and J. Xu, "Comparison of compression performance of hevc working draft 5 with avc high profile," JCT-VC, Tech. Rep., February 2012.

[2] B. Bross, W.-J. Han, G. J. Sullivan, J.-R. Ohm, and T. Wiegand, "High efficiency video coding (hevc) text specification draft 6," JCT-VC, Tech. Rep., February 2012.

[3] F. Li, G. Shi, and F. Wu, "An efficient vlsi architecture for 4x4 intra prediction in the high efficiency video coding (hevc) standard," in *Image Processing (ICIP), 2011 18th IEEE International Conference on*, sept. 2011, pp. 373 –376.

[4] Y. Liu, "Analysis of coding tools in hevc test model (hm 1.0) intra prediction," December 2010, http://www.h265.net.

NTAV/SPA 2012
27-29TH SEPTEMBER, 2012, ŁÓDŹ, POLAND
NEW TRENDS IN AUDIO AND VIDEO / SIGNAL PROCESSING ALGORITHMS, ARCHITECTURES, ARRANGEMENTS AND APPLICATIONS

Encoding of the Structured Light Stripe Pattern Video Sequences

Andrzej Buchowicz, Władysław Skarbek
Institute of Radioelectronics
Warsaw University of Technology
ul. Nowowiejska 15/19, 00-665 Warszawa
e-mail: {A.Buchowicz W.Skarbek}@ire.pw.edu.pl

Piotr Garbat
Institute of Microelectronics and Optoelectronics
Warsaw University of Technology
ul. Koszykowa 75, 00-662 Warszawa
e-mail: P.Garbat@imio.pw.edu.pl

ABSTRACT — **The analysis of the compression methods of the stripe patterns video sequences used in structured light technique is presented in this paper. It has been shown that the illumination of the 3-D scene be the sequentially repeating stripe patterns introduces additional correlation to the structured light video sequence. The experimental results presented in this paper confirm that this correlation can be effectively utilized for the video encoding.**

I. INTRODUCTION

Structured light is one of the techniques used for construction of 3-D scene models out of its 2-D views [1], [2]. The structured light 3-D model capturing system is composed of a computer beamer/projector and a video camera. The analysed scene is illuminated by the projector with a sequence of fixed stripe patterns. Images of the scene illuminated with stripe patterns are registered by a video camera. The scene depth information can be derived by the analysis of the stripe pattern distortions caused be the object shape. This analysis is a compute-intensive process and often can not be performed by the capturing system in real time. In such cases the sequence of stripe patterns images must be stored for off-line processing or transmitted to other systems to be processed. Since such sequence is represented by huge amount of data, the high capacity storage devices or high throughput transmissions channels are required. It is desirable to reduce the sequence bitrate to lower the requirements on storage devices or transmission channels. The analysis of the efficiency of video compression techniques applied to the sequence of the stripe patterns images is presented in this paper.

A. *Video coding*

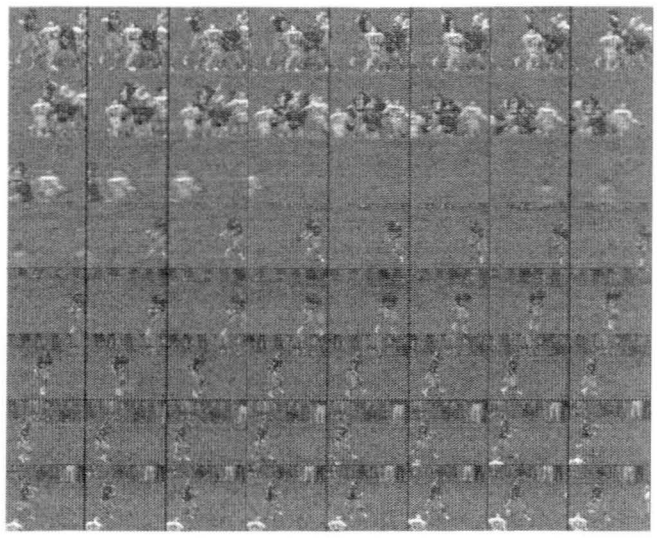
The video sequence is the representation of a natural scene sampled temporally and spatially. The goal of a video coding is to reduce the amount of data required to represent the video sequence while preserving its quality. The video compression is achieved by the exploitation of the temporal and spatial redundancy in a video sequence. Virtually all video coding standards are based on the motion compensated DCT and DCPM hybrid coding algorithm [3]. The MPEG-4 AVC/H.264 standard [4] is currently the most often used in the area of video coding. It is based on the hybrid coding algorithm like its many predecessors. However many enhancements have been added to the classic algorithm to improve its coding efficiency [5] [9].

The MPEG-4 AVC/H.264 coding standard has been significantly extended in its newer editions. The scalable extension (SVC) [10], [11] allows partial transmission and decoding of the encoded video sequence. It provides the functionality of graceful video quality degradation desirable in many applications e.g. video streaming. The video content adaptation to the varying transmission channel parameters can be easily achieved with this functionality. The scalable video bitstream is composed of two or more hierarchically organized layers. The lowest/base layer carries the lowest quality video content. The upper/enhancement layers contain data required to reconstruct video in a better quality. The quality of the reconstructed video sequence depends on the number of the received and decoded layers.

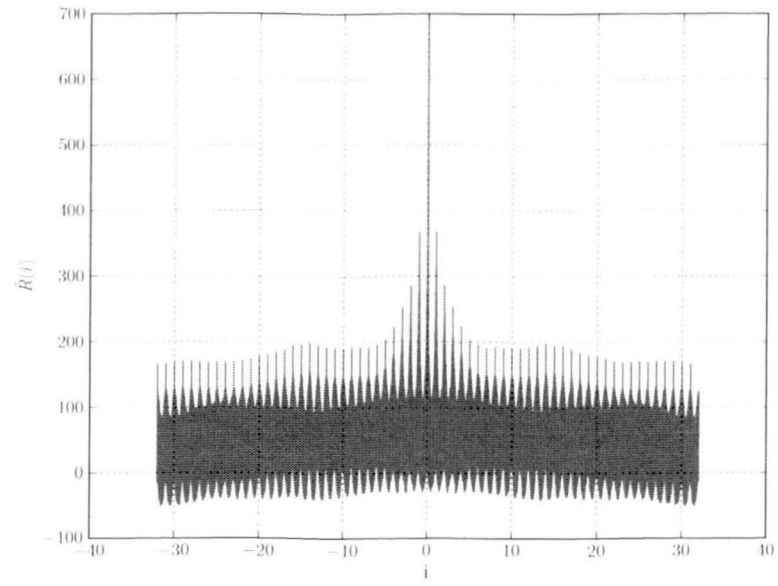
The multiview extension MVC [12] of the MPEG-4 AVC/H.264 standard is related to the 3-D video becoming a hot topic in multimedia technologies last years. It allows encoding of multiple video sequences representing the same scene viewed from different positions. The sequences are captured by multiple video cameras, which must be precisely aligned in space and synchronized in time. Stereovision is an important special case with two images, for the left and the right eye. Currently available displays [13] allow visualisation of the stereo-paired video with 3-D depth impression. However special glasses must be worn in most cases to perceive the scene depth. The rapidly developing autostereoscopic displays technology allows 3-D visualisation without glasses. Multiview autostereoscopic displays [13] will allow a practical implementation of the Free Viewpoint Television concept [14].

The MVC extension efficiency has been achieved by exploitation of both the spatio-temporal redundancy within each view and the inter-view redundancy. The inter-view prediction structures have been proposed in [15], [16]. The MVC extension coding gain achieved with the inter-view prediction highly depends on the acquisition system characteristics (distance between cameras, lens distortions, illumination variations etc.). The bitrate reductions up to 50% have been reported, however it is at a level of $20\% \div 30\%$ in most cases [12].

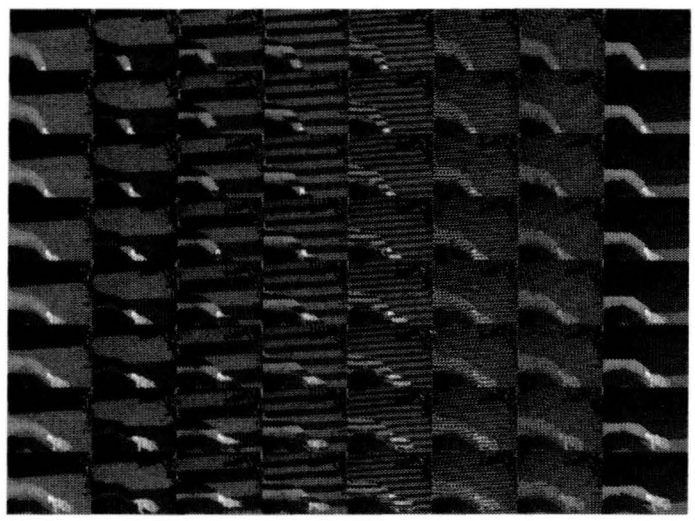
The structure of the MVC bitstream is similar the SVC one. It consists of the 'base view' encoded only with the use of the



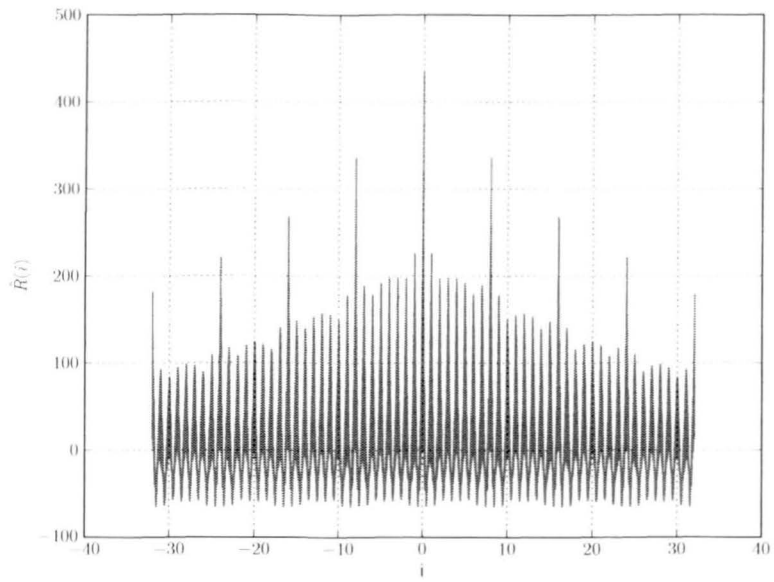
(a) 64 frames of a standard video sequence



(b) Estimation of the autocovariance function of a standard video sequence



(c) 64 frames of a structured light video sequence



(d) Estimation of the autocovariance function of a the structured light sequence

Fig. 1. Comparison of standard and structured light video sequences

within-view spatio-temporal predictions and the 'enhancement views' encoded with the use of within-view and inter-view predictions. The 'base view' may be H.264/AVC compatible encoded to enable its decoding by legacy devices.

II. Stripe pattern video sequences

Digital video sequences, due to regular spatio-temporal sampling pattern, exhibit high spatial and temporal correlation. The values of neighbouring pixels within given frame as well as values of corresponding pixels in neighbouring frames are strongly dependent on each other. It can be confirmed by analysis of the video sequence auto-covariance function. Its estimation will be calculated in the following way:

1) luminance samples of a 3-D video sequence $y_{3D}[m,n,f]$ are converted into a 1-D vector $y_{1D}[k]$ of luminance samples:

$$y_{1D}[k] = y_{3D}[m,n,f] \quad k = m + n \cdot N + f \cdot M \cdot N \quad (1)$$

where M, N are respectively number of columns and rows in each frame, and F is the number of analysed frames, $m = 0, \ldots, M-1$, $n = 0, \ldots, N-1$ are respectively column and row indexes, $f = 0, F-1$ is frame index, $k = 0, \ldots, F \cdot M \cdot N - 1$ is a luminance sample index in the 1-D vector,

2) calculate estimation $\hat{R}[i]$ of the auto-covariance function using FFT/IFFT

$$\hat{R}[i] = \mathcal{IFFT}\left\{\mathcal{FFT}\left(y_{1D}[\cdot] - \overline{y_{1D}[\cdot]}\right)^2\right\} \quad (2)$$

3) the resulting function is scaled by dividing the pixel displacement i by number of pixels in each frame $M \cdot N$ to better visualise inter frame correlation

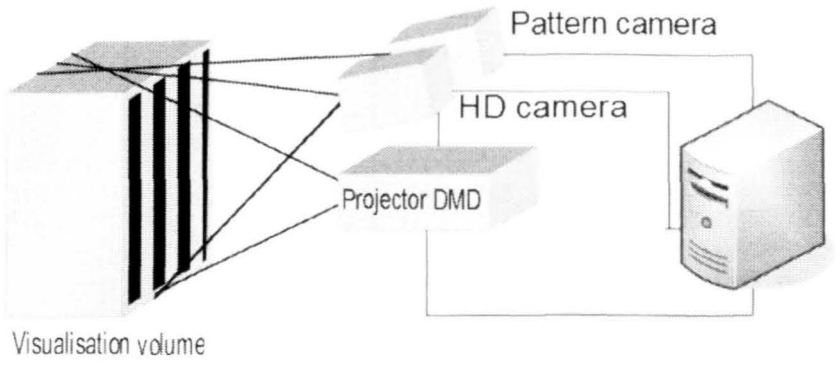
The video test sequence *Football* of CIF resolution (Fig. 1.a) has been used as an example of a standard video test sequence. The estimation of its auto-covariance function is shown in Fig. 1.b. The maxima at every frame are clearly visible.

The structured light video sequences represent scene illuminated by sequentially repeated stripe patterns (Fig. 1.c). The sequences used in the experiment have been captured with the acquisition system presented in Fig. 2. It consists of the projector-camera units, the processing unit and the measurement volume. The projection and detection units works with crossed axis configuration. Projection module perform the Gray Code based pattern projection. It is composed of an light source, a DMD (Digital Mirror Device) and a DMD controller. DMD is used to generate variable patterns and to project the patterns. DMD is a semiconductor based optical switch integrated with micro mirrors. A pattern image consisting of 1 and 0 could be obtained by switching each mirror to on/off state. The the DMD LightBEAM 4500 device is used to display the structured light patterns. This device provides XGA (1024x768) resolution with frame rates of 6Hz-5000Hz for binary patterns. For the second component, image acquisition, a Pattern Camera fast CCD camera model Pike AVT 032F is used. This is an CCD monochrome camera, capable of capturing up to 200 fps, with VGA (640x480) resolution, and FireWire interface. The camera is synchronized by signal from DMD device. The second camera in detection module capture color images with FullHD (1920x1080) resolution. A sequence of Gray Code patterns was selected for this implementation. The 8 binary patterns are generated and uploaded into projection unit memory. With all the files loaded in the DMD memory, the system is ready to start projecting and capturing the patterns. The pattern camera working in trigger mode. Each projected pattern activates image acquisition operation.

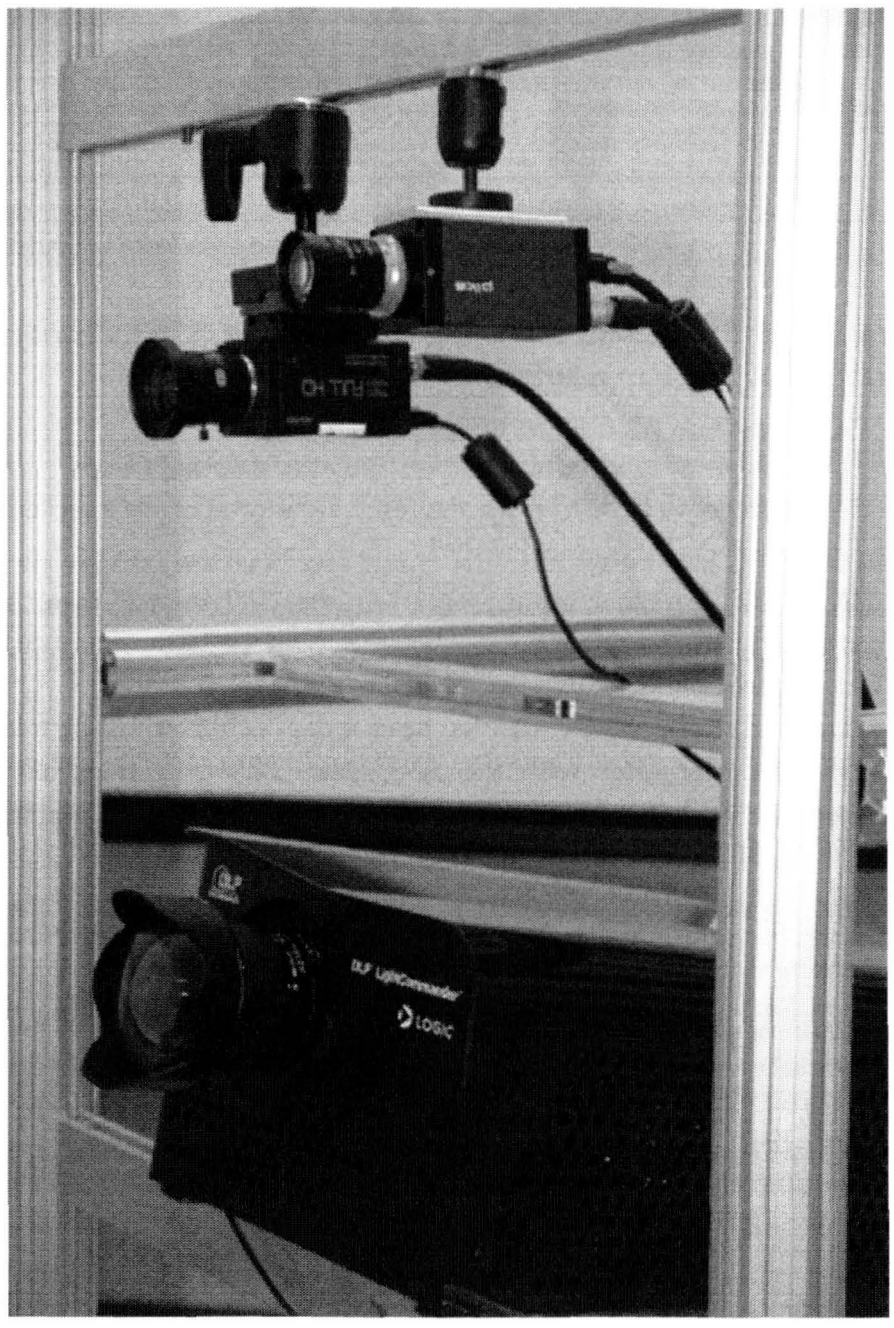
The repeating scene illumination introduces additional correlation to the video sequence. The estimation of the structured light video sequence auto-covariance function (Fig. 1.d) has additional maxima at every eight frame, besides maxima at every frame. The former maxima are much higher than the latter ones.

III. Experimental results

The structured light video sequence has been split into 8 subsequences to exploit its properties discussed in the previous section. Every eight frame, related to one of the eight stripe patterns, has been moved into its corresponding subsequence (Fig. 3.a). The high temporal correlation has been expected with each subsequence, since it contains only images of the scene illuminated only with the one stripe pattern. Additionally inter sequence correlation caused by the same scene composition has been expected. The temporal correlation within each subsequence and correlation between subsequences is similar to the correlations observed in the multiview sequences. The subsequences have been processed by the MVC encoder to check if these correlations can be effectively utilized. The bitstreams representing each subsequence have been then assembled into one MVC bitstream, which has been decoded by the MVC decoder. Reconstructed subsequences have been joined maintaining proper frames order into one sequence. For comparison the entire structured light video sequence has been



(a)



(b)

Fig. 2. General scheme (a) and prototype (b) of the proposed 3D aquisition setup.

processed with the use of standard, single-view, non-scalable encoder and decoder (Fig. 3.b).

The MVC reference software JSVM 8.5 [17] and the MPEG-4 AVC/H.264 reference software JM-18.2 [18] have been used in the encoding experiment. The structured light video sequence shown in Fig. 1.c with 328 frames has been used. It has been split into 8 subsequences (41 frames in each one) for the MVC encoding. GOP composed of 8 frames with

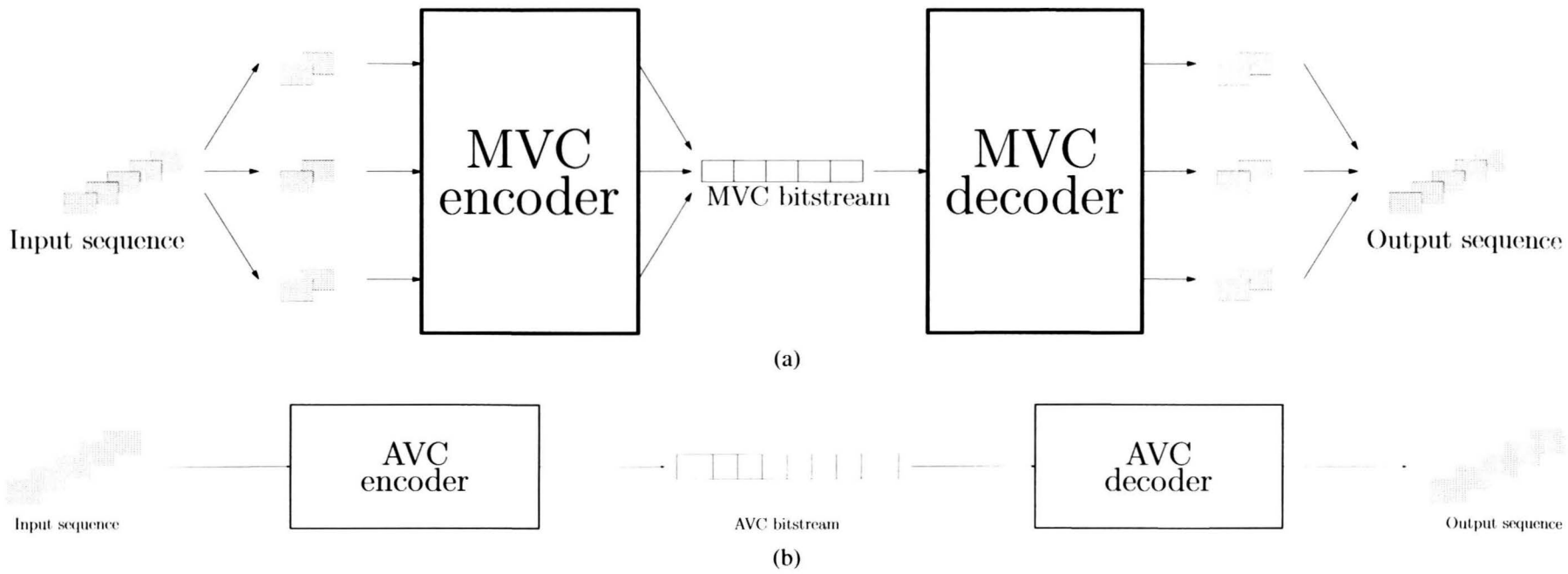


Fig. 3. Structured light video sequence encoding with the use of a) MVC, b) AVC

an anchor frame at the beginning has been used. The following inter-view prediction structures have been used:

- inter-view prediction disabled
- inter-view prediction enabled only for anchor frame
- all frames encoded with the inter-view prediction enabled

In the AVC encoding only first frame has been encoded in the IDR mode, different number of B frames between P frames has been used. The PSNR have been used as a quality measure. The experimental results are presented in Fig. 4.

Generally the better results have been achieved with the MVC encoder than with the AVC one. However there are almost no differences between inter-view prediction structures used in the MVC encoding: the PSNR with inter-view prediction enabled for all frames (MVC ALL) are minimally better the the results for the inter-view prediction enabled only for anchor frames (MVC ANCHOR), which are in turn minimally better than the results with inter-view prediction disabled. The best performance of the AVC encoder has been achieved in configuration with one B frame between I/P frames (AVC IBP). The results with two B frames (AVC IBBP) has been slightly worse, whereas with no B frames (AVC IPPP) significantly worse with respect to the best AVC configuration.

IV. Conclusions

The structural light video sequences encoding options have been examined in this paper. It has been observed that sequential illumination patter introduces additional correlation to these sequences. This correlation is similar to the one observed in multiview sequences. The experimental results have shown that significantly better coding efficiency can be achieved if the structured light video sequence is split into subsequences and MVC encoded with respect to the results obtained with the entire sequence processed with the standard H.264/AVC encoder. However the efficiency gain has been ensured mainly by the temporal correlation within each subsequence. The inter subsequence prediction minimally affects the overall efficiency.

The PSNR based on global image statistics has been used as a quality measure in all experiments. It will be replaced with a measure related to the stripes detection precision in further work.

Another area of the future work will be development of the more efficient representation of the structured light video sequences. The preliminary results show that 8 consecutive stripe pattern images can be combined into one code image also known as an index image. This image has similar properties to the depth image. It is expected that code images representation of the structured light video sequence will better preserve the 3-D scene geometry information than analysed in this paper compressed structured light video

References

[1] P. Fechteler and P. Eisert, "Adaptive color classification for structured light systems," *IET Journal on Computer Vision*, vol. 3, no. 2, pp. 49–59, June 2009.

[2] D. Scharstein and R. Szeliski, "High-accuracy stereo depth maps using structured light," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, 18-20 June 2003, pp. I–195 – I–202.

[3] Y. Q. Shi and H. Sun, *Image and Video Compression for Multimedia Engineering. Fundamentals, Algorithms, and Standards. Second Edition.* CRC Press, 2008.

[4] *Information technology - Coding of audio-visual objects - Part 10: Advanced Video Coding*, ISO/IEC 14496-10:2008.

[5] I. E. Richardson, *The H.264 Advanced Video Compression Standard, 2nd Edition.* Wiley, 2010.

[6] J.-B. Lee and H. Kalva, *The VC-1 and H.264 Video Compression Standards for Broadband Video Services.* Springer, 2008.

[7] D. Marpe, T. Wiegand, and G. J. Sullivan, "The H.264/MPEG4 advanced video coding standard and its applications," *IEEE Communications Magazine*, vol. 44, no. 8, pp. 134 – 143, August 2006.

[8] G. J. Sullivan and T. Wiegand, "Video compression from concepts to the H.264/AVC standard," *Proceedings of the IEEE*, vol. 93, no. 1, pp. 18 – 31, January 2005.

[9] T. Wiegand, G. J. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 560 – 576, July 2003.

[10] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the scalable video coding extension of the h.264/avc standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 9, pp. 1103 – 1120, September 2007.

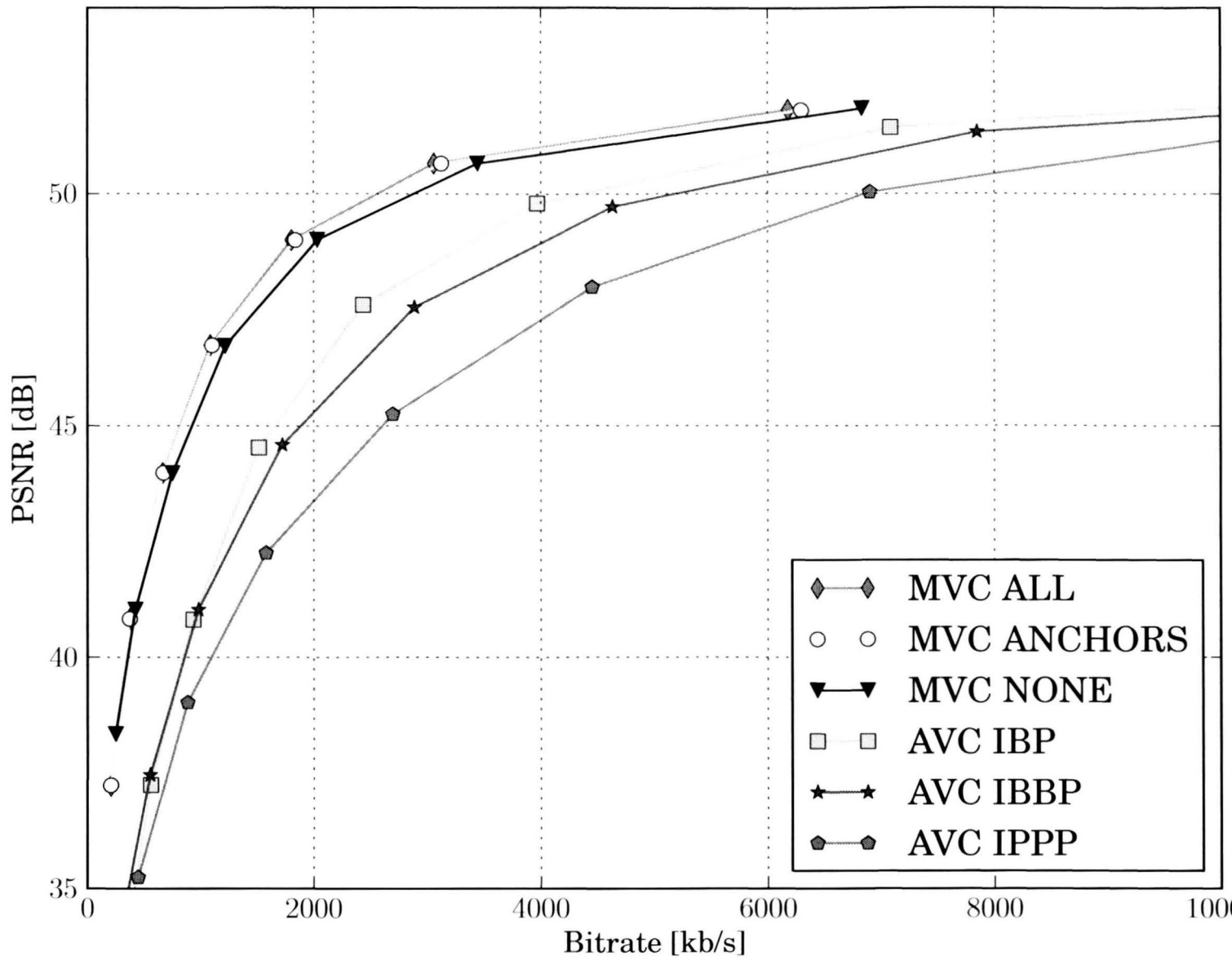


Fig. 4. Encoding experiment results

[11] H. Schwarz and M. Wien, "The scalable video coding extension of the H.264/AVC standard," *IEEE Signal Processing Magazine*, pp. 135 – 141, 2008.

[12] A. Vetro, T. Wiegand, and G. J. Sullivan, "Overview of the stereo and multiview video coding extensions of the H.264/MPEG-4 AVC standard," *Proceedings of the IEEE*, vol. 99, no. 4, pp. 626 – 642, April 2011.

[13] H. Urey, K. V. Chellappan, E. Erden, and P. Surman, "State of the art in stereoscopic and autostereoscopic displays," *Proceeding of the IEEE*, vol. 99, no. 4, pp. 540 – 555, April 2011.

[14] M. Tanimoto, M. P. Tehrani, T. Fujii, and T. Yendo, "Free-viewpoint tv," *IEEE Signal Processing Magazine*, vol. 28, no. 1, pp. 67 – 76, January 2011.

[15] P. Merkle, A. Smolic, K. Mller, and T. Wiegand, "Efficient prediction structures for multiview video coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 11, pp. 1461–1473, November 2007.

[16] P. Merkle, K. Mller, A. Smolic, and T. Wiegand, "Efficient compression of multi-view video exploiting inter-view dependencies based on h.264/mpeg4-avc," in *IEEE International Conference on Multimedia and Expo (ICME'06)*, Toronto, Ontario, Canada, July 2006.

[17] *Text of ISO/IEC 14496-5 :2001/FDAM 15 MVC Reference Software*, ISO/IEC JTC 1/SC 29/WG 11/w10897, December 2009.

[18] K. Shring, *H.264/AVC JM Reference Software*, Fraunhofer Institut fr Nachrichtentechnik, Heinrich Hertz Institute. [Online]. Available: http://iphome.hhi.de/suehring/tml/

Thermal Optimalization of a Stereovision Electronic Travel Aid for the Blind

Bartosz Ostrowski, Mariusz Felczak, Paweł Pełczyński
Institute of Electronics
Technical University of Lodz
Łódź, Poland
bartosz.ostrowski@p.lodz.pl; mariusz.felczak@p.lodz.pl; pawel.pelczynski@p.lodz.pl

ABSTRACT — **The article deals with the thermal optimization of an electronic device board containing the video-and-sound interface for an electronic travel aid (ETA) for assisting the blind in independent mobility. Because the device would be worn as a pair of eyeglasses, the heat should cause as little discomfort to the wearer as possible. In order to optimize the placement of the electronics components on the defined PCB, a special evolutionary, two-step calculation algorithm was developed by the authors.**

KEYWORDS — placement optimization, heat conduction, convection, stereovision, blindness

I. INTRODUCTION

Optimum placement of devices on a PCB plays a very important role for heat dissipation to the ambient. Each electronic component, as a heat source, thermally interacts with the others. This makes electronic component arrangement important for heat dissipation and reliable operation of electronic systems. Apart from that, decreasing the housing temperature even by a single Celsius degree considerably increases the comfort of using any wearable interface.

A lot of methods for optimisation of the positioning of electronic components on PCBs can be found in the literature. The most frequently used assessment criterion is certainly the maximum circuit temperature, but the optimal reliability has been used as well [2][3]. Kos applies heuristic methods for hybrid modules [3]. Electronic devices of this type usually have a well conducting thermal substrate. The author compares the heuristic method with the gradient one. The first one is more suitable for CAD applications than second which has large computational requirements. For the thermal coupling between the components several papers took into account the thermal wake function[4]-[6], i.e. calculating the influence of hot air flow on downstream placed components. In order to optimize an electronic devices' position TFPA (thermal force-directed placement algorithm) was also used [7]. Each chip in the MCM (multi-chip module) pushes other ones to find equilibrium. Forces used for chips are based on heat conduction analogy. For the problem of optimal placement of power chips on the MCMs the modified fuzzy force-directed placement method has also been proposed [9]. Some authors perform permutations of electronic devices on PWBs (printed wiring board) [8]. The aim is to optimize reliability and wire ability of a considered electronic system using Ordered best-first search algorithm, which builds up the linear permutation from front to rear of one module.

In this paper a genetic algorithm will be presented, which is a different approach to the methods published in the literature. A numerical model for electronic component placement optimization on a PCB includes a genetic algorithm and a thermal solver. The genetic algorithm searches arrangements that will be checked by the thermal solver, if they satisfy the requirements of the designer VII. A genetic algorithm is a technique of a solution search based on natural evolution. This kind of solution search is recommended when a function describing an analysed problem is not known, but the solution space can be determined. The genetic algorithms have been described in many textbooks [11]-[14].

II. THE PROTOTYPE OF AN ELECTRONIC TRAVEL-AID SYSTEM FOR THE VISUALLY IMPAIRED

Institute of Electronics at the Technical University of Lodz for several years has been conducting research and development of various Electronic Travel Aid (ETA) systems, i.e. devices for assisting the visually impaired in safe and independent mobility (www.naviton.pl). One of the research projects was the development of special glasses equipped with stereovision cameras, inertial sensors and audio playback capability.

The working principle of the constructed device is to acquire a sequence of stereovision images that are on-line transmitted to a computer. An especially designed application determines the so called depth maps (2.5D images) that are further used for detecting scene objects identified as obstacles. Each detected obstacle, depending on its location in the scene, its size and type, is characterized by a unique audio stream. Those audio signals are being transmitted to the sound module device and played back in stereo headphones. By these means the blind user receives a stream of "auditory pictures" of the environment. The sound generating algorithm combines the data obtained form the segmentation algorithm and the data computed form signals recorded from a linear acceleration sensors and an angular velocity sensor that are located on user's head

The ETA glasses consist of:

- Two VRmC-3 digital cameras from VRMagic that are equipped with separate USB serial buses
- A microprocessor that counts the time interval elapsed from the interruption caused by image acquisition module till the sampling of a signal from linear acceleration sensors,
- A sound card with a USB bus, headphone and microphone ports, which are used for communication between the user and the system,
- A four-port USB 2.0 hub that allows to connect to a computer using a single cable. (Figure 1)
- Stereo headphones integrated with the casing (Figure 2.)

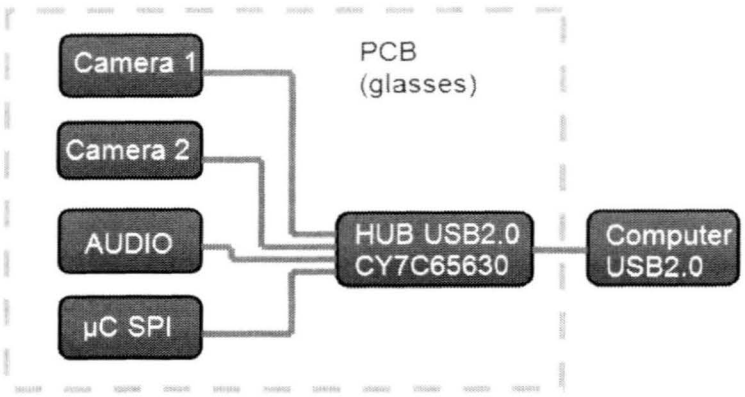


Figure 1. Block diagram.

With the information collected from the end users, we were able to determine the size and shape, which should characterize the device so that it is generally accepted by the blind. It was determined that the device should have a similar shape to large sunglasses, low weight and should not cause discomfort to the wearer.

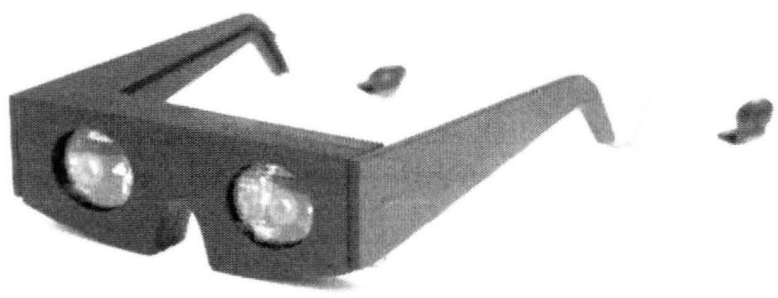


Figure 2. Environmental sound imaging glasses prototype

III. THERMAL PROBLEMS IDENTYFICATION

It has been found that heat dissipated from the vision-acoustic interface can cause discomfort for a person using the system. The device draws its energy from a USB computer port. Despite the small current which is taken, during the system operation, about 3.1 W of power is converted into heat.

For the placement optimization only the devices which dissipate the highest heat quantities were used.

TABLE I. HEAT SOURCES

Lp.	Name	P [mW]	size [mm]
1	Left camera	625	6 x 14
2	Right camera	625	6 x 14
3	USB Concentrator	1300	8 x 8
4	Microprocessor	150	8 x 8
5	Sound card module	425	7 x 7

In order to decrease the time consumption by the optimizing algorithm, the elements used in the thermal model have size equal to 1x1mm. Apart from that, the PCB area used for optimization is limited to a discrete grid Fig. 3ab.

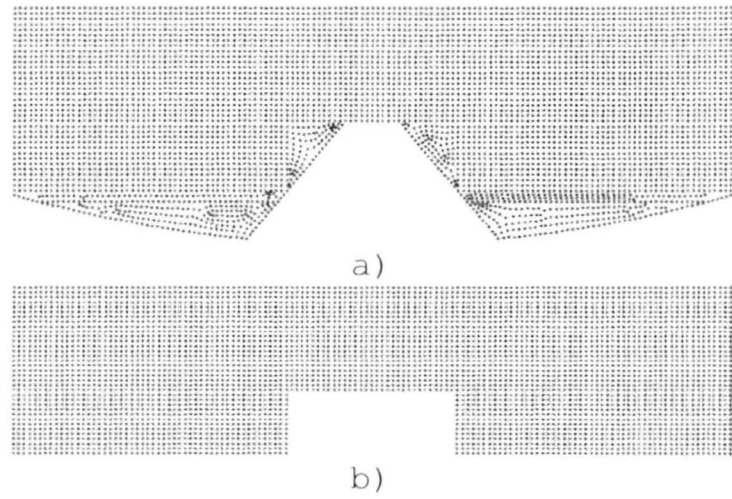


Figure 3. Net raster of PCB a) actual, used while thermal simulations b) limited, used for element placement

In the model thermal conduction throught the substrate is included. The optimization is done for the FR-4 laminate which is typical material used in electronic applications (thickness 1.5mm, thermal conductivity λ=0.25 W/m·K). It is assumed that the ambient temperature is equal to 20°C. In the optimization process only elements 3,4 and 5 could change the position. Elements 1 and 2 are the digital cameras, the position of which on the PCB can not he changed. This results from a fixed camera base (the distance between the optical axes of the cameras) set to 80mm for the project).

IV. THERMAL OPTIMISATION OF ELECTRONIC COMPONENT PLACEMENT

As it is mentioned in the Introduction for the optimization process a genetic algorithm has been used. For that goal special program written in Delphi 7 has been created. For numerical simulation of temperature distribution ANSYS 11 has been used.

A genetic algorithm consisting of two levels was proposed. First, a preliminary optimization which assumed that there are twenty-six positions on the PCB where electronic devices can be placed. The aim of this part of optimization was to place the devices in near optimal positions, which serve as the starting positions for the second level of optimization (permutational coding has been used [2]-[4]). In the second level, the electronic components can be placed on nodes of a 1mm grid (Fig. 3).

In the preliminary part of the algorithm permutational coding is used. Each individual consists of one chromosome that describes the positions of the electronic components. The

algorithm using an objective function performs the adaptation assessment. In order to get the value of the temperatures the thermal solver is used. The thermal model is created in the ANSYS® program in batch mode. The information concerning the electronic component positions is send in a text file to ANSYS® which solves the Kirchhoff–Fourier equations (thermal conduction in the substrate and the devices) and Navier–Stockes equations (forced convective cooling) [15] using the finite element method [16]. Temperature values are returned in a text file to the main program written in DELPHI 7. Using ranking method [17], individual adaptation assessment is performed. Two the best individuals are crossed. If there is no improvement individuals are mutated. The preliminary optimization is finished after N = 200 generations or if there is no solution improvement during M = 100 last generations, then the second level of optimization is started.

The algorithm searches the device positions near those found in the first part. At this level of the optimization, change of the devices position is coded (not the position itself) as a binary sequence (27 bits). Each devices position is represented by 3 bits. The most significant bit (of each 3 bits) decides if any change of the device position is done (1) or not (0), while two the last significant bits decide about direction change n + 1, m(00), n – 1, m(01), n, m + 1(11) and n, m – 1(10) (Fig. 2.). The optimization is conducted using the previously described algorithm and the objective function $f(T_{max}, T_{avg})$. At this level two the best individuals are chosen using ranking method and then crossed. If there is no improvement individuals are mutated (values of two randomly chosen genes are negated).

V. Results of iteration

The optimization gave the placement with average temperature equal to 21.67°C and maximum temperature equal to 24.04°C. The result was found after 186 iterations.

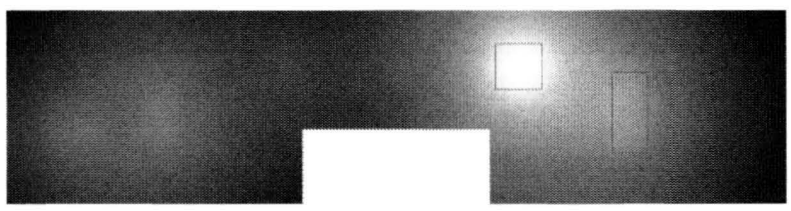


Figure 4. Temperature distribution on PCB and heat sources after optimal component placement

TABLE II. Results of optimization iterations

Iteration number	Maximum temperature [°C]	Average temperature [°C]
start	27.03	22.56
24	25.35	21.93
131	24.28	21.85
186	24.04	21.67

The proposed algorithm enabled to obtain quite easy and fast positions of components for which average temperature is equal to 21.67°C and the maximum temperature equals to 24.04°C. These values are good enough for the user to feel comfortably during the use of the system, which is designed for aiding the visually impaired in independent travel people.

It is also possible to model and thermally optimize various electronic systems using the software prepared during the research.

In the future, the authors plan to compare the obtained results with the Forced Directed Method [7].

VI. Acknowledgment

This work has been supported in part by the Ministry of Science and Higher Education of Poland research grant no. NR02-0083-10 in years 2010-2013.

VII. Bibliography

[1] M. Felczak., B. Więcek. "Application of genetic algorithms for electronic devices placement in structures with heat conduction through the substrate" MICROELECTRONICS RELIABILITY, 2011, 51, nr 2, s. 453-459; ISSN: 0026-2714 13,50/0,00/27,00p

[2] M. D. Osterman and M. Pecht: "Component placement for reliability on conductively cooled printed wiring boards" Journal of Electronic Packaging, 1989, vol.111, p.149-156.

[3] A Kos. and G. De Mey: "Thermal placement in hybrid circuits - a heuristic approach" Active and Passive Electronic Components, 1994, vol.17, p.67-77.

[4] S. S. Kang: "The thermal wake function for rectangular electronic modules" Journal of Electronic Packaging, 1994, vol.116, p.55-59.

[5] R. A. Wirtz and P. Dykshoorn: "Heat transfer from arrays of flat packs in a channel flow" Proceedings 4th IEPS conf., Baltimore, p.318-326.

[6] R. A. Wirtz: "Forced air cooling of low profile air cooling" in "Air cooling technology for electronic equipment" (ed.: S. J. Kim and S. W. Lee), CRC Press, New York, 1996, p.92-102.

[7] J. Lee, "Thermal Placement Algorithm Based on Heat Conduction Analogy", IEEE TRANSACTIONS ON COMPONENTS AND PACKAGING TECHNOLOGIES, VOL. 26, NO. 2, JUNE 2003

[8] J. Lee, J. H. Chou, S. L. Fu, "Reliability and wirability optimizations for module placement on a convectively cooled printed wiring board", INTEGRATION, the VLSI Journal 18 (1995) 173 186

[9] Y.J. Huang, M. H. Guo, S.L. Fu, "Reliability and routability consideration for MCM placement", Microelectronics Reliability 42 (2002) 83–91

[10] Kos and G. De Mey: "Neural computation for optimum power hybrid design" International Journal of Electronics, 1994, vol.76, p.681-692.

[11] Michalewicz Z. "Genetic Algorithms + Data Structures = Evolution Programs", 3rd ed.. New York: Springer, 1996

[12] Gen M., Cheng R., "Genetic algorithms and engineering optimization" New York: J.Wiley & Sons, 2000.

[13] Herrera F., Verdegay J.L. "Genetic algorithms and soft computing", Heidelberg : Physica-Verlag, 1996.

[14] Goldberg D. E., "Genetic algorithms in search, optimization, and machine learning" WNT Warsaw, 2003

[15] J. H. Lienhard IV J. H. Lienhard, "A Heat Transfer Textbook", Cambridge, MA: Phlogiston Press, c2002

[16] Gallagher, Richard H. "Finite element analysis : fundamentals", Englewood Cliffs : Prentice-Hall, 1995

[17] F. Herrera, "Genetic algorithms and soft computing", Physica-Verlag, pp. 51-68, April 1996

[18] P. Svasta, H. Carstea, M. Rangu, A. Avram; "Implementation of a Genetic Algorithm for Components Optimal Placement in Electronic Packaging", 26th International Spring Seminar on Electronics Technology May 8 - 11,2003, Stara Lesna, Slovak Republic

[19] Felczak M ., Więcek B. "Application of genetic algorithms for electronic devices placement in structures with heat conduction through the substrate" Microelectronics Reliability 51 (2011), nr 2, p. 453-459.

SESSION 8:
DSP, HARDWARE AND APPLICATIONS

Novel Law Discovery Perceptrons with Improved Efficiency of the Network Learning

Jarosław Majewski
University of Technology and Life Sciences
Faculty of Telecommunication & Electrical Engineering
ul. Kaliskiego 7, 85-796 Bydgoszcz, Poland
e-mail: jaromaj@utp.edu.pl

Ryszard Wojtyna
University of Technology and Life Sciences
Faculty of Telecommunication & Electrical Engineering
ul. Kaliskiego 7, 85-796 Bydgoszcz

and the Academy of Information Technology
ul. Fordońska 246, 85-959 Bydgoszcz, Poland
e-mail: woj@utp.edu.pl

ABSTRACT — **The issue of effective learning specific neural networks capable of creating symbolic description of rules governing a set of empirical data is considered. In the field of our interests are atypical perceptrons suitable for implementing partial-rational or polynomial functions that describe the data set. A novel factor of the presented approach is an attempt to improve effectiveness of the perceptron learning by making proper transformations of the partial-rational or polynomial functions. These transformations enable to eliminate from the learning process operations on complex numbers as well as time consuming operations connected with using activation functions of the ln(.) and exp(.) type. Such an approach has proved to be a successful way to improve efficiency of the network training in the sense of increasing the learning speed and robustness. The paper presents the proposed transformations used to modify one-dimensional partial-rational as well as one-dimensional polynomial expressions. Perceptron schemes resulting from these expressions are also shown. Moreover, we discuss the applied method suitable for learning the networks and demonstrate the achieved training effects.**

KEYWORDS — neural networks, perceptron training, rules governing numerical data, symbolic description methods

I. INTRODUCTION

In some situations, information about an object of our interests is available only in the form of empirical data. In such cases, there is sometimes a need to create a symbolic description of rules governing the given data set. Creating description of this type is frequently called law discovery approach. So far, the law discovery is based mainly on using partial-rational or polynomial expressions as tools describing relations between elements of a given data set.

For a long time, expressions of this type are applied in different fields of our life. As an example, one can mention mathematical operations of interpolation. Another example relates to measuring physical quantities in the presence of many environmental factors influencing the measurements. In the latter case, the question of a proper calibration of the used measurement instruments appears. This is connected with a complex character of relationships describing the considered empirical data. One of possible solutions of this problem is applying atypical neural networks to determine parameters of complex multidimensional partial-rational or polynomial forms used to describe relations between the measured quantities and the unwanted affecting factors. In the paper, we present a method based on applying the above mentioned technique. Novel neural networks and ideas of more effective learning the networks are presented.

Using neural networks to find rules governing the considered empirical data creates new possibilities to solve this problem and one observes a growing interest in applying such a technique. The role of the neural network is to determine coefficients of the partial-rational or polynomial expressions used as mathematical models of rules governing the data set. One of the main advantages of using neural networks to solve the law discovery problem is a possibility to determine the partial-rational-function or polynomial coefficients while learning the network.

However, to realize the law discovery task, specific neural networks, in our case perceptrons, must be developed to implement the used description of the data rules. The type of the resulting perceptrons depends on the form, in which the function describing the given empirical data set is expressed [1], [3-6], [10-13], [15]. Moreover, proper methods have to be applied to train the network. In general, the network learning is not an easy task. A lot of problems appear in networks with activation functions of an ln(.) type. This results, among others, from difficulties

connected with calculations of the ln(.) functions when dealing with negative arguments. Furthermore, in many cases there is a need to perform operations on complex numbers during the learning process. The problem with the complex numbers is critical when using partial-rational expressions with complex nodes (zeros of denominators) to describe rules governing the data set.

In this paper we propose, as already mentioned, a new way to solve the problem under consideration. Our method is based on proper function transformations and results in proper forms of the used partial-rational or polynomial expressions describing a given data set. The proposed transformations lead to eliminating the ln(.) type activation functions as well as operations on complex numbers from the perceptron training procedure. As a consequence, it is possible to avoid the above discussed difficulties with the perceptron learning and make the training process more effective.

Analytical description of the proposed transformation technique applied to both the partial-rational and polynomial expressions is given in section 2. Section 3 presents the resulting special type perceptrons. The applied training algorithm used to determine coefficients of the given relationships describing the data set is discussed in section 4. Examples of the achieved learning results concerning one-dimensional functions are shown in section 5 whereas section 6 includes concluding remarks.

II. TRANSFORMATIONS OF RATIONAL-FUNCTION AND POLYNOMIAL FORMS LEADING TO THE PROPOSED PERCEPTRONS

Our considerations concerning the proper partial-rational relationship form are aimed at obtaining perceptrons whose learning efficiency could be better than that achievable in the previously published neural networks [3-6], [11-13], [15], applied to solve the law discovery task. Start with the expression:

$$y = a_0 + \frac{w_0}{x + b_0} + \sum_{i=1}^{h} \frac{w_i (x + a_i)}{x^2 + u_i x + b_i} = \frac{c_0 + \sum_{i=1}^{q} c_i x^i}{d_0 + \sum_{i=1}^{p} d_i x^i} \tag{1}$$

where the following condition holds:

$$q \le p \le 2h + 1 \tag{2}$$

In (1), x and y are independent and dependent variables, respectively, and all parameters take real values. The relationship (1), as well as the resulting perceptron, was proposed in [11]. Its disadvantage is a necessity to use activation functions of exp(.) and ln(.) type, which creates the problem of arithmetic operations with negative arguments of the ln(.) activation functions, appearing during the network learning.

To eliminate this drawback, we propose to convert (1) to the form given by:

$$y = a_0 + \frac{w_0}{x + b_0} + \sum_{i=1}^{h} \frac{w_i}{x + 2b_{R_i} + \dfrac{b_{R_i}^2 + b_{I_i}^2}{x}} + \sum_{i=1}^{h} \frac{w_i a_i}{(x + b_{R_i})^2 + b_{I_i}^2}, \tag{3}$$

where the below given condition has to be fulfilled

$$h \ge \frac{p-1}{2} \tag{4}$$

One can show that (3) is another form of (1) if the new parameters introduced to (3) are connected with the parameters of (1) by:

$$b_i = b_{R_i}^2 + b_{I_i}^2, \quad u_i = 2b_{R_i} \tag{5}$$

The performed conversion is based on replacing the real coefficients b_i in (1) by complex numbers with the real part denoted by b_{Ri} and the imaginary by b_{Ii}. The expressions (1) and (3) concern the partial-rational functions and can be implemented by a proper perceptron presented in the next section. The superiority of the form (3) over the form (1) is that we avoid the ln(.) operations on negative values of their arguments.

As shown in [15], using reciprocal operators we can express polynomial expressions by means of the partial-rational relationships. In our case, this means that (3) can be converted to a polynomial form and can be implemented by another perceptron including reciprocal activation functions, which is presented in the next section. Assuming that the a_0 in (3) is equal to zero, i.e. that order of the denominator of (1) is higher than order of the numerator of (1), the polynomial form resulting from inverting (3) is given by:

$$z = \frac{1}{\dfrac{w_0}{x + b_0} + \sum_{i=1}^{h} \dfrac{w_i}{x + 2b_{R_i} + \dfrac{b_{R_i}^2 + b_{I_i}^2}{x}} + \sum_{i=1}^{h} \dfrac{w_i a_i}{(x + b_{R_i})^2 + b_{I_i}^2}} = e_0 + \sum_{i=1}^{p} e_i x^i \tag{6}$$

where the following inequality holds:

$$p \le 2h + 1 \tag{7}$$

Learning perceptrons implementing the expressions (3) and (6) we observed that it is reasonable to increase the number the parameters of these functions that are the subject of changes during the learning process. This is because in this way a chance of finding the best solution of the training increases. Even though for this reason the network learning may get longer, we have decided to make step towards enlarging the number of the variable parameters in the training process. The proposed forms, with the increased parameter number, concerning the partial-rational and polynomial expressions are given be (8) and (9), respectively.

$$y = a_0 + \frac{c_0 w_0}{c_0 x + c_0 b_0} + \sum_{i=1}^{h} \frac{c_{1_i} w_i}{c_{1_i} x + 2c_{1_i} b_{R_i} + \frac{c_{1_i} c_{2_i} \left(b_{R_i}^{\ 2} + b_{I_i}^{\ 2} \right)}{c_{2_i} x}} + \sum_{i=1}^{h} \frac{c_{3_i}^{\ 2} w_i a_i}{\left(c_{3_i} x + c_{3_i} b_{R_i} \right)^2 + c_{3_i}^{\ 2} b_{I_i}^{\ 2}}, \tag{8}$$

$$z = \frac{1}{\frac{c_0 w_0}{c_0 x + c_0 b_0} + \sum_{i=1}^{h} \frac{c_{1_i} w_i}{c_{1_i} x + 2c_{1_i} b_{R_i} + \frac{c_{1_i} c_{2_i} \left(b_{R_i}^{\ 2} + b_{I_i}^{\ 2} \right)}{c_{2_i} x}} + \Lambda} \Lambda \frac{}{+ \sum_{i=1}^{h} \frac{c_{3_i}^{\ 2} w_i a_i}{\left(c_{3_i} x + c_{3_i} b_{R_i} \right)^2 + c_{3_i}^{\ 2} b_{I_i}^{\ 2}}} \tag{9}$$

III. THE PERCEPTRON SCHEMES

The proposed new perceptrons implementing the expressions (8) and (9) are shown in Fig. 1 and Fig. 2, respectively. For clarity reasons, in both figures a more

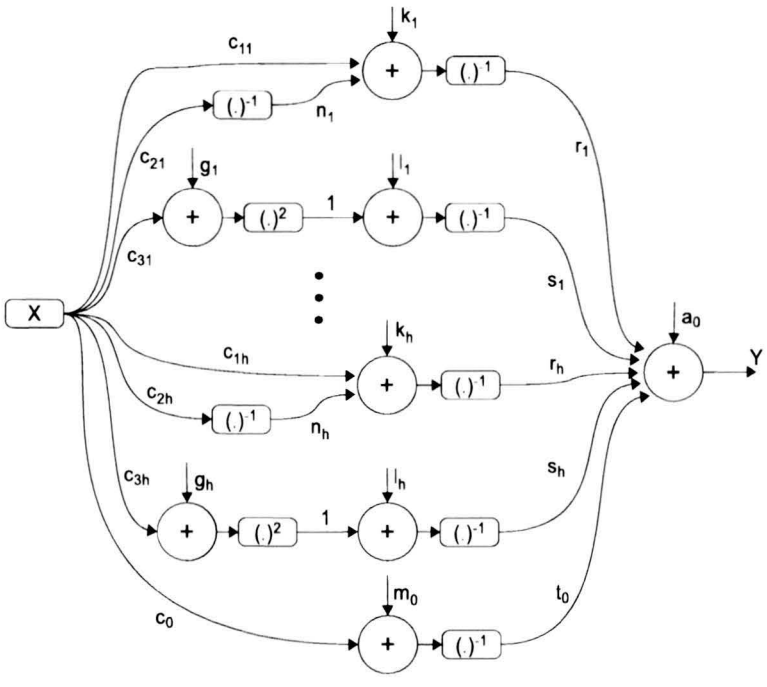


Fig. 1. Simple perceptron-type neural network implementing the partial-rational expression of (8)

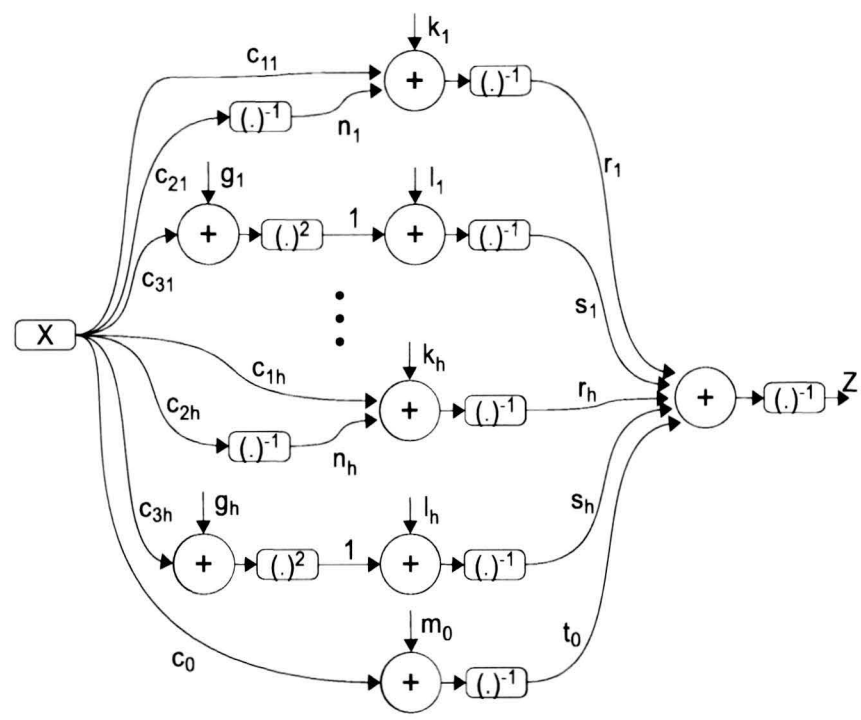


Fig. 2. Simple perceptron implementing the polynomial form given by (9)

compact notations have been introduced. These notations are related to that of (8) and (9) as follows:

$$t_0 = c_0 w_0, \quad m_0 = c_0 b_0,$$
$$s_i = c_{3i}^{\ 2} w_i a_i, \quad l_i = c_{3i}^{\ 2} b_{Ii}^{\ 2},$$
$$g_i = c_{3i} b_{Ri}, \quad r_i = c_{1i} w_i,$$
$$k_i = 2c_{1i} b_{Ri}, \quad n_i = c_{1i} c_{2i} (b_{Ri}^{\ 2} + b_{Ii}^{\ 2}).$$

The presented perceptrons differ only in the applied activation function at their outputs. In case of the scheme of Fig. 1, at the network output node only a summation operation is performed, while in that of Fig. 2, an additional reciprocal operator is placed.

IV. LEARNING ALGORITHM

The law discovery realized by means of the proposed perceptrons shown in Figs. 1 and 2 needs the coefficients of the underlying partial-rational or polynomial functions to be determined in the way of training the networks. The algorithm applied to train our perceptrons, denoted by BP-CM-BFGS, is based on the idea presented in [14]. This is a Back-Propagation (BP) version of the curvilinear line-search memory-less BFGS algorithm proposed by four authors, i.e. Broyden, Fletcher, Goldfarb and Shanno. The abbreviation BFGS is from first letters of names of the four authors. The BP-CM-BFGS method combines the main advantage of second-order algorithms, i.e. fast convergence, with high scalability characteristic of BP algorithms. Due to these highly desirable features, the

method is well suited for learning different types of neural networks. As a result, there is no need to reconstruct the algorithm when dealing with networks of different sizes or different activation functions. The curvilinear line-search feature improves the algorithm convergence while the memory-less feature reduces the required capacity of the computer memory applied to learn the network.

V. LEARNING EXAMPLE

To verify practical operation of the BP-CM-BFGS algorithm with the use of the transformations proposed in section 2, a one-dimensional example of learning the perceptron of Fig. 1 is demonstrated. A subject of our demonstration is the partial-rational function described by:

$$y = f(x) = \frac{2}{x^2 - 4x + 8} + \frac{2}{x^2 - 10x + 30} = \\ = \frac{4x^2 - 28x + 76}{x^4 - 14x^3 + 78x^2 - 200x + 240} \quad (10)$$

The coefficients of (10) present values that should be achieved in the way of training our perceptron, able to implement the general expression of the form (1).

The considered 4-th order function given by (10) has four complex nodes (denominator zeros). These nodes create two pairs of complex conjugate poles of the form:

$$\delta_1 = -2 + j2,\ \delta_2 = -2 - j2,\ \delta_3 = -5 + j\sqrt{5},\ \delta_4 = -5 - j\sqrt{5} \quad (11)$$

In the training procedure, the expression given by (8), being the modified form of (1), is used. The form (8) results from the proposed function transformations discussed in section 2. In (8), the coefficients a_0, m_0, c_0, t_0 are set to be zero because in the target function (10) the corresponding coefficients are equal to zero as well.

The prepared training vector includes 16 pairs of numbers (x, $y=f(x)$), where the x variable takes values in the range from 0.3 to 10.

The established value of a training error was 0.000003. This value was used as indictor of the end of our learning process. The calculated weights resulting from the performed training are:

g_1 = -1.1543, g_2 = 2.6438,
k_1 = 0.2950, k_2 = -1.1352,
l_1 = 1.3135, l_2 = 1.3956,
n_1 = 0.2147, n_2 = 1.2304,
r_1 = -0.0046, r_2 = 0.0014,
s_1 = 0.6570, s_2 = 0.5580,
c_{11} = 1.2350, c_{12} = 0.8325,
c_{21} = 1.4381, c_{22} = 0.8885,
c_{31} = 0.5740, c_{32} = -0.5281.

The obtained training results require certain comments. In the presented training example, the coefficients b_{I1} and b_{I2} of (8), related to the achieved weights l_1 and c_{31} and also weights l_2 and c_{32}, take, for instance, the values b_{I1} = 1.9967 and b_{I2} = 2.2372. The same coefficients, however, related to the got in the way of training k_1, n_1 and c_{11} and also k_2, n_2 and c_{12} weights, take considerably different values, equal to b_{I1}' = 0.3265 and b_{I2}' = 1.0948, respectively. A similar situation, i.e. a great difference is observed between parameter values of b_{R1} and b_{R2} of (8) calculated on the basis of g_1 and c_{31} and also g_2 and c_{32} as well as values worked out on the basis of k_1 and c_{11} and also k_2 and c_{12} parameters resulting from the training. In the first case, the obtained values are b_{R1} = -2.0110 and b_{R2} = -5.0068 while in the second case are b_{R1}' = 0.1194 and b_{R2}' = -0.6818.

The discussed differences may be caused by two reasons. The first one is low precision of the obtained training results and the second an excess of paths in the trained network. Because in our case the network learning error is very low (high learning precision), the second possibility is much more likely than the first one. This conclusion is additionally confirmed by the fact that the weights r_1, r_2, being the final ones relating to two paths in our network, are much lower than weights of the remaining paths, i.e. s_1, s_2. Eliminating from the obtained final results of the training the paths connected with r_1, r_2 implies that as correctly calculated values we can accept only the ones concerning the coefficients b_{I1}, b_{I2}, b_{R1}, b_{R2}. The other coefficients, whose values were also modified while learning the perceptron, should be interpreted and regarded as excess parameters, useless as final results but needed in the training process to improve its effectiveness.

Table 1 presents both target (left column) and trained values (right column) of poles of the function (10). In Table 2, the target function (first row) and that obtained as a result of the performed training (second row) are shown.

TABLE 1. POLES OF THE FUNCTION (10)

Target poles	**Received poles**
$\delta_1 = -2 + j2$	$b_1 = b_{R_1} + jb_{I_1} = -2.0110 + j1.9967$
$\delta_2 = -2 - j2$	$b_2 = b_{R_1} - jb_{I_1} = -2.0110 - j1.9967$
$\delta_3 = -5 + j\sqrt{5} = -5 + j2.2361$	$b_3 = b_{R_2} + jb_{I_2} = -5.0068 + j2.2372$
$\delta_4 = -5 + j\sqrt{5} = -5 - j2.2361$	$b_4 = b_{R_2} - jb_{I_2} = -5.0068 - j2.2372$

TABLE 2. THE EXAMPLE FUNCTION (10)

Target and trained function
$y = \frac{2}{x^2 - 4x + 8} + \frac{2}{x^2 - 10x + 30}$

$$y_r = \frac{1.9944}{x^2 - 4.0221x + 8.0312} + \frac{2.0011}{x^2 - 10.0135x + 30.0728}$$

VI. CONCLUSIONS

The paper is devoted to the problem of learning perceptron-type neural networks that can implement partial-rational or polynomial functions creating a symbolic model of data, knowledge of which we have only in an empirical form. Novel perceptrons (Fig. 1 and Fig. 2) and proper transformations (section 2) of the rational-function and polynomial forms leading to the proposed perceptrons have been shown. A new step towards simplification and improving effectiveness of learning our perceptrons, which are capable of discovering laws governing the empirical data, have been carried out in the paper. The achieved effectiveness improvement means, among others, extending the space of functions whose parameters can be determined in the process of learning the networks. The obtained simplification of the learning procedure results from the possibility of changing complex values of the function coefficients, while training the network, without performing calculations on complex numbers. Unlike in the case published in [11], the approach described in this paper allows also to avoid ln(.) operations on negative numbers. Our further work in this area is oriented at solving problems connected with infinite-value effects, which sometimes appear during the network training, especially when the order of the used functions is high.

REFERENCES

[1] R. Durbin and D. Rumelhart, Product Units: *A Computationally Powerful and Biologically Plausible Extension to Backpropagation Networks*, Neural Computation, Vol. 1, pp. 133–142, 1989.

[2] B. Falkenhainer and R. Michalski. *Integrating quantitative and qualitative discovery in the abacus system*, in Y. Kodratoff and R. Michalski, *Machine learning: An artificial intelligence approach, Volume* III, pages 153–190, Morgan Kaufmann, San Mateo, CA, 1990.

[3] K. Saito and R. Nakano, *Law discovery using neural networks*, Proc. of the 15th International Joint Conference on Artificial Intelligence (IJCAI97), pp. 1078-1083, 1997.

[4] K. Saito and R. Nakano, *Discovery of relevant weights by minimizing cross-validation error*, 4th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD2000), pp. 372–375, 2000.

[5] R. Nakano and K. Saito, *Discovery of a set of nominally conditioned polynomials*, Int. Conference on Discovery Science, *LNAI 1721*, pp. 287–298, 1999.

[6] R. Nakano and K. Saito, *Discovery of Nominally Conditioned Polynomials using Neural Networks, Vector Quantizers and Decision Trees,* Springer-Verlag, Lecture Notes in Computer Science, Vol. 1967/2000, pp. 325–329, 2000

[7] A. B. Tickle, R. Andrews, M. Golea, J. Diederich, *The truth will come to light: Directions and challenges in extracting the knowledge embedded within trained artificial neural networks*, IEEE Trans. on Neural Networks, Vol. 9, No. 6, November 1998.

[8] L. M. Fu., *Knowledge Discovery by inductive neural networks*, IEEE Trans. On Knowledge and Data Engineering, Vol. 11, No. 6, November/December 1999.

[9] A. Ismail and A. P. Engelbrecht, *Paining product units in feedforward neural networks using Particle Swarm Optimization,* In: Development and Practice of Artificial Intelligence Techniques, V.B. Bajic, D. Sha (eds), Proceedings of the International Conference on Artificial Intelligence, Durban, South Africa, pp. 36-40, 1999.

[10] T. Washio, H. Motoda, Y. Niwa, *Discovering admissible simultaneous equation models from observed data*, LNCS 2167, Springer-Verlag, Berlin Heidelberg, pp. 539–551, 2001.

[11] J. Majewski, R. Wojtyna, "*Taking laws out of trained neural networks*", IEEE Workshop SPA 2010 (Signal Processing – Algorithms, Architectures, Arrangements and Applications), pp. 21–24, 2010.

[12] J. Majewski, R. Wojtyna, *Extracting symbolic function expressions by means of neural networks*, Springer-Verlag, series: Advances in intelligent and soft computing, pp. 323–330, 2010.

[13] J. Majewski, R. Wojtyna, *Special neural networks for finding symbolic relationships between numerical data*, Elektronika, Nr 5/2011.

[14] M. S. Apostolopoulou, D.G. Sotiropoulos, I.E. Livieris and P. Pintelas, *A memoryless BFGS neural network training algorithm,* 2009 7th IEEE International Conference on Industrial Informatics, pp. 216–221, INDIN 2009.

[15] J. Majewski, R. Wojtyna, "*Implementing polynomial expressions by means of reciprocal-function-based neural networks*", IEEE Workshop SPA 2010 (Signal Processing – Algorithms, Architectures, Arrangements and Applications), pp. 22–26, 2011.

NTAV/SPA 2012
27-29TH SEPTEMBER, 2012, ŁÓDŹ, POLAND
NEW TRENDS IN AUDIO AND VIDEO / SIGNAL PROCESSING ALGORITHMS, ARCHITECTURES, ARRANGEMENTS AND APPLICATIONS

Fuzzy Clustering Algorithms for Imprecise Database Queries

Automatic Generation of Fuzzy Sets Based on the Natural Distribution of Data

Anna Kowalczyk-Niewiadomy
Institute of Mechatronics and Information Systems
Technical University of Lodz
Lodz, Poland
anna.kowalczykniewiadomy@gmail.com

Adam Pelikant
Institute of Mechatronics and Information Systems
Technical University of Lodz
Lodz, Poland
adam.pelikant@p.lodz.p

ABSTRACT — **Nowadays database management systems are one of the most critical resources in every company. Despite advanced possibilities of SQL, relational database management systems do not support flexible query conditions. The problem of extending database systems with natural language expressions is a matter of many research centers. The basic idea of presented research is to extend an existing query language and make database systems able to satisfy user needs more closely. This paper mostly deals with gaining imprecise information from relational database systems. Presented concept is based on fuzzy sets and automatic clustering techniques that allow built membership function and fuzzy queries processing. Implementation of fuzzy logic on database systems extends traditional SQL language with new mechanisms and new features, so the existing relational database systems will be more flexible, queries more intelligent and similar to ordinary communication methods.**

KEYWORDS — fuzzy logic; fuzzy set; fuzzy clustering; FSQL; fuzzy queries; imprecise queries; FCM; FCMdd; mountain clustering

I. INTRODUCTION

The second half of the twentieth century ushered in rapid development of technology, especially in information technology. The growing demand for storing and processing huge data sets has resulted in evolution of database management systems, that currently represent a critical resource of most companies. Those systems are designed to ensure the cohesion and safety of stored data and their principal objective is to search large data sets efficiently. After years of research in the seventies and eighties, IBM developed Standard Query Language called briefly SQL. Despite advanced possibilities of SQL, it is restricted to the precise communication only. In most business applications, querying precise values or using standard sharp relationships and traditional methods of data aggregation is absolutely sufficient. However, in some cases a standard SQL language, which is based on three-value logic, is not flexible enough. For example, if one is looking for cheap accommodation, or wish to buy a house that costs around €100 000, it is impossible to get results that will satisfy him, by means of traditional precise query language. Both of presented queries use natural language features that are used in everyday life. Traditional SQL is not feasible to build a query that supports such imprecise expressions. Imprecision in such context should not be seen as a drawback, but on the contrary, it allows expressing true needs, preferences and evaluation.

Query results for apartment for about €100.000 can be interpreted as entirely satisfactory, if the price for the apartment is located between €90.000 and €110.000, acceptable to some extent, when the price is not far away from this range and others totally unacceptable. The acceptable range is flexible, so that we get information about houses which cost €85.000. Those examples present that the traditional database query language needs to be extended with natural language elements. The main point for modeling imprecise queries is fuzzy logic and fuzzy sets theory proposed in 1965 by Lotfi A. Zadeh[1].

II. THE CURRENT STATE OF KNOWLEDGE AND RELEVANCE OF RESEARCH.

Over the years, traditional methods of searching information based on the precise conditions are more often replaced by methods based on fuzzy logic elements. The first fuzzy query language was presented by Takahashi in 1991[2]. Two years later Takahashi published the full theory of two languages: calculus query language and fuzzy algebra query language [3]. In the eighties the problem of fuzzy database were investigated by: Zamenkowa[4], Chang Ke[5], Buckles and Petry[6][7].

In the early nineties, thanks to the rapid information technology development, we could notice first implementations of fuzzy query systems. In France, P.Bosc and O. Pivert designed SQLf – fuzzy language which allows getting imprecise information from database [8].

Almost at the same time, in Poland professor S. Zadrożny and professor J. Kasprzyk from Systems Research Institute at Polish Academy of Science (PAS), presented FQUERY system

for Access database. FQUERY consists of tools that enable user building queries with fuzzy values, relations, linguistic modifiers and quantifiers. Currently, mentioned scientists from PAS work on linguistic summaries of databases problem and publish their achievements together with P.Bosc and O. Pivert from Malaga[9].

After the year 2000, there were presented newer solutions based on today's leading database management systems. For example, the research team led by Dr. Jose Gomes Galino of the University of Malaga, has developed a system FSQL for Oracle 7/8, available on the Internet [10]. In Poland, Technical University of Poznan [11] and Silesian University of Technology [12] designed their own fuzzy systems SQLf and Fuzzy Logic Management System.

Despite many implemented systems for query languages, this branch of science is still being investigated and requires extensions to cope with the growing demands. So far, solutions based on the fuzzy sets theory contain strong constraints on the design stage of fuzzy sets. The developed systems are based on rigidly defined membership functions, and therefore require cooperation with expert's knowledge.

The basic idea of our research is to redesign a fuzzy structured query language system by extending traditional SQL (Oracle 11g) with natural language expressions. Due to the fact that few fuzzy query systems already exist in Polish as well as foreign research centers, it is worth to emphasize that the innovative element of this work is development of universal, multi-dimensional algorithms, which automatically generate fuzzy sets, based on the real data distribution. There is no need to use expert knowledge while original algorithms based on fuzzy clustering methods are implemented. In addition, conducting a comprehensive analysis of standardization issues and the labeling process enabled implementation of an intelligent module responsible for the allocation of labels according to the automatically generated fuzzy sets. There are some arguments to build such solution. The work of branch experts generates additional high costs. In many cases the meaning of label changes as a result of data distribution changes. For example prices of apartments usually grow up, so that meaning of cheap and expensive flat changes as well. Additionally in times of crisis praises can rapidly drop. Any of these states requires the help of experts, which can be avoided if the proposed solution is used. The approach (in more detail discussed in Chapters V-VIII) is a completely new idea in the fuzzy SQL language issues.

III. Fuzzy Clustering Algorithms

Data clustering is a process of assigning a set of objects into groups (called clusters) so that the objects in the same cluster are more similar (in some sense or another) to each other than to those in other clusters. Each data point belongs to a cluster to a degree of membership grade. This paper reviews three of the most representative clustering techniques: Fuzzy C-Means, Fuzzy C-Medoids clustering and Mountain clustering. All techniques were implemented and tested against automatic fuzzy sets generation problem. Fuzzy clustering methods together with the author's algorithm and the trapezoidal membership function allowed generating fuzzy sets based on the actual data distribution.

A. *Fuzzy C-means Clustering (FCM)*

Let X be a set of n patterns described by $X = \{x_1, x_2, \dots, x_n\}$. Let c be an assumed number of clusters. $C = \{c_j | \, 1 \le j \le c\}$ is the set of centers. The notation $u_{ij} (1 \le i \le n, 1 \le j \le c)$ indicates the degree of membership of the i-th sample to the j-th prototype. The membership matrix U is limited to values between 0 and 1. However, the summation of degrees of belongingness of a data point to all clusters is always equal to unity (1).

$$\sum_{j=1}^{c} u_{ij} = 1; \; 1 \le i \le n \tag{1}$$

The Fuzzy C-means method was proposed in 1973 by Dunn[13] and modified in 1981 by Bezdek [14]. The algorithm is based on clusters search in a data set, such that an objective function (2) of distance measure is minimized. The squared distance is weighted by the m-th power of the membership in cluster j.

$$J_m(U, c) = \sum_{i=1}^{n} \sum_{j=1}^{c} u_{ij}^m \left\| x_i - c_j \right\|^2 \tag{2}$$

$$1 \le m \le \infty$$

In most cases, this distance measure is chosen as the Euclidean norm (4), but it can be Manhatan (3), Chebyshev (5), Minkowski (6) presented in the paper [15]. One should remember that the result of clustering depends on kind of selected metric.

As an exemplary distance measures we can mention: Manhattan distance defined as (3):

$$d_1(x_i, x_{ij}) = \sum_{k=1}^{d} \left| x_{i,k} - x_{j,k} \right| \tag{3}$$

Euclidian distance defined as (4):

$$d_2(x_i, x_j) = \left(\sum_{k=1}^{d} \left(x_{i,k} - x_{j,k} \right)^2 \right)^{\frac{1}{2}} \tag{4}$$

Chebyshev distance defined as (5):

$$d_\infty(x_i, x_j) = \lim_{m \to \infty} \left(\sum_{k=1}^{d} \left| x_{i,k} - x_{j,k} \right|^n \right)^{\frac{1}{n}} \tag{5}$$

The Minkowski distance (6) is a metric at Euclidean space which can be considered as a generalization of both the Euclidean distance and the Manhattan distance.

$$d_n(x_i, x_j) = \left(\sum_{k=1}^{d} \left(\left| x_{i,k} - x_{j,k} \right| \right)^n \right)^{\frac{1}{n}} \tag{6}$$

The (7) and (8) are mandatory conditions for equation (2) to reach its minimum.

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}}} \tag{7}$$

$$c_j = \frac{\sum_{i=1}^{n} u_{ij}^{m} x_i}{\sum_{i=1}^{n} u_{ij}^{m}} \tag{8}$$

The algorithm works iteratively through the preceding two conditions until there is no more improvement noticed. FCM calculates cluster centers and the membership matrix U using the steps presented at Figure 1.

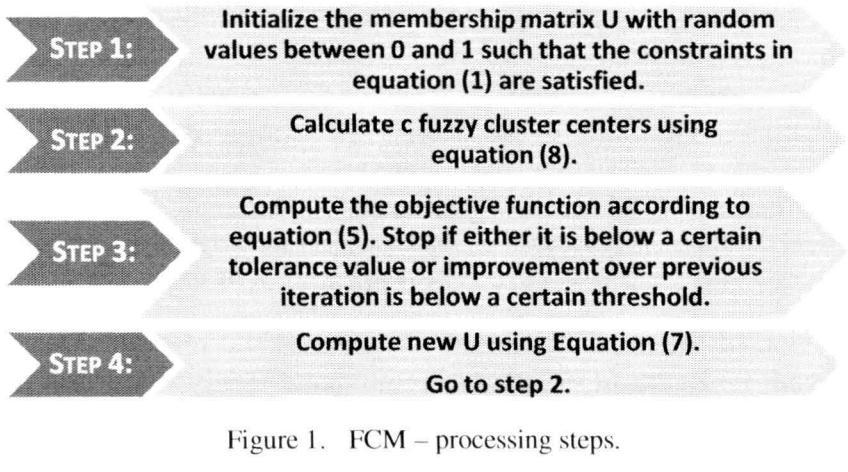


Figure 1. FCM – processing steps.

The main advantage of the FCM algorithm is high performance and low hardware requirements. Unfortunately, this algorithm has three major drawbacks. First, the final distribution of objects between clusters strongly depends on the assumed number of clusters. Second, the performance of FCM depends on the initial membership matrix values. It is advised to run the algorithm for several times, every time starting with new values of membership grades for data points. Third, the algorithm is sensitive to disrupted data (e.g. singular point).

In order to solve this problem, instead of calculating mean we can search the most centrally located cluster point called medoid. In this way, the occurrence of the singular point in the cluster will not cause significant disruptions.

B. Fuzzy C-Medoids Clustering (FCMdd)

Fuzzy C-Medoids Clustering, relies on the basic idea of Fuzzy C-means clustering (FCM) with the difference of calculating cluster centers. The change has a significant influence on the efficiency of the algorithm. Instead of searching means (calculated as a simple arithmetic formula) we need to process several steps over the neighbor points to find medoids. The improvement (minimization) of the criterion function (9) is much more complex and expensive. The notation $r(x_i, v_j)$ indicates dissimilarity between the x_i sample and v_j medoid.

$$J_m(V;X) = \sum_{i=1}^{n} \sum_{j=1}^{c} u_{ij}^{m} r(x_i, v_j) \tag{9}$$

Membership matrix (u) is calculated according to equation (10):

$$u_{ij} = \frac{\left(\frac{1}{r(x_j, v_i)} \right)^{\frac{1}{m-1}}}{\sum_{k=1}^{c} \left(\frac{1}{r(x_j, v_k)} \right)^{\frac{1}{m-1}}} \tag{10}$$

Figure 2 presents basic steps of the FCMdd algorithm.

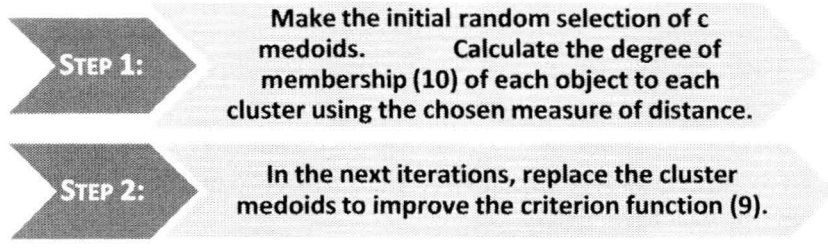


Figure 2. FCMdd – processing steps.

C. Mountain Clustering

The mountain clustering approach is a method to find cluster centers based on a density measure called the mountain function. It is based on three main steps. The first one involves forming a grid on the data space, where the intersections of the grid lines constitute the potential cluster centers. The second step entails construction of a mountain function representing data density measure. The height of the mountain function at a point $v \in V$ is equal to (11):

$$m(v) = \sum_{i=1}^{N} e^{-\left(\frac{\|v - x_i\|^2}{2\sigma^2} \right)} \tag{11}$$

where x_i is the i-th data point and σ is an application specific constant. The third step involves selection of the cluster centers by sequentially reducing the mountain function. This is done by modification of the mountain function to the form represented by equation (12):

$$m_{new}(v) = m(v) - m(c_1) e^{-\left(\frac{\|v - c_1\|^2}{2\beta^2} \right)} \tag{12}$$

IV. Cluster Validation Problem

The problem of data clustering is quite complex, what is mainly caused by wide potential of methods usage. Depending on the situation there is a need to use different types of algorithms, so it is difficult to impose a universal method. One of the main subjects in data clustering is evaluation of the result of clustering algorithms (cluster validation). More precisely, the cluster validation problem is to find an objective criterion to determine how good a partition generated by a clustering algorithm is. Since most clustering algorithms require a pre-assumed number of clusters, a validation criterion to find an optimal number of clusters would be very beneficial.

The first validation was associated with the FCM partition coefficient proposed by Bezdek[14], defined by the equation (13):

$$I_{PC} = \frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^2 \tag{13}$$

To produce a better clustering performance we find optimal cluster numbers for $\max_{2\le c\le n-1} I_{PC}$.

Partition entropy was also proposed by Bezdek for the Fuzzy C-Means algorithm and it is defined by the following equation (14):

$$I_{PE} = -\frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}\log_2 u_{ij} \tag{14}$$

To produce better clustering performance we find optimal cluster numbers for $\min_{2\le c\le n-1} I_{PE}$.

In 1991 Xie and Beni[16] proposed a validation index based on compactness and separation defined as (15):

$$I_{XB} = \frac{\sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^m \left\|x_j - v_i\right\|^2}{n \min\limits_{i\neq j} i,j \left\|x_j - v_i\right\|^2} \tag{15}$$

In 2011 Rubio, Castillo and Melin[17] compared the most commonly used indices such as I_{PC}, I_{PE}, I_{XB} and proposed its own (18) proving its greatest effectiveness.

$$I_{MPE} = \sum_{i=1}^{c}\sum_{j=1}^{n} {u_{ij}}^2 \log_2 u_{ij} \tag{16}$$

$$D_{M_k} = \sum_{\substack{i,j=1 \\ i\neq j}}^{k} \left\|M_i - M_j\right\|^2, k = 1, \dots, c \tag{17}$$

$$I_{RCM} = I_{MPE} + D_M \tag{18}$$

During the research **all presented above validity indexes were implemented** and used in the process of fuzzy sets generation.

V. FUZZY SQL SYSTEM DESIGN

The main idea of our research was to design and implement system, which extends traditional SQL with natural language expressions. The great difference between proposed solution and already existing similar systems is fully automated generation of membership functions and fuzzy sets [18]. In addition to this, the automatic labeling module is also novelty[19][20]. The project consists of three main modules.

First one concerns gathering an input data, and preparing database model for the main processing (Figure 3).

The second one is a set of tools responsible for fuzzy query processing:

- query parser, which decomposes the fuzzy query into understandable by the system and database elements;
- fuzzy sets generator – a complex module using fuzzy clustering methods, novel fuzzy sets algorithms, validity indexes, T/S norms implementations etc.;
- labels assignment process – each fuzzy set is addressed by appropriate label or labels;
- fuzzy conditions, operators and aggregate functions executor.

The third element of the project – the client application is responsible for construction fuzzy queries (tree view and text) and presentation of the results.

A. Data preparation

In order to collect a sufficient data set, there was a need to prepare (design, implement and publish) an exemplary website which provides tutorials for web graphics (3dStudioMax, After Effects, Photoshop). Website traffic statistics are a good source of natural input data. In our project the Google Analytics (GA) was used as a statistical data warehouse. GA is a free, online tool mainly used to analyze websites statistics. GA is a powerful tool with variety of functionalities accessible via web browser and what was very useful via Java API.

One of the most important elements of the research was to prepare software in order to communicate easily with the GA. By use Java programming language, Java GA API (JGA), and JDBC the integration with the Oracle database was provided. Figure 3 presents basic GA integration steps.

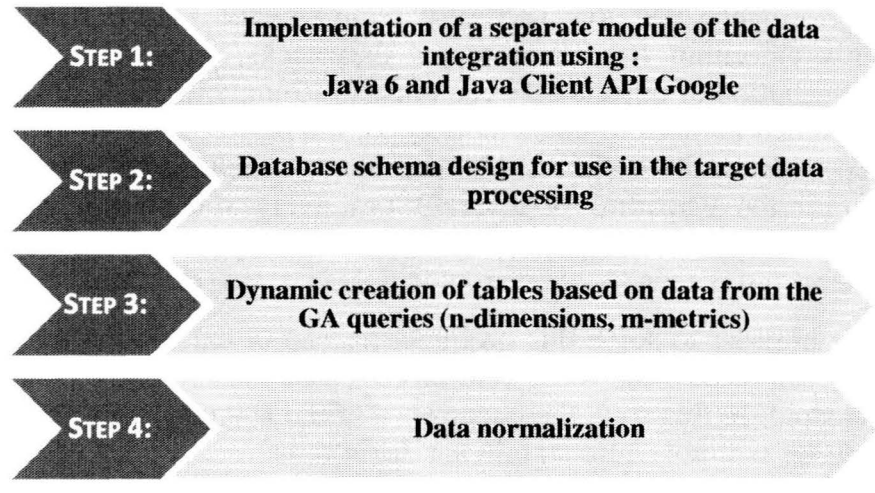


Figure 3. Data preparation diagram.

It is worth to emphasize that data prepared for fuzzy clustering algorithms is normalized to the <0, 100> range in order to eliminate the problem of scale, negative values and to ensure the integrity of the generated fuzzy sets. Standardization refers to both input data as well as a range of labels. In this way, it is possible to assign labels by percentage match in order to avoid context dependency. For example query using "high temperature" expression in context of the weather, boiling water or melting metals is completely different. Data normalization eliminates this problem.

GA enables lots of important statistics about website traffic. It is impossible to discuss all of them, so we focused on most commonly used like metrics: **visits**, **pageviews**, **visit duration**,

avg time on page, **bounce rate**, **%new visits**. Such data can be analyzed in the following dimensions: **the date** (hour, day, month, year), **location** - source of visits (continent, country, city), **the type, version and parameters of a web browser** (IE, FF,) **language** etc.

VI. Labeling Module

Our fuzzy SQL interpreter enables natural language expressions, so labeling module was designed and implemented. Firstly, it was necessary to define labels with the gradation of "strength" of each label (appropriate thresholds was required). Labels of the same type (e.g. short, average, tall) are combined into sets. Each label set is assigned to the attribute e.g. the attribute "time on site" can be short or long ant the attribute "number of visits" can be small, average or big. The process of assigning labels to fuzzy set causes some difficulties connected to following issues:

- Each attribute may have different number of labels;
- Each clustering process can generate different number of clusters.

Presented issues were coped in the implemented algorithm.

VII. Fuzzy Query Processing

In order to process the fuzzy query and generate fuzzy sets automatically, the fuzzy clustering methods, label-attribute and operator logic were implemented. For each distinct attribute to get most relevant results, the algorithms run several times, each time starting with new clustering parameters (different number of clusters, distance measures and start points). After that the validity indexes (discussed in paragraph IV) are calculated and the best result is chosen for selected attribute/attributes. Next, the results of clustering method are processed by novel algorithms based on membership functions in order to generated fuzzy sets with triangle and trapezoid membership functions.

With such prepared data the intelligent module responsible for label allocation according to the automatically generated fuzzy sets and labeling criteria, generates a set of pair-value assignments. Finally generated data collection contains fuzzy set to label matches eg. label_small » fuzzy_set_#1, label_medium » {fuzzy_set_#2 and fuzzy_set_#3} etc.

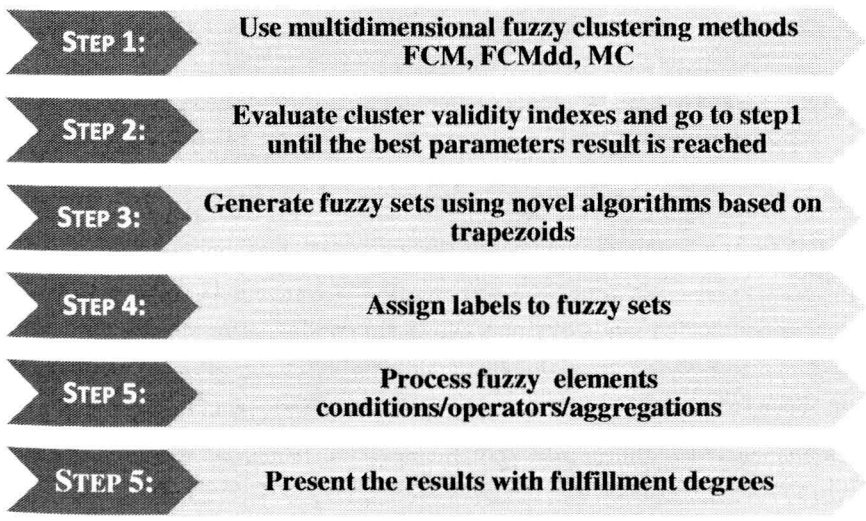


Figure 4. Multidimensional data processing.

After that all the operations for current attribute based on T-Norms and S-Norms are processed according to query logic with respect to supplied order and operators priorities. In case of query with multiple arguments, output of processing of single arguments are joined with respect to order and operators and the final result is built from the composition of each attributes output. The mid results are combined by the use of primary keys or ROWID from primary flat SQL query. Finally the client application presents query results ordered by membership degree (Figure 5).

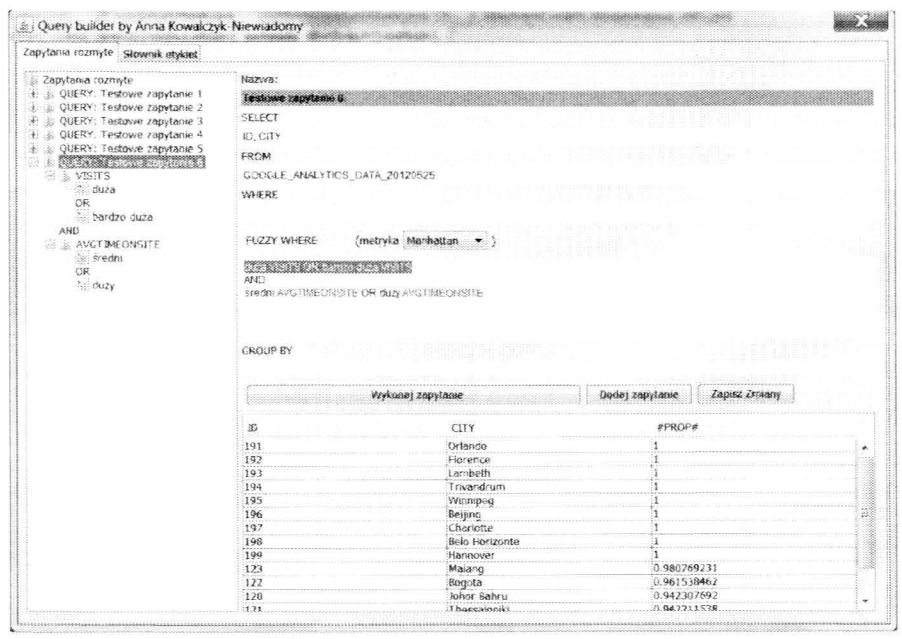


Figure 5. Query builder with exemplay result.

VIII. Examples

SQL queries containing imprecise conditions look quite similar to traditional queries. SQL query is extended by FuzzyWhere , FuzzyHaving clauses. Figure 6 present fuzzy query syntax, containing standard and fuzzy conditions, logical operators (or/and/not) and fuzzy aggregation conditions.

```
SELECT {EXPRESSIONS, AGGREGATE FUNCTIONS AGG }
FROM {TABLENAME/S, VIEW/S}
WHERE {CONDITIONS}
-- FUZZYWHERE <FUZZY CONDITIONS AND OPERATORS IN
NATURAL LANGUAGE >;
FUZZYWHERE
FUZZYLABEL1(ATTRIBUTE1) AND FUZZYLABEL2(ATTRIBUTE1)
OR FUZZYLABEL1(ATTRIBUTE2) AND NOT
FUZZYLABEL3(ATTRIBUTE3)
GROUP BY {GROUP CONDITION}
HAVING <CONDITIONS>
FUZZYHAVING  FUZZYLABEL1(AGG)
```

Figure 6. Imprecise query syntax..

In this section, the autors present the difference between traditional SQL queries and imprecise FSQL. For example if one is searching countries with new visitors interested in web, the query may look like listed on Figure 7.

```
SELECT COUNTRY,
AVG (PERCENTNEWVISITS), AVG (AVGTIMEONPAGE)
FROM WEB_DATA_TST
WHERE PERCENTNEWVISITS > 90 AND AVGTIMEONPAGE > 200
GROUP BY COUNTRY
```

Figure 7. Traditional query.

The result for query is presented on Figure 8.

COUNTRY	AVG(PERCENTNEWVISITS)	AVG(AVGTIMEONPAGE))
USA	97	225
India	96	325
Philippines	93	218
United	92	308
Germany	92	243
Romania	90	224

Figure 8. Traditional query's results.

Similar query in FSQL language may look like listed on ...

```
SELECT COUNTRY,
AVG (PERCENTNEWVISITS), AVG (AVGTIMEONPAGE)
FROM WEB_DATA_TST
FUZZYWHERE
LARGE(PERCENTNEWVISITS)AND LONG(AVGTIMEONPAGE)
GROUP BY COUNTRY
```

Figure 9. Imprecise query.

The result set is not restricted by crisp conditions so it contains more values presented on

COUNTRY	AVG(PERCENTNEWVISI	AVG(AVGTIMEONPAGE)	SATIS
USA	225	97	0.97
India	325	96	0.96
Romania	224	90	0.95
Philippines	207	96	0.95
Germany	217	92	0.94
UK Kingdom	216	95	0.94
Brazil	198	94	0.94
Myanmar	196	90	0.93
Poland	196	94	0.93
Argentina	211	90	0.92
Canada	209	89	0.90
Tanzania	186	90	0.82

Figure 10. Imprecise query's results.

SUMMARY

This paper, presents a novel approach to the problem of imprecise information retrieval from database systems. SQL standard does not provide any mechanism for solving such task. Recently, fuzzy SQL languages have become a very interesting scope of research. Most of current implementations are based on strictly defined membership function and require expert knowledge about threshold degree for specific data type. This article presents an idea of gaining imprecise and incomplete information from database by novel algorithms. Most important points of the carried out research are: natural data set preparation based on real website traffic; fuzzy query processing algorithms located on database server implementation and frontend application implementation.

Main idea of the proposed algorithms is fuzzy clustering mechanisms with automatically generated fuzzy sets. All the processing is done by use of smart clustering combined with validity indexes to reach the best results. The process of generation membership function can be executed on demand, triggered by events or executed by scheduler. That feature gives opportunity to adopt to data distribution changes dynamically. In addition to this, the intelligent labeling mechanism together with own parser assigns labels defined in natural language to generated fuzzy sets. The designed system is able to execute fuzzy conditions and aggregations and can be combined with standard SQL. Currently the integration with SQL is based on Java frontend client application but in future it can be provided as an extension of standard SQL. As a summary it can be said that presented idea of fuzzy sets generator together with query parser and intelligent labeling mechanism enables retrieval of data for query written in meta-natural language (fuzzy query with smart labels).

REFERENCES

[1] Lotfi A. Zadeh, "Fuzzy sets," *Information and Control*, vol. 8, no. 3, pp. 338-353, 1965.

[2] Y. Takahashi, "A fuzzy query language for relational databases," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 21, pp. 1576-1579, 1991.

[3] Y. Takahashi, "Fuzzy database query languages and their relational completeness theorem," *IEEE Transactions on Knowledge and Data Engineering*, vol. 5, pp. 122-125, 1993.

[4] M Zamenkova and A Kendel, "Implementing imprecision in information systems," *Information Sciences*, no. 37(1-3), pp. 107-141, 1985.

[5] S.K. Chang and J.S. Ke, "Translation of fuzzy queries for relational database systems," *IEEE Transactions on Pattern Analysis and Machine Inteligence PAMI-1*, pp. 281-294, 1979.

[6] B. P. Buckles and F. E. Petry, "A fuzzy reprezentation of data fo relational database," *Fuzzy Sets and Systems*, no. 7, pp. 213-226, 1982.

[7] B. P. Buckles, F. E. Petry, and H. S. Sachar, "A domain calculus for fuzzy relational databases," *Fuzzy Sets and Systems*, no. 29, pp. 327-340, 1989.

[8] P. Bosc and O. Pivert, "SQLf: A Relational Database Language for Fuzzy Querying," *IEEE Transactions on Fuzzy Systems*, vol. 3, no. 1, pp. 1-17, luty 1995.

[9] P. Bosc et al., "On advances in soft computing applied to databases and information systems.," *Fuzzy Sets and Systems 196: 1-3*, 2012.

[10] Jose Galindo. A Fuzzy Query Language. [Online]. http://www.lcc.uma.es/~ppgg/FSQL/

[11] Krzysztof Dembczyński, Dominik Przybył, and Piotr Kalinowski. (2006) SQLf_j. [Online]. http://calypso.cs.put.poznan.pl/projects/sqlf_j/pl/index.php?page=intro

[12] Bartosz Dziedzic, Bożena Małysiak, and Dariusz Mrozek, "Interpreter wyrażeń rozmytych stosowanych w składni języka SQL," in *BDAS*, Ustron, 2008.

[13] J.C. Dunn, "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters," *Journal of Cybernetics*, vol. 3, pp. 32-57, 1973.

[14] J.C. Bezdek, "Pattern Recognition with Fuzzy Objective Function Algoritms," *Plenum Press*, 1981.

[15] Hui Xiong, Guoxing Zhan, Junjie Wu, and Zhongzhi Shi, "Distance Measures for Clustering Validation: Generalization and Normalization.," *Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1249-1262 , September 2009.

[16] X. L. Xie and G. Beni, "A validity measure for fuzzy clustering," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, no. 13, 8, pp. 841-847, Aug 1991.

[17] E. Rubio, O. Castillo, and P. Melin, "A new validation index for fuzzy clustering and its comparisons with other methods.," in *Systems, Man, and Cybernetics (SMC), 2011 IEEE International Conference.*, Anchorage, AK, 2011, pp. 301-306.

[18] Adam Pelikant and Anna Kowalczyk, "Implemntation of automatically generated membership functions based on grouping algorithms ," in *The International Conference on "Computer as a Tool"*, Warsaw, 2007.

[19] Adam Pelikant and Anna Kowalczyk-Niewiadomy, "Fuzzy queries in relational databases.," in *System Modelling and Control*, 2009.

[20] Adam Pelikant and Anna Kowalczyk-Niewiadomy, "Algorytm etykietowania analizujący rozmyte zapytania w metajęzyku naturalnym.," in *Bazy Danych Aplikacje i Systemy*, Ustron, 2011.

Route to Chaos Due to Power Source Frequency

Valentin Stefanescu
Electronique et Commande des Systèmes Laboratoire,
EA 3649 (ECS-Lab/ENSEA), Cergy-Pontoise, France and
Faculty of Electronics, Telecommunications
and Information Technology,
Politehnica University of Bucharest, Romania

Madalin Frunzete
Electronique et Commande des Systèmes Laboratoire,
EA 3649 (ECS-Lab/ENSEA), Cergy-Pontoise, France and
Faculty of Electronics, Telecommunications
and Information Technology,
Politehnica University of Bucharest, Romania

Jean-Pierre Barbot
Faculty of Electronics, Telecommunications
and Information Technology,
Politehnica University of Bucharest, Romania and
ALIEN-INRIA

Bogdan Florea
Electronique et Commande des Systèmes Laboratoire,
EA 3649 (ECS-Lab/ENSEA), Cergy-Pontoise, France

ABSTRACT — **In this paper, the behaviour of an electric hybrid dynamical circuit that can exhibit chaotic behaviour under the influence of the power source sampling is studied. A simple oscillating circuit that incorporates a nonlinear element is selected and a switch in the power source that will act as the discrete component is added. The variance of the chaotic behaviour throughout the frequency spectrum was observed; based on this, a method of determining low usable frequencies that will keep the systems stability is proposed. The study is done via numerical analysis and computer simulations.**

INDEX TERMS — multi-cell chopper, alpazur oscillator, route to chaos, sampling

I. INTRODUCTION

Electronic circuits have been studied throughout time from many aspects like communication [?], cryptography [?], control theory [?], music [?], meteorology and so on. Their behaviour is usually easily simulated and understood. There are situations though when the circuit behaviour is chaotic in real life given a certain set of circumstances. When studying a chaos related problem, you can take several approaches. One would be to identify the source causing chaotic behaviour, then either eliminate the chaos by adjusting parameters or avoiding chaotic behaviour by limiting certain parameters that can lead to chaos. Another approach would be to mathematically remodel the system in such way that it does not present any possibility of having bifurcations and not worry about chaos at all.

Since the system cannot always be remodelled and not all parameters are easily reachable and modifiable, the simplest course of action would be to avoid the problem areas. This paper proposes a solution to avoid a specific type of chaotic behaviour that appears in a simple electronic circuit with two choppers as power sources.

The second section is presenting the goal of the paper and the theoretical aspects necessary to better understand what has been done. After this, the analysis method, known as route to chaos, has been described and presented in the third section for this particular case; also some simulation results are presented. Finally some conclusions and ideas for further research are given.

II. PROBLEM STATEMENT AND THEORETICAL BACKGROUND

II-A. Problem statement

The study starts with a given circuit model running in a stability point (the circuit presents no chaotic behaviour at all) with all parameters considered ideal (ideal components, ideal power source). Than, one will notice that by replacing the ideal sources with a model resembling the output signal given by a chopper, the whole system can run into chaos at certain frequencies. The goal is to eliminate chaotic behaviour. There are many approaches that can be applied: either use a classical power engineer's solution - increase the frequency of the chopper (power source) which will generate a more stable output signal and thus reduce the possibility of chaos [?] (which can be proven not to be entirely true), or a more physics related approach may be used by tweaking the internal components of the load in such way that chaos is avoided. The second solution is viable but only in certain situations where the internal parameters of the load can actually be tuned.

What this paper suggests is a third solution: a correlation between the frequency of the power source sampling and the chaotic behaviour of the whole system is determined and than applied to select the frequencies for the power source in order to avoid chaotic behaviour.

II-B. Theoretical model

The model used was an alpazur oscillator. This was selected because it is a 2-state piecewise-smooth [?] system and because it's simple in structure which makes it easier to analyse chaotic behaviour. This system has been studied throughout time [?], [?]. It was proven that it can be chaotic

throughout time [?], [?]. It was proven that it can be chaotic in certain theoretical situations even if all the signals and components used are ideal. The system can be described by Fig. 1:

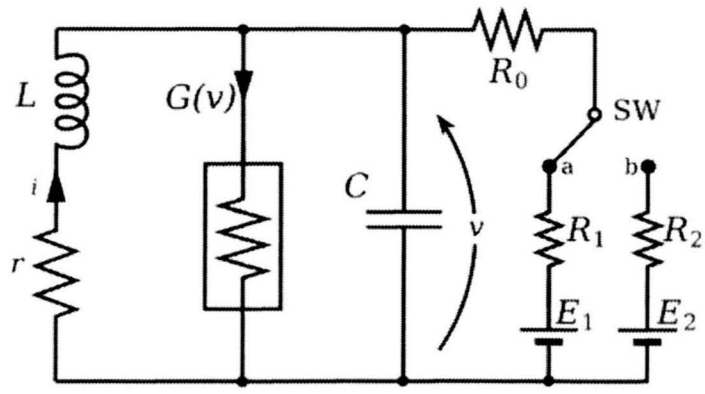


Fig. 1. 2-state alpazur oscilator circuit

The system in Fig. 1 is composed by 3 elements: the resonant elements, inductor and capacitor, the non-linearity that gives us the chaotic behaviour and the power source and command area (switch). The process relies on the balance of energy between the energy given by the power source (energy variation achieved by activating the switch) and the energy that is first accumulated than released by the non-linear element (the non-linearity acts as both resistance and power source - negative resistance). Since the analysis is done on a theoretical model, the non-linearity has been mathematically modelled by the equation:

$$g(v) = -a_1 v + a_3 v^3$$

The chaotic behaviour of the system in Fig. 1 is determined by the switching law applied to the switch *SW* that controls the passing from one state to another. Once set, this law cannot be modified throughout the experiment.

The whole system can be described than by the following set of equations:

$$\begin{cases} L\frac{di}{dt} = -ri - v \\ C\frac{dv}{dt} = i - g(v) + \frac{E_j - v}{R_0 + R_j} \end{cases} \quad withj = 1, 2$$

The modelling of the system requires a set of variable changes and notations. Also, the main interest is to properly determine i and v as the parameters that define the state of the system at a given time. Therefore the following notation and variable changes define the state of the system:

$$X = \begin{pmatrix} x \\ y \end{pmatrix}, \text{ where } x = i\sqrt{L} \text{ and } y = v\sqrt{C}$$

The rest of the variable changes are:

$$\tau = \frac{t}{\sqrt{LC}} \qquad r = r\sqrt{\frac{C}{L}} \qquad b = a_1\sqrt{\frac{L}{C}}$$
$$c = \frac{3a_3}{C}\sqrt{\frac{L}{C}} \quad A_j = \sqrt{\frac{L}{C}}\frac{1}{R_0 + R_j} \quad B_j = \sqrt{L}\frac{E_j}{R_0 + R_j}$$

In order to further simplify the equations, one can select $a_1 = 1$ and $a_3 = -1/3$. After applying the transformations and replacing the above the following set of equations is obtained:

$$\begin{cases} \frac{dx}{d\tau} = -rx - y \\ \frac{dy}{d\tau} = x + (1 - A_i)y - \frac{1}{3}y^3 + B_i \end{cases} \quad \text{where } i = 1, 2$$

The two positions of the switch take the system into two different states. The states are shown in the schematic on Fig. 2.

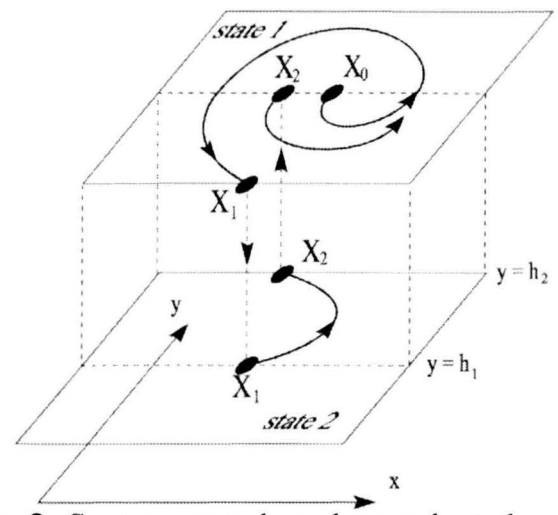


Fig. 2. System state based on selected source

The sources have been replaced with signal generators that produce a square signal at different frequencies. The amplitude of the oscillation is bewteen $-U$ and U of the respective source as shown by the equation:

$$E_{i(generated)} = E_{i(ideal)} + v_{sqare}$$
$$v_{sqare} = E_{i(ideal)}\frac{4}{\pi}\sum_{k=0}^{\infty}\frac{\sin(2k-1)2\pi fT}{2k-1}$$

where:

T – signal period
f – applied frequency

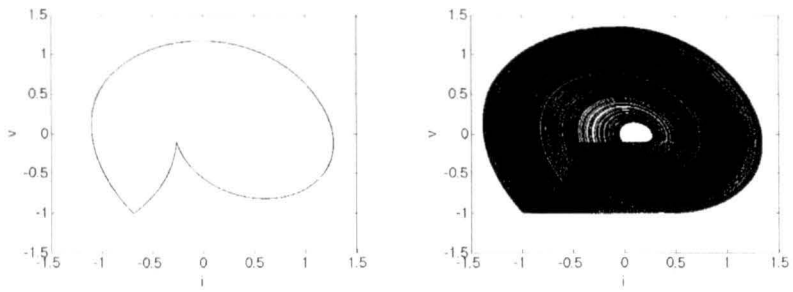
T and f represent in fact parameters of the simulation. They are configured as symulink environment variables that control the square signal generator.These values do not influence in any way the way the switching law of *SW* (Fig. 1). The simulation itself being numerical, has an internal frequency given by the simulation step parameter in Matlab environment. It is very important that the source frequency f is not greater than the $\frac{1}{2}$ of the simulation internal frequency to respect the Nyquist condition.

II-C. Behaviour of the ideal model

The model described above, exhibits a chaotic behaviour in certain situations. If all the elements are considered to be ideal, for a specific set of component values in the system and for two specific switching thresholds [?], [?], [?], bifurcations [?], [?] appear and the system will exhibit chaotic behaviour. In order to show this chaotic behaviour, a dependency between v - the tension applied to the capacitive

element and i - the current through the inductive element will be plotted. By doing this, one of two situations can be encountered:

- a stable system (see Fig. 3(a))
- a chaotic system (see Fig. 3(b))



(a) Stable system (b) Chaotic system

Fig. 3. System behaviour

In order to achieve the desired results, the following set of values are proposed for the components in the circuit:

$$E_1 = 143V \quad E_2 = -100V$$
$$R_0 = 40\Omega \quad R_1 = 460\Omega \quad R_2 = 10\Omega$$
$$L = 1mH \quad C = 1mF \quad r = 10\Omega$$

These values are used to obtain the stable circuit. Also, the two power sources are considered ideal (infinite sampling frequency).

The only aspect not described in the circuit schema is the command of the switch. The command is done based on two thresholds with respect to the capacitor voltage. The inferior threshold is set at $-1V$ and it is used as a commutation level when the circuit is driven by the source E_1 and it reaches this level, the switch moves on to the other source, E_2. It stays there until the voltage level rises to $-0.1V$ (the superior threshold) and then it goes back to the first source again. If the circuit is stable, the switches will always take place in the same spot on the graphic thus appearing as one constant path but when chaos appears, the result will be represented by multiple paths on the plotted dependency between i and v.

III. ROUTE TO CHAOS DUE TO CHOPPER FREQUENCY

A method is required that allows for a better analysis of the degree of chaos of the system. It is known that the system starts in a stability point but the variation of the source frequency will generate chaos. One can observe that the behaviour of the circuit is modified drastically when the frequency is changed even with a small amount. In order to observe the evolution throughout the spectrum, a representation called *route to chaos* [**?**], [**?**], [**?**] is needed. Such an example of route to chaos is given in Fig. 4. This will help determine at what frequencies the system is stable and where chaos appears.

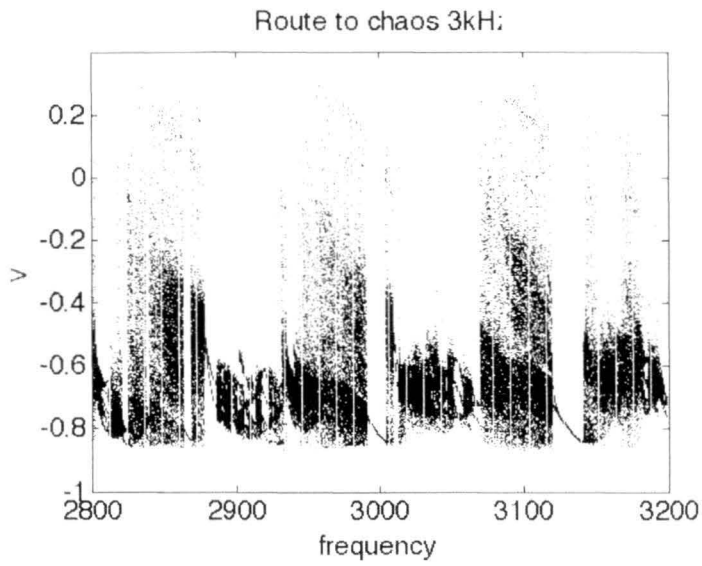


Fig. 4. Route to chaos

It is known that the route to chaos has been used in literature before with some very well written studies published by Mario di Bernardo [**?**] but all those studies were focused on either continuous systems or on discrete time systems. The proposed system is a hybrid dynamical system and so none of the existing methods in literature are applicable in the given situation. Therefore, a new method of chaos quantification is suggested and applied. The idea is to record the switching points on the higher level. Then, these points are placed on a vertical bar according to the current at which the switch was recorded (each i will have multiple v values). This way, the same number of points will be plotted for each simulation (each frequency step will have an identical number of switching points) but the dispersal of the points will differ based on the chaotic behaviour of the circuit. A stable system will have all the points overlapped (the multiple v values corresponding to the given i will have the same value or very close to each other) and they will appear as one single point in the graphical representation.

In the above example (see Fig. 4) one can observe that we have some areas of stable behaviour around $3kHz$. The same phenomena happens at other frequencies as well (similar results were recorded for frequencies of $4.6kHz$ and $6kHz$ - Fig. 5). This is the first indication that there are usable lower frequencies that allow the given system to be stable.

In order to clarify how the above graphics have been obtained, the systems behaviour in the designated areas will be analysed (see Fig. 6).

Here, one can observe the pass from a stable cycle to a chaotic one in a gradual manner. Variations show that it is possible to make adjustments quite easily because the system is not too sensible to the frequency (a variation of up to a few hertz is allowed) and so this frequency area can be used as a guide line for determining an optimal working frequency. It is also noticeable that in the graphics displaying the route to chaos some regions of particular interest are found: single

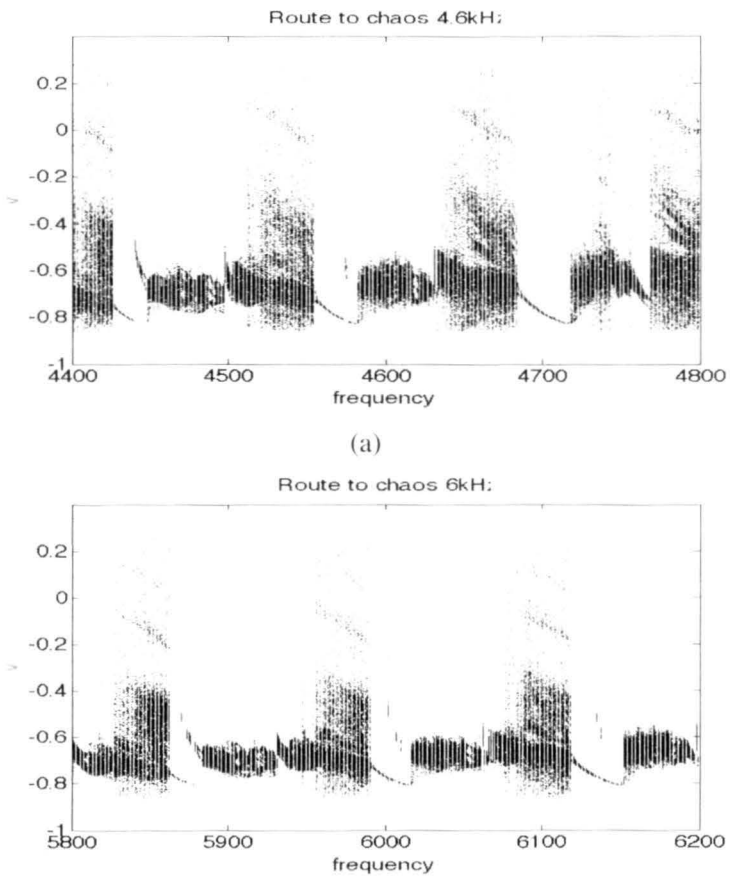


Fig. 5. Route to chaos - other regions of interest

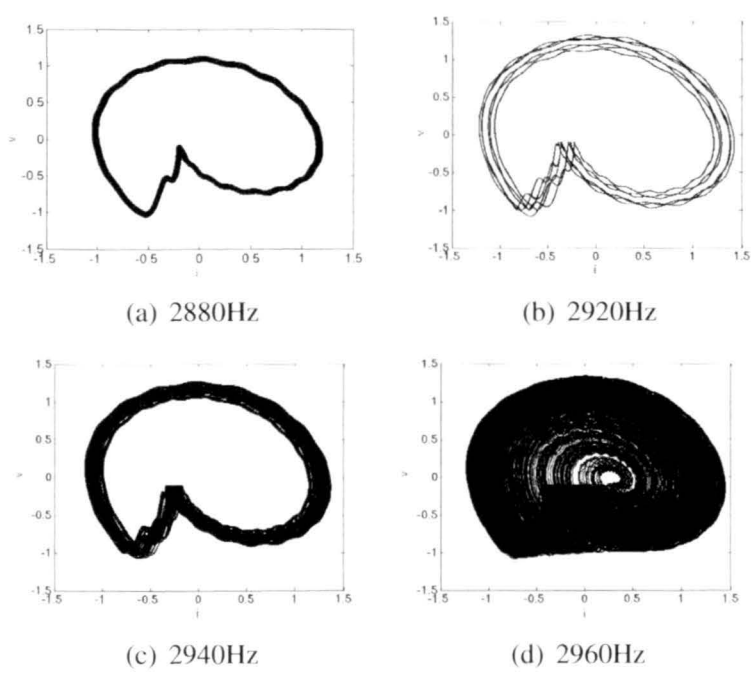


Fig. 6. $3khz$ Valley

period regions - one continuous thin line going from left to right; period doubling [**?**] regions - two thin lines that either converge when coming from a chaotic region or diverge when going towards chaos regions; chaos regions that vary in dimension.

At higher frequencies ($4.6kHz$) the area of usable frequencies is wider that at $3kHz$ and it keeps growing with the frequencies. This does not mean that chaos will eventually disappear. Chaos can be noticed at some very high frequencies so that proves you cannot just increase the frequency to eliminate chaos.

III-A. Preliminary analysis and a possible solution

Just proving, in numerical simulations, that lower frequencies can be used to achieve stability of a system is not enough from an engineering point of view. The main goal, in this section, is to figure out a way to determine these frequencies without prior testing. It is also mandatory to be able to provide the correct frequency areas without heavy computation. Any adjustment needed should not require a drastic increase in the control system's load that could prevent it from working properly in real time.

To do so, one could suggest computing the *FFT* of the voltage on our load using an ideal source model (treat it as a simple signal source). This will be used as a guideline for the proposed determination process. To better understand how the values of the *FFT* were obtained, we must imagine that the square signal generators were replaced with constants in the simulation and the voltage value on the alpazur load was recorded as a waveform for a period of time. This waveform was obtained for the ideal source case via numerical simulation (no actual physical measurements were done). This would suggest that, in the future, if this method is to be applied it is necessary to have some prior knowledge of the circuit we want to set in order to build a virtual model.

This method was selected because, in the end, this resumes to the study of a frequency behaviour and a FFT analysis is usually the best way to do it (Fig. 7).

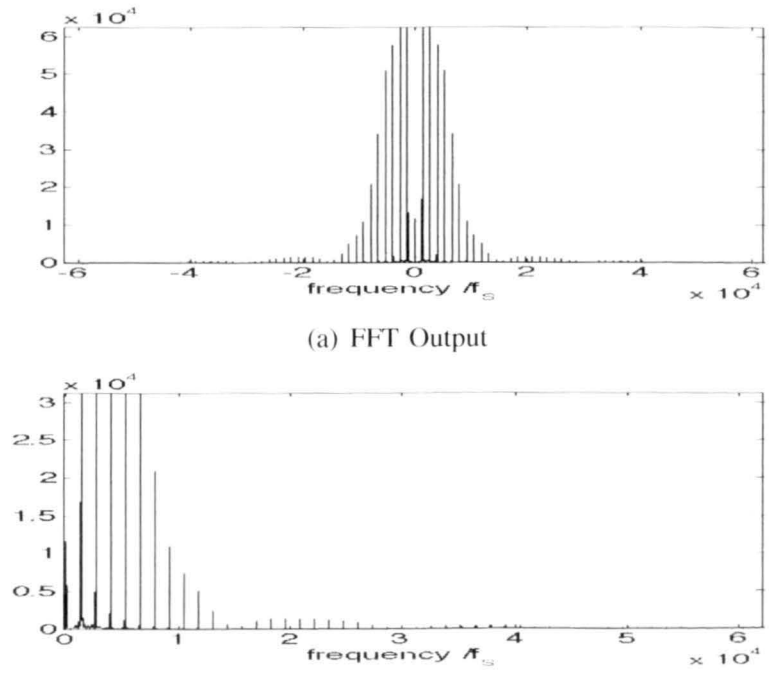


Fig. 7. Output FFT zoom

It is easily noticed that the FFT analysis provides us with some areas of interest that, for the purpose of better understanding the phenomena, shall be called *valley areas* (they present no impulses). These are the areas where the

method suggests to look for the stable frequencies. In theory, the system should be stable in all the areas between impulses. In fact a frequency can be found lower than $3kHz$ (in this case) that will work but determining that frequency and using it it will be very difficult if not impossible as the system will be very sensible to variations making adjustments very hard and not usable in a real life scenario.

The advantage of using the FFT is that the calculus is done quite fast and it can be done before setting up the system. So that the regions of interest will be known prior to the determination process.

IV. CONCLUSIONS

The proposed method proves that there is a quick and simple solution in solving the chaos problems due to source frequency in situations similar to the one presented. This may not be the only solution but the results are convincing.

Further research is necessary to observe the behaviour of the circuit from other aspects like: if the two sources work at different frequencies - does this reduce or augment the chaotic behaviour? A study of the behaviour of this method in a real life environment is also necessary to further acknowledge the validity of the numerical analysis. During the simulations there were assumed a limited amount of parameters for the components that may not cover all real life situations (for example, a thermal analysis can be done). Perhaps a self adapting control mechanism is necessary to compensate for minor imperfections in a real life model. As future development goes, this is a starting point that, in theory, should improve the stability of a system that can be affected by chaos and keep the costs of the devices used to a minimum.

ACKNOWLEDGMENT

The work has been funded by the Sectoral Operational Programme Human Resources Development 2007-2013 of the Romanian Ministry of Labour, Family and Social Protection through the Financial Agreement POSDRU/88/1.5/S/61178.

V. REFERENCES

[1] V. Grigoras, V. Tataru, and C. Grigoras, "Chaos modulation communication channel: A case study," in *Proc. Int. Symp. Signals, Circuits and Systems ISSCS 2009*, 2009, pp. 1–4.

[2] M. S. Baptista, "Cryptography with chaos," *Physics Letters A*, vol. 240, no. 1-2, pp. 50 – 54, 1998.

[3] W. Perruquetti and J-P. Barbot, *Chaos in automatic control*, CRC Press, Taylor & Francis Group, 2006.

[4] G. Chen and B. Han, "An audio scrambling degree measure based on information criteria," in *Proc. 2nd Int Signal Processing Systems (ICSPS) Conf*, 2010, vol. 1.

[5] M. Di Bernardo, C. Budd, Champneys, A.R., and P. Kowalczyk, *Piecewise-smooth Dynamical Systems - Theory and Applications*, Springer, 2008.

[6] T. Kousaka, T. Ueta, and H. Kawakami, "Bifurcation of switched nonlinear dynamical systems," vol. 46, no. 7, pp. 878–885, 1999.

[7] F. Angulo, M. di Bernardo, and G. Olivar, "Switching control of limit cycles in planar systems: a nonlinear dynamics approach," in *Proc. CDC Decision and Control 43rd IEEE Conf*, 2004, vol. 5, pp. 4897–4902.

[8] V. Moreno-Font, L. Benadero, and A. El Aroudi, "Non-smooth bifurcations in a 1-d piecewise linear model of a single inductor two-output dc-dc switching converter," in *Proc. IEEE Int Circuits and Systems (ISCAS) Symp*, 2011, pp. 2725–2728.

[9] G. Olivar, F. Angulo, and M. di Bernardo, "Hopf-like transitions in nonsmooth dynamical systems," in *Proc. Int. Symp. Circuits and Systems ISCAS '04*, 2004, vol. 4.

[10] M. Di Bernardo and S.J. Hogan, "Discontinuity-induced bifurcations of piecewise smooth dynamical systems," 2010.

[11] M. Di Bernardo, M.I. Feigin, S.J. Hogan, and M.E. Homer, "Local analysis of c-bifurcations in n-dimensional piecewise smooth dynamical systems chaos," 1999.

[12] D. Arroya-Almanza, A. Pisarchik, and F. Ruiz-Oliveras, "Route to chaos in a ring of three unidirectionally coupled semiconductor lasers," , no. 99, 2012, Early Access.

[13] P. S. Bodger, G. D. Irwin, D. A. Woodford, and A. M. Gole, "Bifurcation route to chaos for a ferroresonant circuit using an electromagnetic transients program," *IEE Proceedings-Generation, Transmission and Distribution*, vol. 143, no. 3, pp. 238–242, 1996.

[14] D. Cafagna and G. Grassi, "Bifurcation and chaos in the fractional chua and chen systems with very low order," in *Proc. IEEE Int. Symp. Circuits and Systems ISCAS 2009*, 2009, pp. 2846–2849.

[15] M. di Bernardo, F. Garefalo, L. Glielmo, and F. Vasca, "Analysis of chaotic buck, boost and buck-boost converters through switching maps," in *Proc. th Annual IEEE Power Electronics Specialists Conf. PESC '97 Record*, 1997, vol. 1, pp. 754–760.

[16] Huang Anshan, "Period doubling and period plus one law-chaotic phenomena in similar chua's circuit," in *Proc. , China. Conf. Int Circuits and Systems*, 1991, pp. 736–739.

Introductory Study on the Discrete Trigonometric Transform (DTT) Wavelet-Like Decomposition-Reconstruction Procedure

Przemysław Korohoda
AGH University of Science and Technology, Chair of Electronics
30 Mickiewicza, 30-059 Krakow, Poland
e-mail: korohoda@agh.edu.pl

Adam Dąbrowski
Poznan University of Technology, Division
of Signal Processing and Electronic Systems,
Chair of Control and System Engineering, Faculty of Computing
11 Strzelecka, 60-965 Poznań, Poznan, Poland
e-mail: adam.dabrowski@put.poznan.pl

ABSTRACT — **In the paper the results of investigations indicating computational efficiency of Discrete Trigonometric Transforms (DTTs) used to design filters by means of the complete wavelet-like decomposition/reconstruction procedure, are presented. The study indicates flexibility when comparing DTT selection from the set of 16 transforms, and the sampling type used. The results, obtained for the block length up to N=1024, should encourage applications of the suggested solutions to all areas where the classical wavelet filters are utilized, while the reconstruction error at the Peak Signal to Noise Ratio (PSNR) level of 37.5 dB can be accepted to balance the benefit of reduction of computations up to 68 times, obtained after including the Hamming window in the filtering process. The comparisons are based on the error measure derived from the signal covariance matrix.**

KEYWORDS — trigonometric transforms; downsampling; upsampling; halfband filtering; generalized convolution

I. INTRODUCTION

A filtering scheme commonly known as the Mallat decomposition and reconstruction procedure [15] is particularly well suited to approximate the Discrete Wavelet Transform along with its inverse [23, 24], which is often described as the time-frequency analysis tool. The frequency part of such approach is typically interpreted with the Discrete Fourier Transform [19, 21]. However, the DFT domain for real signals is symmetric and as such, for a given block length, say N, offers only half of N independent frequency values. The Discrete Trigonometric Transforms (DTTs) [1, 3] do not need to preserve such symmetry and therefore they have all N values to be used independently. However, the DTTs do not have the classical shifting property [16, 20], which makes filtering in the time domain not as straightforwardly represented as in the case of the DFT. Therefore, in [16] a concept of symmetric convolution was presented and described in detail, introducing relevant formulas for all 16 DDTs, and in [20] the profound overview of the developed solutions, also referring to the multiplication-convolution property, was provided. Regarding the presented study, the consecutive efforts leading to the possibility of using the DTTs interpretation for the Mallat scheme may be found in the series of publications. In [4] the generalized convolution was suggested, further developed in [6, 7, 8, 9], and in [10, 11] some introductory results indicating possible DTT-based approach, but with filtering performed in the transform domain, to the discussed scheme were presented. In [12] the efficiency of only decomposition stage of the procedure with the use of the DTT-based time domain windowed filters was positively verified. In this paper, the results indicating a possible efficient application of the DTT-based wavelet-like complete scheme, are presented. The study is a natural extension of the work described in [13].

II. DTTS AND THE GENERALIZED CONVOLUTION

In total, there are 16 DTTs: eight cosine (i.e., DCTs) and eight sine (i.e., DSTs). Among them there are eight even (e) and eight odd (o) transforms [1, 3, 6]. For simplicity, all of these transform versions are henceforth referred according to the numbers listed in Table 1.

TABLE I. NUMBERING OF THE DTTS

DTT	number	DTT	Number
DCT Ie	1	DST Ie	9
DCT IIe	2	DST IIe	10
DCT IIIe	3	DST IIIe	11
DCT IVe	4	DST IVe	12
DCT Io	5	DST Io	13
DCT IIo	6	DST IIo	14
DCT IIIo	7	DST IIIo	15
DCT IVo	8	DST IVo	16

In case of the DTTs the convolution-multiplication property of the DFT [17, 18] can be replaced by a somehow generalized form in such a way that the multiplication of the transform domain characteristics is equivalent to the so-called generalized convolution realized in the signal (time) domain [4, 5, 6], i.e.

$$\mathbf{y} = \mathbf{A}^{-1}(\mathbf{Ax} \circ \mathbf{Ah}) = \mathbf{x} \otimes \mathbf{h} = \mathbf{Hx} \qquad (1)$$

where $\mathbf{x}$, $\mathbf{h}$, and $\mathbf{y}$ are vectors representing signal blocks of length N, $\mathbf{A}$ is the transform matrix, symbol $\circ$ represents the element-by-element multiplication of vectors or matrices and $\otimes$ is the generalized convolution operator. $\mathbf{H}$ is the matrix that depends on $\mathbf{h}$ and $\mathbf{A}$. For the DFT this is known as the circular convolution matrix. However, for other linear transforms expression (1) represents the generalized convolution concept.

On the part of the first author this work was supported by the AGH University of Science and Technology grant no. 11.11.120.766.

As described in [8, 9, 12] and particularly in [13], the number of multiplications necessary to obtain the result of **H** times **x** can be considerably reduced for the DTTs with the use of the relevant windowing technique, represented by the window matrix **W**

$$\mathbf{y} \cong (\mathbf{W} \circ \mathbf{H})\mathbf{x} \,. \tag{2}$$

However, as indicated in (2), for the DTTs and the ideal half-band filters such windowing inevitably introduces errors.

III. Decomposition-reconstruction scheme

Figure 1 shows the considered one-level decomposition-reconstruction scheme, also known as the analysis-synthesis procedure. Here, $\mathbf{y}_L$ and $\mathbf{y}_H$ represent the low- and high-frequency halves of the transform coefficients, respectively.

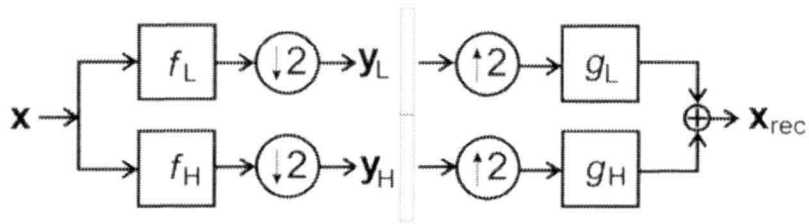


Figure 1. One level of the wavelet-like decomposition procedure used in the paper.

Low-pass and high-pass filtering, represented by the square shaped blocks, can be either performed directly in the signal (time) domain or, equivalently, in the transform domain [11, 12]. When realized in the time domain with the low-pass and high-pass half-band filters defined in the DTT domain, the generalized convolution is involved. Important features of this scheme are downsampling and upsampling operations, performed in the signal domain. Note, that a similar scheme, but with the selection of samples realized in the transform domain, were presented independently in [2] and [14] and in the preceding authors' papers. However, the difference of our approach is important as explained in [10].

The whole procedure depicted in Figure 1 can be formally described by a matrix, comprising all necessary operations [10, 11].

IV. Error measure

Apart from windowing as described by (2), there might be some other sources of errors. The complete procedure, as depicted in Figure 1, does not guarantee perfect reconstruction, because such feature depends on selection of decomposition and reconstruction filters. The wavelet filters [21, 22, 23, 24] are precisely designed to preserve the signal. In the presented approach, the only assumptions are: a) the filters have the ideal half-band characteristics defined in the given DTT domain, b) the reconstruction and decomposition filters are the same. Therefore, to compare efficiencies of various options the introduced root mean square errors (RMSEs) evaluated with the following formula [3]

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\mathrm{trace}\left((\mathbf{X}-\mathbf{Y})\mathbf{R}_\rho(\mathbf{X}-\mathbf{Y})^T\right)} \tag{3}$$

have been compared. In (3) **X** and **Y** are matrix operators, which are the subject of comparison, and $\mathbf{R}_\rho$ is the covariance matrix computed for the signal, that is assumed to be the first order Markov process with zero mean (μ=0), unit standard deviation (σ=1), and the intersymbol correlation ρ [3]

$$\mathbf{R}_\rho : R(m,n) = \rho^{|m-n|} \,. \tag{4}$$

In this paper, in some comparisons, the RMSE is replaced with the Peak Signal to Noise Ratio (PSNR), expressed in dB and computed with the assumption that the span of the signal range is 6σ [12].

V. Varying the DTT and sampling style selection

A fact that only four out of the sixteen DTTs offer perfect reconstruction with the use of the considered procedure has already been indicated in [10, 11]. However, similar shapes of the basis functions for various DTTs prompted investigation of the introduced errors, if the reconstruction filters are defined with some other transform than the decomposition filters. Another question is related to the manner of the sampling, which may be performed in four possible ways: even-even, odd-even, even-odd, odd-odd (for the low-pass and the high-pass paths, respectively). Figure 2 presents the obtained results showing that the transforms No. 1, 3, 9, and 11, i.e., DCT Ie, DCT IIIe, DST Ie and DST IIIe form close relationships, and may be used interchangeably to define the decomposition and reconstruction filters. However, the RMSE values are also pretty small for the remaining pairs of the DTTs. Note, that on diagonals of the patterns in Figure 2, there are results for the same transform used in both stages of the procedure. In general, in spite of the fact that some differences are visible, selection of the sampling type does not affect the results considerably.

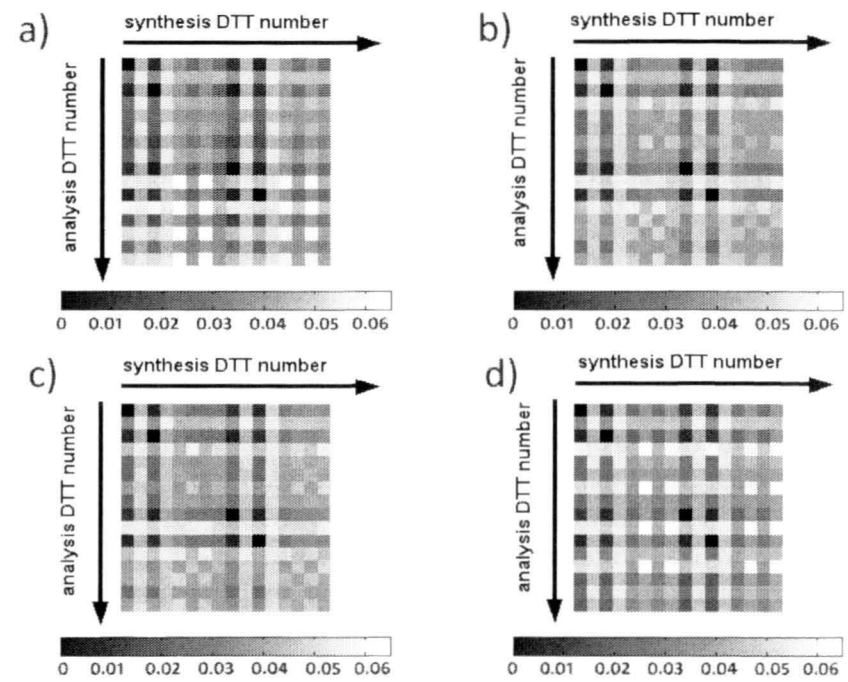


Figure 2. Theoretical RMSE results (for assumed ρ=0.95 and N=256), presented in the gray levels form, obtained for cross-selection of the DTTs and four possible options of down-sampling: a) even-even, b) odd-even, c) even-odd, d) odd-odd (first is the low-pass path). In each case, the up-sampling was the same as the down-sampling.

VI. THE BASIS FUNCTIONS

To provide comparisons with the well known wavelet functions, an example of the basis functions set was computed as described in [12] for the 4-level scheme, but only the low frequency transform coefficients were processed in deeper levels. The results are presented in Figure 3, Figure 4, and Figure 5. Figure 6 shows the basis functions indicating the consequences of using the Hamming window at all levels except the last one.

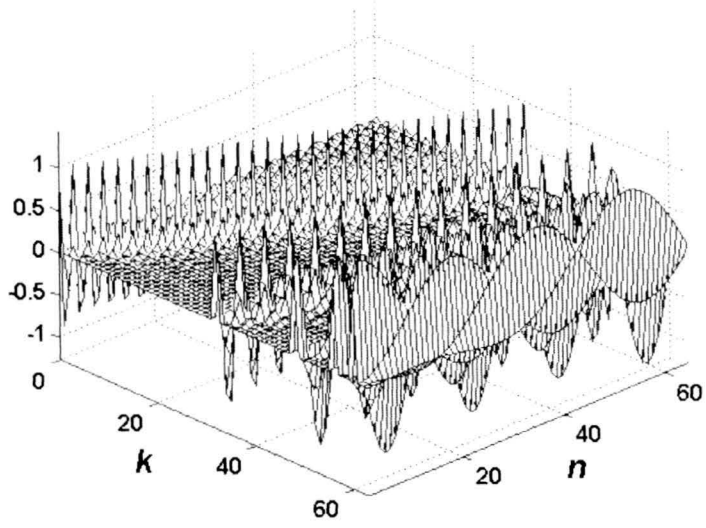


Figure 3. Complete set of basis functions for the 4-level analysis-synthesis procedure computed for the DCT IIIe (DTT no. 3), N=64, n is the index in the primary domain.

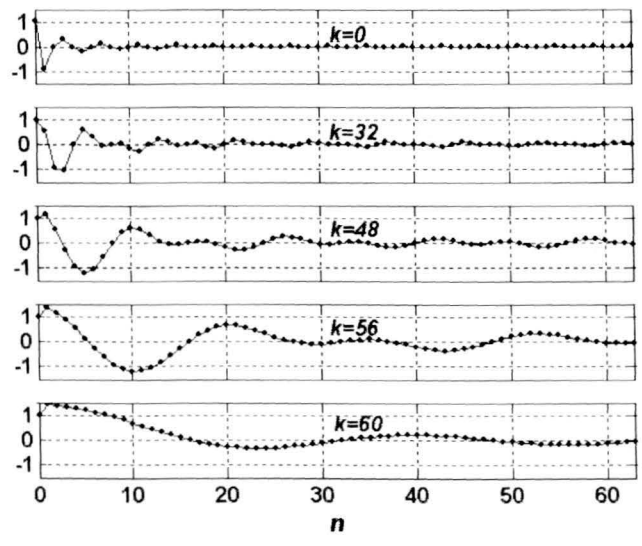


Figure 4. Selected basis functions from Figure 3.

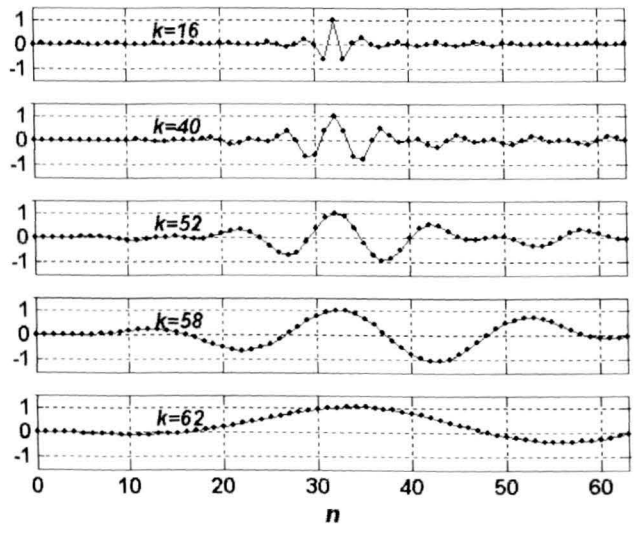


Figure 5. Selected basis functions from Figure 3.

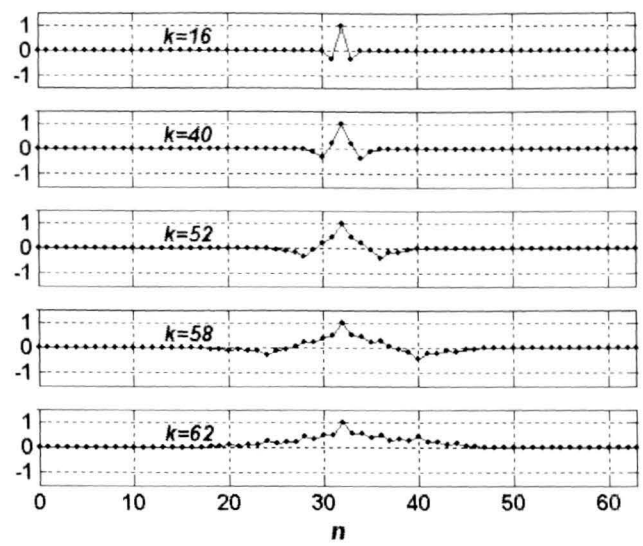


Figure 6. Selected basis functions for the same procedure as for Figure 3, but with additionally used Hamming window with length W=5 - compare with Figure 5.

The length of the window was selected to be very short, to provide hard conditions for the reconstruction and to exploit the windowing in lower levels, where the length of the signal block is reduced by the factor of 2 at each level. Note, that all functions exhibit local preferences, as one typically expects in case of the wavelet-like analysis, and the harsh windowing did not affect such feature.

VII. COMPARATIVE RESULTS FOR SELECTED OPTIONS

To provide some overview of the procedure performance, the RMSE values were computed for different N values and all DTTs, corresponding to the diagonals of the patterns in Figure 2. In [12] one can find the equivalent results obtained for the decomposition stage only. Figures 7 to 15 make it possible to compare the following results: obtained without windowing, for the rectangular window, and for the Hamming window, used as described in [12], and for two assumed correlation values: for ρ=0.95 and ρ=0.80.

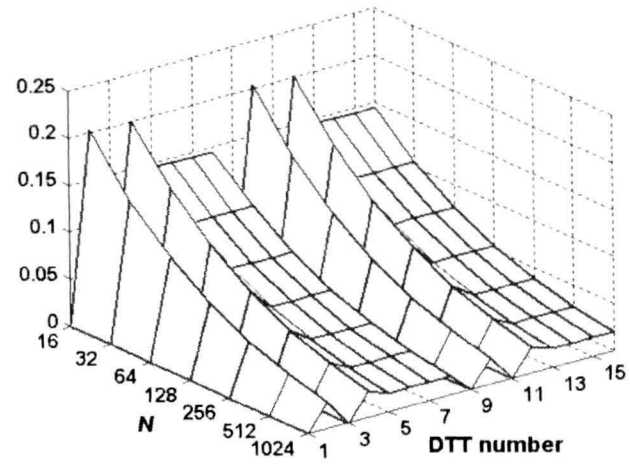


Figure 7. RMSE, comparing the original and reconstructed signal, computed from theoretical formula (3) for ρ=0.95, for all 16 DTTs and the range of block length values, N.

Figures 16, 17, and 18 compare reconstruction accuracy for the selected transforms: DCT IIIe, which guarantees the perfect reconstruction without windowing, and DCT IIe, which does not. Level and type of distortions with windowing are similar for both transforms. Note also the hardly noticeable distortions for the DCT IIe used without windowing – Figure 17.

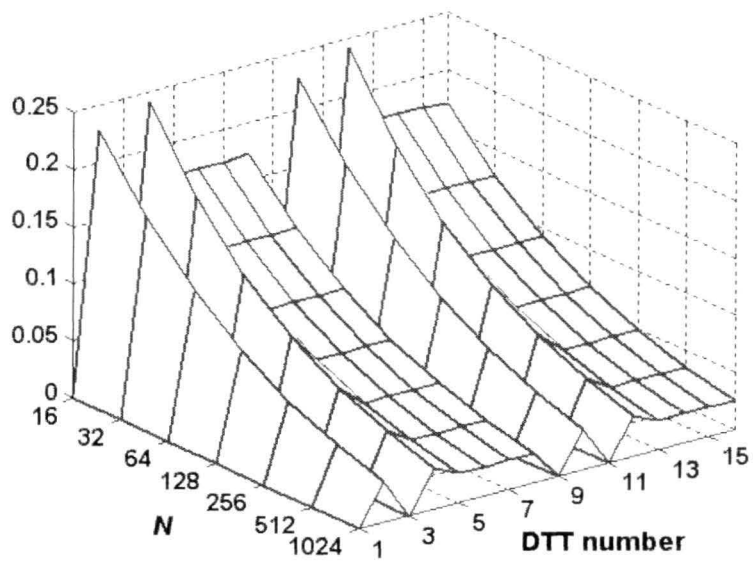


Figure 8. RMSE comparing the original and reconstructed signal, computed from theoretical formula (3) for ρ=0.80, for all 16 DTTs and the range of block length values, N.

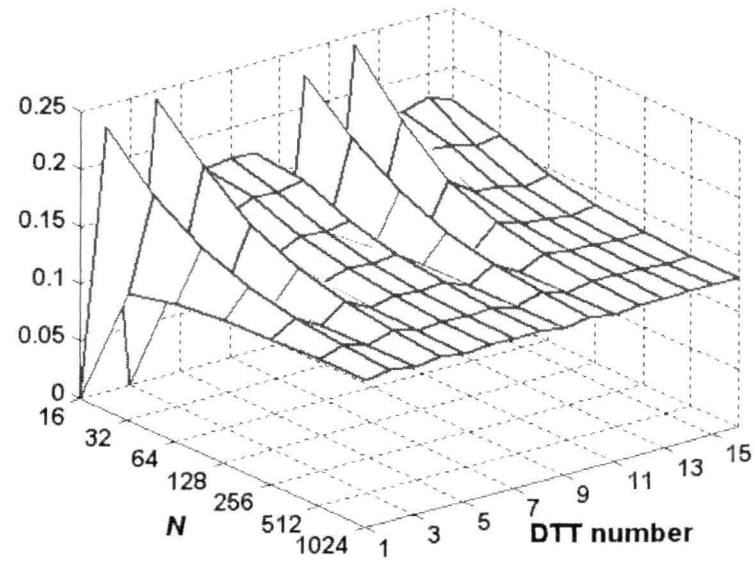


Figure 11. RMSE comparing the original and reconstructed signal, computed from theoretical formula (3) for ρ=0.80, for all 16 DTTs and the range of block length values, N, while using the rectangular window of length W=15.

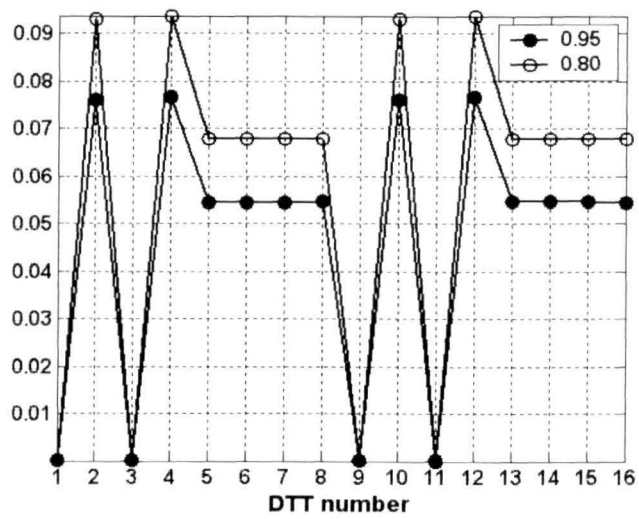


Figure 9. Comparative selection of the theoretically computed RMSE values (3), presented in Figure 8 and Figure 9, for ρ=0.95 and ρ=0.80, while N=128.

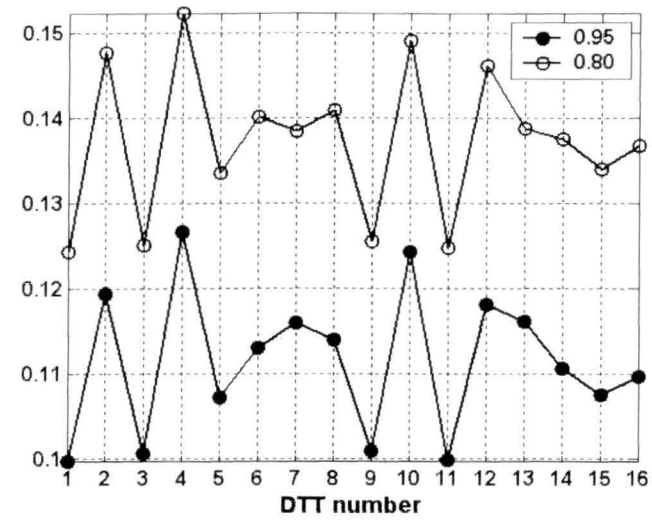


Figure 12. Comparative selection of the theoretically computed RMSE values (3), presented in Figure 10 and Figure 11, for ρ=0.95 and ρ=0.80, while N=128.

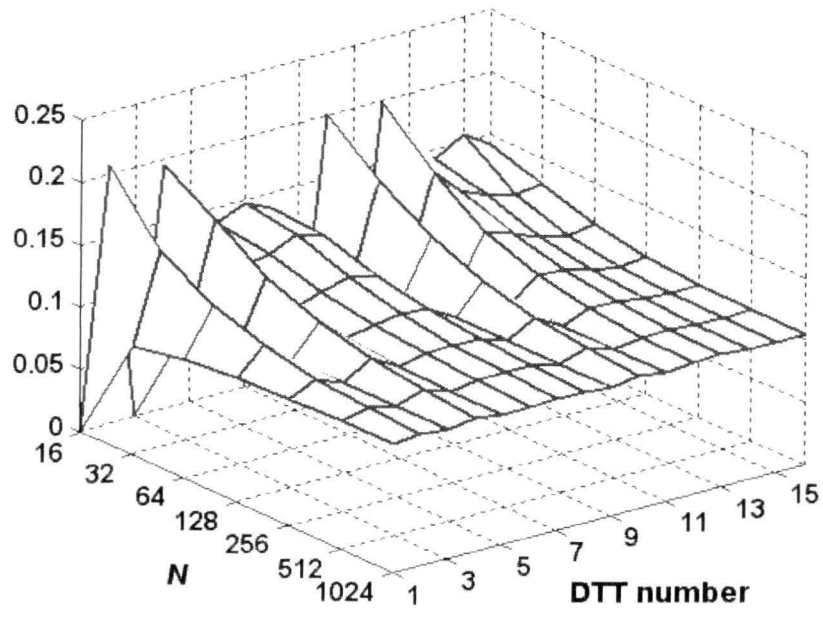


Figure 10. RMSE comparing the original and reconstructed signal, computed from theoretical formula (3) for ρ=0.95, for all 16 DTTs and the range of block length values, N, while using the rectangular window of length W=15.

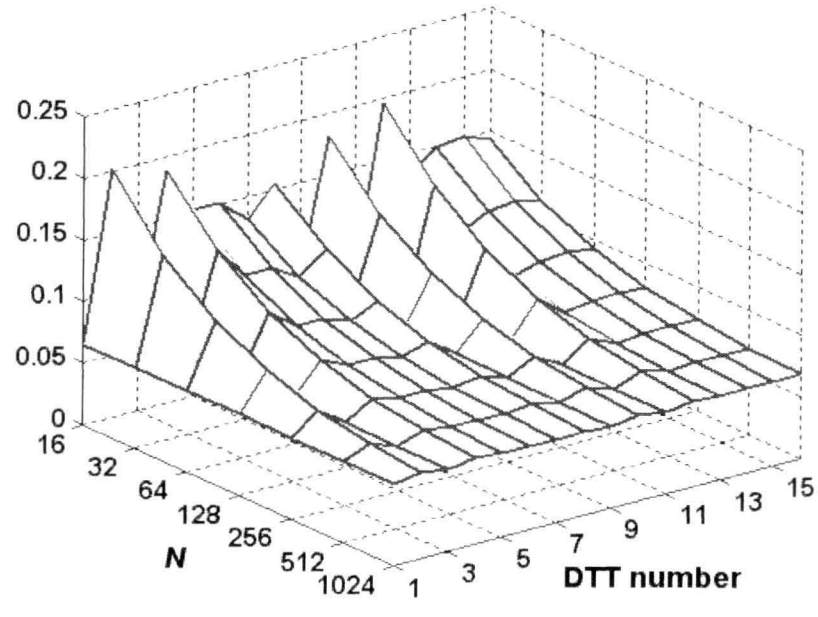


Figure 13. RMSE comparing the original and reconstructed signal, computed from theoretical formula (3) for ρ=0.95, for all 16 DTTs and the range of block length values, N, while using the Hamming window of length W=15.

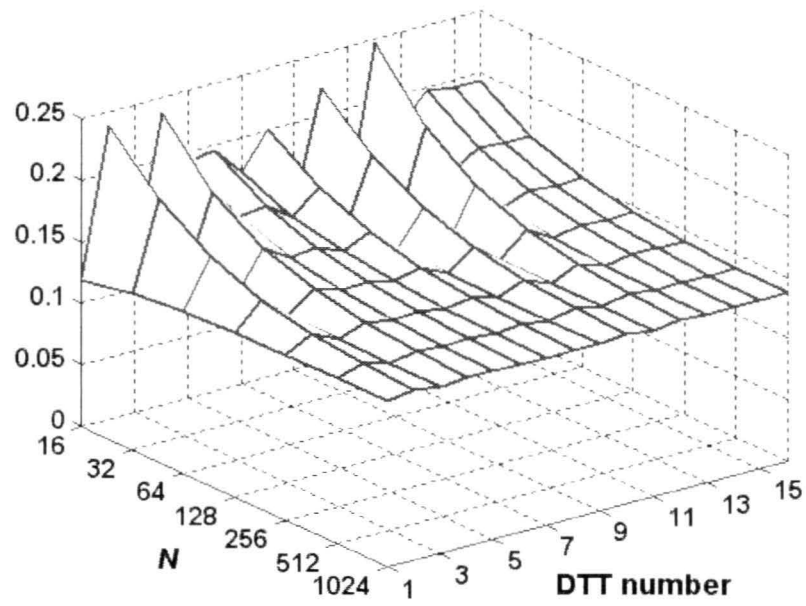


Figure 14. RMSE comparing the original and reconstructed signal, computed from theoretical formula (3) for ρ=0.80, for all 16 DTTs and the range of block length values, N, while using the Hamming window of length W=15.

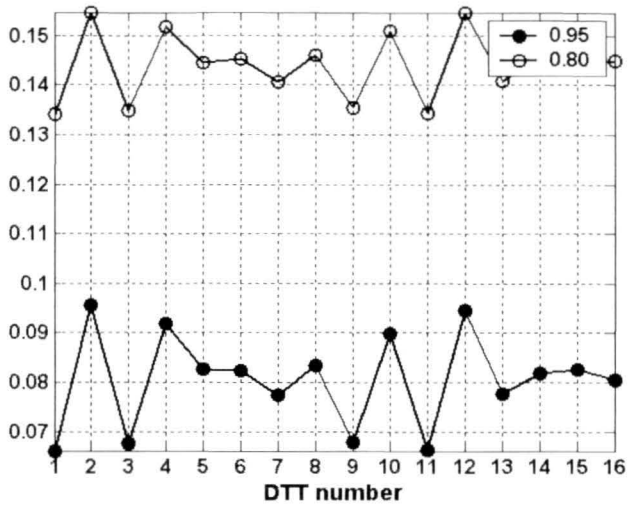


Figure 15. Comparative selection of the theoretically computed RMSE values (3), presented in Figure 13 and Figure 14, for ρ=0.95 and ρ=0.80, while N=128.

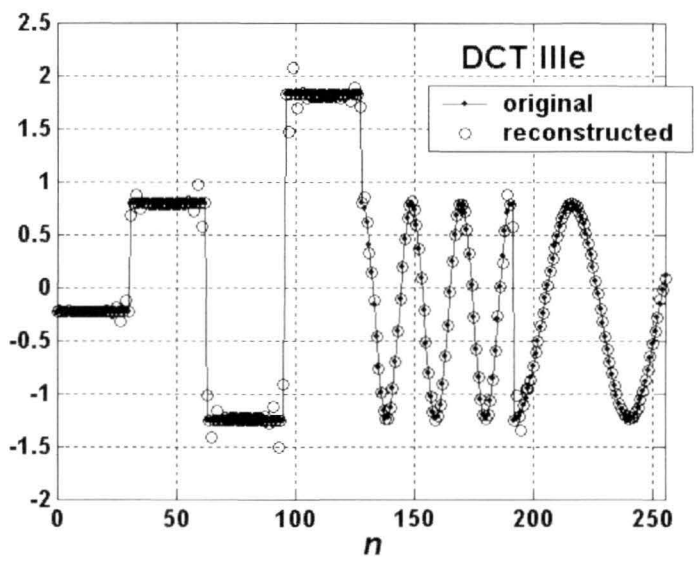


Figure 16. The original test signal (N=256, ρ=0.954, μ=0, σ=1) and the reconstructed signal, after using the wavelet-like procedure designed for the DCT IIIe (DTT no. 3) with use of the Hamming window, W=15. RMSE(theory)=0.0642, RMSE(signal)=0.0573, PSNR=40.4dB.

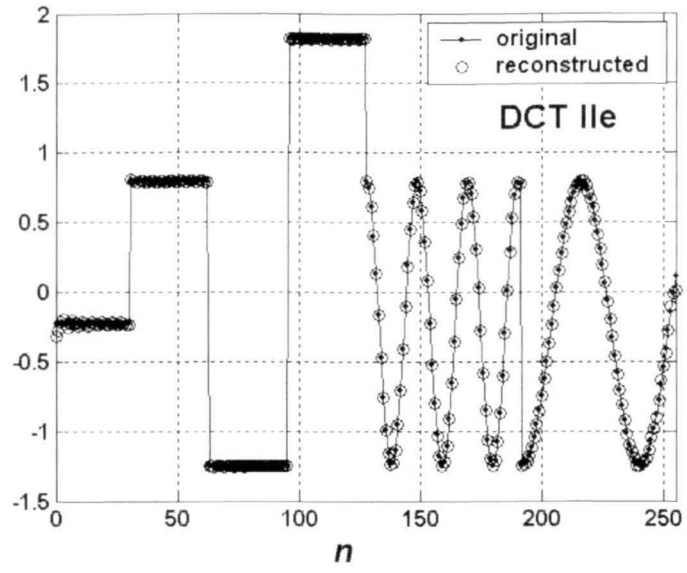


Figure 17. The original test signal (N=256, ρ=0.954, μ=0, σ=1) and the reconstructed signal, after using the wavelet-like procedure designed for the DCT IIe (DTT no. 2), without windowing. RMSE(theory)=0.0542, RMSE(signal)=0.0112, PSNR=54.6dB.

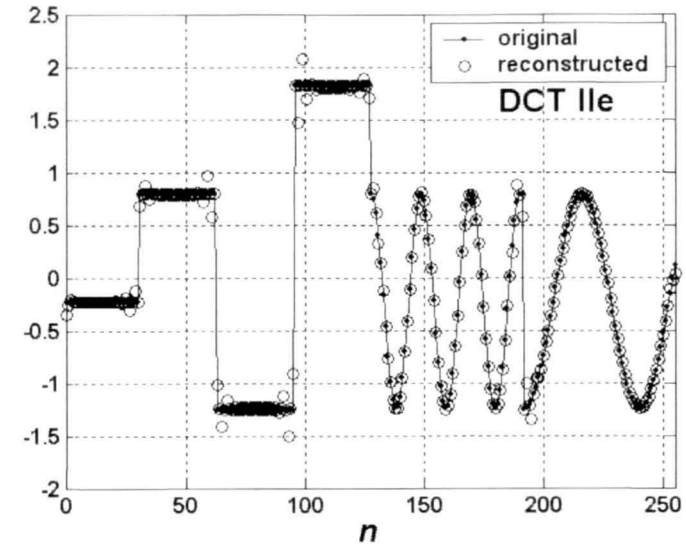


Figure 18. The original test signal (N=256, ρ=0.954, μ=0, σ=1) and the reconstructed signal, after using the wavelet-like procedure designed for the DCT IIe (DTT no. 2) with use of the Hamming window, W=15. RMSE(theory)=0.0800, RMSE(signal)=0.0581, PSNR=40.3dB.

VIII. Concluding Comments

The presented results, even if being only examples obtained for selected transforms, can be considered as general for the whole DTT family, because similar graphs are obtainable for all relevant DTTs. This remark is valid for example for Figure 2, in which the length of the block has a noticeable impact on the RMSE level. Thus this becomes serious for short blocks but the shown patterns are very much alike. Note that Figures 9, 12, and 15 provide the RMSE values computed along diagonals of similar graphs to those presented in Figure 2, but for N=128. Figures 16, 17, and 18 should be also considered as examples indicating that all DTTs can be used efficiently to provide filters for the decomposition-reconstruction scheme. According to the presented results, the signal intersymbol correlation value plays an important role when the Hamming window is applied, if compared with the rectangular window, and is hardly noticeable for the windowless application. For ρ=0.95 and N=128 or larger, the Hamming window with length W=15 makes it possible to reconstruct the signal with RMSE around 0.08 for all DTTs, which is equivalent to PSNR=37.5 dB. For

N=1024 such window means reduction by 68 times in the amount of computations as compared to the direct matrix based realization of the procedure. Such performance suggests possible applications in the areas of classical signal or image compression, watermarking, or feature extraction.

REFERENCES

[1] N. Ahmed, K. R. Rao, Orthogonal Transforms for Digital Signal Processing, Springer-Verlag, 1975.

[2] N. Basumallick, S. V. Narasimhan, A Discrete Cosine Adaptive Harmonic Wavelet Packet and Its Application to Signal Compression," Journal of Signal and Information Processing, vol.1, 2010, pp.63-76.

[3] V. Britanak, P. Yip, K. R. Rao, Discrete Cosine and Sine Transforms – General Properties, Fast Algorithms and Integer Approximations. Elsevier, Academic Press, 2007.

[4] P. Korohoda, A. Dąbrowski, "Generalized primary domain interpretation of product filtering of digital signals in the transform domain," Proceedings of Signal Processing, (Poznań, Poland), vol. 51, no. 3, Sept. 2003, pp.75-80.

[5] P. Korohoda, A. Dąbrowski, "Fast filtering by generalized convolution related to discrete trigonometric transforms," Proceedings of the IEEE workshop on Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA 2007), Poznań, Poland, Sept. 2007, pp.57-62.

[6] P. Korohoda, A. Dąbrowski, "Generalized convolution a s a tool for the multi-dimensional filtering task," Multidimensional Systems and Signal Processing, vol. 19, 2008, pp.361-377.

[7] P. Korohoda, A. Dąbrowski, "Digital filtering by discrete trigonometric transforms realized with generalized convolution," Elektronika, vol. 48, no. 4, 2008, pp.95-98.

[8] P. Korohoda, A. Dąbrowski, "Lowpass filtering with filters defined in the discrete trigonometric transform domains," Proceedings of the IEEE workshop on Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA 2009), Poznań, Poland, Sept. 2009, pp.66-69.

[9] P. Korohoda, A. Dąbrowski, "Discrete trigonometric transforms (DTT) filters," Elektronika, vol. 51, no. 3, 2010, pp.45-49.

[10] P. Korohoda, A. Dąbrowski, "Experimental study of wavelet-like decomposition based on filtering in domains of discrete trigonometric transforms," Proceedings of the IEEE workshop on Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA 2010), Poznań, Poland, Sept. 2010, pp.16-20.

[11] P. Korohoda, A. Dąbrowski, "Wavelet-like decomposition based on filtering in domains of discrete trigonometric transforms - case study," Elektronika, vol. 52, no. 5, 2011, pp.85-89.

[12] P. Korohoda, A. Dąbrowski, "Experimental study on the Discrete Trigonometric Transform (DTT) wavelet-like decomposition stage with windowed filters," Proceedings of the IEEE workshop on Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA 2011), Poznań, Poland, Sept. 2011, pp.33-38.

[13] P. Korohoda, A. Dąbrowski, "Wavelet-like decomposition stage with windowed filters defined for the Discrete Trigonometric Transforms (DTTs)," Electrical Review, vol. 88, no. 6, pp.30–35.

[14] D. Majorkowska, A. Ţariov, "Procedures of multilevel 2D data decomposition and reconstruction with wavelet-like packets," (In Polish), Elektronika, vol. 48, no. 2, 2008, pp.48-52.

[15] S. G. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 11, July 1989, pp.674-693.

[16] S.A. Martucci, "Symmetric Convolution and the Discrete Sine and Cosine Transforms," IEEE Trans. on Signal Processing, vol. 42, no. 5, May 1994, pp.1038-1051.

[17] S. K. Mitra, Digital Signal Processing: A Computer Based Approach, McGraw-Hill, 1998.

[18] J. G. Proakis, D. G. Manolakis, Digital Signal Processing: Principles, Algorithms, and Applications, Prentice Hall, 3rd Ed. 1996.

[19] K. R. Rao, D. N. Kim, J. J. Hwang, Fast Fourier Transform: Algorithms And Applications, Springer, 2010.

[20] N. Roma, L. Sousa, "A tutorial overview on the properties of the discrete cosine transform for encoded image and video processing," Signal Processing, vol. 91. no. 11, 2011, pp.2443-2464.

[21] G. Strang, T. Nguyen, Wavelets and Filter Banks, Wellesley College, 2nd Ed. 1996.

[22] P. P. Vaidyanathan, Multirate Systems and Filter Banks, Prentice Hall, 1992.

[23] M. Vetterli, C. Herley, "Wavelets and Filter Banks: Theory and Design," IEEE Transactions on Signal Processing, vol. 40, no.9, pp.2207-2231, 1992.

[24] M. Vetterli, J. Kovaćević, Wavelets and Subband Coding, Prentice Hall, 1995.

Effectiveness of Lattice Factorization of Two-Channel Orthogonal Filter Banks

Mykhaylo Yatsymirskyy
Institute of Information Technology
Technical University of Łódź
Łódź, Poland
mykhaylo.yatsymirskyy@p.lodz.pl

Kamil Stokfiszewski
Institute of Information Technology
Technical University of Łódź
Łódź, Poland
kamil.stokfiszewski@p.lodz.pl

***Abstract* — In this paper the authors show that for sequential processors the usage of lattice structures reduces almost by the factor of two the computational time of the filtering process performed by two-channel orthogonal filter bank with respect to the direct method of calculating the convolution. In addition authors give the derivation and perform the analysis of the lattice factorization of two-channel orthogonal filter banks, which can also be easily extended to more general biorthogonal filter banks' case.**

Keywords* — *orthogonal filters; lattice structures; wavelet transform

I. Introduction

Wavelet transform is widely used in data compression, system identification, signal approximation and interpolation, image processing and recognition, as well as in watermarking [1-4]. This makes the research on improvement of the algorithms of the wavelet transform calculation very intense [5-8]. Construction of the wavelet transform relies on the successive repetition of the basis analysis stages for forward transformation or the basis synthesis stages for the inverse transformation in a pyramid or parallel scheme manner. The classical model of the single basis stage of the considered transformation is a two-channel biorthogonal filter bank, like the one shown in Fig. 1.

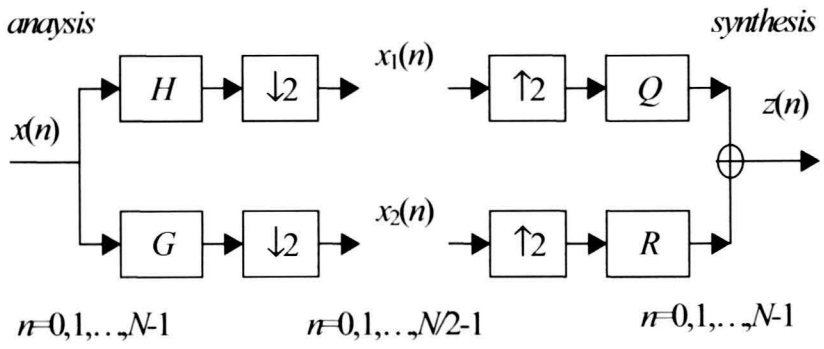


Figure 1. Scheme of a single analysis and synthesis steps of the wavelet transformation

Units H, G, Q and R represent finite-response linear filters $H = h_0, h_1, ..., h_{K-1}$, $G = g_0, g_1, ..., g_{K-1}$, $Q = q_0, q_1, ..., q_{K-1}$ and $R = r_0, r_1, ..., r_{K-1}$, where K is a natural even number. Units $\downarrow 2$ represent twofold element downsampling of the output signal after the filtering analysis stage and units $\uparrow 2$ represent the corresponding twofold element upsampling realized by proper insertion of zeros to the input signal of the filtering process' synthesis stage. Under the perfect reconstruction constraint i.e. $z(n) \equiv x(n)$ [9], the considered filters fulfill a set of conditions given by the following system of $(K-1)$ equations:

$$\left\langle (-1)^k h_{K-1+2m-k} q_k \right\rangle = \sum_{k=0}^{K-1} (-1)^k h_{K-1+2m-k} q_k = \pm\delta(m),$$

where $\delta(0) = 1$ and $\delta(m) = 0$ for the remaining values of m, $m = 0, \pm 1, \pm 2, ..., \pm(K/2-1)$. Moreover the impulse responses of the Q and R filters constitute the modulated copies of H and G filters' responses: $q_k = \pm(-1)^k g_k$, $r_k = \mp(-1)^k h_k$, $k = 0,1,.., K-1$.

An important class of biorthogonal filters consists of the filters, for which there is an additional relationship between the analysis and synthesis filter pairs: $g_k = (-1)^k h_k$, $k = 0,1,...K-1$. For the effective implementation of the two-channel orthogonal filter banks lattice structures are often used. There are two basic variants of such structures known. In the first variant for filters of length K, where K is a natural even number, there are $K/2$ $\mathbf{D_k}$ vector rotation and unit delay operations, as well as one additional operation of two-point Haar transform calculation [3], denoted below in Fig. 2 by $\mathbf{H_2}$.

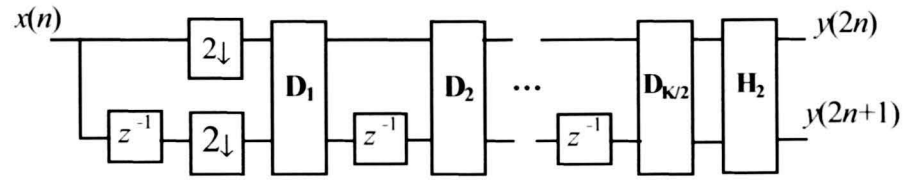


Figure 2. Scheme of a first variant of lattice structure for two-channel orthogonal filter bank calculation

For the second, the more effective, of the considered variants, the essential effectiveness improvement comes form the fact of reduction of the mentioned $\mathbf{H_2}$ operation [10,11], Fig. 3.

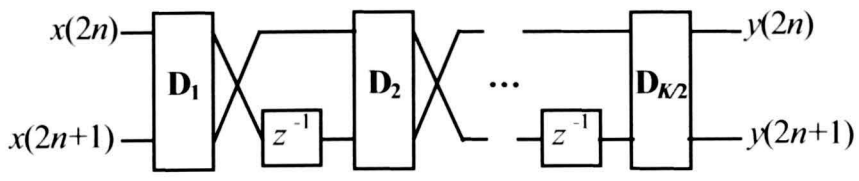


Figure 3. Scheme of a second variant of lattice structure for two-channel orthogonal filter bank calculation

Here the blocks $\mathbf{D_k}$ and $\mathbf{H_2}$ are given by the following relationships:

$$\begin{bmatrix} y_0 \\ y_1 \end{bmatrix} = \mathbf{D_k} \begin{bmatrix} x_0 \\ x_1 \end{bmatrix}, \quad \mathbf{D_k} = \begin{bmatrix} c_k & s_k \\ s_k & -c_k \end{bmatrix}, \quad \mathbf{H_2} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}.$$

Taking into consideration the above discussion it's justified to undertake the research of the computational and accuracy effectiveness of the lattice structure implementation of two-channel orthogonal filter banks.

II. FORWARD AND INVERSE LATTICE STRUCTURES

The forward lattice structure shown in Fig. 4 consists of $K/2$ stages, where K is a natural even number. Each of those stages consists of $N/2$ base operations $\mathbf{D_k}$, where N is a natural even number which determines the structure length [11]. In the first stage the input of each $\mathbf{D_k}$ operation is made of a pair of the input signal samples x_{2i} and x_{2i+1}, $i = 0,1,..., N/2-1$. Operations in all of the successive stages are the structural copies of their preceding stages. Each pair of the consecutive stages are connected with each other by a cyclic N - point downward shift. The outputs of the last stage constitute the output signal y_i, $i = 0,1,..., N-1$ of the considered structure. Fig. 3 shows the example of the forward lattice structure for $K = 6$ and $N = 8$. Here the auxiliary variables t_1 and t_2 are introduced in order to maintain the structural consistency of the structure's downward shift.

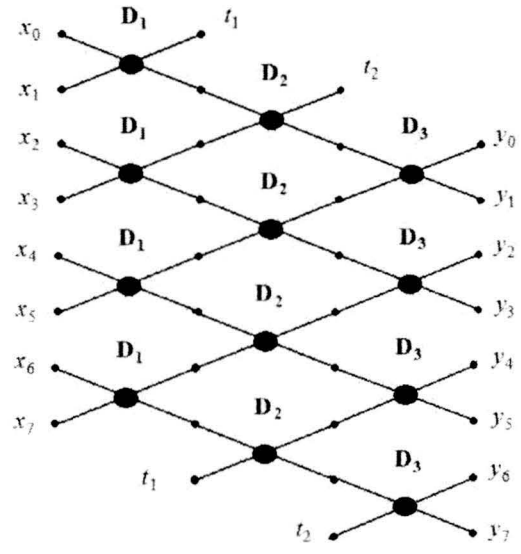


Figure 4. Data flow diagram of the lattice structure for $K = 6$ and $N = 8$.

A forward lattice structure consists of two-point operations denoted by $\mathbf{D_k}$

$$\begin{bmatrix} y_0 \\ y_1 \end{bmatrix} = \mathbf{D_k} \begin{bmatrix} x_0 \\ x_1 \end{bmatrix}, \quad \mathbf{D_k} = \begin{bmatrix} a_k & b_k \\ c_k & d_k \end{bmatrix}, \tag{1}$$

where k denotes the stage's index.

If $\det(\mathbf{D_k}) = a_k d_k - b_k c_k \neq 0$, then for each base operation $\mathbf{D_k}$ there exists its inverse counterpart $\mathbf{D_k}^{-1}$,

$$\mathbf{D_k}^{-1} = \frac{1}{\det(\mathbf{D_k})} \begin{bmatrix} d_k & -c_k \\ -b_k & a_k \end{bmatrix}, \tag{2}$$

for which the following equality holds $\mathbf{D_k}\mathbf{D_k}^{-1} = \mathbf{D_k}^{-1}\mathbf{D_k} = \mathbf{I}$, where $\mathbf{I}$ is 2 x 2 point identity matrix. From Fig. 4 and 5 one can conclude that the transformation realized by the considered lattice structure can be described by the following expressions:

$$y_{2i} = \sum_{p=0}^{K-1} h_{K,K-1-p} x_{2i+p}, \quad y_{2i+1} = \sum_{p=0}^{K-1} g_{K,K-1-p} x_{2i+p}, \quad i = 0,1,..., N/2-1 \tag{3}$$

where $h_{K,p}$ and $g_{K,p}$, $p = 0,1,..., K-1$ are filters' H_K and G_K impulse responses respectively, i.e. two convolutions with a decimation are being calculated. If each of the base operations $\mathbf{D_k}$ is orthogonal, i.e. it's inverse operation's matrix is equal to the considered forward operation's matrix transpose: $\mathbf{D_k}^{-1} = \mathbf{D_k}^{\mathrm{T}}$, then the resulting structure is the orthogonal lattice structure. On the other hand, if one would use non-orthogonal base operations then the resulting structure would become biorthogonal lattice structure.

The simplest way to construct an inverse lattice structure can be realized by reversing the forward structure. This can be achieved by replacing all of the forward base operations by their respective inverse counterparts, and performing calculations in accordance to the forward structure, but in a reverse direction.

III. FACTORIZATION OF THE TWO-CHANNEL ORTHOGONAL FILTER BANK

Let's consider two-channel filter bank $H_K = \{h_{K,k}\}$ and $G_K = \{g_{K,k}\}$, which fulfills the orthogonality condition (perfect reconstruction condition for two-channel orthogonal filter bank),

$$g_{K,p} = (-1)^p h_{K,K-1-p}, \quad p = 0,1,..., K-1, \tag{4}$$

$$\left\langle h_{K,p} h_{K,p+2m} \right\rangle = \sum_{p=0}^{K-1} h_{K,p} h_{K,p+2m} = \delta(m), \quad m = 0,1,..., K/2-1 \tag{5}$$

where $\delta(0) = 1$ and $\delta(m) = 0$ for $m \neq 0$ [1,2]. It's known [12], that such a filter pair can be factorized into a forward lattice structure shown in Fig. 4 and 5. In order to that one may use symmetrical $\mathbf{C_k}$ and asymmetrical $\mathbf{S_k}$ base operations. In the first case forward and inverse base operations are identical

$$\mathbf{C_k} = \mathbf{C_k}^{-1} = \begin{bmatrix} a_k & b_k \\ b_k & -a_k \end{bmatrix}, \quad \det(\mathbf{C_k}) = -a_k^2 - b_k^2 = -1, \tag{6}$$

In the second case forward base operation $\mathbf{S}_k$ and the inverse base operation $\mathbf{S}_k^{-1}$ have the following form

$$\mathbf{S}_k^{-1}=\begin{bmatrix} a_k & -b_k \\ b_k & a_k \end{bmatrix},\quad \mathbf{S}_k=\begin{bmatrix} a_k & b_k \\ -b_k & a_k \end{bmatrix},$$
$$\det(\mathbf{S}_k)=\det(\mathbf{S}_k^{-1})=a_k^2+b_k^2=1. \tag{7}$$

For the purposes of the filter bank factorization each of the mentioned orthogonal base operations, either (6) or (7), can be chosen. However, taking into consideration implementational issues of the analysis and synthesis stages of the considered filters, it's worth to choose the symmetrical variant (6). In such case directly from the form of the forward lattice structure the following equation results, which gives the recursive dependence between impulse responses for the filter pairs H_M, G_M and H_{M-2}, G_{M-2}

$$\begin{bmatrix} h_{M,M-1} & h_{M,M-2} & \cdots & h_{M,1} & h_{M,0} \\ g_{M,M-1} & g_{M,M-2} & \cdots & g_{M,1} & h_{M,0} \end{bmatrix}= \tag{8}$$
$$\mathbf{C}_{\mathbf{M/2}}\begin{bmatrix} g_{M-2,M-3} & g_{M-2,M-4} & g_{M-2,M-5} & \cdots & g_{M-2,0} & 0 & 0 \\ 0 & 0 & h_{M-2,M-3} & \cdots & h_{M-2,2} & h_{M-2,1} & h_{M-2,0} \end{bmatrix},$$

where $M=K,K-2,\ldots,4$.

Now the solution to the considered factorization problem can be formulated in the following way: for given orthogonal filters H_M and G_M one has to find the elements of the base operation $\mathbf{C}_{\mathbf{M/2}}$, and also the elements of the filters H_{M-2} and G_{M-2}, so that they fulfill equation (8). In order to do that let's multiply the equation (8) on the right hand side by $\mathbf{C}_{\mathbf{M/2}}^{-1}$

$$\begin{bmatrix} g_{M-2,M-3} & g_{M-2,M-4} & g_{M-2,M-5} & \cdots & g_{M-2,0} & 0 & 0 \\ 0 & 0 & h_{M-2,M-3} & \cdots & h_{M-2,2} & h_{M-2,1} & h_{M-2,0} \end{bmatrix}=$$
$$\begin{bmatrix} a_{M/2} & b_{M/2} \\ b_{M/2} & -a_{M/2} \end{bmatrix}\begin{bmatrix} h_{M,M-1} & h_{M,M-2} & \cdots & h_{M,1} & h_{M,0} \\ g_{M,M-1} & g_{M,M-2} & \cdots & g_{M,1} & h_{M,0} \end{bmatrix}. \tag{9}$$

After making proper multiplications on the right hand side of equation (9) one receives the following system of equations

$$a_{M/2}h_{M,1}+b_{M/2}g_{M,1}=0,\quad a_{M/2}h_{M,0}+b_{M/2}g_{M,0}=0, \tag{10}$$
$$b_{M/2}h_{M,M-1}-a_{M/2}g_{M,M-1}=0,\quad b_{M/2}h_{M,M-2}-a_{M/2}g_{M,M-2}=0. \tag{11}$$

In order to reduce the number of equations in the system (10)-(11) we will use the orthogonality condition (perfect reconstruction condition) (4), (5). Then from the condition (4) one receives the following set of equations: $g_{M,0}=h_{M,M-1}$, $g_{M,1}=-h_{M,M-2}$, $g_{M,M-2}=h_{M,1}$ and $g_{M,M-1}=-h_{M,0}$. Plugging these equations into (10)-(11) leads to (12)-(13)

$$a_{M/2}h_{M,1}-b_{M/2}h_{M,M-2}=0,\quad a_{M/2}h_{M,0}+b_{M/2}h_{M,M-1}=0, \tag{12}$$
$$b_{M/2}h_{M,M-1}+a_{M/2}h_{M,0}=0,\quad b_{M/2}h_{M,M-2}-a_{M/2}h_{M,1}=0. \tag{13}$$

It can be seen that the pairs of equalities (12) and (13) repeat. Therefore we can resign of one of the chosen pair e.g. (13). Continuing further, from the pair of equalities (12) we obtain the following dependencies

$$\frac{a_{M/2}}{b_{M/2}}=\frac{h_{M,M-2}}{h_{M,1}}=-\frac{h_{M,M-1}}{h_{M,0}}$$

which show, that the only possible way of satisfying each of the equations in (12) is to make the following equalities hold $h_{M,M-2}h_{M,0}=-h_{M,M-1}h_{M,1}$. The last of these equalities holds automatically for orthogonal filters, because it is the component of the previously given condition (5). It's easy to verify this last statement, since for $K=M$ and $m=M/2-1$ from (5) one obtains the following equality $h_{M,M-2}h_{M,0}+h_{M,M-1}h_{M,1}=0$. Therefore one of the equations in (12) can be eliminated. If we eliminate the second equation, then the first equation along with the dependency between the elements of the base operation given in (6) leads us to the final system of two simultaneous equations with two unknowns

$$a_{M/2}h_{M,1}-b_{M/2}h_{M,M-2}=0,\quad a_{M/2}^2+b_{M/2}^2=1. \tag{14}$$

Solving the above system we find the desired elements of the $\mathbf{C}_{\mathbf{M/2}}$ operation

$$b_{M/2}=\pm h_{M,1}/\sqrt{h_{M,1}^2+h_{M,M-2}^2},\quad a_{M/2}=b_{M/2}h_{M,M-2}/h_{M,1}. \tag{15}$$

We then use operation $\mathbf{C}_{\mathbf{M/2}}$ to reduce by two elements the lengths of the considered filters with the help of (9). Continuing recursively this process of reduction of the lengths of the considered filters for $M=K,K-2,\ldots,4$, we finally find the base operations $\mathbf{C}_k$ for each of the stages, except for the first stage of the considered factorization. The elements of the base operation of the first stage can be obtained from the equality which holds for $\mathbf{D}_1$

$$\begin{bmatrix} c_1 & d_1 & 0 & 0 \\ 0 & 0 & a_1 & b_1 \end{bmatrix}=\begin{bmatrix} g_{2,1} & g_{2,0} & 0 & 0 \\ 0 & 0 & h_{2,1} & h_{2,0} \end{bmatrix} \tag{16}$$

what completes the factorization.

IV. Effectiveness Comparison

If one performs the analysis stage directly with the use of (3), then he or she needs K multiplications and $K-1$ additions in order to calculate a single output element. If on the other hand one uses the lattice structure with base operations (6) and (7), each consisting of four multiplications and two additions, then the calculation of a single output element requires K multiplications and $K/2$ additions. This means that the lattice structure reduces almost by a factor of two the number of additions without the change of the number of multiplications with respect to the direct convolution implementation, in which, additionally, the downsampled elements are still calculated.

The number of multiplications in the lattice structure can be further reduced by introducing simplified base operations on all of it's stages, except the first one [9]. Let's suppose we do that. In order to carry this out we divide $\mathbf{C}_k$ operations by a_k

$$\mathbf{C}_k^{'} = \mathbf{C}_k / a_k = \begin{bmatrix} 1 & b_k / a_k \\ b_k / a_k & -1 \end{bmatrix} = \begin{bmatrix} 1 & b_k^{'} \\ b_k^{'} & -1 \end{bmatrix}, \; k = 2,3,..., K/2 \; .$$

Then, in the first stage we use the below base operation, which compensates the above divisions by a_k in the subsequent stages:

$$\mathbf{D}_1^{'} = r\mathbf{D}_1 = r\begin{bmatrix} a_1 & b_1 \\ c_1 & d_1 \end{bmatrix} = \begin{bmatrix} a_1^{'} & b_1^{'} \\ c_1^{'} & d_1^{'} \end{bmatrix}, \quad r = a_2 a_3 ... a_{K/2} \; .$$

This leads to the possibility of the reduction of a number of multiplications made for a calculation of a single output element to the amount of $K/2+1$, preserving at the same time the desired number of $K/2$ additions, which is equivalent to the computational complexity of a lifting scheme structure [7]. This leads to the conclusion that the lattice structure with simplified base operations allows for almost twofold reduction of the number of multiplications as well as additions in comparison to a process of calculation of all needed elements of the output signal with the use of convolution (3). The additional advantage of the lattice structure is the possibility to perform calculations in accordance with an in-place scheme, i.e. without using any additional memory resources.

V. EXPERIMENTAL RESULTS

In order to verify the effectiveness of the lattice structures three basic procedures for the realization of the synthesis step of the two-channel orthogonal filter banks were implemented. Namely: 1) DBF (two-channel filter bank), which uses formulas (3) directly; 2) L_ DBF which implements a symmetrical base operations lattice structure; 3) L_ DBF_2, which implements the lattice structure with simplified base operations. Moreover, additional auxiliary functions where implemented, intended for the realization of the mentioned synthesis filters and the calculation of the two-channel orthogonal filter bank in the form of a lattice structure. The tests were made for different types of orthogonal filters on a typical PC system. The obtained results are well represented on the example of the Daubeshies filters. Tab. 1 shows the impulse responses of the Daubeshies filters of length 6 (DB6) [2].

TABLE I. IMPULSE RESPONSES OF DB6 FILTERS

i	0	1	2	3	4	5
$h_{6,k}$	0.332671	0.806892	0.459878	-0.135011	-0.0854413	0.0352263
$g_{6,k}$	0.0352263	0.0854413	-0.135011	-0.459878	0.806892	-0.332671

The elements of the general and simplified base operations resulting form the considered factorizations with the use of expressions (9), (15) and (16) are shown in Tab. 2 and 3 respectively.

TABLE II. THE ELEMENTS OF THE GENERAL BASE OPERATIONS

k	a_k	b_k	c_k	d_k
1	0.924508	0.381162	-0.381162	0.924508
2	0.479286	0.877659	0.877659	-0.479286
3	-0.105301	0.994440	0.994440	0.105301

TABLE III. THE ELEMENTS OF THE SIMPLIFIED BASE OPERATIONS

k	a_k	b_k	c_k	d_k
1	-0.0466592	-0.019237	0.019237	-0.0466592
2	1.00000	1.83118	1.83118	-1.00000
3	1.00000	-9.44381	-9.44381	-1.00000

Tab. 4 contains the results of the performed effectiveness tests for the mentioned procedures.

TABLE IV. PROCESSOR'S CALCULATION CYCLE COUNTS FOR A CHOSEN DIMENSIONS N OF THE INPUT DATA

N	DBF	L_DBF	L_DBF_2	DBF / L_DBF	L_DBF / L_DWT_2	DBF / L_DWT_2
256	19956	10553	9935	1.89	1.06	2.01
1024	79707	41588	39659	1.92	1.05	2.01
4096	319864	167128	159641	1.91	1.05	2.00

As it can be seen the L_ DBF lattice structure has a significant advantage (almost twofold) over the trivial algorithm (3) DBF. The considered advantage is substantially greater than one might have expected, since in this case, form the theoretical point of view, reduction of the arithmetical operations is of the factor of one fourth. On the other hand the usage of the lattice structure L_ DBF_2 with simplified base operations did not match the expected result. This comes form the fact that in this case theoretical estimation indicates that L_DWT_2 procedure should have needed about one forth less cycles to execute than the L_ DBF procedure, while in fact both of these procedures have been executed in an almost equal number of processor's cycles (with a difference of about 5%). Such result is explicable, since the actual effectiveness of implementation of the considered algorithms is obviously influenced not only by the estimated number of arithmetical operations, but also by the way the auxilary implementational operations are performed, as well as by the way the data transfer is actually organized.

In addition to efficiency, also the accuracy of the calculations has been investigated. The obtained results are perfectly consistent with the general principle stating that the reduction of the calculation accuracy error comes along with

the reduction of the number of arithmetical operations needed to perform a given task. Namely, in MSE terms sequential execution of the analysis and synthesis stages with the use of the lattice structure has had a substantially smaller (over 100 times) calculation accuracy error in comparison to a method of direct computation.

VI. CONCLUSIONS

The investigation was undertaken of the effectiveness of the lattice structure, which for filters of the impulse response of length K consists of $K/2$ vector rotation base operations and $K/2$-1 unit delay operations. The considered structure makes an almost twofold reduction of the number of arithmetical operations in comparison to a direct computation method using the convolution. In case of the lattice structure with simplified base operations it also makes an almost twofold reduction of the number of arithmetical operations in comparison to a direct computation of the convolution, and in terms of computational complexity, is equivalent to a lifting scheme structure. The performed experimental studies have shown the lattice structure reduces almost by the factor of two the computational time of the filtering process performed by two-channel orthogonal filter bank, and thus can be successfully used in construction of signal processing software for sequential processing systems, including PC ones.

REFERENCES

[1] T.P. Zieliński, "From theory to digital signal processing," WKŁ, Warszawa, 2009, (in Polish).

[2] P.J. Fleet, "Discrete wavelet transformations: An elementary Approach with applications, " John Wiley & Sons, Inc., New Jersey, 2008

[3] G. Strang and T. Nguyen "Wavelets and filter banks," Wellesley-Cambridge Press, 1999.

[4] P. Lipiński, "Watermarking software in practical applications, " Bulletin of the Polish Academy of Sciences: Technical Sciences, 2011, vol.59, no 1., pp. 21-25.

[5] T. Cooklev, "An efficient architecture for orthogonal wavelet transforms, " IEEE Signal Processing Letters, 2006. vol. 13., no 2, pp. 77–79.

[6] J. T. Olkkonen and H. "Olkkonen, Discrete lattice wavelet transform, " IEEE Transactions on Circuits and Systems-II: Express briefs, 2007, vol. 54, no 1, pp. 71–75.

[7] I. Daubechies and W. Sweldens, "Factoring Wavelet Transform into Lifting Steps, ", The Journal of Fourier Analysis and Applications, 1998, vol. 4. no 3. pp. 245–267.

[8] T. C. Denk and K. K. Parhi, "VLSI architectures for lattice structure based orthonormal discrete wavelet transforms," IEEE Transactions on Circuits and Systems-II, Analog and digital signal processing, 1997, vol. 44, no 2, pp. 129–132.

[9] M. Yatsymirskyy, "A novel matrix model of a two channel bank of orthogonal filters," Methods of Applied Computer Science, 2011, no 1 (26), pp. 205-212, (in Polish).

[10] P.P. Vaidyanathan and P.Q. Hoang, "Lattice structures for optimal and robust implementation of two-channel perfect reconstruction QMF banks," IEEE Transaction on Acoustics and Signal Processing, 1988, vol. 36, no 1, pp. 81–94.

[11] M. Yatsymirskyy, "Lattice structures for synthesis and implementation of wavelet transforms," Journal of Applied Computer Science, 2009, vol. 17, no 1, pp. 133–141.

[12] M.M. Yatsymirskyy and O.R. Shaleva, "Simplified lattice factorization of two-channel orthogonal filter bank," Modeling and Information Technologies, no. 64, 2012, pp. 149-156, (in Ukrainian).

Mechatronic Setup for Measurement of Dynamical Properties of Electronic Gyroscopes

Marcin Kaczmarski
Institute of Automatic Control, Technical University of Łódź
Łódz, Poland
marcin.kaczmarski@p.lodz.pl

Paweł Pełczyński
Institute of Electronics, Technical University of Łódź
Łódz, Poland
pawel.pelczynski @p.lodz.pl

ABSTRACT — **The aim of the presented work was the design and realization of a mechatronic setup for measurements of dynamical properties of electronic gyroscopes. A motivation for the project was a need for identification of MEMS gyroscope parameters, not disclosed in their data sheets. The built setup is capable of producing controlled, angular excitations, necessary for measuring dynamical properties of MEMS gyroscopes. It registers gyroscope response synchronized with the excitation. An example of identification of gyroscope parameters is also presented. Results of the measurements are used to calibrate a currently developed local navigation system for visually impaired pedestrians.**

KEYWORDS-COMPONENT — MEMS, gyroscope, angular rate, gyro frequency response, parameter identification

I. INTRODUCTION

Accelerometers and gyroscopes are used in a number of specialized equipment and systems, for example in air navigation instruments [1]. They gained real popularity after a new technology of their realization, called Micro-Electro-Mechanical Systems (MEMS) was developed. It allowed for cheap, massive scale production and opened new areas of applications. Initially MEMS accelerometers were mainly used as the sensors for impact in triggering automotive airbags. Soon, however, accelerometers and gyros, seized market of consumer electronics. Today we can find them in everyday life equipment, such as mobile phones, cameras, gaming consoles and in medical devices, and even in sports equipment. They provide signals for optical image stabilization, proper positioning of image on the display, motion tracking or pedestrian step count [2].

Gyroscopes are generally divided into two groups: directional and angle rate gyros, but in MEMS technology often the second type is realized. They measure the angular velocity of the objects rotating around the axis. They are vibrating sensors using the so-called Coriolis effect [1]. They measure the change of vibration plane of the mass, caused by its rotation. The mass is etched in poly-silicon and placed in silicon framework in such a way, that it can move only in one direction. When the weight moves towards the outer edge, it is accelerated by Coriolis force. The resulting displacement is estimated with use of a capacitor, in which one electrode is attached to the ground, and the second is fitted on the mobile frame. As a result of the Coriolis force the elements of this structure are moving, which changes the capacity of the capacitor, which is then measured.

Modern MEMS gyros are fabricated as hybrid integrated circuits, containing complete measurement systems. An example of digital MEMS gyroscope is shown in Fig. 1. It contains the following components:

- MEMS sensor – vibrating mass with variable capacitors,
- MEMS driver – generator of mass excitation signal,
- Signal conditioning and ADC – signal amplifier, demodulator, low-pass filter and analog-to-digital converter (ADC),
- Temperature sensor – sensor used for proper device calibration and compensation of temperature drift,
- Control logic – a logic used for signal processing, module calibration and self-testing,
- Serial bus interface – an interface for data transmission, typically I2C or SPI serial interface.

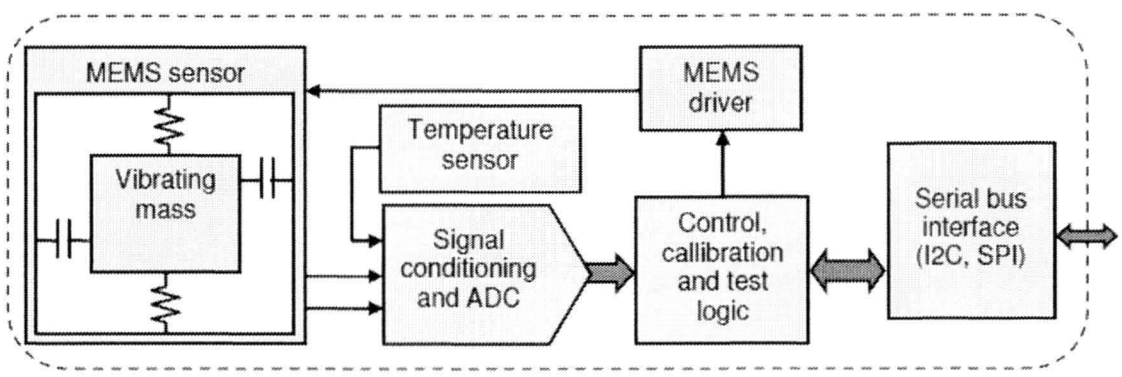


Figure 1. Block diagram of the digital MEMS gyroscope.

Integrating more MEMS sensors for two or three axes rotation rate sensing is a common case.

Due to the dynamical properties of the vibrating mass and the delay introduced by an amplifier, filter, ADC and digital logic, an integrated gyroscope features specific frequency response, both in amplitude and in phase. It is rarely presented in the device data sheet. Most often the only cutoff frequency is done. There is a need to identify gyro dynamic properties for a number of specialized applications.

The motivation for furthermore gyro identification was a project of local navigation system for the visually impaired pedestrians, carried out in the Institute of Electronics in the Technical University of Łódź [3]. The main role of a designed system is detection of obstacles and producing warning signals. Tracking egomotion parameters of a stereo camera, placed on a head of system user, helps to reliably track obstacles in the reference frame associated with it. This task is known as the local navigation in the so called body frame (b-frame) [1]. Egomotion is estimated on the basis of visual information: scene depth obtained from stereoscopy and the optical flow in consecutive images of a sequence [4],[5]. The readings of electronic gyroscopes help to predict the optical flow of image feature points reducing false estimations and lowering computational effort of image feature fitting [6]. It requires perfect synchronization of gyroscope signals and images, which is not possible without the identification of gyro dynamical properties [7].

II. DESIGN OF MEASUREMENT SETUP

For measuring the dynamic properties of gyroscopes a laboratory setup was designed and developed. Its functional diagram is shown in Fig. 2. The setup photo can be seen in Fig. 3. Buehler DC motor was used as a system for measuring static and dynamic parameters of gyroscopes. It is connected with a planetary gear reducing its speed. The speed reduction ratio is 24,7:1. The engine is equipped with an incremental encoder with a resolution of 400 pulses per revolution. In addition, the measurement is performed using quadrature counter that allows to multiply the number of pulses to 1600 and also measures the number of pulses of the encoder. The motor is controlled via the mcIO-C3 servo produced by miControl. It allows to control the DC motor supply voltage to 24V. The encoder or Hall sensor can serve as a feedback information of the position or velocity [8]. It has four additional, digital inputs, one digital output and a single analog input for a typical industrial measuring range of ± 10V. Built-in analog-to-digital converter with a resolution of 10 bits is capable of measurements with an accuracy of about 20mV at 1LSB. The communication with the mcIO-C3 module is realized using the CAN 2.0A interface. CAN OPEN protocol, implemented inside the controller, allows to use it in various control systems. The maximum data transfer rate in this interface is 1Mb/s. With a length of the identification field equal to 11 bits and the maximum transmission speed, it takes 110μs to send a frame of 8 bytes data. The implemented protocol requires handshake, which takes 3ms. This ensures maximum speed of data exchange equal to 330 frames per second. Direct readout of angular speed using a transmission confirmation leads to an error, whose value can be determined from the equation (1):

$$error_{velocity\_reading} = \frac{motor\_velocity\left[\frac{rot}{\min}\right] \cdot 360\left[\frac{\deg}{rot}\right]}{k \cdot 60 \cdot 10^3 \left[\frac{ms}{\min}\right]} \quad (1)$$

where: k - motor reduction gear=24,7.

For example, for a given angular speed of the engine equal to $1500\left[\frac{rot}{\min}\right]$ the error equals to $0{,}364\left[\frac{\deg}{ms}\right]$.

This necessitated the development of independent, parallel system, coupled with the encoder engine, whose role was a measurement of both the velocity of the converter mounted on the motor shaft, as well as the angular velocity ω of a gyroscope under test at regular intervals [9]. It is built on the basis of Atmel's microcontroller - AT90CAN128. It has the ability to work with low voltage in the range from 2.7 to 5.5 V and built-in I2C interface, so it could be adapted to communicate with the gyroscope.

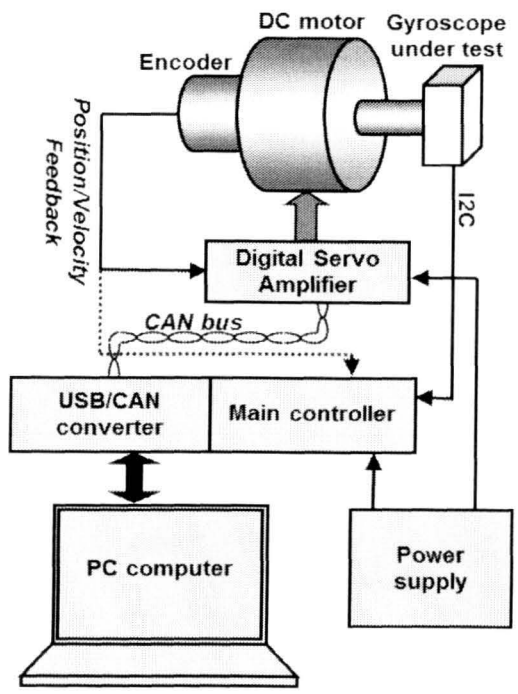


Figure 2. Schematic diagram of measurement setup.

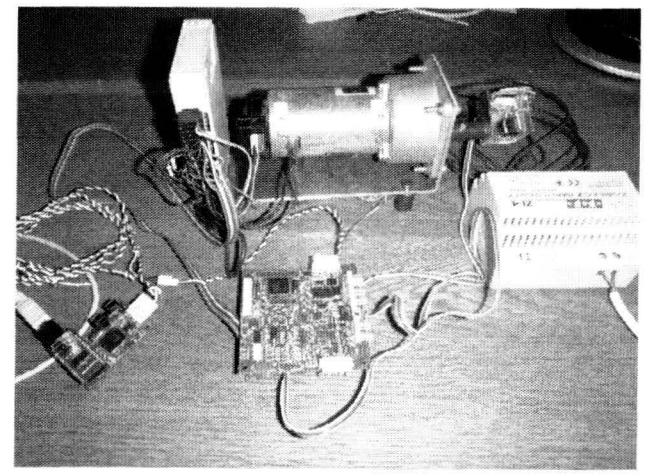


Figure 3. View of the measurement setup.

A specialized 32-bit quadrature counter LS7366 was used for the motor encoder pulse counting. It communicates with the microcontroller via SPI interface.

Communication with the master PC was made by FTDI FT245RL system. This is the USB to parallel port converter. Virtual Com Port (VCP) drivers have been applied on the PC side, which allowed to obtain transmission speed over 300kB/s, and simplified access to incoming data.

III. MEASUREMENTS REALIZED BY THE SETUP BUILT

The developed measurement setup allows to perform the following experiments investigating the static and dynamic properties of gyroscopes:

- the answer to the linear increase of angular rate,
- angular rate step response,
- amplitude and phase frequency response of the gyroscope.

During the estimation of the angular rate measurement error, the gyro is excited in the form of a slow linear rise of angular velocity in the range 0-650[°/s]. During the test the angular velocity of the motor shaft and the rotational speed of each gyroscope axis is recorded. The collected data are used to determine gyro velocity error as a function of engine speed. The step response registration helps to determine the dynamics, the nature of the gyroscope response during rapid changes of the measured velocity. The gyro step response shape permits choice of an appropriate model of the investigated system. It could be proportional model, inertial one with delay or oscillative. The excitation starts from zero shaft speed and reaches its maximum value equal to: 500, 1000, 1500, 2000 or 2500 [rot/min]. The experiment includes both step from zero to a maximum value of angle rate, and the return to the initial, zero state.

More detailed information on the chosen model can be obtained by examining the amplitude and phase frequency response. In the constructed setup the excitation was a sinusoidal signal with an amplitude of the engine shaft velocity equal to: 500, 1000, 1500, 2000, 2500, 3000, 4000 [rot/min] and the frequency range was 0.1Hz to 10Hz for each of the specified amplitude.

IV. MEASUREMENT OF MG1101 GYROSCOPE PARAMETERS

The developed test bench served to investigate the properties of the MG1101B gyroscope produced by GYRATION [10]. Gyroscopes of this type are characterized by small size and low power consumption (Fig. 4.). In addition, they ensure high measurement dynamic range of ± 500°/sec. They integrate analog-to-digital converters for two-axis sensing. They are extremely lightweight and communicate with external system via 2-wire serial interface (I2C). Low drift, high temperature stability and internal shock mounting make them high accuracy angular measurement devices.

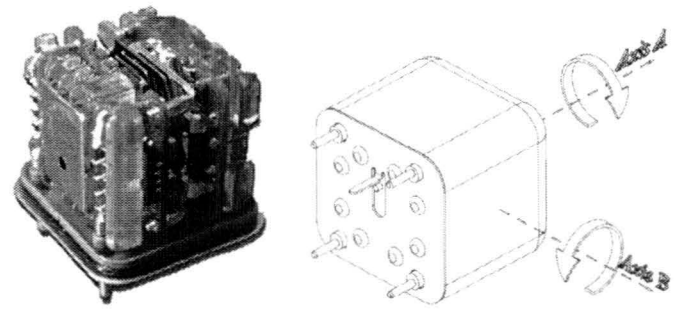
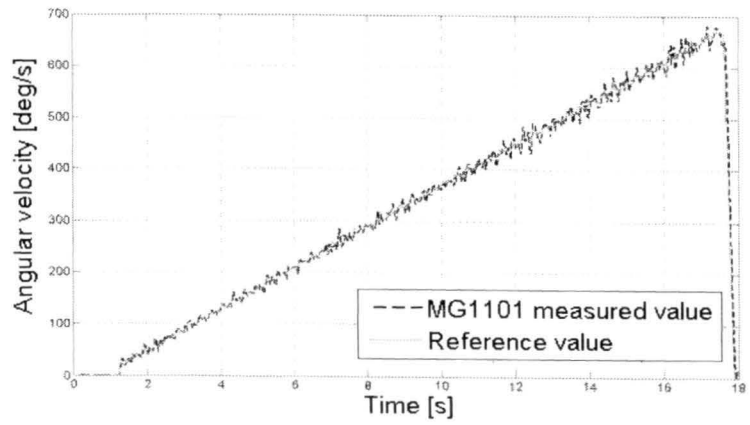


Figure 4. Internal view and rotation axes of MG1101 gyroscope (reprint from MG1101 datasheet).

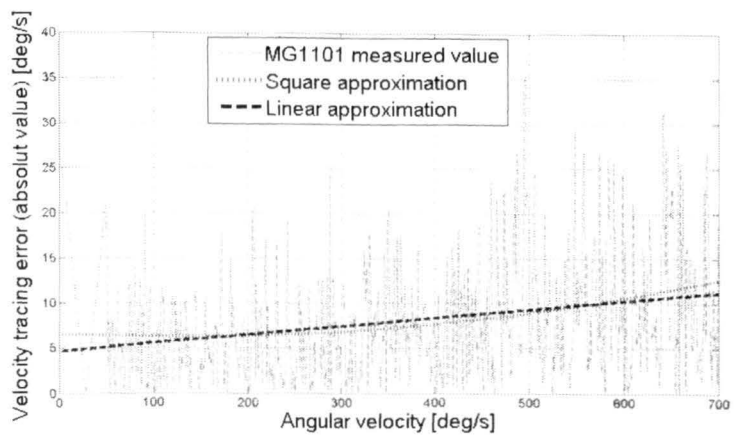
A. *Estimation of the gyroscope response to the linear increase of angular rate*

The first stage of research was to determine the gyro response to a linear increase of angular rate. For this purpose the gyroscope has been mounted to the DC motor shaft, so that its axis of rotation was parallel to the axis of rotation of the motor. Linear increase of rotational speed was obtained by a series of speed setting commands, issued at intervals of 0.01s. The obtained measurement result is shown in Fig. 5. In the first plot a comparison of the set and measured velocity is shown. The second plot shows magnitude error of the measurement.

The gyroscope features good linearity of the output value. For demanding applications, like the Kalman filtering in fusion of different data for navigation purposes [11],



a)



b)

Figure 5. MG1101 gyroscope answer to linear velocity increase a), and its measurement error magnitude b).

the dependence of measurement error on the angular velocity can be approximated by linear or square function. For linear approximation the following dependence was found:

$$error_{velocity}\left[\frac{^\circ}{s}\right] = 0.0094 \cdot velocity\left[\frac{^\circ}{s}\right] + 4.6338 \quad (2)$$

In the case of more precise, square approximation the following equation was found:

$$error_{velocity}\left[\frac{^\circ}{s}\right] = 1.9535 \cdot 10^{-9} \cdot velocity^2 + \\ -0.0051 \cdot velocity + 6.604 \quad (3)$$

B. Estimation of gyroscope step response

In order to determine the gyro response to a step change of angular velocity an experiment was conducted in which the motor controller generated a rapid change of its angular rate from 0 up to the maximal value. After two seconds the engine was stopped, which returned the value to 0. A unit step excitation allows to identify the system transmittance model and to estimate its parameters. The tests were conducted for velocity values equal to 121, 243, 364, 500, 607 [°/s]. An example of sample response waveform is shown in Fig. 6. The observation of a shape of MG1101 gyroscope response to the signal change from zero to the maximal value did not give a definite answer about the type of the system. The oscillation was observed, but it could be the effect of possible oscillations of the excitation. Due to this uncertainty a response to more reliable, rapid break of motor rotor was analyzed. This excitation can be treated as almost ideal velocity step. The step response posed visible oscillations, that quickly disappeared. They reached almost 4% of the overall step. Normalized system step response of angular velocity is shown in Fig. 7. In all the tests a slight delay in the gyro response signal could be seen. The measured delay is constant and does not exceed the value of 50ms. It may be related to the time needed to convert the gyroscope response to the digital signal. The maximum sampling rate of the gyroscope is 30Hz, which causes a minimum delay on data output about 33ms. An additional delay could be introduced by some internal filtering of the signal. The obtained results are with accordance with technical datasheet of the device [10].

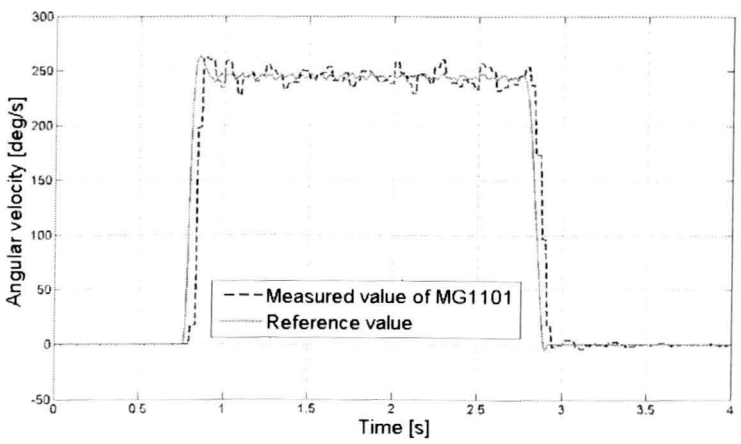


Figure 6. Response to a step change of velocity 250°/s.

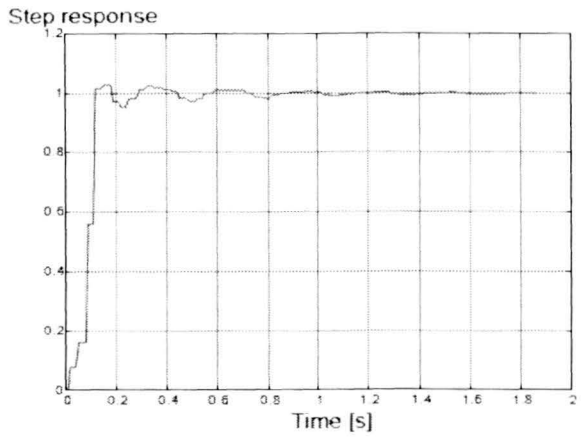


Figure 7. The normalized gyroscope response to a step change of angular velocity.

C. Estimation of amplitude and phase frequency response of the system

More detailed information about the parameters of investigated object is obtained on the basis of the frequency characteristics. In determining the frequency amplitude response, the engine was subjected to sinusoidal voltage changes. The frequency range was from 0.1Hz to 10Hz. The amplitude and phase delay of the gyro answer were observed. The series of these measurements (Fig. 8.) delivered data to estimate frequency response of the gyroscope. Amplitude and phase responses for different excitation amplitudes are shown in Fig. 9. When analyzing these characteristics it can be seen, that in the range up to approximately 2 Hz,

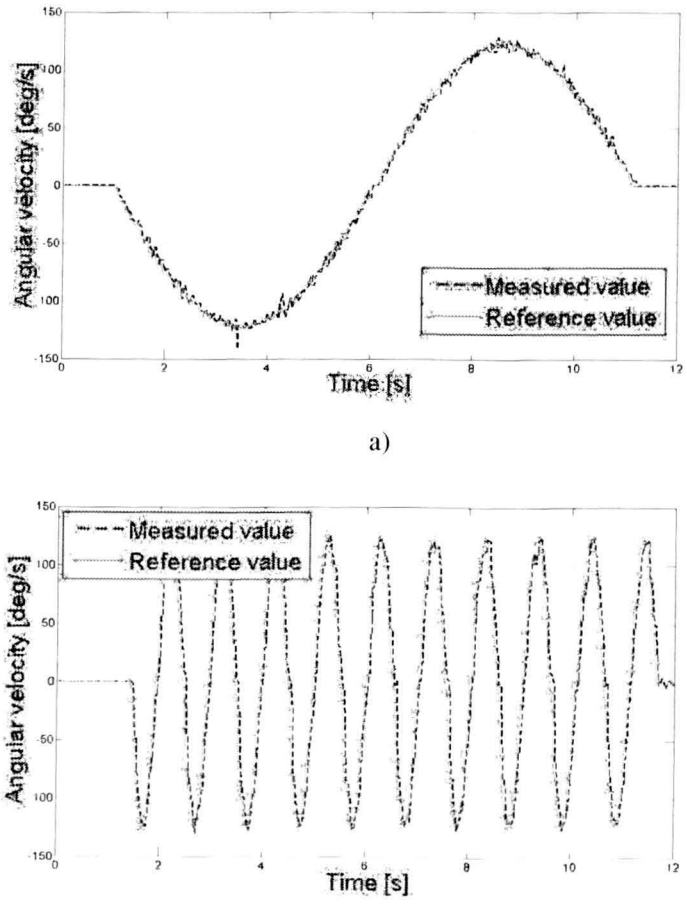
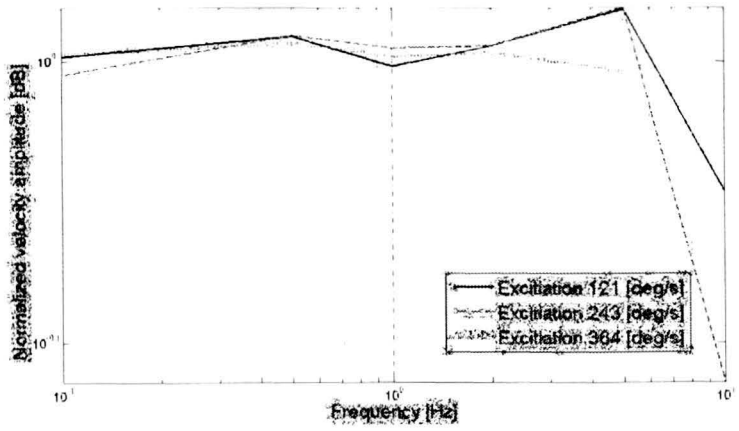


Figure 8. Examples MG1101 gyroscope answer to sinusoidal excitation a) 0.1Hz, b) 1Hz.

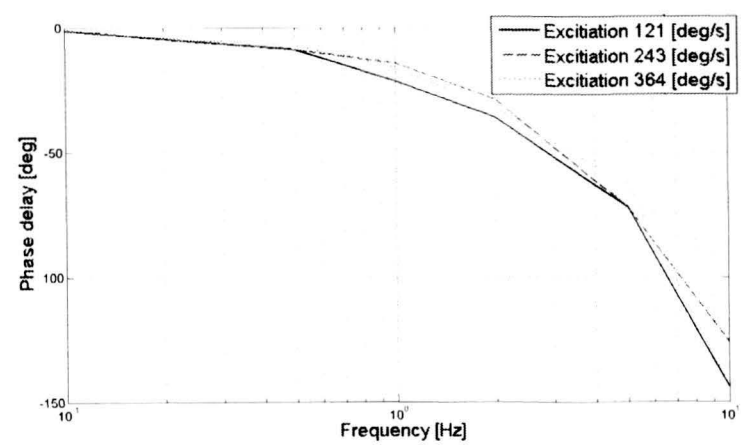
the ratio of the output signal amplitude to the input signal is at the level of unity. In this frequency range the system reliably measures the angular velocity. At a frequency of 5 Hz a small raise of the output can be observed. It is typical for the oscillatory system, which confirms the observations of the step response. The resonant frequency is about 5Hz. Estimated phase responses show the presence of delay, which is related to signal processing by internal gyroscope circuits. This introduces a rapidly increasing phase. For the frequency equal to 10Hz, the measured signal was practically inverted with respect to the reference signal.

D. *The analysis of the gyroscope characteristics and parameters*

The investigation of static and dynamic properties of MG1101 gyroscope shown that it has a wide velocity measurement range up to ±700[°/s], although with the increase of the measured value the error increases, as shown in Fig. 5.b). The measured values fluctuate around the true value, but the mean error is close to zero. The examination of the dynamic properties revealed the occurrence of oscillations at the transient states. A thorough analysis of its frequency response confirms, that the studied gyroscope can be modeled using an oscillating block.



a)



b)

Figure 9. Magnitude a), and phase b) frequency response obtained for different amplitudes of excitation.

An ARX model was identified for the gyroscope under test using the least sum of squares. The identification was performed using MATLAB computing environment with the MATLAB System Identification Toolbox. The dependence describing the behavior of the studied system is given by equation (4):

$$A(q)\cdot y(t) = B(q)\cdot u(t) + e(t) \tag{4}$$

where the identified parameters take the values:

$$A(q) = 1 - \frac{0.3429}{q} + \frac{0.1414}{q^2} - \frac{0.3566}{q^3} + \frac{0.2081}{q^4} \tag{5}$$

$$B(q) = -\frac{0.1042}{q} + \frac{0.5194}{q^2} - \frac{0.673}{q^3} + \frac{0.9058}{q^4} \tag{6}$$

To verify the identified model it was compared to the step response and frequency response of the gyro. The results are illustrated in the Fig. 10. and Fig. 11. Good representation of the device characteristics confirms the correctness of the applied mathematical description.

V. CONCLUSIONS

The design and realization of a mechatronic setup for measurements of dynamical properties of electronic gyroscopes was presented. It allows for identification of MEMS gyroscope parameters, not disclosed in their data sheets. The built setup is capable of producing controlled, angular excitations, and measuring step response, linear excitation response and frequency response. An example of an identification of MG1101 gyroscope parameters was presented and discussed. The identification of this gyro was utilized in the project of local navigation system for visually impaired pedestrians, [12-14].

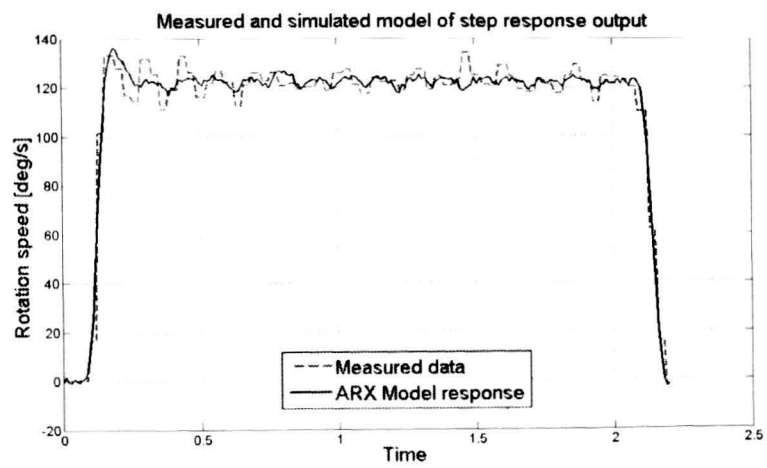


Figure 10. A comparison of real and modelled step response of measured gyroscope.

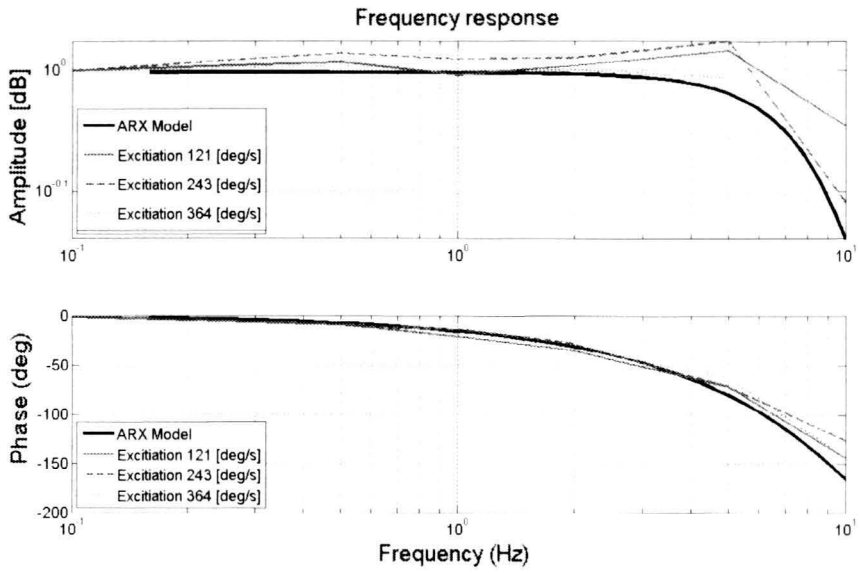


Figure 11. A comparison of real and modelled magnitude and pahse frequency response of measured gyroscope.

ACKNOWLEDGMENT

This work has been supported in part by the Ministry of Science and Higher Education of Poland research grant no. NR02-0083-10 in years 2010-2013.

REFERENCES

[1] D. H. Titterton, J.L. Weston, "Strapdown Inertial Navigation Technology," Progress in Astronautics and Aeronautics, MIT Lincoln Laboratory, 2nd. Edition, 2004.

[2] M. Veth, J. Raquet, "Fusion of low-cost imaging and inertial Sensors for navigation," NAVIGATION - Journal of the Institute of Navigation, Vol. 54, No. 1. 2007.

[3] M. Bujacz, P. Skulimowski, P. Strumiłło, "Sonification of 3D scenes using personalized spatial audio to aid visually impaired persons," International Conference on Auditory Display, Budapest, Hungary, 2011.

[4] A. R. Bruss, K. P. Horn, "Passive navigation," Computer Vision, Graphics, and Image Processing 2,11983, pp. 3-20.

[5] P. Skulimowski, P. Strumillo, "Refinement of depth from stereo camera ego-motion parameters," Electronics Letters, Vol. 44, Issue 12, June 2008, pp. 729 – 730.

[6] L. Armesto, J. Tornero, M. Vincze. Fast "Ego-motion estimation with multi-rate fusion of inertial and vision," The International Journal of Robotics Research, Vol. 26, No. 6, 2007, pp. 577-589.

[7] E. M. Foxlin, M. Harrington, Y. Altshuler, "Miniature six-DOF inertial system for tracking HMDs," Proc. SPIE Vol. 3362, 1998, pp. 214-228.

[8] M. Kaczmarski, G. Granosik, "Rehabilitation robot RRH1," Archive of Mechanical Engineering, 2011, Vol.58 nr 1 pp.103-113.

[9] M. Pytasz, G. Granosik, "Modelling and control of wheeler the hyper mobile robot," Proceedings of the 13th IEEE/IFAC International Conference on Methods and Models in Automation and Robotics, Szczecin, Poland 2007.

[10] Gyration, Inc., "MG1101 technical datasheet," www.gyration.com, 2005.

[11] E. S. Jones, S. Soatto, "Visual-inertial navigation, mapping and localization: a scalable real-time causal approach," International Journal of Robotics Research, Vol. 30, No. 4, 2011, pp. 407-430.

[12] P. Baranski, M. Polanczyk, P. Strumillo "Fusion of data from inertial sensors, raster maps and GPS for estimation of pedestrian geographic location in urban terrain," Metrology & Measurement Systems, Vol. XVIII, No 1., 2011, pp. 145-158.

[13] B. Ostrowski, P. Pełczynski, R. Danych, „A Hardware interface of visual-acoustic travel aid system for the blind," Przegląd Elektrotechniczny, 2011, pp. 130-132. In polish: „Sprzetowy interfejs wizyjno-akustycznego systemu wspomagajacego niewiadomego w samodzielnym poruszaniu się".

[14] P. Pełczyński, B. Ostrowski, D. Rzeszotarski, "Motion vector estimation of a stereovision camera with interial sensors," Metrology and Measurement Systems, 2012, year XIX, nr 1, pp. 141-150.

Authors' Index

NTAV/SPA 2012

AUTHORS' INDEX

NTAV/SPA 2012 Scientific Committee Chairmen:

prof. Andrzej Materka
Chairman

Prof. Materka is the director of the Institute of Electronics at the Lodz University of Technology.

He was the Dean of the Faculty of Electrical and Electronic Engineering, for two terms from 2002 till 2008.

He has published more than 200 technical articles and 6 monographs.

His research interests include analog circuit testing, semiconductor device modeling, medical electronics, digital signal and image processing, artificial neural networks, database systems, design of brain-computer interfaces and blind support systems.

Prof. Materka is a Senior Member of the Institute of Electrical and Electronics Engineers (IEEE).

prof. Adam Dąbrowski
Vice-Chairman

Prof. Dąbrowski is a full professor in digital signal processing at the Department of Computing, Poznan University of Technology, Poland.

His scientific interests concentrate on: digital signal and image processing (filtering, signal separation, multirate and multidimensional systems, wavelet transformation), multimedia, biometrics, visual systems, and processor architectures.

He is the author or a co-author of 4 books and over 300 scientific papers.

Currently he is the Chairman of the Signal Processing (SP) and Circuits & Systems (CAS) Chapters of the Poland IEEE Section.

Organizing Committee Chairmen:

prof. Paweł Strumiłło
Chairman

Prof. Strumiłło is the head of the Medical Electronics Division at the Institute of Electronics, Technical University of Lodz (TUL), Poland.

His research interests include processing and analysis of biomedical signals and images, applications of artificial neural networks, and electronic systems aiding the disabled.

He has published over 100 scientific papers, coauthored several teaching scripts and recently a book on electronic navigation systems for the visually impaired.

Prof. Strumiłło is a Senior Member of the IEEE Computational Intelligence Society.

dr. Michał Bujacz
Vice-chairman

Dr. Bujacz is an Assistant Professor at the Institute of Electronics, Technical University of Lodz (TUL), Poland.

His research interests include sonification, psychoacoustics, audio synthesis and processing, spatial audio, HRTFs, virtual reality and teaching. He is working on the development of an electronic travel aid for the blind and establishing a new Audio Engineering laboratory at TUL.

He obtained his Ph.D. in 2011 and has published over 20 research papers.

Dr. Bujacz was one of the founders and the first vice-president of the IEEE Student Branch at the Technical University of Lodz.

Proceedings of the Joint Conference NTAV/SPA 2012
New Trends in Audio and Video Signal Processing: Algorithms, Architectures, Arrangements and Applications
27–29th September 2012 Łódź, Poland

NTAV-SPA 2012 Conf ID 21037

ISBN: 978-83-7283-502-4

Lodz University of Technology Press
Printed by QuickDruk

ISBN: 978-83-7283-502-4
9 788372 835024

SPONSORS:

Ministry of Science and Higher Education
Republic of Poland

NATIONAL INSTRUMENTS™

Foundation for the Development of Radiocommunications and Multimedia Technologies

TOMTOM