PROCEEDINGS OF THE
MULTIDISCIPLINARY DOCTORAL SCHOOL
2010-2011 ACADEMIC YEAR
FACULTY OF INFORMATION TECHNOLOGY
PÁZMÁNY PÉTER CATHOLIC UNIVERSITY
BUDAPEST
2011

Faculty of Information Technology
Pázmány Péter Catholic University

Ph.D. PROCEEDINGS

PROCEEDINGS OF THE MULTIDISCIPLINARY DOCTORAL SCHOOL 2010-2011 ACADEMIC YEAR FACULTY OF INFORMATION TECHNOLOGY PÁZMÁNY PÉTER CATHOLIC UNIVERSITY BUDAPEST

July, 2011

Pázmány University ePress
Budapest, 2011

© PPKE Információs Technológiai Kar, 2011

Kiadja a Pázmány Egyetem eKiadó
2011
Budapest

Felelős kiadó
Dr. Fodor György
a Pázmány Péter Katolikus Egyetem rektora

Cover image by Bálint Péter Kerekes, A passive probe array and the electronic depth control probe array implanted to the rat sensory cortex (S1 Trunk region). This arrangement was used in the thalamocortical interactions experiment.

A borítón Kerekes Bálint Péter ábrája látható: Egy passzív szonda hálózat és egy mély elektronikus vezérlő szonda hálózat beültetve egy patkány szenzoros agyterületébe (S1 főrégió) Ez az elrendezés volt használva egy thalamocorticalis interakció kísérletben.

HU ISSN 1788-9197

Contents

Introduction

It is our pleasure to publish this annual proceedings again to demonstrate the genuine multidisciplinary research done at our Jedlik Laboratories by the many talents working in our Interdisciplinary Doctoral School.. Thanks are also due to the supervisors and consultants, as well as to the five collaborating National Research Laboratories of the Hungarian Academy of Sciences and the Semmelweis Medical School. The collaborative work with the partner Universities, especially, Katolieke Universiteit Leuven, Politechnico di Torino, Technische Universitat in München, University of California at Berkeley, University of Notre Dame, Univetsidad Sevilla, Universita di Catania is gratefully acknowledged..

As an important development of this special collaboration, we were able to jointly complete the second year with the Semmelweis Medical School a new undergraduate curriculum on Molecular Bionics, the first of this kind in Europe.

We acknowledge the many sponsors of the research reported here. Namely,

- the Hungarian National Research Fund (OTKA),
- the Hungarian Academy of Sciences,
- the National Development Agency (NIH),
- the Gedeon Richter Co.,
- the Office of Naval Research (ONR) of the US,
- IBM Hungary,
- NVIDIA Ltd.,
- Eutecus Inc., Berkeley, CA,
- Morphologic Ltd., Budapest,
- Analogic Computers Ltd., Budapest,
- AnaFocus Ltd., Seville, and

some other companies and individuals.

Needless to say, the resources and support of the Pázmány University is gratefully acknowledged.

Budapest, July 2011.

TAMÁS ROSKA
Head of the Jedlik Laboratory

PÉTER SZOLGAY
Head of the Doctoral School

Towards combining cortical electrophysiology and fMR measurements

Bálint Péter Kerekes
(Supervisor: phd. md. István Ulbert)
bkerekes@cogpsyphy.hu

Abstract- **The goal of our research is to make a system, which can be used in fMR-EEG experiments. First of all we need to know very well the functionality, and the behavior of a region of the brain to have a good principle in the planned experiments. We used multisilicon probes, and microelectrode arrays with electronic depth control for these tests. Meantime we planned a data transmission system, which meets the fMR conditions.**

Index Terms- **electronic depth control; silicon microprobes; fMR-EEG; data transmission**

I. INTRODUCTION

In this study we are in the designing phase of an EEG system which will be compatible with fMR. There are many problems in this issue. We need to make a system which can work in a 3T magnetic field without any artefacts, or with artifacts what can be eliminated later in the data processing phase. I have looked through the existing systems in the scientific literature [1-8], but every one of these systems are struggling with some of these problems: the great magnetic field which can indicate unwanted currents in the system, the vibrations, the amplification and the digitalization needs to be near the electrodes, the power supplies of these systems, the synchronization of the fMR and EEG systems, and so one. Not to mention the speed of the imaging technique in fMR [9]. We are planning experiments in deep sleeping humans [10-12], especially concentrating on the somatosensory cortex[13], but to know more of the oscillations, and the functions of this region we made animal tests in anesthetized rats to get familiar with the subject[14-15]. We investigated the cortico-cortical thalamo-cortical interactions, and made laminar analysis of the somatosensory cortex in slow wave sleep [16-21].

II. MATERIALS AND METHODS

A. Probes

- *Silicon probe* (Fig. 1.) [15]: The length of the silicon probe is 12mm, with a 7mm long part that can be inserted in the tissue, 280 µm wide, and 80 µm thickness. 24 square shaped and 100 µm spaced platinum recording sites were exposed at the end of the shaft. Bonding pads were designed at the other end of the device in the form of 200 µm x 200 µm SiO_2-Pt micro grids. The recording sites were electronically connected to the bonding pads via 4 µm wide and 300 nm thick conductive paths made out of Pt. The probe and a 26 pole Preci-dip connector (used for packaging) were connected through a printed circuit board [15].

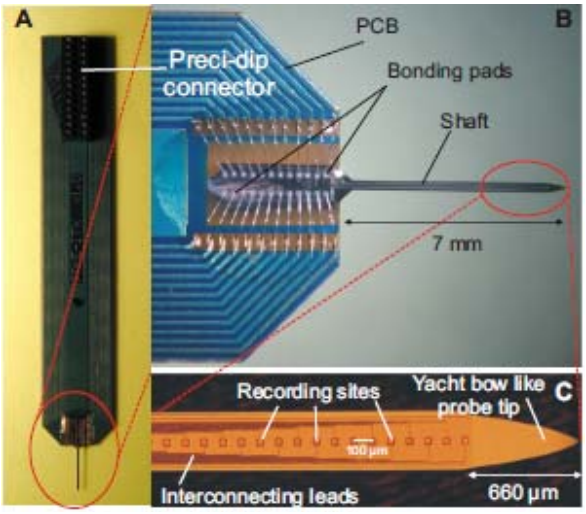


Figure1. Outlook of the silicon probe [15]. A- The PCB with the 26 pole connector. B- The silicon chip with the Al bonding. C- The tip and the PT contacts of the shaft.

- *Microelectrode array* (Fig. 2.) [14]: A novel two dimensional silicon-based electrode array was developed in the framework of the NeuroProbes EU project. The electrode array is equipped with electronic depth control (EDC) system in order to select up to 32 active recording sites from the 2052 electrode sites without moving the array, the sites are separated by 40 µm in two rows. The array consists of four shanks (8mm long each) in a comb-like structure. The electrodes can be electronically switched to the eight output lines in 2x2 groups like in a tetrode configuration, any combination of two tetrodes can be selected on each shank. The complete system consists of the electrode array, switching matrix, front-end electronics, conditioning, multiplexing and interface electronics and the control software.

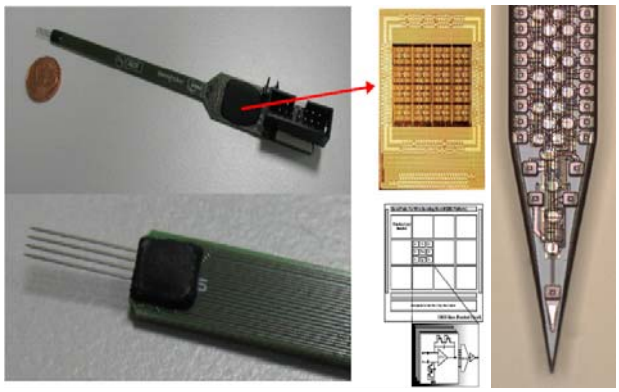


Figure 2. Overview of the 2D electrode array [14] with analog front-end. Analog front-end and structure on the right. Tip section of the shank of the 2D probe on the left.

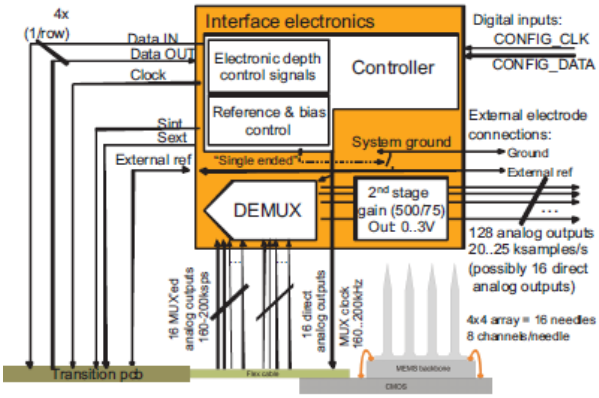


Fig.3 Sketch of the interface box.

B. *Data transmission*

In the fMR-EEG experiments we need a system which made of minimally magnetizeable materials. The A/D converter (Fig. 4), preamplifier (head stage), the amplifier, the data transmission and the power supply of these systems, optical USB link, fiber- ribbon- and USB cables.

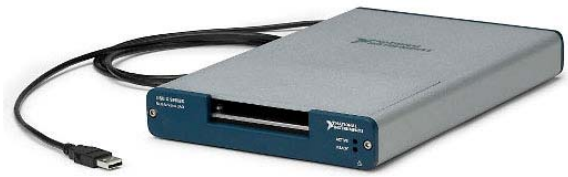


Fig.4 NI usb-6353 X series data acquisition device.
http://sine.ni.com/ds/app/doc/p/id/ds-151/lang/en

The planned MR compatible system (Fig. 5.) consists of a Head Stage (Gain 10x, Dc, 24 channel + reference), that connects with the Main Amplifier (Gain 100x, 0.1 Hz-6kHz band, ±3.2 V Lithium battery) through ribbon cable. After the amplification step the next is the A/D conversion with the above mentioned device (24 channel, 16 bit/ 20-40kHz/ channel), the A/D converter is supplied by a +14.4 V Lithium battery. The A/D converter sends a +5 V power supply and the digital signals to the Optical USB links sender side (USB to optical conversion in the sender side, send the data through fiber optic cables and a back conversion on the receiver side). The optical USB receiver sends the signals to a computers USB port, and we can save the signals with a costume made LabWiev software (.cnt format). The systems elements from the point of the fiber optic cables will be out from the scanner room.

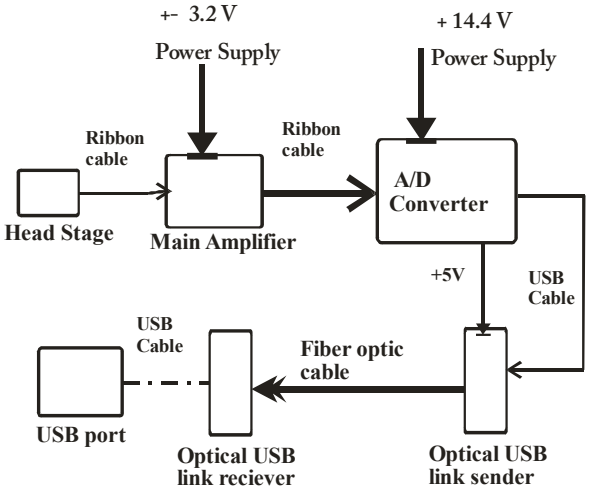


Figure 5. The sketch of the planned MR compatible Data transmission system.

C. *Implantation*

Impedance measurement: The impedances of the silicon electrode were measured before and during acute implantation. A two electrode setup was used in vitro in physiological saline solution (0.9% NaCl). The impedance was measured at 1 kHz. The average site impedance and the standard deviation was 1067.36±99.91 kΩ, but this was managed with an electrochemical etching process, where +2V was applied through the electrode sites immersed in physiological saline solution for 10 sec. The average impedance of the probe sites dropped this way to 659.04±59.47 kΩ. With longer activations the impedances could reach 200-300 kΩ.

Surgery: Probes were inserted in the neocortices of rats, mostly starting through the S1 trunk region reaching below the thalamus and in the primary motor cortex. Surgeries were performed under ketamine/xylazine anesthesia (ketamine: 75 mg/kg, xylazine: 5mg/kg). The probe assembled on the PCB was fixed to a stereotaxic device which was used for the insertion through the intact dura. The data was preamplified (g=10 gain, band pass filtered between DC and 100 kHz) and amplified (g=100 gain, band pass filtered between 0.1 Hz and 6 kHz).

In the situation of the microelectrode array the electrode selection was sent to the probe via a hardware controller using the NeuroSelect software which provides a graphical user interface for managing all versions of NeuroProbes microarrays with electronic depth control.

III. RESULTS

A. Silicon electrode

The analysis of the somatosensory cortex was measured with laminar silicon probes. The probe provided good quality local field potential (LFP), multi-unit (MUA) and single-unit (SUA) activity recordings. The slow oscillation (SO) [16-27] rhythm was observed in anesthetized rats. The SO alternates between a depolarized („up-state") and a hyperpolarized („down-state") state (Fig. 6.). The LFP of the cortical „down-state" (characterized by neuronal silence) is negative superficially, and has a positive polarity in the deep cortical layers.

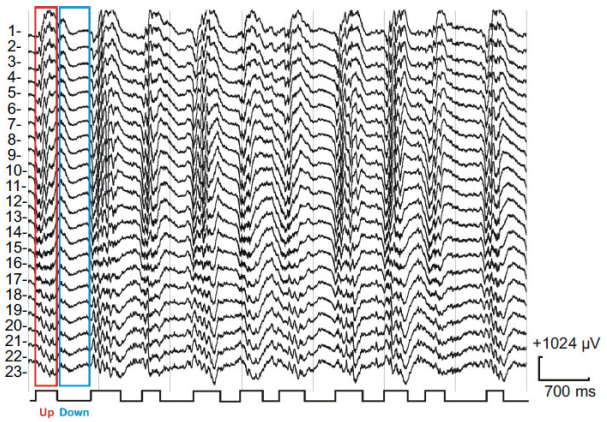


Figure 6. LFP with silicon probe in rat brain S1Trunk region

From the experiments there are significant evidences of the SO-s cortical origin. We made offline analysis of the datas we measured, for example current source analysis (CSD) to show the spatiotemporal changes of the transmembrane current sinks and sources, and because the probes design fits the needs (equidistant contact spacing, perpendicularly implanted to the laminar brain structures and the probe embrace the whole cortex) the CSD profile (Fig. 7.) of the cortex in SO can be made.

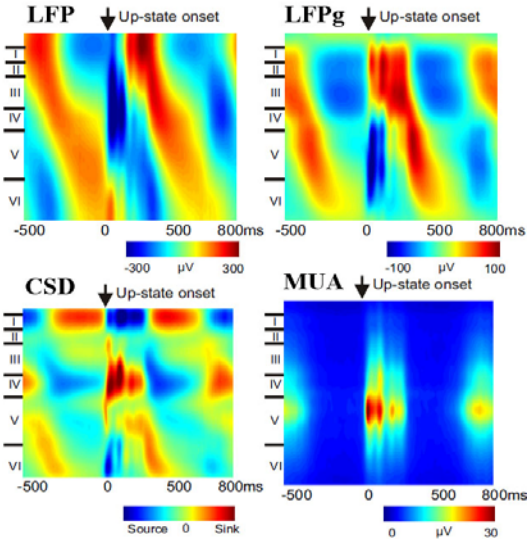


Figure 7. Up-state locked averages LFP, LFPg, CSD, MUA average maps.

B. Microelectrode array

We made comparisons between a passive electrode array, and the EDC microelectrode array. In this experiment a 4mm passive probe were implanted in the motor cortex, and the EDC probe into the S1 trunk region and the underlying thalamic nuclei (Fig. 8.).

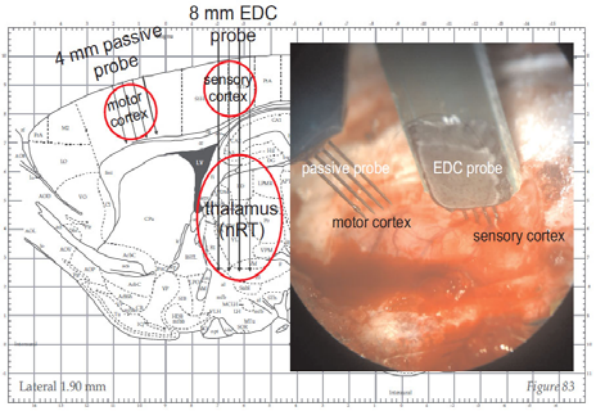


Figure 8. The targeted brain structures in the left, the implantation in the right

We measured good quality LFP, MUA (Fig. 9.) and SUA. The signals were analyzed forth offline. We wanted to have some infomations of the correlations of different regions of the brain. We compared some combinations from the gathered datas: -cross-correlations of the „up-state" onsets in different cortex layers from the same shank; -cross-correlations of the „up-state" onsets in the same layers from different shanks both located in the cortex; -cross-correlations of the „up-state" onsets from the same shank and different channels in the thalamus; -cross-correlations of the „up-state" onsets of two different shanks and the channels were in the cortex, and in the thalamus (Fig. 10.). A significant 40 ms lead of the thalamic onsets was measured.

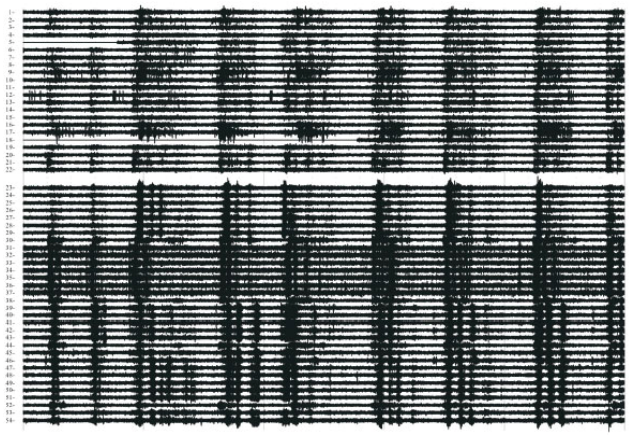


Figure 9. MUA passive in motor (1-24 channels), EDC probe in thalamus below S1 Trunk region (25-54)

All of these experiments were tested with the planned data transmission system too, and it complied with the desired demands for these experiments. We tested some of the planned systems boundaries, and it works well on a doubled sampling

rate than the old data acquisition system we used, even if its power supply were the lithium-ion batteries (for almost 6 hours).

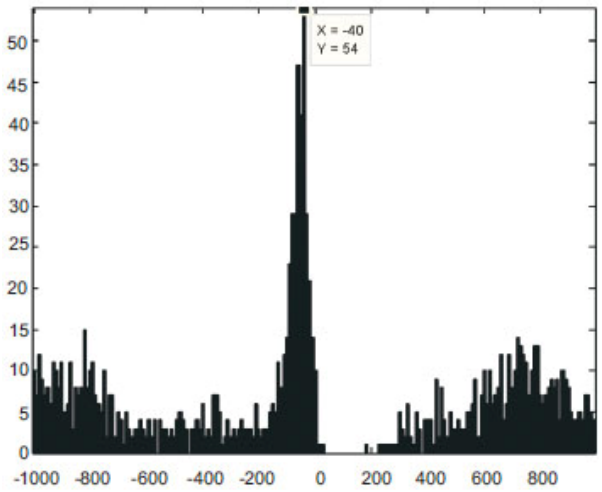


Figure 10. Cross correlogramm of up-state onsets from two different shafts. One in the cortex, one in the thalamus. The thalamus leads by 40 ms.

IV. CONCLUSION

From the experiment we made in the animal models, we could say that now we have a good basis for the planned fMR-EEG experiments, what we can compare. We started to gather the needed instrument for the MR compatible EEG system, and the data transmission system is almost ready. We started to test the system, and it shows good results for future investigations. I learned how to operate the fMR machine, and getting familiar with patient handling, and fMR paradigms for different tasks. I started to plan how to solve the two systems compatibility problems, and read through the literature of the problems of the existing systems.

V. FUTURE PLANS

In the fMR-EEG experiments we need to find solutions for the artefacts (magnetic, jitter, connectivity, susceptibility), the best possibility for the synchronization. We would like to do some experiments to know the susceptibility artefact of the EEG systems pieces. We will test our data transmission systems reliability in MR situations, and find the best placing in MR environment.

ACKNOWLEDGEMENT

The author wish to acknowledge Dr. György Karmos, Dr. István Ulbert, Richárd Fiáth, and Domonkos Horváth from the MTA-PKI for the assistance, and advice, Dr. Lajos Kozák from the MR Research Center Szentágothai Knowledge Center Semmelweis University for the study lectures in MR.

REFERENCE

[1] M. Markus Ullsperger, PhD and P. Stefan Debener, Eds., *Simultaneous EEG and fMRI Recording, Analysis, and Application*. Oxford University Press, 2010, p.^pp. Pages.
[2] C. Mulert and L. Lemieux. (2010). *EEG–fMRI Physiological Basis, Techniqueand Applications*.
[3] M. Scott H. Faro and P. Feroze B. Mohamed. (2010). *BOLD fMRI A Guide to Functional Imaging for Neuroscientists*.
[4] R. B. Buxton. (2009). *Introduction to Functional Magnetic Resonance Imaging Principles and Techniques (2 ed.)*.
[5] P. D. Ingmar Gutberlet. (2009, Did you know ...? MR Correction.
[6] M. Filippi, Ed., *fMRI Techniques and Protocols*. Humana Press, a part of Springer Science, 2009, p.^pp. Pages.
[7] D. Weishaupt, *et al.*, Eds., *How Does MRI Work? An Introduction to the Physics and Function of Magnetic Resonance Imaging*. Springer, 2006, p.^pp. Pages.
[8] S. M. Mirsattari, *et al.*, "EEG monitoring during functional MRI in animal models," *Epilepsia,* vol. 48, pp. 37-46, 2007.
[9] B. Kastler and Z. Patay, *MRI orvosoknak A mágneses magrezonancia orvosi képalkotó eljárásként való alkalmazásának alapelevei*. Budapest Udine: Folia neuroradiologica, 1993.
[10] M. Czisch, *et al.*, "Functional MRI during sleep: BOLD signal decreases and their electrophysiological correlates," *European Journal of Neuroscience,* vol. 20, pp. 566-574, Jul 2004.
[11] M. Czisch, *et al.*, "Acoustic Oddball during NREM Sleep: A Combined EEG/fMRI Study," *Plos One,* vol. 4, Aug 2009.
[12] E. B. Issa and X. Q. Wang, "Altered Neural Responses to Sounds in Primate Primary Auditory Cortex during Slow-Wave Sleep," *Journal of Neuroscience,* vol. 31, pp. 2965-2973, Feb 2011.
[13] A. Devor, *et al.*, "Coupling of total hemoglobin concentration, oxygenation, and neural activity in rat somatosensory cortex," *Neuron,* vol. 39, pp. 353-359, Jul 2003.
[14] H. P. Neves, *et al.*, "Multi-channel neural probes with electronic depth control," presented at the IEEE BioCAS 2010, 2010.
[15] L. Grand, *et al.*, "A novel multisite silicon probe for high quality laminar neural recordings," *Sensors and Actuators a-Physical,* vol. 166, pp. 14-21, Mar 2011.
[16] M. Steriade, *Neuronal Substrates of Sleep and Epilepsy*: Cambridge University Press, 2003.
[17] M. Steriade, *et al.*, "The Slow (4 Hz) Oscillation in Reticular Thalamic and Thalamocortical Neurons: Scenario of Sleep Rhythm Generation in Interacting Thalamic and Neocortical Networks," *The Journal of Neuroscience,,* vol. 13, 1993.
[18] M. Steriade, *et al.*, "Intracellular Analysis of Relations between the Slow (<1 Hz) Neocortical Oscillation and Other Sleep Rhythms of the Electroencephalogram" *The Journal of Neuroscience,* vol. 13, 1993.
[19] M. Volgushev, *et al.*, "Precise long-range synchronization of activity and silence in neocortical neurons during slow-wave sleep," *Journal of Neuroscience,* vol. 26, pp. 5665-5672, May 2006.
[20] M. Steriade, "The corticothalamic system in sleep," *Frontiers in Bioscience,* vol. 8, pp. D878-D899, May 2003.
[21] M. STERIADE and F. AMZICA, "Intracortical and corticothalamic coherency of fast spontaneous oscillations," *Neurobiology,* vol. 93, 1996.
[22] S. Chauvette, *et al.*, "Origin of Active States in Local Neocortical Networks during Slow Sleep Oscillation," *Cerebral Cortex,* vol. 20, pp. 2660-2674, Nov 2010.
[23] R. Csercsa, *et al.*, "Laminar analysis of slow wave activity in humans," *Brain,* vol. 133, pp. 2814-2829, Sep 2010.
[24] M. Massimini, *et al.*, "The sleep slow oscillation as a traveling wave," *Journal of Neuroscience,* vol. 24, pp. 6862-6870, Aug 2004.
[25] S. Sakata and K. D. Harris, "Laminar Structure of Spontaneous and Sensory-Evoked Population Activity in Auditory Cortex," *Neuron,* vol. 64, pp. 404-418, Nov 2009.
[26] T. J. Sejnowski and A. Destexhe, "Why do we sleep?," *Brain Research,* vol. 886, 2000.
[27] M. Steriade, *et al.*, "A Novel Slow (<1 Hz) Oscillation of Neocortical Neurons in vivo: Depolarizing and Hyperpolarizing Components," *The Journal of Neuroscience,* vol. 13, 1993.

Complex Electrophysiological Analysis of the Effect of Cortical Electrical Stimulation in Humans

Tóth Emília
(Supervisor: Dr. István Ulbert)
totem@digitus.itk.ppke.hu

***Abstract*— Direct cortical electrical stimulation (DCES) is frequently performed in concurrence with electrocorticogram recording for functional mapping (or electrical stimulation mapping-ESM) of the cortex and identification of critical cortical structures. In medically refractory epilepsy surgical candidates intracranial electrodes are necessary to localize the epileptogenic focus prior to surgical resection. This electrodes are used to record the underlying brain activity and also for electrical stimulation of the cortex. Electrical stimulation mapping (ESM) is the gold standard for identifying functional and pathological areas of the brain. Although the procedure remains unstandardized, and limited data support its clinical validity nevertheless, electrical stimulation mapping for define language areas has likely minimized postoperative language decline in numerous patients, and has generated a wealth of data elucidating brain-language relations [3]. Our aim was to study another way of cortical stimulation, so called single pulse electrical stimulation (SPES) to map pathological and functional networks in the brain.**

***Index Terms*- component biomedical signal processing, electrodes, electrocorticography, epilepsy, in vivo, human**

***Abbreviations*- ESM=electrical stimulation mapping; SPES=single pulse electrical stimulation; CT=computed tomography; DCES= direct cortical electrical stimulation; CCEP=cortico-cortical evoked potential; BA=Brodmann area; ROC curves=receiver operating characteristic curves**

I. Introduction

Mapping of functional areas in the human brain is crucial in epilepsy and tumor surgery. There are several non-invasive methods to identify eloquent cortices, such as functional Magnetic Resonance Imaging or Positron Emission Tomography, but the gold standard is direct high frequency cortical electrical stimulation. In this study we used single pulse electrical stimulation evoked late responses to map language and motor networks and to better understand the electrophysiological mechanisms of the cortico-cortical evoked potentials.

Single pulse electrical stimulation is a new method to investigate the cortico-cortical connections in vivo in the human language, motor and sensory system which can provide insight into the mechanisms of higher-order cortical functions and the connections between the functional areas [1]. When using a crown configuration, a handheld wand bipolar stimulator may be used at any location along the electrode array. However, when using a subdural strip, stimulation must be applied between pairs of adjacent electrodes due to the nonconductive material connecting the electrodes on the grid. Electrical stimulating currents applied to the cortex are relatively low, between 2 to 4 mA for somatosensory stimulation, and near 15 mA for cognitive stimulation. The functions most commonly mapped through DCES are primary motor, primary sensory, and language. The patient must be alert and interactive for mapping procedures, though patient involvement varies with each mapping procedure. Language mapping may involve naming, reading aloud, repetition, and oral comprehension; somatosensory mapping requires that the patient describe sensations experienced across the face and extremities as the surgeon stimulates different cortical regions.[2]

High frequency electrical stimulation is the gold standard in neurosurgery for mapping brain functions, but the exact mechanisms behind the effect and parameters used need to be further studied. There is also some risk associated with the

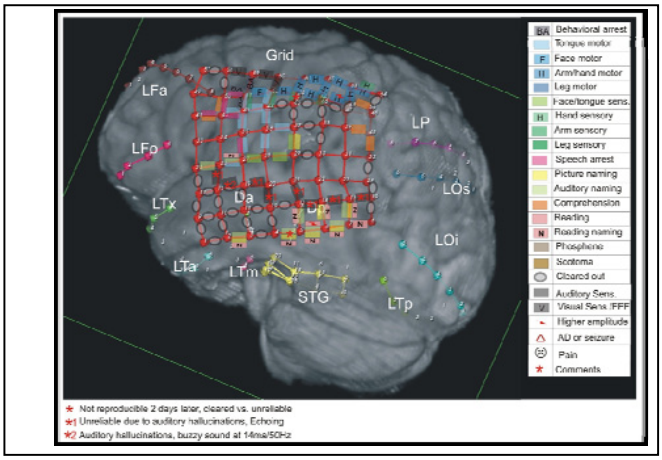


Figure 1. Reconstructed MRI picture with the implanted electrode array, colored lines represent the functions revealed with ESM.

stimulation, due to its proepileptic effect and the limits imposed by the fact that the cortex has to be exposed using some type of surgery.

Our aim with this study was to find other ways to map functionial networks in the brain, but using a less invasive method. Single pulse electrical stimulation (0,5Hz) is much less invasive in terms of seizure generation, but the distribution of the evoked potentials may reveal the intracortical pathways between cortical regions.

II. METHODS

A. Clinical electrodes and recordings

The electrode implantations and recordings, along with ESM and SPES took place at two well established epilepsy surgical centers in Budapest (National Institute of Neuroscience) and New York (North Shore-LIJ Health System). Patients were implanted with intracranial subdural grid, strip, and in some cases depth electrodes for 5–10 days. They were monitored to identify the seizure focus, at which time the electrodes were removed and, if appropriate, the seizure focus was resected. Continuous intracranial EEG was recorded with standard recording systems with sampling rates 1000 or 2000 Hz. The microelectrodes were implanted in eleven cases, perpendicularly to the cortical surface to sample the width of the cortex. This 24 contact laminar electrode has been described previously [4]. Differential recordings were made from each pair of successive contacts to establish a potential gradient across the cortical lamina.

B. Functional Stimulation Mapping

For localization of functional cortical areas, electrical stimulation mapping was carried out according to standard clinical protocol (bipolar stimulation: 2–5 s, 3–15 mA, 20–50 Hz). Areas were defined as expressive language sites when stimulation resulted in speech arrest. When stimulation resulted in a naming deficit based on auditory or visual cues, or an interruption in reading or comprehension, the area was deemed a nonexpressive language site. Sensory and motor areas were identified when stimulation caused movement or changes in sensation.

C. Cortical Electrical Stimulation and Cortico-Cortical Evoked Potentials.

Following implantation of intracranial electrodes, patients were monitored for epileptic activity and during this time, CCEP mapping was performed using single-pulse stimulation. Systematic bipolar stimulation of each pair of adjacent electrodes was administered with single pulses of electrical current (3 mA-15 mA, 0.5 Hz, 0.2-ms pulse width, 20-25 trials per electrode pair). The associated evoked responses (CCEPs) were measured at all other electrode sites. The current amplitude of 10 mA activated the maximal number of neuronal elements without epileptic afterdischarges or other clinical signs. The 2 seconds interstimulation interval was used to minimize the effect of overlapping evoked responses and to leave enough restitution time for the cortex. Patients were awake and at rest at the time of CCEP recording

D. Analysis of CCEPs.

Electrophysiological data analyses were performed using Neuroscan Edit 4.5 software (Compumedics) and own developed MATLAB scripts. Evoked responses to stimulation were divided into 2-s epochs (-500 ms to 1,500 ms) time-locked to stimulation pulse delivery. The CCEP consists of two usually negative peaks termed N1, timed at ~10–30 ms, and N2, which exhibits a broader spatial distribution and occurs between 70 and 300 ms [1]. To quantify the magnitude of the CCEPs in the time window of the N2, the data were low-pass filtered (30 Hz), and baseline correction (−450 to −50 ms) was performed. The SD was computed for each electrode separately using all time points in the -450 to -50 time window, CCEPs were considered significant if the N2 peak of the evoked potential exceeded the baseline amplitude by a threshold of ±6 SD as determined from the receiver operating characteristic (ROC) curves.

E. Electrode localization

To co-register the electrodes to anatomical structures, we used sophisticated imaging techniques, developed by our co-operational research team. We used intraoperative pictures and a postoperative CT scan to localize the electrodes in the skull. This was co-registered to a high resolution preoperative MRI where we could precisely localize the anatomical structures. Using these scans and freely available softwares (Bioimagesuite, Freesurfer, FSL, AFNI) we developed a semi – automated co-localizing each electrode to the underlying Brodman area of the brain. Determination of the seizure onset zone was performed by epileptologists [5].

F. Patients

Twenty-two patients (ages 6–53 years, 28±14.84, ten females) with medically intractable focal epilepsy were enrolled in the study after informed consent was obtained. These procedures were monitored by local Institutional Review Boards, in accordance with the ethical standards of the Declaration of Helsinki.

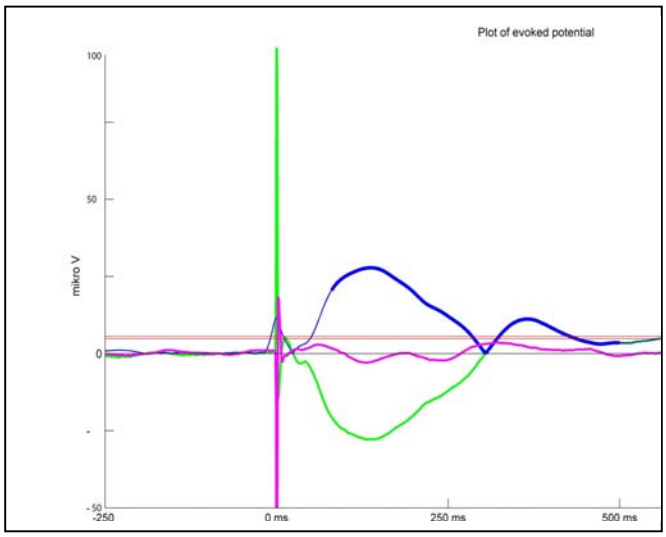


Figure 2. This figure shows averaged responses time locked to the bipolar stimulation artefact (-250-600ms). Green line is the significant response, blue line is the absolute value of the significant response, pink line is a non significant response and the red horizontal line is the threshold for the two responses.

III. RESULTS

A. *Analysis of the significant signal features.*

Due to the artifact caused by the stimulation we only focused on the N2 response, which semmed very reliable and reproducible. The variance of both time and amplitude of the N2 peak was high, but it the largest number of peaks occurred around 150 ms, and showed quasi-normal distribution, with two smaller deflections at around 180 -190 ms and 210-250ms. Analysis of 892 peaks, the average latency was 152.84 ms, with 58.7 ms standard deviation.

B. *Create a graph.*

A significant evoked response indicates the relationship between the electrodes which were stimulated and which showed the significant response. Significant CCEPs were converted to a distance matrix and transformed to a graph using multidimensional scaling

On the one hand the result shows that the functional areas which are close to each other are tightly connected (above somatosensory cortex BA40, BA3, BA2; visual cortex BA17, BA18, BA19 and motor cortex BA6, BA4). On the other hand, those regions which are physically more distant from each other seemed also connected, such as Broca's (BA 45) and Wernicke's (BA 21, BA20, BA22) area. Using this methodology we tried to map as many areas of the brain as possible, to be able to map all the connections between regions which were covered with electrodes.

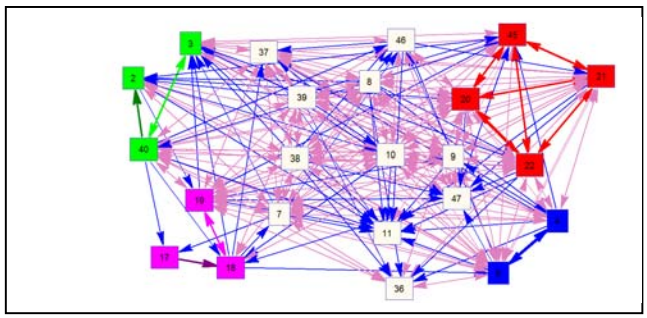


Figure 3. Significant CCEPs were converted to a distance matrix and transformed to a graph using multidimensional scaling. Numbers in squares represent Brodmann areas and lines represent connections. Functional networks are color coded: green sensory, pink visual, red language, blue motor. Lines color coded: thin light pink bidirectional, thin blue unidirectional, darker lines between the elements of functional networks is unidirectional, same color is bidirectional. Stimulating electrodes over Broca's area showed significant responses in electrodes part of the language network as defined with functional stimulation mapping. Responses to stimulation of the primary motor cortex revealed connections to major hubs involved in motor processing.

C. *Analysis of changes taking place in cortical layers.*

After processing the data from the laminar microelectrode and the implanted macroelectrodes, it can be concluded that after the stimulus, there is a decrease in the power of 15-100 Hz frequency band, and the stimulus elicit deactivation in the middle cortical (3th-5th) layers. This finding is in correlation with previous animal studies, which showed wide band decrease in oscillatory power after stimulation was induced.

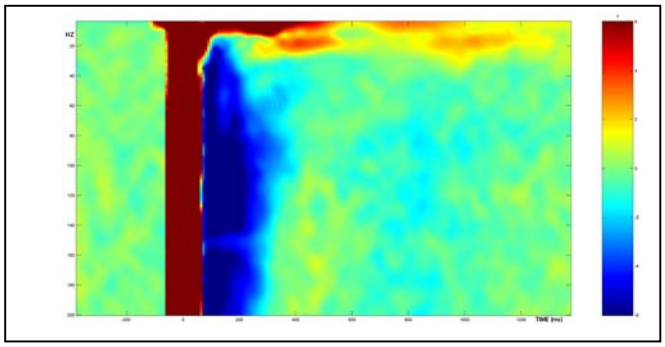


Figure 5. Time frequency analysis of data derive from subdural electrodes, blue color indicate the power decrease in the 15-200 Hz frequency band, red color indicate decrease.

IV. FURTHER AIMS

To verify the results, we need to refine the automatic detection method to increase significant response detection reliability. We also need higher number of patients involved in the study to increase the statistical significance of the study.

V. CONCLUSION

The results suggest that single pulse electrical stimulation evoked potentials may reveal connections of functional areas and functional networks of the human brain. Other studies also report that direct cortical stimulation has a suppressive effect on fast cortical activity and epileptic spikes [7], or can help to clarify the size of the area to be removed[8].

We conclude that single pulse electrical stimulation is a promising technique in delineating eloquent cortex and might be a useful tool to identify pathological networks.

REFERENCES

[1] R. Matsumoto, D.R. Nair, E. LaPresto, I. Najm, W. Bingaman, H. Shibasaki, and H.O. Lüders, "Functional connectivity in the human language system: a cortico-cortical evoked potential study.," Brain, vol. 127, (no. Pt 10), pp. 2316-30, Oct 2004.

[2] L. Schuh and I. Drury, "Intraoperative Electrocorticography and Direct Cortical Electrical Stimulation.," Seminars in Anesthesia vol. 16, pp. 46-55, 1996.

[3] M.J. Hamberger, "Cortical language mapping in epilepsy: a critical review.," Neuropsychol Rev, vol. 17, (no. 4), pp. 477-89, Dec 2007.

[4] I. Ulbert, E. Halgren, G. Heit, and G. Karmos, "Multiple microelectrode-recording system for human intracortical applications.," J Neurosci Methods, vol. 106, (no. 1), pp. 69-79, Mar 2001.

[5] D. Kovalev, J. Spreer, J. Honegger, J. Zentner, A. Schulze-Bonhage, and H.J. Huppertz, "Rapid and fully automated visualization of subdural electrodes in the presurgical evaluation of epilepsy patients.," AJNR Am J Neuroradiol, vol. 26, (no. 5), pp. 1078-83, May 2005.

[6] C.J. Keller, S. Bickel, L. Entz, I. Ulbert, M.P. Milham, C. Kelly, and A.D. Mehta, "Intrinsic functional architecture predicts electrically evoked responses in the human brain.," Proc Natl Acad Sci U S A, Jun 2011.

[7] M. Kinoshita, A. Ikeda, R. Matsumoto, T. Begum, K. Usui, J. Yamamoto, M. Matsuhashi, M. Takayama, N. Mikuni, J. Takahashi, S. Miyamoto, and H. Shibasaki, “Electric stimulation on human cortex suppresses fast cortical activity and epileptic spikes.,” Epilepsia, vol. 45, (no. 7), pp. 787-91, Jul 2004.

[8] A. Valentín, G. Alarcón, M. Honavar, J.J. García Seoane, R.P. Selway, C.E. Polkey, and C.D. Binnie, “Single pulse electrical stimulation for identification of structural abnormalities and prediction of seizure outcome after epilepsy surgery: a prospective study.,” Lancet Neurol, vol. 4, (no. 11), pp. 718-26, Nov 2005..

Visual Feature Detection and Classification for Banknote Detection on Low Resolution Image Flows

Mihály Radványi
(Supervisor: Dr. Kristóf Karacs)
radmige@itk.ppke.hu

***Abstract*— In countries where banknotes cannot be distinguished by their size visual information is crucial for determining the proper value of a given banknote. In this paper I present an algorithm that helps blind and visually impaired people in recognizing banknotes using the Bionic Eyeglass. The algorithm extracts special visual features of banknotes in order to determine their value and orientation using topological description.**

Keywords – banknote, bionic, CNN, eyeglass, feature, orientation

I. INTRODUCTION

The Bionic Eyeglass [1],[2] is a portable device recently proposed to aid blind and visually impaired people in everyday navigation, orientation and recognition tasks that require visual input. In countries where banknotes cannot be distinguished by their size visual information is crucial for determining the proper value of a given banknote. Recognizing the value of a banknote is one of the few important tasks we have identified based on the feedback from potential users.

In our previous works we have already introduced the concept and the prototype of the Bionic Eyeglass [3] which was built using the Bi-i visual computer [4] as its main computational platform. The Bi-i visual computer can provide enough computing power for complex algorithms and it implements the Cellular Neural/Nonlinear Network – Universal Machine (CNN-UM) [5],[6] and the underlying Cellular Wave Computing principle.

Since the banknotes have different visual features on each side – front side and back side – different algorithms were developed for both cases. In this paper an algorithm for extracting visual features on the front side of the banknotes is discussed. Further on extracting objects a primary classification method is applied during the morphological processing resulting indifferent output masks containing the detected objects. The output objects are then given to another more complex classification and decision that is not discussed in details in this paper [7].

However, by processing an image flow frame by frame robust decisions can be achieved, in cases when there is strong correlation between frames – e.g. on video flow the frames following each other are similar – running the whole algorithm for each frame is a waste of time and processing power. Designing and using a reliable tracking algorithm would increase the performance of the banknote recognition task.

The paper is structured as follows: in Section 2 main types of patterns of banknotes are discussed; Section 3 describes the morphological processing and detection of feature objects; Section 4 introduces a side detection algorithm; Section 5 presents experimental results; Section 6 concludes the paper.

II. PATTERNS ON BANKNOTES

When designing a reliable banknote detector algorithm a proper description and examination of the target object or object sets is needed. In this section front side elements of banknotes are introduced.

Considering Hungarian banknotes two kinds of objects can be distinguished on the front side. There are objects that appear to be the same on each banknote (*static objects*) and objects that change regarding to the actual value of the given banknote (*dynamic objects*). Fig.1. demonstrates the different objects that can be determined on Hungarian banknotes.

Figure 1. Different objects of Hungarian banknotes are shown. Dashed bounding box indicates static objects and continuous the dynamic ones. Objects within green frames are capable for banknote recognizing tasks, while reds are not due to detection

Detailed analysis of front side objects of banknotes identified two types of objects that are diverse enough to be used in banknote recognition tasks. The set of number digits are obvious to be different on each value, however the usage of this object set can be sensitive for processing losses. Losing one digit can easily ruin the proper recognition of banknote values. More robust results were obtained by using the facial portraits among front side objects. Not only the masks of portrait and number objects are important to use during recognition but the mask containing the holes within them too. In the following section detection of different objects is discussed.

III. DETECTING BANKNOTES

A. Banknote detection

As in each image processing task the algorithmic output is really sensitive for the initial steps applied on input images.

Defining an optimal threshold value for binarizing an input frame is always difficult. In the present algorithm a simple adaptive threshold is used to determine a Region of Interest (ROI) that overlaps with areas of banknotes on a given frame. Then a threshold value is calculated using Otsu's method on a candidate object of the ROI that has the highest area. Binarizing an image with that method results in a mask on that banknote background and foreground – objects on banknotes – are separated.

B. Feature extraction

Having a foreground-background separated image does not mean that we have identified all the objects as well. Further steps are needed in order to extract features we are interested in. At first all objects that are connected to the border of the image are removed resulting in a mask containing patterns on banknotes. By applying the same step on the inverted image gives us another binary mask on which holes of banknote pattern objects appear. Fig.2. shows an example for each processing steps mentioned above.

Figure 2. An example for different processing stages. From left to right: input, region of interest, binary image, extracted patterns and extracted holes are shown.

Further examples of banknote patterns and holes are shown on Fig.3. below.

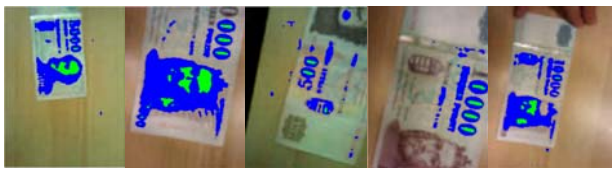


Figure 3. Further examples of detected banknote patterns are shown with blue, holes within them appear in green.

C. Detecting portraits

As it was mentioned before front side portrait objects were found to be ideal for banknote recognition tasks. To extract portrait objects from banknote patterns shown in Fig.3 simple morphological descriptors were used, just like filled area, number of holes, orientation and eccentricity. In cases when portraits were missing from input frames by simply defining portrait object as the ones with highest filled area led us to misclassify some other objects thus further analysis and definitions were needed. Taking into account that number digits are likely to merge together into one large object when using weak threshold values, a more sophisticated criterion was defined for detecting portraits. A few examples of detected portrait masks are shown in Fig.4.

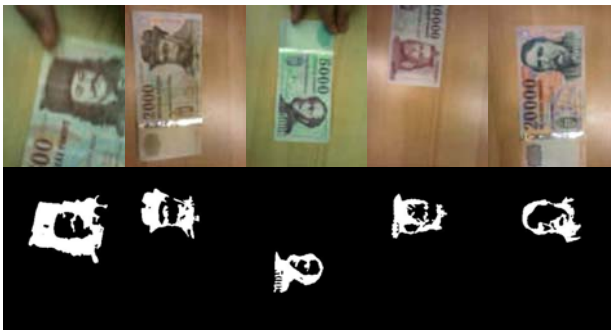


Figure 4. Input images and detected portrait objects are shown .

Although criterions for portrait objects were designed carefully, a few typical misclassifications could be identified. Since the algorithm presented in this paper is not supposed to make a final decision on detected objects, typical misclassifications can be used for our advantage during a later more complex classification and validation method. The following figure shows typical failures.

Figure 5. Typical objects that are found to be portraits during an early classification. The final validation was trained to indentify these typical misclassifications leading mistakes used as advantages.

Since the neural network responsible for final validation [7] was trained for these special cases, misclassifications can be used as advantages. In some cases it can be determined that the backside of banknote is seen on the actual frame, in other cases even the proper value can be extracted from a typical misclassified object.

D. Finding numbers

Although banknote recognition based on portraits showed great results, better performance can be achieved by detecting and identifying other features as well. As it was previously discussed, besides portraits number digits are another type of objects that can be used for recognition tasks.

The detection of numbers is devided into two steps. At first zero digits are found, then number objects are extended resulting in a final mask containing all digits.

Zero digits are found by using the mask of holes. The holes with higher eccentricity, similar size and orientation are selected, enlarged and then used for reconstructing [8] objects they are within. The resulting objects are the filtered and defined as zero digits.

After having zero digits a median line is fit on them and an iterative algorithm starts. The main goal of that algorithm is to go along the median line in both directions and to detect the changes of background and foreground areas. After a given condition – the number of background pixels exceeds the double of previous background area – the algorithm stops and defines two end points for the line. Finally all objects hitting

that line are reconstructed and defined as number mask. In Fig.6 the main steps of that algorithm are shown.

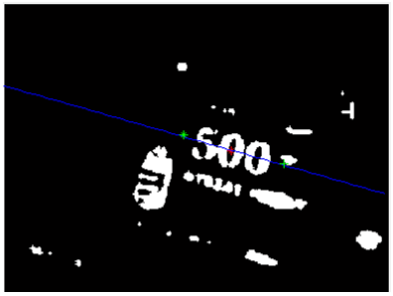


Figure 6. Extending number objects is carried out by going along the blue line starting from the center of a detected zero (red cross) in both directions until a given criterion is met.

By obtaining the mask of digits as well a more efficient decision and recognition can be carried out using both the portrait recognition thread and a number recognition algorithm.

The corresponding UMF (Universal Machines on Flows) diagram is shown on Fig.7.

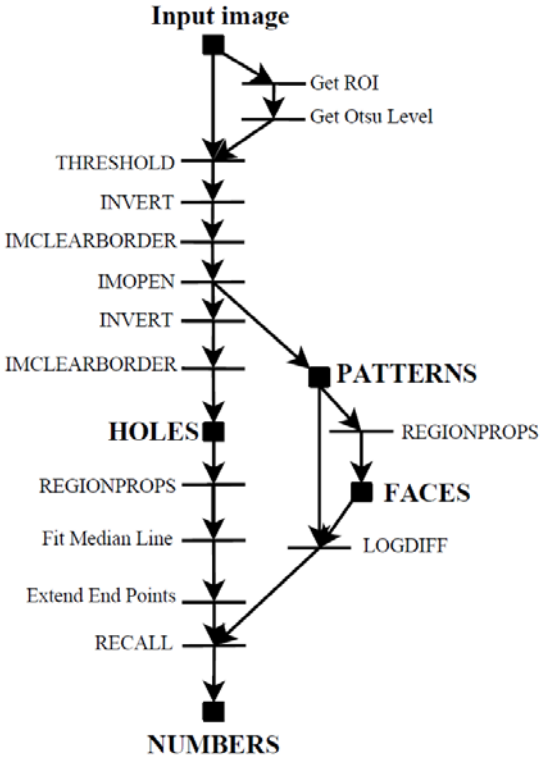


Figure 7. UMF(Universal Machines on Flows) diagram of banknote pattern detector algorithm

IV. Side detection

Banknote recognition based on front side objects was introduced in this paper but in order to use it efficiently one must know in advance which side of the banknote is the camera facing. Assuming that the position of the banknote is known the following method can be used to determine the actual side of the banknote.

By having the coordinates of the four corners of banknotes we are able to define a few regions to be observed. See Fig.8. for details.

Figure 8. Specific areas for side detection. Three areas (A,B,C) are used during detection, the other three (D,E,F) are just marked to help understanding the method.

Regarding to Fig.8. three areas (A,B,C) take part in the processing method, the other three(D,E,F) are indicated to help understanding. The processing starts with observing the middle area – B (or E) – on a binarized image. This area must contain about 98% white pixels. If not the actual frame has to be rotated. At this point the banknote should lie as it is shown on Fig.8. The decision is very simple only the number of black pixels in area A has to be compared to the number of black pixels in area C – or area D to area F respectively. If the upper area (A) contains more black pixels we can be sure that the front side of the banknote appears on the frame, otherwise the backside.

V. Experimental results

Results discussed in this section are from a collaboration of a few other students working on the banknote recognition task of the Bionic Eyeglass

A. Test environment

We used a Samsung Galaxy S mobile phone as the input device and streamed the incoming video flow through wireless connection to an external device. We have developed a protocol, so that either the frontend or the backend part can be exchanged. We have used two backend interfaces: Bi-i, an analogic visual computer [4] and a Matlab environment on a standard PC with the MatCNN toolbox.

B. Human tests

We performed experiments with three visually impaired subjects who were given a few basic instructions about the optimal measurement technique and the 6 Hungarian banknotes. Their task was to tell which note they are holding with the help of our method installed to a mobile device. Each person made three series of tests.

C. Results

Since the output of the presented algorithm is supervised by the neural network based classification, and the typical misclassifications can still be used for decent recognition, only results of the final classification are discussed in this section.

We tested the portrait classificator on 1541 labeled test images. Without defining sub-classes the precision rate ($p = c / b$) was 84.4%, while the accuracy ($a = c / e$) was 81.8%, where c is the number of correctly classified objects, b

is the number of classified objects, and *e* is the number of all objects. Using manual sub-class selection the precision rate was 88.8% and the accuracy was 85.3%. With sub-classes defined by unsupervised learning the precision increased to 92.3% and the accuracy to 86.7%.

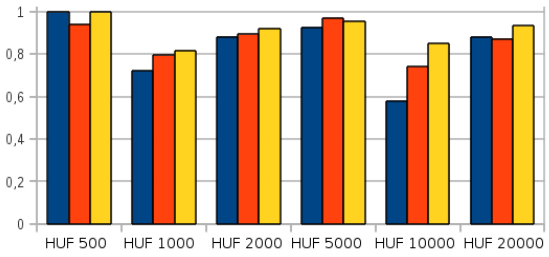


Figure 9.Precision rate of portrait recognition for the individual classes without sub-classes (blue), with manually selected sub-classes (red) and with sub-classes generated by clustering (yellow).

The ensemble decider made a decision 438 times during the nine tests, and 406 of them were correct, which corresponds to a 92.7% accuracy. Based on these votes, the subjects could correctly identify the banknote in 94.4% of the cases.

The most typical cases of error occurred when a part of the banknote was hidden by the hand or finger of the user or when the distance between the camera and the banknote was not enough for accurate detection. As time went by, we have noticed that the ratio of errors reduced, and the subjects have mastered the use of the device.

VI. Conclusions

I presented an algorithm to detect and pre-classify front side objects of banknotes for banknote recognition task. The algorithm extracts portrait and number objects by examining geometrical properties of objects, like filled area, eccentricity, number of holes and orientation. The pre-classified output masks are given away for another more sophisticated classification and validation method that makes the final decision on banknote values using complex Zernike descriptors and a neural network. The algorithms have been tested through a mobile phone interface with blind subjects. The ensemble decider uses multiple features to classify the banknotes, and the subjects rapidly learned the know-how, and could confidently use the system for banknote identification afterwards.

When dealing with video flows, we can greatly improve recognition by making use of the hypothesis about the location of the banknote based on previous frames. By using candidate key-frames that are fully processed, on the following frames we have the possibility to make estimations on the position of the visual features of the banknotes with a simple tracking algorithm.

Acknowledgment

The help of the volunteers from the IT for visually impaired foundation, and the contribution of Mihály Szuhaj are kindly acknowledged.

References

[1] T. Roska, D. Bálya, A. Lázár, K. Karacs, R. Wagner, and M. Szuhaj, “System aspects of a bionic eyeglass,” in Proc. of the 2006 IEEE International Symposium on Circuits and Systems (ISCAS 2006), Island of Kos, Greece, May 21–24, 2006, pp. 161–164.

[2] K. Karacs, A. Lázár, R. Wagner, D. Bálya, T. Roska, and M. Szuhaj, “Bionic Eyeglass: an Audio Guide for Visually Impaired,” in Proc. of the First IEEE Biomedical Circuits and Systems Conference (BIOCAS 2006), London, UK, Dec. 2006, pp. 190–193.

[3] K. Karacs, A. Lázár, R. Wagner, B. Bálint, T. Roska, and M. Szuhaj, “Bionic Eyeglass: The First Prototype, A Personal Navigation Device for Visually Impaired,” in Proc. of First International Symposium on Applied Sciences in Biomedical and Communication Technologies (ISABEL 2008), Aalborg, Denmark, 2008.

[4] A. Zarandy and C. Rekeczky, “Bi-i: a standalone ultra high speed cellular vision system,” IEEE Circuits Syst. Mag., vol. 5, no. 2, p. 36–45, 2005.

[5] T. Roska and L. O. Chua, “The CNN universal machine: an analogic array computer,” IEEE Trans. Circuits Syst. II, vol. 40, pp. 163–173, Mar. 1993.

[6] L. O. Chua and T. Roska, Cellular Neural Networks and visual computing.Cambridge, UK: Cambridge University Press, 2002.

[7] A. Stubendek “Human Like Semantic Models for Object Detection and Classification” PPKE – ITK PhD Proceedings 2011, Budapest.

[8] L. Kék, K. Karacs, and T. Roska. (2007) Cellular wave computing library, version 2.1 (templates, algorithms and programs). [Online]. Available: http://cnn-technology.itk.ppke.hu/Template_library_v3.1.pdf visited on 19-06-2010.

Active Wave Computing Using Large Infrared Proximity Array

Miklós Koller
(Supervisor: Dr. György Cserey)
kolmi@itk.ppke.hu

***Abstract*—We found that the spatial-temporal dynamics of the perceived 3D surface using an infrared proximity array gives a well characterized spatial-temporal dynamics at the output of the CNN wave-computer. For practical reasons, the continuous input is a series of consecutive infrared images, which are the consecutive states of the input stream. The spatial-temporal dynamics of the output can be well characterized by quantitative and qualitative (specific morphology and oscillation frequency) spatial and temporal features. Some known wave-computing template cases [1] were taken as a base, where for various input constants the system responded constant, periodic or chaotic spatial-temporal dynamics. In case of qualitatively similar inputs, similar spatial-temporal signatures were recorded.**

***Keywords*-spatio-temporal dynamic pattern, CNN (Cellular Nonlinear Network), infrared proximity array**

I. Introduction

If we talk about wave-propagation, we must make a distinction between classical waves (propagating in conservative systems, behaving as a closed system) and nonlinear (active) waves. In the case of active waves there is a capital importance: the propagation of the wave is supported by the media, from an energetic point of view. Subsequently the propagating wave does not decay nor the waveform is distorted during the propagation. These waves can not be reflected from the boundaries, nor can they be interfered, but they can be diffracted. When they are colliding with each other, they annihilate each other (except a special kind of them, the trigger wave: after some definition they can merge at collision).

Although in general the classical wave is associated with the word "wave", there are many well-described phenomena of active wave propagation. In chemistry, reaction-diffusion systems can produce waves [2]. As examples from biology, nerve impulse propagation shows a wave-process [3] and the communication method of the amoeba [4] [5] based on waves too. Studying the examples of nature, a unified paradigm has been developed by Chua [6] [7] [8] [9] [10] [11]. Generally speaking: solving some of the computational problems with a system utilizing active waves could be much more obvious [12].

In our experiments we dealt with some aspects of active wave propagation. The necessary computational model was implemented in the form of computer simulation, for which the input was supplied by a special sensor array [13]. Our sensor array is built up from an 8×8 infrared distance-sensor, which was originally developed to analyze the phenomenon of hyperacuity [14] in space [15]. Since our elementary cells measure distance, it is suspected that the input forming reflected light from the measured object can supply additive information from the wave-propagation modifying features of the environment.

In this paper we formulate some statements regarding active wave propagation. We have observed the following: measuring the spatio-temporal dynamic of a 3D surface using this 2D infrasensor-array we can get a well-describable spatio-temporal dynamic on the output of the CNN. For practical reasons, the continual input is built up from different still images, in a consecutive order. The spatio-temporal dynamic of the output can be described with spatial and temporal qualitative and quantitative (defined morphology and oscillation frequency) attributes. Furthermore we have done measurements, in which we analyzed the successively stitched sequence of the individual dynamics. For different still image inputs we get different spatio-temporal dynamic on the output. If we concatenate the different still inputs to a continuous input flow, we can measure the stitched sequence of the previously measured spatio-temporal dynamics.

This paper is organized as follows: In the second section a short introduction is given to the model and the experimental environment. The third section contains the measurements for constant inputs. In the fourth section, analysis of the measurements is presented in detail in case of a dynamic input. The fifth section concludes the results, and finally a discussion takes place.

II. Realization of the computational model, the measuring environment

To analyze the phenomena we have realized the computational model of the Cellular Neural Networks (CNN), as a computer simulator. The input channel was served by a real hardware-element.

Our sensor array is built up from 8×8 pieces of active infrared distance-sensors. Every cell contains an infra light source and a phototransistor. In this way every pixel on the input-picture represents a distance value. The benefits of this kind of input device can come to the front at 3D problems, where it is not necessary to recline upon the uncertainty of the light-shadow image coming from the measured surface.

In the simulated computational model we approximated the solution of the differential-equation with the implicit Euler method. The output function was a modified sigmoid, the parameters can be seen in the (1) formula.

$$y = 2 * (\frac{1}{1 + e^{-2.65*x}} - 0.5) \quad (1)$$

The output of our computational model was fundamentally determined by the applied antisymmetrical feedback (A) template class. The recognition came from the doctoral thesis of Istvn Petrs [16]. The qualitative theory of nonsymmetric feedback template were first exposed in [17].

The applied templates can be seen at the (2) form.

$$\mathbf{A} = \begin{pmatrix} 0 & 0 & 0 \\ s & p & -s \\ 0 & r & 0 \end{pmatrix}, \mathbf{B} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & 0 \end{pmatrix}, \mathbf{z} = z \quad (2)$$

The values of the formal parameters were the following:

$$\mathbf{T1}: \quad s = 1.1, p = 0.9, r = -0.6, b = 1.0, z = 0 \quad (3)$$

$$\mathbf{T2}: \quad s = 1.1, p = 1.2, r = -0.6, b = 1.2, z = 0 \quad (4)$$

$$\mathbf{T3}: \quad s = 1.1, p = 1.0, r = 0.3, b = 1.2, z = 0 \quad (5)$$

$$\mathbf{T4}: \quad s = 1.1, p = 1.4, r = -0.35, b = 1.6, z = -0.2 \quad (6)$$

The initial state of the system was the same as the input picture, and in every iteration the system got the same input-picture. The boundary cells were zero-flux. The time step of our digital simulation was 0.01 time unit. The input scene is built up of two plain surfaces, which were in different distance from the panel. The border line between them had no extension. The distances of the surfaces from the sensor array were determined in such a way, that they will be near the two end-points of the measurements' dynamic-range. The exact placement of the two surfaces will be outlined in the next section.

III. Measurement results in case of constant input

In this section we are going to analyze some cases, when little difference in the inputs can produce qualitatively different spatio temporal output flow. The difference between the inputs stands in the scale of the masking/coverage rate among the nearer surface and the sensor array.

The first selected standalone measurement and its different processing with different templates is as follows: the nearer plain surface, which builds up the positive (black) saturated value, has a half column-wide masking rate with the sensor array.

With template **T1**(3) we perceived alternating white and fluctuating grey rows on the output of the system. The period of the grey rows' fluctuating is 290 iteration.

With template **T2**(4) we perceived similar results as before: in an alternating order there are white and fluctuating grey rows, latter with period of 410 iterations. An additional difference is in the fluctuating rows is the wider color-dynamic range of the grayscale-values.

With template **T3**(5) the first two column of the bottom two row fluctuate, in a 2 column-wide range at right, with a period of 625 iteration. The remaining part of the system is in stable negative saturation region(white color).

With template **T4**(6) the output is in the stable negative saturated region.

The second selected standalone measurement and its different processings with different templates can be seen in (Figure 1). The first subfigure (a) shows the input picture: the nearer plain surface, which makes the positive value (black), has three and a half column-wide masking rate with the sensor array.

With template **T1**(3) (Subplot b) the output array after the 2nd/3rd column is fluctuating with greyscale values, and with a period of 320 iterations.

With template **T2**(4) (Subplot c) we perceived similar results as before: on the output array after the 3rd column there is a greyscale oscillation. But there is two essential differences: the period of the fluctuation is about 500 iterations, furthermore, the output-cells represent values for a more wide dynamic range.

With template **T3**(5) (Subplot d) the left half of the output gets stable positive saturation (black color). The 5th, 6th and 7th columns show fluctuating state; the 8th column is in stable negative saturation (-1 value, white). This kind of fluctuating differs in a qualitative way from the previous templates' wave-rising. The reason is the opposite sign of the south cell value of the feedback template (*A8*). In the previous two cases we perceived wave propagation in vertical direction; but here we perceive it in horizontal direction. The period of the fluctuation is about 650 iteration.

With template **T4**(6) (Subplot e) we get vertically directed wave-propagation again, but only in a 3-column-wide region. The first three column of the output array is in stable positive state; the 4th, 5th and 6th column show wave-propagation with period of 800 iterations; the 7th and 8th columns are in stable negative state.

The third selected standalone measurement and its different processings with different templates is as follows: the nearer plain surface has a little more than five columns wide masking rate with the sensor array.

With template **T1**(3) the whole output array gets in stable state after a few iterations. The states are from greyscale-range.

With template **T2**(4) the first 5 columns get in positive saturated state, the remaining 3 columns' upper 6 rows show a vertical-directed wave-propagation. The period is about 580 iterations. The right 3 cells of the 7th and 8th rows get a stable-greyscale value.

With template **T3**(5) the first five columns of the output is in stable positive saturated state. The 6th, 7th and 8th columns show horizontal-directed wave-propagation, with a period of about 680 iterations.

With template **T4**(6) the first five columns of the output are in stable positive saturated state; the 6th, the 7th and 8th columns show vertical-directed wave-propagation, with a period of about 1080 iterations.

IV. Measurement results in case of dynamic input flow

In the previous section we have analyzed the output of our system for slightly different inputs. With a little, quantitative change in the input we were able to qualitatively change the spatio-temporal dynamics of the output. In the light of this one can state the question: what would be the output if this slightly differing inputs would be stiched after each other, and this flow would be added to the input? Are we able to recognize the previously identified spatio-temporal patterns and dynamics in the same order?

Realizing the fluently changing input, not only the previously showed input pictures were utilized. We measured additively a lot of inputs, which positive region stood between the existing ones. In this way we get approximately a fluent input in time and space. The input pictures were changed after 1000 iterations, roughly imitating the ratio between the speed of a real CNN chip and the speed of an environmental event. In the output flow we were able to recognize the previously, self-standing detected spatio-temporal dynamics. In the followings we will emphasize the differences between the previously showed output-dynamics (the case of still inputs) and the recently measured new ones (the case of changing input flow).

In the case of the first two template (**T1** (3); **T2** (4)) we measured broadly speaking the same output dynamics in the appropriate time-place as we take the appropriate input file. Their qualitative and quantitative descriptors are almost the same.

In the case of the third template (**T3** (5)) we perceived qualitative difference in the spatio-temporal flow (regarding the still inputs). Until the first column does not get an appropriately positive value in the input flow, the whole output field is stabilizing in the negative saturation state (-1, white). When we gave the input as a stable, constant value, the less intensive input was also able to trigger the dynamic oscillation. When our input flow arrived the second and third time proposed input-levels, the spatio-temporal dynamic already represents the wanted qualitative and quantitative descriptor-levels (as we have seen before).

In the case of the fourth template (**T4** (6)) we get back appropriately the previously explored spatio-temporal dynamics, as we used the analogous still inputs.

V. Conclusion

We found that the spatial-temporal dynamics of the perceived 3D surface using an infrared proximity array gives a well characterized spatial-temporal dynamics at the output of the CNN wave-computer. The spatio-temporal dynamics of the output can be well characterized by quantitative and qualitative (specific morphology and oscillation frequency) spatial and temporal features. In case of qualitatively similar inputs, similar spatial-temporal signatures were recorded. Furthermore we analyzed the successively stitched sequence of the individual dynamics. For different still image inputs we get different spatio-temporal dynamic on the output. If we concatenate the different still inputs to a continuous input flow, we can measure the stitched sequence of the previously measured spatio-temporal dynamics.

Acknowledgment

The Office of Naval Research (ONR) and the Hungarian Scientific Research Fund (OTKA) which supports the multidisciplinary doctoral school at the Faculty of Information Technology of the Pázmány Péter Catholic University and the Bolyai János Research Scholarship is gratefully acknowledged. The authors are also grateful to Professor Tamás Roska for the discussions and his suggestions.

References

[1] L. Kék, K. Karacs, A. Zarándy, and T. Roska, "CNN template and subroutine library for cellular wave computing," tech. rep., Report DNS-1-2007, 2007.

[2] A. N. Zaikin and A. M. Zhabotinsky, "Concentration wave propagation in two-dimensional liquid-phase self-oscillating system," *Nature*, vol. 225, pp. 535–537, 1970.

[3] A. C. Scott, "The electrophysics of a nerve fiber," *Rev. Mod. Phys.*, vol. 47, pp. 487–533, Apr 1975.

[4] P. Newell and J. Reissig, "Microbial Interactions (Receptors and Recognition, Series B)," 1977.

[5] B. Goodwin, *How the leopard changed its spots: The evolution of complexity*. Princeton Univ Pr, 2001.

[6] L. Chua, *CNN: A paradigm for complexity*. World Scientific Pub Co Inc, 1998.

[7] L. Chua, "CNN: A vision of complexity," *International Journal of Bifurcation and Chaos in Applied Sciences and Engineering*, vol. 7, no. 10, p. 2219, 1997.

[8] R. Dogaru and L. O. Chua, "Edge of chaos and local activity domain of the gierer-meinhardt CNN," *Int. J. of Bifurcation and Chaos*, vol. 8, no. 12, pp. 2321–2340, 1998.

[9] R. Dogaru and L. O. Chua, "Edge fo chaos and local activity domain of fitzhugh-nagumo equation," *Int. J. of Bifurcation and Chaos*, vol. 8, no. 2, pp. 211–257, 1998.

[10] R. Dogaru and L. O. Chua, "Edge of chaos and local activity domain of the brusselator CCN," *Int. J. of Bifurcation and Chaos*, vol. 8, no. 6, pp. 1107–1130, 1998.

[11] L. O. Chua, "Passivity and complexity," *Circuits and Systems I: Fundamental Theory and Applications, IEEE Transactions on*, vol. 46, pp. 71–82, Jan. 1999.

[12] T. Roska, "Cellular wave computers for brain-like spatial-temporal sensory computing," *Circuits and Systems Magazine, IEEE*, vol. 5, no. 2, pp. 5–19, 2005.

[13] M. Koller and G. Cserey, "CNN computational abilities of large infrared proximity arrays," in *Cellular Nanoscale Networks and Their Applications (CNNA), 2010 12th International Workshop on*, pp. 1–4, IEEE, 2010.

[14] K. Lotz, L. Boloni, T. Roska, and J. Hamori, "Hyperacuity in time: a CNN model of a time-coding pathway of sound localization," *Circuits and Systems I: Fundamental Theory and Applications, IEEE Transactions on [see also Circuits and Systems I: Regular Papers, IEEE Transactions on]*, vol. 46, pp. 994–1002, Aug. 1999.

[15] Á. Tar, M. Koller, and G. Cserey, "3D Geometry Reconstruction Using Large Infrared Proximity Array for Robotic Applications," in *Mechatronics, 2009. ICM 2009. IEEE International Conference on*, pp. 1–6, IEEE, 2009.

[16] I. Petrás, *Spatio-Temporal Patterns and Active Wave Computing*. PhD thesis, Neuromorphic Information Technology, Interdisciplinary Graduate Program, Analogical and Neural Computing Laboratory, Computer and Automation Institute, Hungarian Academy of Sciences, 2005.

[17] L. O. Chua and T. Roska, "Stability of a class of nonreciprocal cellular neural networks," *Circuits and Systems, IEEE Transactions on*, vol. 37, no. 12, pp. 1520–1527, 1990.

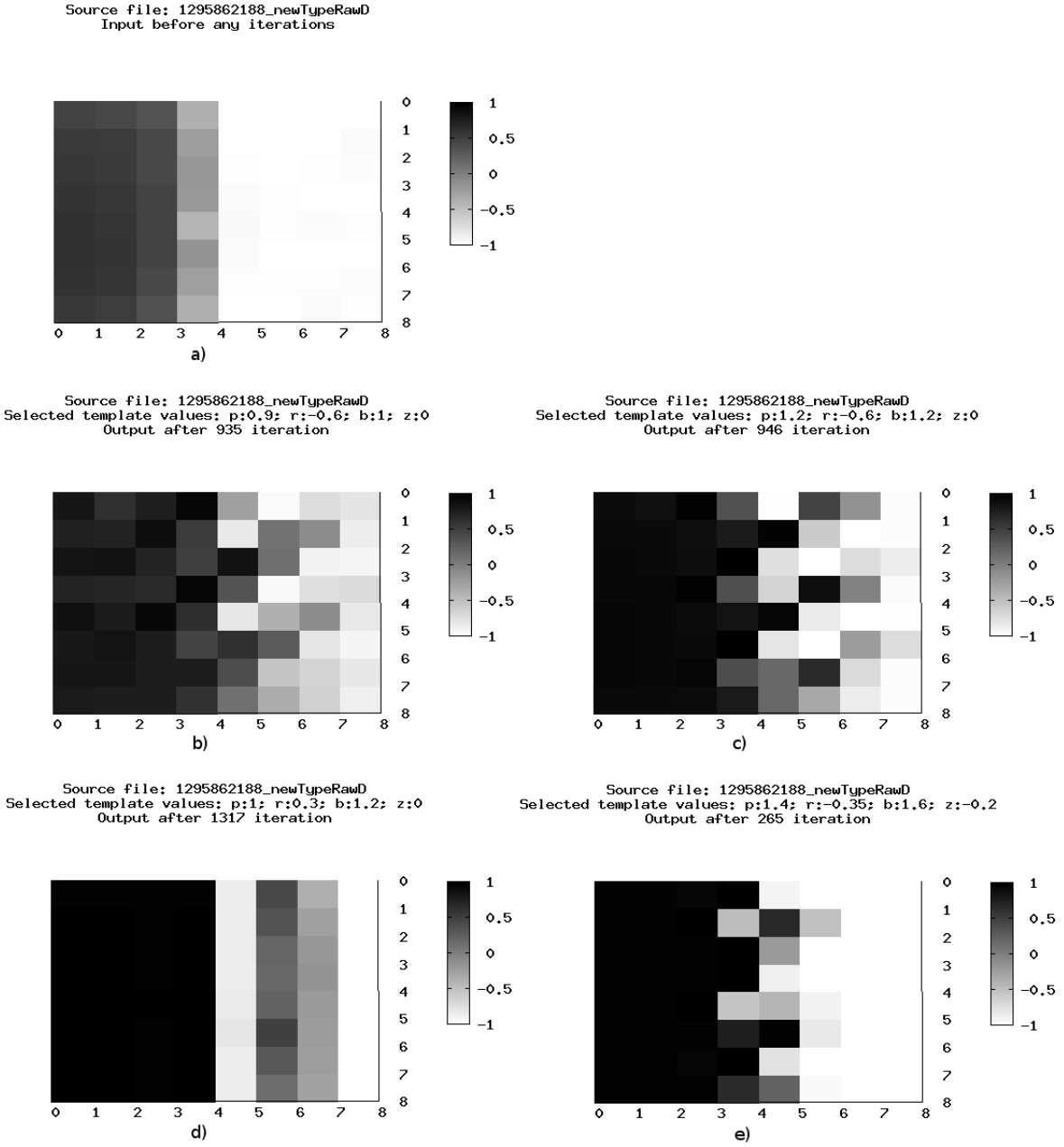


Fig. 1. The second measurement picture, where the input surface has three and a half column-wide masking rate with the sensor array left side: a). The other subfigures show some kind of snapshots from the spatio-temporal output flows, generated with the different templates: b) - with template **T1** (3); c) - with template **T2** (4); d) - with template **T3** (5); e) - with template **T4** (6)

Polyhedron based algorithm optimization method for GPUs and other many core architectures

Ádám Rák
(Supervisor: Dr. György Cserey)
rakadam@digitus.itk.ppke.hu

***Abstract*—In this paper, a new method of a compiler application to many-core systems is introduced. In this method, the source code is transformed into a graph of polyhedrons, where memory access patterns and computations can be optimized and mapped to various many core architectures. General optimization techniques are summarized.**

***Index Terms*—GPU, polyhedrons, hybrid-compiling, many-core, loops**

I. Introduction

New generation hardwares contain more and more processing cores, sometimes over a few thousand, and the trends show that these numbers will exponentially increase in the future. The question is how developers could program these systems and may port already existing implementations on them. There is a huge need for this today as well as in the forthcoming period. This new approach of the automation of software development may change the future techniques of the computing science.

The other significant issue is that GPUs and CPUs are started merging for the biggest vendors (Intel, NVIDIA, AMD). This means that developers will need to handle heterogeneous many core arrays, where the amount of processing power and architecture can be radically different between cores. There are no good methodologies for rethinking or optimizing algorithms on these architectures. Experience in this area is a hard gain, because there seems to be a very rapid (≈3 year) cycle of architecture redesign.

Exploiting the advantages of the new architectures needs algorithm porting which practically means the complete redesign of the algorithms. New parallel architectures can be reached by "specialized" languages (DirectCompute, CUDA, OpenCL, Verilog, VHDL, etc.), for successful implementations, programmers must know the fine details of the architecture. After a twenty years long evolution, efficient compiling for CPU does not need detailed knowledge about the architecture, the compiler can do most of the optimizations. Can we develop as efficient GPU (or other parallel architecture) compilers as the CPU ones? Will it be a two decade long development period again or can we make it in less time?

The specification of a problem describes a relationship from the input to the output. The most explicit and precise specification can be a working platform independent reference implementation which actually transforms the input from the output. Consequently, we can see the (mostly) platform independent implementation, as a specification of the problem.

Parallelization must preserve the behavior in the aspect of specification to give the equivalent results, and should modify the behavior concerning the method of the implementation. Automated hardware utilization has to separate the source code (specification) and optimization techniques on parallel architectures [1].

There are different trends and technical standards emerging. Without the claim of completeness, the most significant contributions are the following: OpenMP [2] - supports multi-platform shared-memory parallel programming in C/C++ and FORTRAN, practically it uses pragmas for existing codes. OpenCL [3] - is an open, standard C-language extension for the parallel programming of heterogeneous systems, also handling memory hierarchy. Threading Building Blocks of Intel [4] - is a useful optimized block library for shared memory CPUs, which does not support automation. One of the automation supported solution providers is the PGI Accelerator Compiler [5] of The Portland Group Inc. but it does not support C++. There are problem-software or language specific implementations on many-core architectures, one of them is a GPU boosted software platform under Matlab, called AccelerEyes' Jacket [6]. Overviewing the growing area, there are successful partially solutions, but there is no universal product and still there are a lot of open problems.

A. Parallelization conjecture

Given $\forall P_1$ non-parallel program, with the time complexity $\Omega(f(n)) > \mathcal{O}(1)$, $\exists P_M$ parallel implementation on M processors with the time complexity $\mathcal{O}(g(n))$ where:

$$\lim_{M,n\to\infty} \frac{f(n)}{g(n)} = \infty \tag{1}$$

From equation 1 we can derive a practical measure of how well an algorithm can be implemented sufficiently on a parallel system:

$$\eta(M) = \lim_{n\to\infty} \frac{f(n)}{g(n)} \quad M \geq \eta(M) \geq \mathcal{O}(1) \tag{2}$$

$\eta(M)$ can be called parallelization efficiency, it can describe in an abstract way how friendly the algorithm is with many-cores, or we can use it on a given implementation too. It can be seen that speedup can be achieved only if $\eta(M) > 1$. Theory indicates that we should set a more practical limit, like:

$$\eta(M) \geq \sqrt{M} \tag{3}$$

Brute-force algorithmic constructions for many-cores usually yield the limit seen in equation 3, we can say the algorithm is practically parallel if this equation holds.

II. Polyhedrons

Because computation time generally centers in loops, it is practical to depict the algorithm as a graph of loops. When we execute an algorithm on a parallel architecture, it usually boils down scheduling loops on cores. In order to optimize and schedule them better, we can convert loops into polyhedrons, where each execution of the loop kernel is a single point of the polyhedron, and the nodes are in a loop variable coordinate system. Polyhedrons, depicted by Π, are generic geometric shapes bounded by flat faces. The exact definition varies by context, but it is the following for our use:

$$\bar{x} \in \mathbb{N}^d \tag{4}$$

$$\mathbf{M} \in \mathbb{R}^{(d+1)\mathbf{x}n} : \mathbf{M} \cdot \begin{bmatrix} \bar{x} \\ 1 \end{bmatrix} \geq \bar{0} \tag{5}$$

$$\mathbb{P} := \left\{ \bar{x} \,\middle|\, \mathbf{M} \cdot \begin{bmatrix} \bar{x} \\ 1 \end{bmatrix} \geq \bar{0} \right\} \tag{6}$$

$$F^{\Pi}_{filter} := \mathbb{P} \rightarrow \{true, false\} \tag{7}$$

$$K_{kernel} := \{\partial_{R1} ... \partial_{Rn}, \mathcal{S}, \partial_W\} \tag{8}$$

Where x is a point of the polyhedron, matrix $\mathbf{M}$ defines the faces of the polyhedron, F^{Π}_{filter} is a plan-time executable function to decide the subset of polyhedron nodes we want to execute. K_{kernel} is the kernel operation we want to run in the polyhedron nodes. In equation 8 $\partial_{R1} ... \partial_{Rn}$ are the memory reads, $\mathcal{S}$ is the sequential arithmetic, and ∂_W is the memory write.

Because of the possibly overlapping memory operations we have to take into account the dependencies between nodes of the polyhedron, and possibly between polyhedrons too. We depict dependency set as $\mathbb{D}$.

$$\mathbb{D} := \{f_i | f_i : \mathbb{P} \rightarrow \mathbb{P}\} \tag{9}$$

$$f_i(\bar{x}) := round\left(\mathbf{D}_i \cdot \begin{bmatrix} \bar{x} \\ 1 \end{bmatrix}\right) \qquad \mathbf{D}_i \in \mathbb{R}^{(d+1)\mathbf{x}d} \tag{10}$$

$$F^{D}_{filter} := \mathbb{D}x\mathbb{P} \rightarrow \{true, false\} \tag{11}$$

A single memory access in the kernel is described in the following:

$$\partial_R, \partial_W : \text{read or write operation} \tag{12}$$

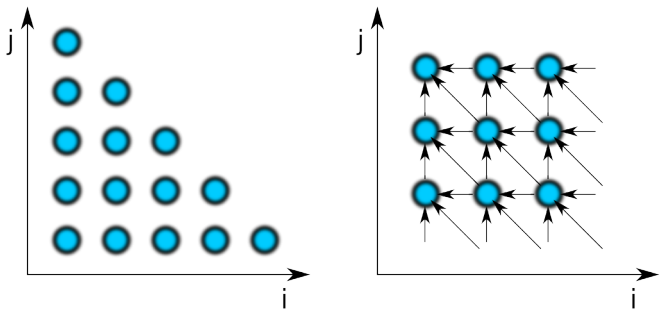
$$g : \mathbb{P} \rightarrow \mathbb{N}^n \tag{13}$$

$$g(\bar{x}) := round\left(\mathbf{A} \cdot \begin{bmatrix} \bar{x} \\ 1 \end{bmatrix}\right) \qquad \mathbf{A} \in \mathbb{R}^{(d+1)\mathbf{x}n} \tag{14}$$

Where a member of $\mathbb{N}^n$ is an n dimensional memory address, and $g(\bar{x})$ is the access pattern.

The following nominations and definitions are used in this paper:

Plan: $\mathcal{P}(\Pi)$ is a possibly parallel walk of the polyhedron, this can be generated in run-time.



(a) (b)

Figure 1: Geometric polyhedron representation of loops in the dimensions of i, j loop variables. Nodes depict the state of the loop, arrows depict the dependencies of the states.

Breakup: $\mathcal{B}(\Pi)$ depicts a decomposition of a polyhedron into another one, containing smaller data-flow.

Plan efficiency: $\eta(\mathcal{P})$ is closely related to parallelization efficiency. It well characterizes the hardware computing resources according to the plan.

Access pattern: $\partial(\mathcal{P}, \Pi)$ depicts the memory access pattern of the polyhedron taking the plan into account.

Access pattern efficiency: $\eta(\partial(\mathcal{P}, \Pi))$ characterizes the efficiency of the utilization of the hardware memory bandwidth.

A. Polyhedrons as loops

The following code sample is equivalent to the polyhedron given in the Figure 1(a):

```
for(int i=0;i<5;i++)
for(int j=0;j<(5-i);j++)
{
  S1(i, j);
}
```

B. Data-flow dependencies in polyhedrons

Data-flow dependencies can be represented by appropriate arrows between the nodes of the polyhedron, an example of equivalent code and its polyhedron (Figure 1(b)) is given in the following

```
for (int i = 0; i < n; i++)
for (int j = 0; j < m; j++)
{
  int sum = array[i][j];
  if (i > 0) sum += array[i-1][j];
  if (j > 0) sum += array[i][j-1];
  if (i > 0 and j > 0)
    sum -= array[i-1][j-1];
  array[i][j] = sum;
}
```

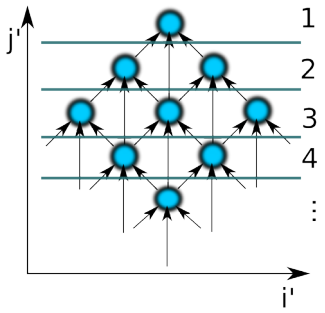


Figure 2: Polyhedrons can be partitioned into parallel slices, where parallel slices only contain independent parts of the polyhedron. For an example, rotating the polyhedron depicted in Figure 1(b), an optimized parallel version can be transformed.

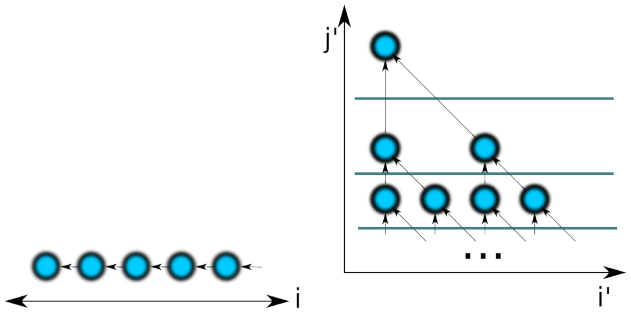


Figure 3: Polyhedron representation of the dot-product example (a). To rearrange the parentheses in associative chains give a possible solution for its parallelization (b).

C. Optimization with polyhedrons

Polyhedron representation of the loops provides a high level geometrical description where mathematical methods and tools can be used. Affine transformations can be applied on polyhedrons. These approaches can be used for multicore optimizations. Polyhedrons can be partitioned into parallel slices, where parallel slices only contain independent parts of the polyhedron. For an example, rotating the polyhedron depicted in Figure 1(b), an optimized parallel version can be transformed, see Figure 2.

For the optimization, the access pattern for cache locality, the appropriate transformations on polyhedron representation for parallelization and the control overhead has to be taken into account. Usually the system has to make decisions in runtime, which increases the control overhead.

III. Problems beyond polyhedrons

A. Dot-product, a simple example

Let us examine the simple dot-product example in the following code comparing to its polyhedron representation in Figure 3.

```
result = 0;
for (int i = 0; i < n; i++)
  result += vector1[i]*vector2[i];
```

Unfortunately, there is no usable polyhedron transformation for its parallelization at all. In this example, data-flow dependencies force a strict order of the execution. These dependencies are connected through an associative addition operator. The solution of this problem here is to rearrange the parentheses in associative chains.

B. Predictive lossy compression

A flowchart of a more difficult example, a simple method for predictive lossy compression can be seen in Figure 4. In this method, there is simple prediction from the previous element through a quantizer-dequantizer pair which applies division and multiplication. Its dataflow is linear. The parallelization of this algorithm is hard because there are strong data-flow dependencies, non-invertible and non-associative in critical paths.

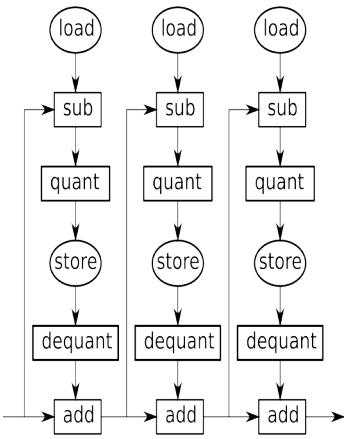


Figure 4: In this predictive lossy compression method, there is simple prediction from the previous element through a quantizer-dequantizer pairs which applies division and multiplication. The parallelization of this algorithm is hard because there are strong data-flow dependencies, non-invertible in critical paths due to the feedback loops and non-associative.

Experimental evidence suggests that number theory can solve this problem. In this case, dependencies can be transformed to additions of remainders, which is associative. This solution is very convoluted and non-intuitive.

IV. High-level hardware specific optimizations

Computational complexity usually concentrates in loops. Loops can be represented by polyhedrons, these are important building blocks of the program. Branching can be reduced into sequential code, sometimes these parts of the program can be built into the polyhedrons. After eliminating the side

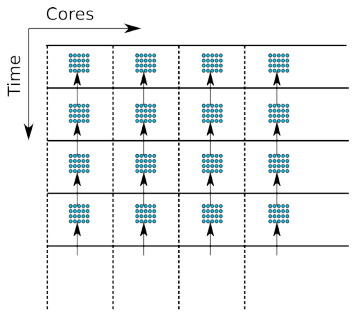


Figure 5: Scheduling polyhedron nodes to the time $\times$ core plain.

effects, the sequential code can be translated to pure data-flows. Applying these methods, only a bunch of polyhedrons - connected together in a data-flow graph - have to be optimized.

A. Kernel scheduling to threads

Kernel execution scheduling is a mapping of the polyhedron nodes to the symbolic plane of time $\times$ core id. Horizontal and vertical barriers can be defined on this plane. The projections of the horizontal barriers are synchronization points of the time axes. The core groups are separated by the vertical barriers. Dependency rules cross horizontal barriers parallel with the vertical barriers (Figure 5). After the barrier separation, the micro-scheduling is usually trivial in the groups.

A Plan is a parallel walk of the polyhedron defined below

$$\mathcal{P}_\Pi : \mathbb{N}x\mathbb{N} \rightarrow \begin{cases} \mathbb{P} \\ \emptyset(\text{NOP}) \end{cases} \tag{15}$$

Indexing of the plan is : $\mathcal{P}_\Pi(t,i)$ where t is time and i is the core id.

Plan efficiency defines the effectiveness of the usage of computing resources according to the plan:

$$\mathcal{P}(t,i) \quad t \in [1;T] \quad i \in [1;I] \tag{16}$$

$$\eta(\mathcal{P}) = \frac{|\mathbb{P}|}{T \cdot I} \tag{17}$$

B. Practical heuristics

1) Time complexity resulting from memory operations: $\mathcal{A}(\Pi)$ is the average size of the memory transfers in a single point of the polyhedron. S is the bandwidth of the memory hardware.

$$\mathcal{A}(\Pi) = \sum_i \frac{|\mathcal{P}_i| \cdot \mathcal{A}(\Pi_i)}{S \cdot \eta(\partial_i)} \tag{18}$$

2) Time complexity resulting from computation: $T(\Pi)$ is the average time complexity of a single point of the polyhedron.

$$T(\Pi) = \sum_i \frac{|\mathcal{P}_i| \cdot T(\Pi_i)}{\eta(\mathcal{P}_i)} \tag{19}$$

3) Typical problem decomposition:

1) Consider the algorithm as a set of loops and find the computationally complex loops
2) Analyze polyhedron geometry in the loop variable space
3) Discover dependences in-loop and between loops
4) Eliminate as many dependencies as possible
5) Quantify $\Pi, \partial, \mathcal{P}$
6) Estimate speed based on η
7) Optimize: transform $\Pi, \partial, \mathcal{P}$ to increase speed

V. Conclusion

Single polyhedron optimization can not solve the huge problem of parallelization, but based on our experiments it seems to be a promising basis. Determining more data-flow dependence operator primitives provides more handy tools to tune architecture specific algorithms. Problems can be treated more easily in their simplest forms, already optimized forms are usually hard to handle. There are quite many heuristics for hardware optimization, sometimes not even the manufacturer knows them completely. We are seeing exponential growth in core number (according to Moore's Law), very soon only $\eta(M) > \mathcal{O}(1)$ algorithms will be practical to implement for huge sizes.

The given model is designed for many core programming and the theoretical aspects were derived from practical experience on GPUs and FPGAs. The model can be easily extendable to FPGAs, Cell BE and CPU clusters.

Acknowledgment

The support of NVIDIA Professor Partnership Program and the Bolyai János Research Scholarship is gratefully acknowledged. The authors are also grateful to Professor Tamás Roska for discussions, his suggestions and his never ending patience.

References

[1] A. Rak, G. Feldhoffer, G. Soós, and G. Cserey, "CPU-GPU Hybrid Compiling for General Purpose: Case Studies," in *Proceedings of 12th IEEE CNNA - International Workshop on Cellular Nanoscale Networks and their Application*, 2010.
[2] L. Dagum, R. Menon, and S. Inc, "OpenMP: an industry standard API for shared-memory programming," *IEEE Computational Science & Engineering*, vol. 5, no. 1, pp. 46–55, 1998.
[3] A. Munshi, "The OpenCL specification version 1.0," *Khronos OpenCL Working Group*, 2009.
[4] J. Reinders, "Intel threading building blocks," 2007.
[5] M. Wolfe, "Implementing the PGI Accelerator model," in *Proceedings of the 3rd Workshop on General-Purpose Computation on Graphics Processing Units*, pp. 43–50, ACM, 2010.
[6] AccelerEyes, "Jacket: a GPU engine for MATLAB," 2009.

Transformation of Algorithmic Representations of the Fast Level-set Method between Virtual Machines

Gábor J. Tornai
(Supervisors: Dr. Tamás Roska and Dr. György Cserey)
torgaja@digitus.itk.ppke.hu

***Abstract*—This paper considers the model task of finding closed curves and surfaces in arbitrary dimension with the well-known level set framework. The problem is to find an optimal algorithmic representation of this task on different local-parallel architectures or to be more abstract on different parallel-cellular virtual machines. There are three main proposals of this paper. First, transforming optimally the sequential version of the algorithm or the task itself is not sufficient, the initial conditions must also be transformed. Second, running the algorithm considering the previous statement the required iteration number is a constant depending on the chosen initial condition. Third, a virtual machine which models GPUs is given. We have made investigations on three different machines: CPU, CNN-UM, and GPU. The theoretical results concordant with the implementations showed that on a local-parallel architecture the iteration number can be smaller than 4.**

***Keywords*-level set, algorithm mapping, many core, parallel architectures, GPU, CNN-UM, virtual machine**

I. Introduction

In many applications of practical interest, we often wish to find or trace closed curves or (hyper)surfaces around a region. This task is well handled with the level set framework [1]. The family of these methods are used in computational geometry, fluid mechanics, computer vision and materials science [2]. In this paper I worked with a subset of these tasks, where the exact time evolution of the model is out of our interest but the steady state solution of the given PDE should be approximated. This subset has especially great importance in computer vision, object detection and image processing.

This work was inspired by the technology response to emerging physical constraints, the CNN paradigm, and a sequential algorithmic representation of the above mentioned subset task [3]. It is clear now that many core locality is not an option but a must. However, with this technology shift a lot of problems emerge as well. First, algorithmic complexity for Turing machines, which described the behavior of μ-recursive functions, fails. Second, how can a given problem be adopted to a specific machine? Third, which machine shall be chosen for a specific problem class to be optimal? Finally, how can optimality be defined? The second question can be answered partially by the following statement. Transforming an optimised sequential algorithm to any other machine does not lead to an efficient algorithm.

Another inspiration came from the above mentioned efficient rule based fast level set algorithm. The invention in this algorithm was the dimensional reduction of the computational needs and of course the rule based computation. I have adopted the rule based approach for two different machines. One is the CNN-UM virtual machine [4] and the other is the GPU [5].

The paper is organised as follows. In the second section after the precise subclass of the tasks was defined, a brief summary from the used machines is given. The third section explains the machine specific algorithms. The fourth section summarises the results and then a conclusion follows.

A. Interesting observations and glimpse of the main results

If one investigates the CNN-UM transformation [4] and the GPU based adoptation [5] some interesting observations can be made. First, on CNN-UM time requirement for one step *does not depend on the front size*, and the *number of iterations* for reaching the stable state *decreases asymptotically* as the size of the initial front (near the zero level set) increases. Second, until we have not run out of the GPU memory (bandwidth and size), the *runtime was independent from the problem size* and it followed a linear characteristics. These transformed algorithms are described in the III-B and III-C subsections respectively.

On locally connected many core architectures, the required initial condition differs from the classical sequential one which is optimized for reducing the dimensionality of the problem. (2) The number of steps are controlled with the machine specific initial condition. In this setup the maximum number of steps can be smaller than four. (3) Until a task fits into the memory (bandwidth and size) of the machine the computations may be hidden with the memory transactions, then the growth is linear with respect the problem size.

II. Background

A. Formulation of the investigated subset problem

The original level set method was first described in 1987. It has become an efficient tool so far giving a natural solution to many problems. This is done by embedding the curve $\gamma(s,t)$ to be found in a higher dimensional function, and represent it as the zero level set. The basic model for the level set evolution is a simple one: $\frac{\partial \phi}{\partial t} + \vec{F} \cdot \nabla \phi = 0$ Where F can be any arbitrary function and the underlying curve is the zero level set of ϕ that is going to be denoted as the level set function.

According to the considerations in [3] one might omit the exact solution of the underlying PDE and use a rule based approach. Let us assume that ϕ is defined over a domain D $\in R^K$ $(K \geq 2)$. One can define two sets namely L_{in}, L_{out} as follows:

$$L_{in} = \{\mathbf{x} | \phi(\mathbf{x}) < 0 \; and \; \exists \mathbf{y} \in N(\mathbf{x}) \; that \; \phi(\mathbf{y}) > 0\} \quad (1)$$

$$L_{out} = \{\mathbf{x} | \phi(\mathbf{x}) > 0 \; and \; \exists \mathbf{y} \in N(\mathbf{x}) \; that \; \phi(\mathbf{y}) < 0\} \quad (2)$$

$$N(\mathbf{x}) = \{\mathbf{y} \in D | \sum_{k=1}^{K} |y_k - x_k| = 1\} \; \forall \mathbf{x} \in D \quad (3)$$

The level-set function itself is an approximated signed distance function near the zero level set:

$$\phi(x) = \left\{ \begin{array}{ll} -3, & \text{if } \mathbf{x} \text{ is inside } \gamma \wedge \mathbf{x} \notin L_{in} \text{ inner points} \\ -1, & \text{if } \mathbf{x} \in L_{in} \\ 1, & \text{if } \mathbf{x} \in L_{out} \\ 3, & \text{if } \mathbf{x} \text{ is outside } \gamma \wedge \mathbf{x} \notin L_{out} \text{ outer points} \end{array} \right\} \quad (4)$$

Having these two sets, the motion of the zero level set can be obtained by investigating only the sign of the force field. In this way replacing elements from one set to the other the desired motion is received.

B. Machines

1) CPU: Until the beginning of the past decade CPU was the dominant way of computing. The reason was the classical scaling down. Fortunately, the computational theoretical constructs, was laid down by Turing and Kleene introducing and analizing the Turing-machine and the equivalent constructs. Computational complexity classes such as P, NP are definite.

2) CNN-UM: The CNN like architectures are handled with the Cellular Nonlinear Network-Universal Machine (CNN-UM) or equivalently the Universal Machine on Flows (UMF) [6] construct. This construct is Turing equivalent. Unfortunately, computational complexity may depend heavily on input flows. However, an input independent worst case can be given. As an example let us consider a binary image where connected black pixels are denoted as object. In this case a recall operation depends heavily on object shape, shadow depends on position.

3) GPU: In the case of GPU a good virtual machine model capacle of describing algorithms is still has been looked for. The physical model is symmetrical hierarchical processor arrays with embedded hierarchical memory organization. The OpenCL language is the only well-defined abstract model. However, it is unsuitable for quantitative analysis.

Let us consider a tree graph with the following topology. The root element has arbitrary but fixed number of children going to be denoted as sharing nodes. The sharing nodes have arbitrary but fixed number of children these will be referred to as computing nodes. The edges are weighted. The weights are the same between the root and the sharing points, so is for the next level. The root and the sharing nodes contain a tape which can be accessed with delay. The ammount of the delay is the edge weight. The root tape is infinite, the tapes of the sharing nodes are bounded (arbitrary but fixed). The computing nodes are modified Turing machines. The properites are the followings: (1) Turing machines of a sharing node share their states with infinite speed (2) have the same partial function (3) each machine has its own tape (4) can make state transition at the same time iff their state is the same (5) can read and write the tapes in their sharing node and also the root tape.

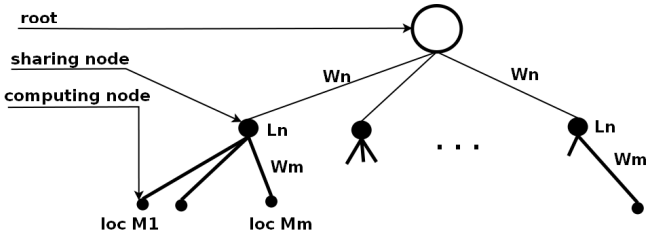


Figure 1. This graph abstracts the architecture of a GPU. It has one root element, n piece of sharing nodes and $n \cdot m$ leaves. The leaves are going to be referred to as computing nodes. Each computing node is a Turing machine (loc M_1 ... loc M_m) with finite length tape. Turing machines of a sharing node shares their states among each other. They step at once if, and only if they are in the same state. Only one machine is selected otherwise. The sharing nodes contains finite tapes only the root element has an infinite tape.

From pure comutational complexity point of view this is equivalent with a Turing machine. The great difference as mentioned in the I-A subsection is the exponentially increasing amount of elements in a device.

III. Comparing the task related algorithms

By investigating the sequential algorithm, four different stages can be separated and a few notes can be made. The four well-distinct operations are as follows: update L_{out}, update L_{in}, clean L_{out}, clean L_{in}. First note, there are two dependencies namely update L_{out} shall precede clean L_{in} and similarly update L_{in} shall come before clean L_{out}. There are no data dependencies among the elements of the active front so a loop operation can be parallelized easily and effectively.

A. CPU

First I must empasize again that this subtask is an *approximation* of the Level-Set PDE only the steady state solution is desired. Motion of the implicitely represented curve can be obtained by manipulating the two sets (L_{in} and L_{out}). At every set element the sign of the force field (F) is checked. If the sign is bad then one of the following operators is applied (for $x \in L_{in}$ $switch\_out$ and for $x \in L_{out}$ $switch\_in$).

switch_in(x): (1) delete x from L_{out}, (2) add x to L_{in}, (3) set $\phi(x) \leftarrow -1$, (4) compute $F(x)$. Then $\forall y \in N(x)$ if $\phi(y) = 3$ add y to L_{out}, set $\phi(y) \leftarrow 1$, compute $F(y)$.

switch_out(x): (1) delete x from L_{in}, (2) add x to L_{out}, (3) set $\phi(x) \leftarrow 1$, (4) compute $F(x)$. Then $\forall y \in N(x)$ if $\phi(y) = -3$ add y to L_{in}, set $\phi(y) \leftarrow -1$, compute $F(y)$

- initialize arrays for ϕ, F; bidirectionally linked lists for L_{in} and L_{out}.
- Evolving cycles: for $i = 1 : N_a$ do
  1) scan every point $x \in L_{out}$

– if $F(x) > 0$ then $switch\_in(x)$

2) eliminate inner points from L_{in}
3) scan every point $x \in L_{in}$
   – if $F(x) < 0$ then $switch\_out(x)$
4) eliminate outer points from L_{out}
5) check stopping condition

With this algorithmic representation the complexity of the algorithm is $O(l)$, where $l = |L_{out} \cup L_{in}|$. For proof please read [3]. It can be seen that for the majority of the cases $l \approx n^{k-1}$, where n is the number of samples in one dimension or the bounding box of the object, and $\dim(data) = k$.

B. CNN-UM

The CNN-UM algorithm is based on the set theoretic description of the level set function. That is why two additional sets are introduced to cover all the values in ϕ:

$$F_{in} = \{\mathbf{x} \in D | \phi(\mathbf{x}) < 0 \wedge \mathbf{x} \notin L_{in}\} \quad (5)$$

$$F_{out} = \{\mathbf{x} \in D | \phi(\mathbf{x}) > 0 \wedge \mathbf{x} \notin L_{out}\} \quad (6)$$

The algorithm requires five binary images to start level set evolution. Four of the five images are the initial state of ϕ more precisely: F_{in}, L_{in}, L_{out}, F_{out}. The fifth image is the generated force field. Here we are not dealing with how the force field is generated.

1) calculate F
2) *update L_{out}:* This is one major part of the algorithm. First $F_{outmask}$ is computed which is the part to be stepped outward. By means of $F_{outmask}$ new L_{out} is generated using dilation and logical operators. Then $F_{out}$is updated.
3) *clean L_{in}:* Redundant points from L_{in} are eliminated and F_{in} is updated.
4) *update L_{in}:* First F_{inmask} is computed – the parts to be stepped inward. Dilating F_{inmask} and then using logical operators and the next piece of L_{in} is obtained. Then F_{in} is stepped according to L_{in}.
5) *clean L_{out}:* Redundant points of L_{out} are eliminated, lastly F_{out} is refreshed.
6) *stopping conditions:* There are three stopping conditions. If either is fulfilled, the algorithm stops. Otherwise, the algorithm steps again to 1^{st} step. The stopping condition:
   - $F(\mathbf{x}) \leq 0 \, \forall \mathbf{x} \in L_{out}$ and $F(\mathbf{x}) \geq 0 \, \forall \mathbf{x} \in L_{in}$
   - An oscillatory state is reached.
   - The pre-specified number of iteration is reached.

C. GPU

The algorithm operates directly on the level set function. It is divided into subregions among the sharing nodes. Each subregion is distributed further to the computing nodes. This subregions are stored in equally sized pieces. Exact sizes may depend on the physical realization of the machine.

The first difference regarding the other two algorithms is as follows: the two evolution operation (Update L_{out} and Update L_{in}) is done simultaneously in one piece of area, then come the two cleaning operations for the same area. The second main difference is the distributed nature of the data along the sharing node tapes and computing nodes. The algorithmic steps in the computing node level are as follows:

1) initialize
2) visit area points and change the value of the level set function according to the rules
3) visit area points and clean according the local rules
4) recalculate F
5) check stopping condition if not go to second step

The information is used that there are no dependencies between the two update operators and the two cleaning operators as described in the beginning of this section.

IV. Results

A. Investigation of the required number of iteration

First some basic quantity is defined then the required iteration number is derived for a given initial condition configuration.

Definition 1. *Inner distance index (d_i^x) of a given pixel (x) is a positive number denoting the distance of the closest positive pixel according to the connectivity scheme.*

Definition 2. *Outer distance index (d_o^x) of a given pixel (x) is a natural number denoting the distance of the closest negative pixel according to the conectivity scheme.*

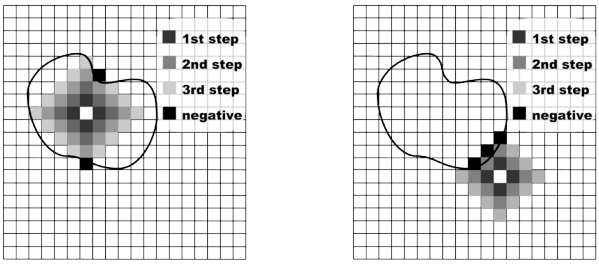
Definition 3. *General distance index (d_g^x) of a given pixel (x) is a natural number denoting the distance of the closest pixel of the opposite sign according to the connectivity scheme.*

Definition 4. *Distance index function, δ of a discretised region D and a given configuration (ϕ) is as follows. $\delta : (x \times \phi) \mapsto d_g^x$, where $x \in D$ and ϕ is the level set function.*

Before we proceed let us take a note. There is a difference between step and iteration. A step is a local event meaning the level set function has moved in-, or outward. An iteration is one repetition of the four stages of the meta algorithm.

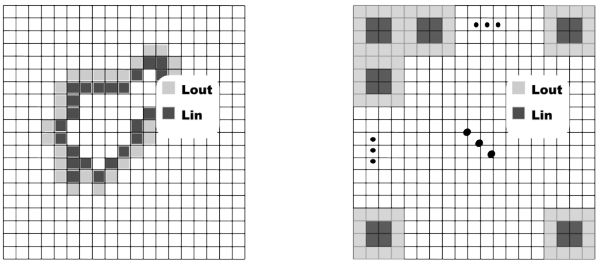
Distance indices have the following meaning: they are the minimum required iteration or equivalently the number of steps for the level set function to change sign. (Again, minimum means: in every iteration the zero level set moves toward the investigated pixel.) It can be seen that only the points inside (outside) the curve have inner (outer) arrival index, but all pixel has general arrival index. It is true that required # Update $L_{out} \leq \max d_o^x \phi(x) > 0$ and # Update $L_{in} \leq d_i^x \phi(x) < 0$ to reach the given pixel. An illustration of the two distance indices can be seen in Fig. 2.

The maximum of δ is the upper bound of steps that the level set function must take. We must make an additional note. A force field that does not make step in every iteration can be constructed easily. However, such force field may not be really useful in the field of image processing – not counting the smoothness regularization term. In the case of a sequential algorithm this can be approximated with Gaussian filtering [7] applied on the reached steady state with a force missing the



(a) inner distance index (b) outer distance index

Figure 2. In this figure the distance indices are illustrated. Distance index is plotted for the central white pixel. The absolute black pixels on (a) are the nearest pixels with opposite sign and so is for (b). It can be seen that the value of the distance index is definite and clear. The grey pixels denote the distance from the central white pixel.



(a) initial condition for sequential machine (b) initial condition for parallel machine

Figure 3. Initial conditions (a) optimized for sequential machine and (b) for parallel machine. In the first case the size of the active front is relative small compared to the whole image. On the parallel one the whole image is an active front so the computing width of the parallel machine can be utilized. The reason is the difference in the $\max(\delta)$ value. On (a) it is proportional with the size of the width of the picture, but on (b) it is one. This makes the difference in the required number of iterations.

curvature dependent part. An equivalent operation can be done extremely fast on a local parallel architecture on the whole image as a preprocessing step, or during the level set evolution. With commonly used force field terms the required number of iterations can be upper bounded by the $\max(\delta)$ quantity.

B. Inital condition transformation

With the δ function the last sentence can be proven in the abstract. The strategy is as follows. First, a short description for the optimal initial condition on a sequential machine is given. Second, initial conditions for local parallel machines with as low $\max(\delta)$ as possible are constructed.

As descibed in the last paragraph of III-A subsection from the complexity point of view an optimal initial condition for a sequential machine has a dimensional reduction effect. If the algorithm visits a wide neighborhood of a surface in a volume then the complexity is proportional with the size of the surface. There are important questions to be answered: 'Is the steady state of the active front inside this given wide band?' and 'How shall be the width choosen to preserve the dimensional reduction nature for tipical resolutions?' lastly 'How shall be the width choosen to incorporate the steady state?'

On a local parallel machine it is totally different. There are a lot of processors working in parallel. This gives the possibility to investigate a greater area of active front simultaneously. So the initial condition may be changed. This is illustrated in Fig. 3. A sequential machine optimized initial condition has a small active front relative to the image. On a parallel machine optimized one the whole area is an active front so the computing width of the parallel machinecan be utilized. This implies that arbitrary initial condition can be chosen. Then why not choose initial conditions with minimal $\max(\delta)$, which is *1*. In two additional iteration can be checked that the active front is in an oscillatory cycle so the last statement of the abstract holds.

We have seen that a parallel machine optimized initial condition has lost the dimensional reduction nature of the sequential one. If the problem size tends to infinity the sequential initial condition outperforms the parallel one, but this is only true if the number of parallel processors stay constant which is obviously not true. The number of processors increases exponentially in the field of digital CMOS designs. To use hundreds or thousands of locally connected processors the sequential algorithms have to be corrupted in a special way to fit into this new paradigm.

V. Conclusion

This paper presented algorithmic representations of the well-known level set algorithm on different parallel-cellular virtual machines. It is shown that transforming the sequential version of the fast level set algorithm or the task itself is not enough, the initial conditions must also be transformed. Moreover running the algorithm considering the previous statement the required iteration time is a constant depending on the chosen initial condition. We have made investigations on three different machines: CPU, CNN-UM, and GPU. Also a virtual machine which models GPUs is given. The theoretical results concordant with the implementations showed that on a local-parallel architecture this constant can be smaller than 4.

References

[1] J. A. Sethian, *Level set methods and fast marching methods: evolving interfaces in computational geometry, fluid mechanics, computer vision, and materials science*. Cambridge University Press, 2000.

[2] G. Sapiro, *Geometric partial differential equations and image analysis*. Cambridge Univ Pr, 2001.

[3] Y. Shi, *Object based dynamic imaging with level set methods*. PhD, Boston Univ. College of Eng., 2005.

[4] G. J. Tornai, G. Cserey, and A. Rák, "Spatial-temporal level set algorithms on CNN-UM," in *International Symposium on Nonlinear Theory and its Application, (NOLTA 2008)*, pp. 696–699, 2008.

[5] G. J. Tornai and G. Cserey, "2D and 3D level-set algorithms on GPU," in *Cellular Nanoscale Networks and Their Applications (CNNA), 2010 12th International Workshop on*, pp. 1–5, IEEE, 2010.

[6] L. O. Chua and T. Roska, *Cellular neural networks and visual computing, Foundations and applications*. Cambridge University Press, 2002.

[7] A. Witkin, "Scale-space filtering," in *Proceedings of the Eighth international joint conference on Artificial intelligence-Volume 2*, pp. 1019–1022, Morgan Kaufmann Publishers Inc., 1983.

Object and pedestrian detection with monocular camera system

Tamás Fülöp
(Supervisor: Dr. Ákos Zarándy)
fulop.tamas@itk.ppke.hu

***Abstract*—Object detection and recognition is one of the most difficult and unsolved problems in computer science. This report shows my object detection and recognition algorithms which are able to recognize pedestrians in an optimal environment. The pedestrian recognition is one of the most important tasks of image processing. Robust algorithms are not yet completed. My work is a step towards a working algorithm.**

***Index Terms*—monocular camera system, image processing, object detection, recognition**

I. Introduction

Object detection is one of the most difficult image processing problems and it is a key initial step of all security-surveillance and other monitoring algorithms. Many algorithms exist, which try to solve segments of the problem, but their efficiency is not good enough, and they had many constraints [6]-[14].

One of the most difficult question is how we can develop robust, environment independent algorithms. Many monitoring systems work on stationary platform, where the segmentation is a simple foreground-background separation, which is illumination and background model dependent.

Moving platform video analytics is a more difficult problem. We cannot make it reliable just in time background model. These models need a large computation capacity (segmentation, correspondence and decision).

The active sensor technologies, like LIDAR are expensive, while cameras are cheap. Many monitoring applications must to told: what the object is. A video system can give more than binary information about objects.

Increased numbers of active devices cause noise level growing, thanks to the interference between devices. The active systems need better filter algorithms and other algorithm development to avoid noise and false alarm.

Using passive devices, like cameras, can be a confirmation of the active devices.

In this semester I worked on a real time monocular pedestrian detection system, which can be implementable on shared architecture too, and I start to look around, the possible cooperation strategy with other sensor systems.

II. BASICS

A. General object detection method on moving platform

Segmentation is the key initial step of all object detection algorithms. On stereo camera system can be used various corresponding techniques, but it needs large computation capacity, and more place to use it, but we can extract 3D information about objects.

The monocular stationary camera platforms have got an extensive literature. Using monocular systems is a constraint, because we cannot extract exact 3D information without motion and other additional information. This is a simple prediction about the space with more preliminary information needed for prediction e.g.: what the background is, what the possible objects are, or what we want to detect.

The basic foreground-background separation based systems are using a reference image, when no moving objects are on scene. It needs a stable camera, with constant illumination conditions and a good background model. The algorithms are usually a simple frame differencing between captured and a background image. Where the algorithm finds difference between images, there is a possible object. These algorithms are illumination dependent, and there have many adaptation problems. It can be neutralized with special methods, like that systems, which working with local pixelrelation [1]. The problem with this and other systems is, how we can built a stable background model fast enough , because the optical flow can be spoil on moving platform. Unfortunately, this model is not applicable for moving platform detection.

Other way, when we use segmentation techniques on split images. Many segmentation techniques exist with different strengths. When the segmentation method cuts an object in half it causes problems, because we cannot correspond the parts automatically.

After segmentation and using any corresponding techniques, we can start the recognition process. It can be a dummy process, when we known the geometry of an objects

and simple superpose it.

As we shown, we need a suitable procedure for detection, but the monocular system cannot generate clear a result, so the result will be depending on many knowledge about the scene and any background information. I would like to show that how my algorithm works, how can I extract information from an image.

1) Capturing

This step is one of the most important. The image with some compression methods (like JPEG) can leave compression artifacts, that can raise errors on following steps, like the segmentation boundary.

Very important thing is the color fidelity of an image. Sometimes it is enough to use black and white pictures, but as it has been observed, the color information can be useful for preliminary decision. HSV color scheme was used in this case, because the V channel carried the light value, like a B&W image.

2) Segmentation

Segmentation is an initial step of working on an image. Detecting objects is a hard and not a solved problem in monocular stationary camera system too.

Typical segmentation problems:

- Light and shadow,
- Contrast,
- Color difference,
- Texture difference,
- Twinkle,
- Noise.

I found many segmentation methods, with different efficiency (clustering type, histogram-based, edge detection based, region growing, level set, graph partitioning method). Some of them are enough fast, but more of them are too slow to use for real time segmentation. The common problem with all of them: they are not are able to make adaptive segmentation, which is quite important for a noisy and dynamically changing environment. Most of the methods have found edges. There were strongly depending on light conditions. We must to see that, this is the trade of, or need automatic parameter changing if possible.

3) Hypothesis generation

The segmentation has got an error. When we start the hypothesis generation on over segmented picture, that can decrease the number of problems. After it, we can correspond the small items that can be dependent on the edge, color, texture and from other properties. The algorithm can get additional correspondence to make a better result, when the method makes correspondence between frames in a movie. This is tracking, when the algorithm tries to find the segmented parts which move together.

4) Detection

Algorithm needs a decision, how to associate the different segments. The method is not trivial, it depends on many things, such as size, texture, color, brightness, but more sophisticated algorithm can correspond to the parts with their edges. The orders should present a serious problem, however: the correspondence is not safe.

5) Recognition

After detecting structures, we want to know what ther are. Recognition is important when we want to make a reaction when something happens, because some object must generate alarm, some of them should not. Recognition is not trivial, because it depends on the quality of the detection steps. The clearer the detection, the more precise the recognition is.

The recognition is usually a slow process. Algorithms can be faster with a Region Of Interest (ROI) selection method, when some properties signalizing the algorithm.

B. Present techniques to recognizing pedestrians

The pedestrian detection is an important task in most security systems, especially in driving aid systems. A number of related surveys exist with different focus, how to recognize pedestrian: MIT, Daimler AG with Iteris (USA), Volvo with Mobileye (Netherland), many universities, etc.

Some algorithms are working with stereo camera system, and others working as a monocular camera system. The presently working algorithms are not efficient enough , the true positive result is between 60% and 90%. More reliable and widely used the infrared camera systems, which working well only in night.

These algorithms use a different strategy for recognizing a pedestrian. Some algorithms use special phenomena like motion patterns of walking. Other methods are based on a good image segmentation and use generative models (appearance of pedestrian class in terms of its class conditional density function) or discriminative models (Bayesian maximum-a-posteriori decision by learning parameters of a discriminator function between the pedestrian and non-pedestrian classes) for recognizing. [6]-[14] The generative models are faster, but they have got several application constraints.

III. WAY TO RECOGNIZE A PEDESTRIAN

The difficulty of pedestrian detection is the environment itself. We mustu recognize the objects and select a part of them. Altough everybody thinks that, a pedestrian is a well defined object in space, but the truth is , a walking person has got a number of properties (size, color (skin, dress, depending from light conditions), position). Some elements are predictable (domain and ratio of size, domain of color), but many different object can be “pedestrian like”.

A. Capture

RGB color scheme is a usual way for capturing picture, but many ROI operations are working better and easier in HSV, where the color information is represented in a ring. The RGB to HSV transformation is trivial. My algorithm is working with HSV.

Firstly I made some recordings by car. I captured a color movie with a Philips Web Camera and a b&w movie with Eye-RIS system. Unfortunately, both movies had a great quantity of noise. Color movie had got huge number of artifact, while the captured b&w movie had got brightness problem, because Eye-RIS camera system can't change their integration time automatically.

I started my work with color movies. Firstly, the Hue channel can be used for fast ROI selection. Some article show, the Hue of skins is the same: race and brightness independent [4][5]. I found that, the hair color behave similarly. When I mapped the possibilities and much bottleneck of my own movies I have started to find other solutions. Daimler AG published their pedestrian corpus for free academic use in 2009. [3] It contains large resolution (~20000 frames) uncompressed b&w movie with typical situation, so I have started to use it for my work.

Fig. 1. Result of graph based segmentation technique.

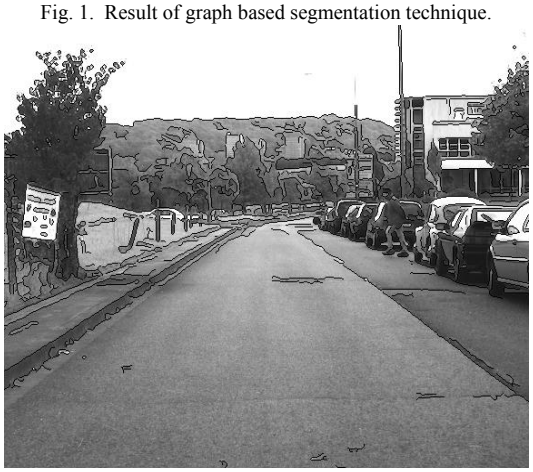


Fig. 2. Canny-edge detection can recognize most of real edges. Combine with graph based segmentation is makes better result. The black lines the result of edge detection

B. Segmentation

I tried many segmentation methods. I tried to find the fastest and the best. The speed is an important aspect in my work, because I need real time recognition. I tried a different way of segmentation algorithms, but finally I used a fast, graph based algorithm [2], which can be implementable on many core processor chips and FPGA too. This algorithm is fast enough, and well-parameterized. Fig 1. show a segmented image with this algorithm, where the car goes ahead on a road.

I found a good parameter set-up for the algorithm, which can detect the most objects, but always has got problem with small parts. Other way, more sensitive setup causes too much small parts, because the basics of the algorithm cannot handle well the gap of edge. So I used a less sensitive setup with a correction. Before using segmentation method the algorithm made a canny filter based edge detection for smaller parts. My method sign the edges as different color points as different, as Fig 2 shows. After it, graph based method can select the not detected regions.

I used MATLAB for testing, and I get an over-segmented picture.

Fig. 3. The result of segmentation process. The pedestrian is on road.

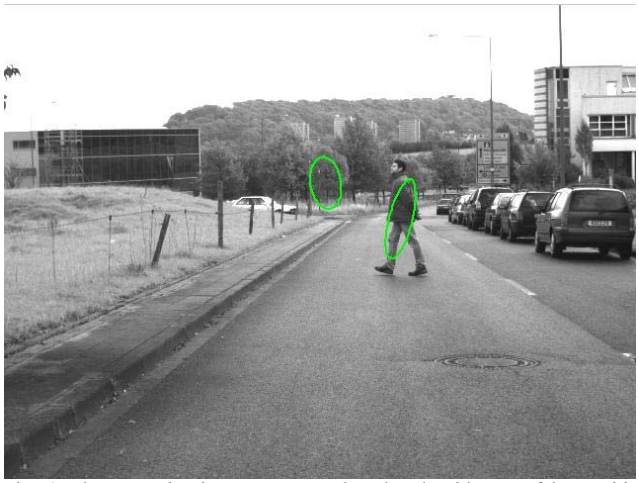


Fig. 4. The recognize is now success, but the algorithm get false positive result.

C. Corresponding

My algorithm using a simple corresponding technique. I examined how a pedestrian built up, what is the typical ratio. If the algorithm found possible correspondable points, it will be signed it for themselves for detection phase. Corresponding is now the bottleneck of my algorithm, because it need to built a neighborhoods to know their environment and possible correspondence.

D. Detection

My detection is blurred sometimes with the correspondence. Detection firstly makes a ROI selection under Corresponding too. When correspondence find a too big structure (like sky, big house, road) it signed as a typical background objects.

E. Recognition

One of the most important thing, how can the algorithm make fast decision. The generative pedestrian model has got many well recognizable sign, as their shape. I can easily extract informations from unique or corresponded space as X-Y size, edges, orientation. The ratio of a human is near 2:1. This ratio is depending on from the position of walking and the physique of human. The human leg position determine the orientation

IV. Results

The segmentation is working, but the false alarm rate is too high at present. This is the reason why my corresponding technique is not efficient enough for the different parts. Now the algorithm needs either a better segmentation set-up for less segments, either a better (and faster) segmentation technique.

The generative recognition works well on properly segmented images, which shows that the most critical point is the segmentation-correspondence. I made some modifications on the algorithm, because the segmentation result depends on the distance of objects, but it is not enough. A human being is segmented as arms, body, legs, head. I can detect the properties of these parts of body, and I can signal well.

My working generative model based result corresponds to the newly developed results of the company Mobileye, which Volvo builts in their cars. The company founders used a generative style model for decision making. [6] This system seems to be working poperly and now everybody can buy it alone or with a car in a Volvo. I have spoken with the main engineer of Volvo Hungária Ltd. who informed me that this system does not work all time, because it has got a very narrow range of operation. It is calibrated for only the most critical situations.

V. Conclusion

Passive systems are capable of being used for detection and recognition, and my own algorithm shows a good solution. The Volvo is a market leader car manufacturing company with their safety systems. Their roadmap shows that their pedestrian safety project runs till 2015 indicating that, this is a really significant problem, without any trivial solutions.[15]

VI. Future work

The algorithm needs to be repaired. An Other way would be the passive sensor as I have shown, but it cannnot detect any information about space, so it need to be examined how some cooperation strategy cen be built between active or other passive sensors and camera systems. This is also not a trivial question because it needs some research about possible cooperation strategy, hierarchy.

Acknowledgment

The author is also grateful to Ákos Zarándy for the discussions and his suggestions. The author expresses their thanks to Furukawa Electric Institute of Technology for all its support.

References

[1] Tamás Fülöp and Ákos Zarándy Real-time moving object segmentation algorithm implemented on the Eye-RIS focal plane sensor-processor system

[2] Pedro F. Felzenszwalb and Daniel P. Huttenlocher, Efficient Graph-Based Image Segmentation , International Journal of Computer Vision, Volume 59, Number 2, September 2004

[3] Daimler Pedestrian Detection Benchmark Dataset – M. Enzweiler and D. M. Gavrila. "Monocular Pedestrian Detection: Survey and Experiments". IEEE Trans. on Pattern Analysis and Machine Intelligence, vol.31, no.12, pp.2179-2195, 2009.

[4] Scott Leahy, Face Detection on Similar Color Photographs, Scott Leahy's work, Stanford University 2003

[5] H. Andersen, and E. Granum , Estimation of the illuminant colour from human skin colour M. Störring,. IEEE International Conference on Face & Gesture Recognition, pages 64-69, Grenoble, France, March 2000. IEEE Computer Society

[6] Amnon Shashua. Yoram Gdalyahu. Gaby Hayun , Pedestrian Detection for Driving Assistance Systems: Single-frame Classification and System Level Performance. (Founder of mobile-eye)

[7] M. Enzweiler, D. M. Gavrila, Monocular Pedestrian Detection: Survey and Experiments,IEEE Trans. On Pattern Analysis and Machine Intelligence, Volume 31, No 12, Dec 2009

[8] D. M. Gavrila, Multi-cue Pedestrian Detection and Tracking from a Moving Vehicle, International Journal of Computer Vision, Vol 73, 2007

[9] M. Enzweiler, P. Kanter, D. M. Gavrila, Monocular Pedestrian Recognition Using Motion Parallax, Intelligent Vehicles Symposium Eindehoven 2008

[10] M. Enzweiler, D. M. Gavrila, A Mixed Generative-Discriminative Framework for Pedestrian Classification

[11] S. Munder, D. M. Gavrila, An Experimental Study on Pedestrian Classification, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol 28. No. 11, Nov. 2006

[12] C. Pai, H. Tyan, Y. Liang, H. Mark Liao, S. Chen, Pedestrian detection and tracking at crossroads, The Journal of The Pattern Recognition, Vol. 37, 2004

[13] C. Curio, J. Edelbrunner, T. Kalinke, C. Tzomakas, Werner von Seelen, Walking Pedestrian Recognition, IEEE Transactions on Intelligent Transportation Systems, Vol. 01, Sep. 2000

[14] D. M. Gavrila, Pedestrian Detection from a Moving Vehicle, Proceedings ot the European Conference on Computer Vision, Dublin

[15] Volvo Car Corporation

Collision avoidance for UAV using visual detection

Tamás Zsedrovits
(Supervisors: Dr. Tamás Roska and Dr. Ákos Zarándy)
zseta@digitus.itk.ppke.hu

***Abstract*— One of the missing critical on-board safety equipment of the Unmanned Arial Vehicles (UAVs) is the collision avoidance system. In 2010 we launched a project to research and develop an SAA system for UAS. As the system will be on-board in a small aircraft we have to minimize the weight, the volume, and the power consumption. The acceptable power consumption is 1-2W and the mass of the control system is maximum 300-500g. Here we present the concept of a visual input based See and Avoid (SAA) system. This paper introduces the long range visual detection algorithm.**

Index Terms — UAV, See and Avoid, long range visual detection.

I. INTODUCTION

The usage of the Unmanned Aerial Vehicle (UAV) are increased in many fields of the life (surveillance tasks, fire-fighting, meteorological observation, remote monitoring, and telecommunications etc.). However, most of these systems are controlled by humans remotely. One of the key issues that must be resolved to build autonomous UAV is to be able to perform See and Avoid (SAA) functions at an "equivalent level of safety" (ELOS) to manned aircraft while not negatively impacting the existing infrastructure and manned Traffic Alert and Collision Avoidance System (TCAS) [1], [2]. By realizing that, in 2010 we launched a project to research and develop an SAA system for UAS. As the system will be on-board in a small aircraft we have to minimize the weight, the volume, and the power consumption. The acceptable power consumption is 1-2W and the mass of the control system is maximum 300-500g. A purely camera based SAA system would provide cost and weight advantages against radar based solutions currently under research [3], [4]. There is not known purely camera based SAA system yet, the commercial UAVs have mostly radar based systems.

We have reported a hardware-in-the-loop simulator environment in [5], where airplanes were flying in virtual space, but followed real dynamic models in their manoeuvres. This simulation showed the feasibility of the system under ideal or quasi-ideal circumstances.

In this paper, new detection algorithm is introduced. We captured real image sequences with distant airplanes, and modified our algorithm to be able to detect airplanes even with 3.5 pixels diameter only, which enable to detect a Cessna from 3.7 km with a HD camera having 50.4° field of view.

II. SENSE AND AVOID SYSTEM

The aim of our work is to build an autonomous UAS with visual detection based SAA capability. The diagram of the closed-loop flight control system is shown on Fig. 1.

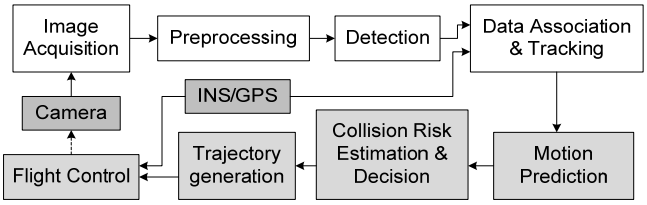


Fig. 1. Diagram of the control system

The inputs of the system are coming from *Cameras* and from *INS/GPS* (Inertial Navigation System/Global Positioning System). Input images are acquired and transformed to the right format by *Image Acquisition*. The high frequency noise and redundant colour information are filtered out by *Preprocessing*. The task of the *Detection* is to determine the angle of attack and orientation data of the approaching object.

The relative position of the target, $p_c(t)$, can be expressed in the camera frame as follows:

$$p_c(t) = \left[p_{c,x}(t)\ p_{c,y}(t)\ p_{c,z}(t)\right]^T \quad (1)$$

Assuming pinhole camera model the location of the target on the image plane can be computed as follows:

$$z_c(t) = \frac{f}{p_{c,x}(t)} \begin{bmatrix} p_{c,y}(t) \\ p_{c,z}(t) \end{bmatrix} = \begin{bmatrix} p_{m,y}(t) \\ p_{m,z}(t) \end{bmatrix} \quad (2)$$

where f is the focal length of the camera. The details can be seen in Fig. 2.

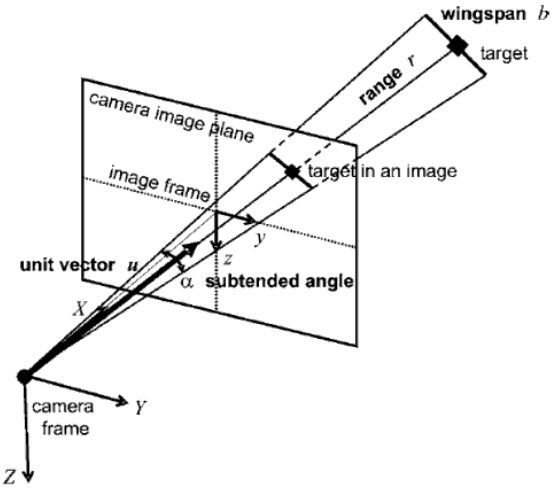


Fig. 2. Projection from camera frame coordinates to image plane

By locating and tracking the intruder on the image plane the image processing unit can determine the direction vector $d(t) = p(t)/\|p(t)\|$ and the subtended angle $\phi(t) = 2\tan^{-1}\left(\frac{b}{2\|p(t)\|}\right)$ under which the target is seen. (The constant b in the formula is the unknown wingspan of the target which is estimated).

The calculated data from *Detection* and the own position and inertial data measured by onboard *INS/GPS* module are combined by *Data Association & Tracking*. According to the combined data the relative motion of the approaching object is estimated with specific Kalman filtering methods [6]. An alert signal is released by *Collision Risk Estimation & Decision* if a risky situation is identified. In a risky situation a modified trajectory is generated by *Trajectory generation*. The avoiding manoeuvre is controlled by *Flight Control*.

III. LONG RANGE DETECTION [7]

In [5] the tests of the image processing algorithm in simulations are presented. It was shown that with this algorithm in the described environment (for detect one intruder aircraft in daylight with clear or cloudy sky when the contrast of clouds are small or medium). In this section the improved algorithm and tests on real videos in long-range situations are presented.

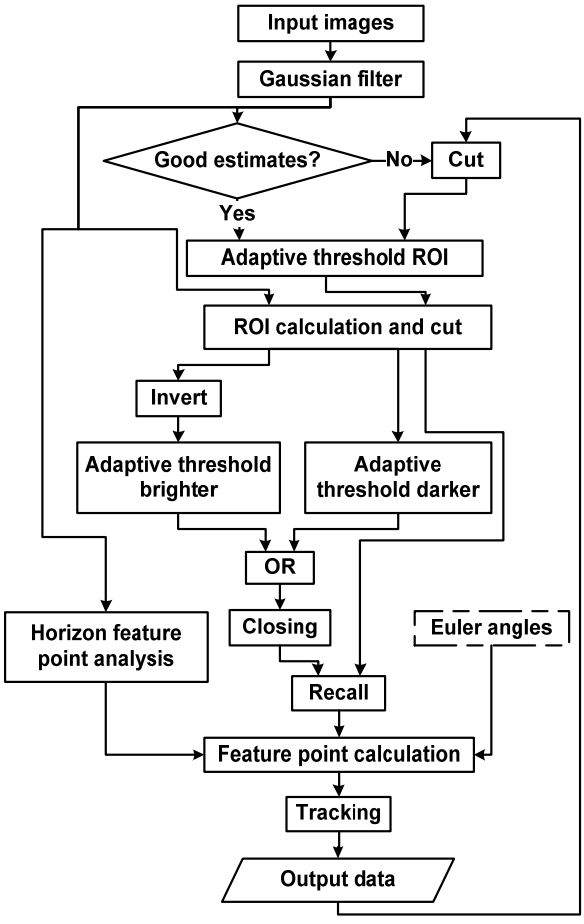


Fig. 3. Diagram of the image processing algorithm

On Fig. 3. the flowchart of the image processing algorithm is shown. The input images of the algorithm are at least 2 megapixel.

A. Preprocessing

As shown in Fig. 3. the first step is a Gaussian low pass filter to filter out high frequency noise. 2D Gaussian filter preserves edges which is important in this application. In this case a 3x3 Gaussian filter is sufficient. The coefficients are calculated according to (3).

$$h_g(n_1,n_2) = e^{\frac{-\left(n_1^2+n_2^2\right)}{2\cdot\sigma^2}} \quad h(n_1,n_2) = \frac{h_g(n_1,n_2)}{\sum_{n_1}\sum_{n_2} h_g(n_1,n_2)} \qquad (3)$$

where n_1 and n_2 are the coordinates and σ is the standard deviation. The next step is a space variant adaptive threshold [8] to filter out the slow transitions in an image (Fig. 4. b). This adaptive threshold is either executed on the entire raw image or on a smaller sub-image of it, depending on whether we have good position estimate or not. To reduce the input image size and speed up the computation, a window containing the intruder airplane according to the previous results is cut. The adaptive threshold results a binary image containing some of the points of the aircraft (Fig. 4. b), plus other points coming from clouds, ground objects, or noise.

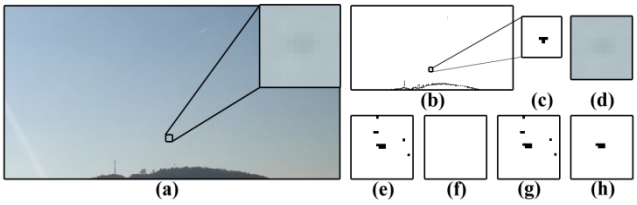


Fig. 4. The steps of the segmentation (ROI size = 24), (a) the central part of the original 1440x1080 pixels image with the enlarged area contains the aircraft, b) result of the adaptive threshold c) the enlarged area contains the aircraft on the adaptive threshold image (binary ROI) (d) coloured ROI (e) darker pixels (f) brighter pixels (g) OR operation and closing (h) segmented aircraft

B. Segmentation

On the adaptive threshold image the centroid [8] coordinates of the objects are calculated. The calculated centroid coordinates are the centre points of the Region of Interest (ROI) windows. There are two types of ROIs one on the adaptive threshold image (binary ROI (Fig. 4.c)) and one on filtered input image (coloured ROI (Fig. 4.d)). The size of the ROI is determined by the previously calculated wingspan size plus 20 pixels in each direction. The next steps of the algorithm are calculated only on ROI images to speed up the calculation and lower the power consumption.

The approaching aircraft is composed by darker and brighter pixels than the background. Therefore, two adaptive thresholds are used to get the pixels of the aircraft (Fig. 4.e, f). After the combination of the two results with the binary OR operation, a binary closing [8] is run to connect the found pixels (Fig. 4. g). After the closing a binary reconstruction operation is applied based on the binary ROI image to filter out noise remaining after the adaptive thresholds. The shape of the

detected aircraft is given by the result of the reconstruction (Fig. 4. h).

C. Tracking

Our camera is attached to the nose of the Unmanned Aerial Vehicle (UAV). If our plane is carrying out some manoeuvre the calculated position values have to be corrected to eliminate the effect of our ego motion. Euler angles [9] are provided by the *INS/GPS* module can be used to calculate these corrections, but these Euler angles are many times imprecise and in some cases they are not provided at all. The position and the orientation of the horizon is used by Horizon feature point analysis to correct the calculated position coordinates.

After this step the positions according to each ROI are collected and are given to Tracking. Multi Target Tracking Library from Eutecus Inc. is used [10]. The algorithm consists of four main steps (Fig. 5.): 1) Estimation: Using the track data gathered previously, the set of measurements are estimated, 2) Distance calculation: the distances in the proper metrics between the estimation and the input measurements are calculated, 3) Data association/ gating: the measurements and estimations are assigned to each other with a given threshold, 4) Correction/ track management: the estimated variables are corrected based on the measurement assigned to them, non-assigned tracks become subject to deletion and new tracks are initialize using non-assigned measurements.

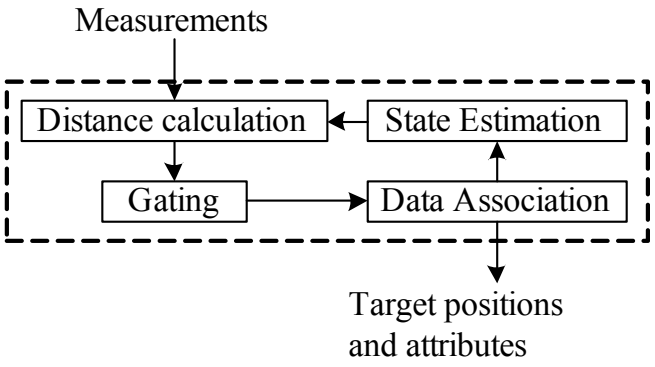


Fig. 5. Diagram of the tracking algorithm

For the estimation the library provides first, second and third order steady state Kalman filtering methods. We used second order 4D Kalman filtering method with optimal Kalman filter parameter and transient handling. The state variables was the two coordinates of the centroid of the object and the two size of the bounding box of the object with a given weight. Based on the found tracks, the position coordinates and the subtended angles are calculated.

D. Detection performance

We demonstrate the detection performance through an example, by detecting a remote Cessna. The camera was on ground and was fixed. We had estimated the relative position of the Cessna based on the landmarks. According to this estimation the Cessna was 3.7 km to the camera. In the video this aircraft was only 3.5 pixels and the size of the aircraft coincide with our range estimation (5).

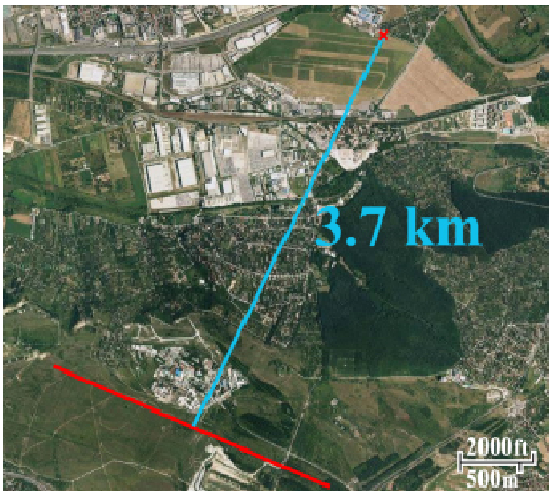


Fig. 6. In the image we marked the position of the camera with a red x, the route of the recorded aircraft with a red line, and the distance with blue.

The resolution of the camera was 1440 x 1080 pixels, the size of the sensor was 4.8mm (1/3 inch), the focal length was 5.1 mm and the field of view was 50.4°.

$$Field\ of\ View = 2 \cdot \tan^{-1}\left(\frac{sensor\ width}{2 \cdot focal\ length}\right) \quad (4)$$

The length of the aircraft was about 8m and the wingspan was 11m. From the size data, the field of view and the resolution we can get the estimated distance.

$$\text{distance} = \frac{8}{\tan\left(\frac{\text{size in pixels}}{1440}\text{FoV}\right)} \quad (5)$$

In Fig. 7. the central part of one video frame is shown and the detected aircraft is enlarged. We tracked 16 tracks with gate of 30 pixels, so the maximum distance of the estimated and the measured point in Euclidian norm was 30 pixels. The average velocity of the detected aircraft is 60m/s, from 3.7km it is around 27 px/s, so it is 1 px/frame and we could have some estimation error too.

The fade in time was 8 frames so for a given track in 8 consecutive frames the tracker has to assign a corrected estimation value to say it is a valid track. The fade out time was 20 frames, because of the noisy measurements, so if in 20 consecutive frames there isn't any estimation which is assigned to a given track, the track is deleted.

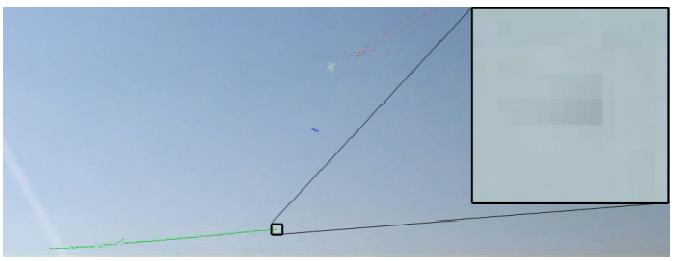


Fig. 7. Central part of processed video frame with track of intruder (dotted green line) and the enlarged pixels of the intruder

IV. CONCLUSIONS

The concept of a Sense and Avoid system has been introduced. In particular, the image processing algorithm for the long range detection of distant aircrafts was shown. The system is capable to identify a Cessna from 3.7 km size using a HD camera with 50.4° field of view. Small objects with size of 3.5 pixels were detected and tracked.

To be able to implement the system on a flyable compact low-power unit, we need to find a many-core architecture, which can efficiently execute the required computations.

ACKNOWLEDGMENT

The ONR Grant (Number: N62909-10-1-7081) is greatly acknowledged.

REFERENCES

[1] DeGarmo, M. T., "Issues Concerning Integration of Unmanned Aerial Vehicles in Civil Airspace," Tech. rep., MITRE Center for Advanced Aviation System Development, 2004.

[2] Cox, T. H., Nagy, C. J., Skoog, M. A., Somers, I. A., and Warner, R., "Civil UAV Capability Assessment," Tech. rep., NASA Dryden Flight Research Center, 2004.

[3] Hutchings, T., Jeffryes, S., and Farmer, S. J., "Architecting UAV sense & avoid systems," *Proc. Institution of Engineering and Technology Conf. Autonomous Systems*, 2007, pp. 1–8.

[4] Fasano, G., Accardo, D., Forlenza, L., Moccia, A., and Rispoli, A., "A multi-sensor obstacle detection and tracking system for autonomous UAV sense and avoid," XX *Congresso Nazionale AIDAA, Milano*, 2009.

[5] T. Zsedrovits, Á. Zarándy, B. Vanek, T. Péni, J. Bokor, T. Roska, "Collision avoidance for UAV using visual detection", *ISCAS 2011*

[6] B. Vanek, T. Péni, T. Zsedrovits, Á. Zarándy, J. Bokor and T. Roska., "Vision only Sense and Avoid system for small UAVs", *AIAA Guidance, Navigation, and Control Conference 2011,* submitted for publication

[7] T. Zsedrovits, Á. Zarándy, B. Vanek, T. Péni, J. Bokor, T. Roska, "Visual Detection and Implementation Aspects of a UAV See and Avoid System Collision avoidance for UAV using visual detection", *ECCTD 2011 (accepted)*

[8] Pratt, W. K., *Digital Image Processing: PIKS Inside*, PixelSoft Inc., Los Altos, CA, 2001.

[9] Stengel, R., *Flight Dynamics*, Princeton Press, 2004.

[10] http://www.eutecus.com/login/DownloadLink.jsp?requestedResource=/ProdServ/InstantVision/mtt_flyer.pdf

BIOMIMETIC PROCESSINGS OF THE OUTER PLEXIFORM LAYER WITH MEMRISTIVE GRIDS

András Gelencsér
(Supervisor: Dr. Themistoklis Prodromakis)
gelan@digitus.itk.ppke.hu

***Abstract*—We present a biology motivated, hexagonal, memristive grid model for early visual processing in this paper. The structure of the network is based on the organization and functions of the outer plexiform layer in the vertebrate retina. We use the novel device, the memristor, to achieve a nonlinear, locally adaptive grid, which dynamics is very similar to a biological synapse. We demonstrate that such a grid is capable to detect efficiently the edges in a grayscale image against the different environmental factors, such as different lighting conditions, disturbing noises or incidental device faults.**

Keywords-Retina, Memristor, Hexagonal Grid, Image Smoothing, Edge Detection

I. INTRODUCTION

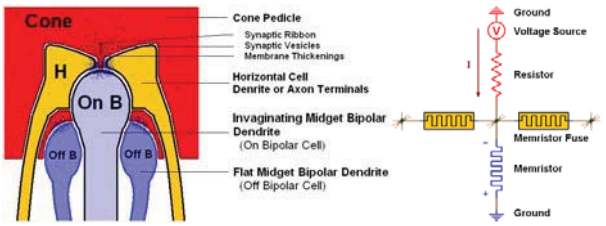
The perfection and efficiency of nature has inspired many researchers and engineers [1], [2] to gaving birth to the biomimetics. One of the first, who identified opportunities of building biologically similar electronic circuits, was Carver Mead in the late 1980's [3]. He coined this design paradigm as "Neuromorphic Engineering". He created the first neurally inspired chips, including the silicon retina [4]. The vertebrate retina is one of the most developed sensory organ. This is a complex neuronal network, which pre-process and compresses the gained information. The inspiration of the retina allows us to do higher order functions as early vision processes [5]. The newest researches have some evidence that the retina has at least ten parallel, specific signals, which ones come from the same visual input [6]. In this paper we introduce a new retinomorphic image processing approach inspired by nature. Our model is based on the outer plexiform layer of the vertebrate retina and mimics its functioning. There are already different models of this problem [4] and exist such solutions [7], but we use completely new memristive grids. This device is highly scalable and it has dynamics behaviour. That enable us a practical realization of highly interconnected, parallel architectures just like in the biological systems.

II. BIOLOGICAL NETWORK

The retina is a very complex system, which is build up from many neurons [8]. Neurobiologists have identified five major classes of neurons divided into about 60 anatomically different types of cells in the mammalian retina [9], distributed across 7 different layers. However, we consider only the outer plexiform layer to have a brief review of this structure, eminently to the connections and signal transmission between the cells.

A. The Outer Plexiform Layer

The horizontal cells play an important role in the synaptic interactions at this layer. These cells are interconnected with electrical interconnections (gap junctions) and they form a lateral network. Additionally, they are not only receiving information from the photoreceptors, but have a feedback mechanism too [10]. This feedback signal pools the information from the horizontal cell network over a wide spatial area of the OPL. The photoreceptor's response is proportional to the ratio between its photo input and the local average of the surround region of retina. The current flow between the horizontal cells diminishes the gradient between the adjacent photo inputs. This means that the OPL does a Gaussian filtering, namely they smooth the input image. According to that, the OPL is the first level of processing the visual information.



(a) Synaptic-triad of OPL (b) Node from the grid

Figure 1. **(a)** A synaptic-triad from the OPL, form by cone pedicle, ON and OFF bipolar cell dendrites and horizontal cell dendrites or axons. There is also some neuro biological elements associated with these triads too. The right side figure **(b)** shows a node from our grid, which we modeled based on the synaptic-triad. The voltage source represents the output signals of photoreceptors. Every node form a synaptic-triad, and they are interconnected with memristive fuses, which imitate the dendrites or axons of horizontal cells. The memristor represents the dendrites of bipolar cells, which are the output of the OPL.

We scrutinize the architecture of the synaptic interconnections in the OPL, to create our model of this layer. The interconnections occur in the lower part of the cone, in the cone pedicle, and they called synaptic-triads. The central element of a triad is an invaginating midget bipolar dendrite (from ON bipolar cell), the lateral elements are invaginating horizontal cell dendrites or axons. There are flat midget bipolar dendrites (from OFF bipolar cell), that make basal junctions at the bottom of the pedicle [11], shows in Figure 1. (a). There is also some biological element associated with the triads, synaptic ribbon, synaptic vesicles and membrane thickening [12].

III. RETINOMORPHIC MEMRISTIVE GRIDS

The vertebrate retina uses a massive parallel network of neurons to pre-process the visual input and retrieve the contours of the objects. This biological process already inspired many linear or nonlinear grids [1], [4]. Efficiently capturing of the edges of an image requires some smoothing on the picture, for eliminating noise and irregularities [13] and there is a so called CMOS Resistive Fuse [14], which is capable for such a task. We use a very similar approach, but with a very significant difference. In place of resistors, we use a novel device, the memristors for our networks, because their more beneficial attributes as high scalability and dynamics behaviour. Their characteristic is suitable for mimicking the dynamics of a biological synapse.

A. Memristive Components

Leon Chua was the first person, who predicted the existence of the 'Memristor' in his 1971 paper [15]. The device was first fabricated by HP Labs in 2008 [16]. The memristor is a kind of resistor with state-dependent dynamic response. The memristance is not an unknown phenomenon in the nature. There is possible with memristive devices to emulate the function of biological synapses due to the similar dynamics [17]. The memristor is a good, non-linear, dynamic device to build a grid, which can similarly work as the OPL in vertebrate retina, but it has a disadvantage. The memristance depend not only the quantity of passing charge, but its direction as well, so the memristor has polarities. In order to avoid the biasing polarity dependency we use memristor fuses [18]. These devices consist of two identical memristores connected in series with reverse polarity. The memristance of the memristive fuse will be independent of the biasing voltage polarity due to the symmetry of the series connection.

B. Correlation with Biology

The visual image is captured by the photoreceptor cells in biological retina. In our model we do not want emulate the exact receptors, but only the effect in translating light stimuli into current bias. This model works only in grayscale range for simplicity, so we represent these cells signalling with equivalent voltage sources. This approximation enables us to feed with the correct signals our memristive grid, which emulates the OPL. Every voltage source connects to a discrete node in our network Figure 1. (b). These nodes represent the triad synapses in the OPL. The voltage sources are the photoreceptor outputs, the memristive grid represents the horizontal cell dendrites and axons, the horizontal cell network. The memristores characterise the output of the OPL, the dendrites of the ON and OFF bipolar cells. Now we disregard from the two separate, ON and OFF pathways to simplify the vertical structure that emulates the OPL.

If we examine the vision system of the more developed species we find that the photoreceptor cells and the related layers underneath follow some sort of hexagonal adjustment [19]. The two tasks, namely gathering efficiently information about the visual word and partition the plane into regions of equal area the best way [20], are very similar: need a set of identical building blocks (sensors) arranged in a regular, hexagonal grid structure on a planar surface. In our grid every node makes contacts with the six neighbouring nodes, so they form a hexagonal network similarly to the biological system Figure 2.

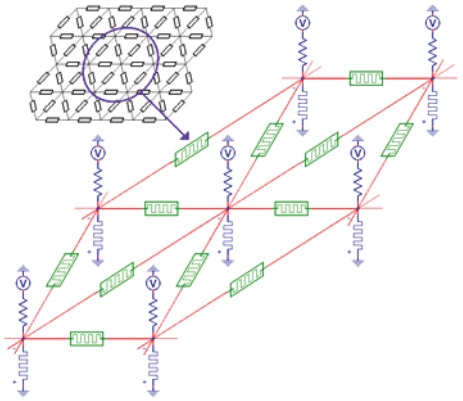


Figure 2. This is a schematic illustration of the hexagonal, memristive mimics the OPL of the retina.

Our solution is similar as in [4], with a significant different: we use dynamics memristive fuses between the nodes in our grid. This hexagonal memristive grid replicates the horizontal cell network, which do the Gaussian filtering in the input image. The detailed function of the grid will describe in the following chapter. This pre-processed information flows through the bipolar cells from the OPL to the IPL. This data is required to locate the edges of the images and more complex attributes such as motion detection and the determination of movement speed and orientation.

C. Biasing the network

Photoreceptor cells generate action potentials with higher or lower frequency compared with their basic spiking. We use spike voltage sources to embody the functioning of those cells in our circuit. The spiking frequency and the amplitude of a spike was modulated arbitrarily in every single voltage source. We upload the perspective pixel wise, as a digital image, to the grid. Every pixel of the input image is represented with a spike voltage source generator. The frequency of the spikes determines the intensity of the pixels, so the image is effectively transcribed into a 2D vector of currents in the hexagonal network Fig. 3.

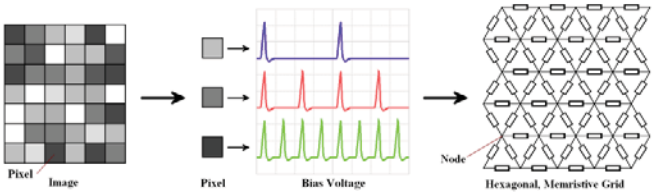


Figure 3. This figure shows the way, how will be the image to electrical current in the hexagonal network. Every pixel of the image are represented with a spike voltage source generator. The frequency of the spikes determine the intensity of the pixels.

A general grayscale image is 8-bit, namely it has 256 shades to visualize an image. For simplicity, we limit the number of intensities (4, 8 or 16 different shades) on our biasing grayscale images. The neurotransmitter release of the photoreceptors is ceased in the light in the vertebrate retina.

That denotes the cones and rods react with hyperpolarisation in case of light stimuli, so the dark means the stimulus for the photoreceptors. According to that, the darker pixels represented with higher, the whiter pixels represented with slower spike frequency.

IV. RESULTS

The results from different environmental conditions presented in this chapter. We run simulations with different lighting conditions and noise. The robustness of the system is also tested.

A. Smoothing and local Gaussian filtering

There exist numerous edge-detection algorithms, but most of them sensitive to noise. Using Gaussian filter for noise suppression is a common method. The Gaussian blur uses a Gaussian function to calculate every pixel in the image. The one dimensional Gaussian function is the following:

$$g(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}}$$

The two dimensional version is a product of two Gaussian functions in each dimension:

$$g(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}$$

where x and y is the distance from the origo and σ is the standard deviation of the Gaussian distribution.

However the uses of Gaussian blur has some disadvantage. It can cause edge position displacement, some less intense edges can vanish and some fake edges can appear. The occurrence of these phenomenon's is diminishable, if we use a local Gaussian filter. We use the very dynamic memristor to achieve this effect. In this case the filter variance is adapting to the local variance of the image and the smoothing gives better result.

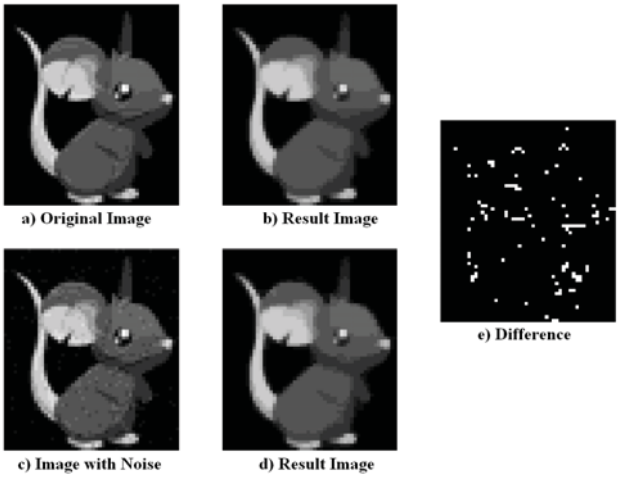


Figure 4. The local gaussian filter effect on the image. **(a)** is the original image, **(c)** is the corrupted one with white noise, **(b)** and **(d)** are the corresponding outputs of the grid, **(e)** is the accentuated difference between the two results.

Every pixel is represented with a current in our circuit. The current can flow away or back through the memristor fuses, according to the amount of the neighbouring current. If the difference is high, then the current flow is higher; if the difference is low, then the current flow is lower. The circuit realize a local Gaussian filter, because of the lateral current flows. That means the input image will smooth out and the possible noise will eliminate as on Fig. 4 can see. (a) is the clear original image, (c) is the original image corrupted with additive white noise with Gaussian distribution. The noise has 0 mean and 0,3 standard deviation (σ). (b) and (d) show the corresponding outputs of the memristive grid, which is measured at the nodes. Between the two results are not any considerable differences, only 3% of the total pixel has one grade intensity mismatch.

B. Edge detection

The information from the OPL enables the detection of the contour of an object, which happens in the IPL in the vertebrate retina. Altough we model only the OPL, there is an opportunity in our system to extract the edges of the input image. Every memristive fuse will have a potential difference between its two nodes, according to the pixel grayscale intensities. Wherever there is an edge in the image the corresponding potential difference among these pixels will be high, so there will flow much current and the change rate of the memristance will be higher of the devices related to the edges as the other devices. It is possible with a threshold mechanism to find all this memristive fuses and the corresponding pixels.

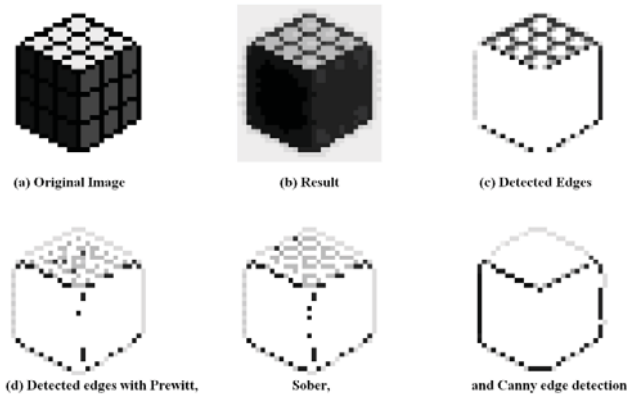


Figure 5. Edge Detection on the Image. **(a)**shows the input image,**(b)** is the output image of the grid. **(c)** shows the calculated edges from the model. **(d)** shows the results some conventional edge detection algorithms (Prewitt, Sober, Canny).

As the Fig. 5 is shown, our hexagonal memristive model is able to find the object boundaries. In this example we can observe some maleficent effect of the smoothing process. The crosses on the Rubik cube darker sides are smoothed out (b). However, our edge detection method found the remaining edges very nicely (c). There is also some result with conventional edge detectors like Prewitt, Sober and Canny for comparison. In this case they produce more imprecise boundaries like our method. Fig. 6 demonstrate that our model can detect the edges despite the brighter or lighter lighting conditions. (a) is 30% lighter (d) is 20% darker input image like the original one. We can observe on (c) and (f), that the edge results converge to the same state in every case.

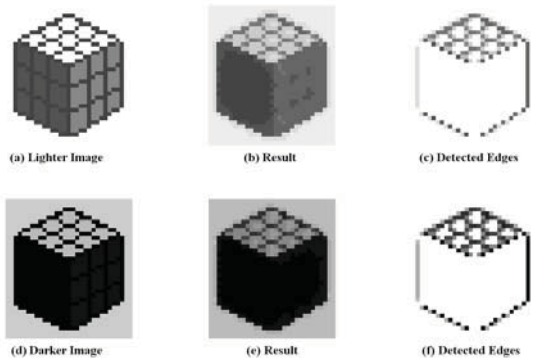


Figure 6. Edge Detection on the Image with different environmental conditions. **(a)** input image in brighter lighting condition, **(b)** output image, **(c)** detected edges on this image.**(d)** input image in darker lighting condition, **(e)** output image, **(f)** detected edges on this image.

C. Fault Tolerance

The memristor is an emerging nanoscale device. This technology is relatively new and the production processes have some imperfections, accordingly the probability of a defect is greater with smaller device size on the wafer. Keeping this in mind, we evaluate our system behaviour, if we have some faulty device in the network. In a perfect scenario all memristive fuses are reliably set with the following conditions: Ron=100Ω, Roff=16kΩ, Rinit=200Ω. In order to test the robustness of our system we assign random initial states to the different percentage of memristive elements. The random initial conditions differ from the original ones in the following mode: The value of a defected Ron could be from 50% to 400% compared to the original one. The value of Roff could be from 62.5% to 125% and the value of Rinit could be from 50% up to 4000% considering the new Ron and Roff values too. One of the results is shown on Fig. 7. There is three case of one biasing image. The (b) show the result of a perfect network. (a) is a network with 25% faulty devices and (c) is a network with 50% faulty devices. (d) shows the difference between the flawless and the 75% correct (e) shows the difference between the flawless and only 50% correct network. We can observe that, the smoothing of the image will decrease, because of the high number of defective memristors.

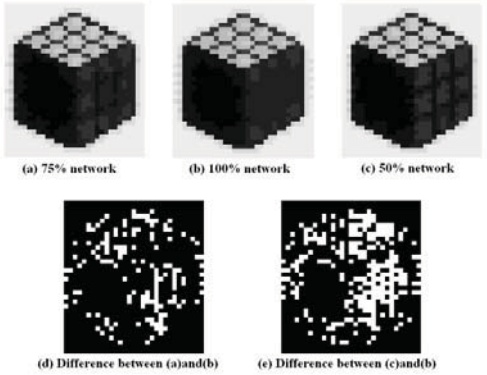


Fig. 7. Results with different level of faulty devices. **(a)** network with 25% faulty devices, **(b)** perfect network, **(c)** network with 50% faulty devices. **(d)** and **(e)** show the difference between the perfect and corresponding faulty network.

V. CONCLUSIONS

We replicated the outer plexiform layer of a biological retina with a novel device dynamics. This model output could also serve as the input of a network which models the inner plexiform layer of the retina. Our future goal is to create a more complex model, so we can mimics the functions of a whole vertebrate retina, and solve more composite problems like detection of movement or object tracking.

REFERENCES

[1] C Koch, J Marroquin, and A Yuille, “Analog "neuronal"networks in early vision,” Proceedings of the National Academy of Sciences, vol. 83, no. 12, pp. 4263–4267, June 1986.

[2] Y. Yao and W.J. Freeman, “Pattern recognition in olfactory systems: modeling and simulation,” in Neural Networks, 1989. IJCNN., International Joint Conference on, June 1989, pp. 699–704 vol.1.

[3] Carver Mead, Analog VLSI and neural systems, Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1989.

[4] Carver A. Mead and M.A. Mahowald, “A silicon model of early visual processing,” Neural Networks, vol. 1, no. 1, pp. 91–97, 1988.

[5] Tomaso Poggio, Vincent Torre, and Christof Koch, “Computational vision and regularization theory,” Nature, vol. 317, pp. 314–319, Sept 1985.

[6] Botond Roska and FrankWerblin, “Vertical interactions across ten parallel, stacked representations in the mammalian retina,” Nature, vol. 410, pp. 583–587, March 2001.

[7] B.E. Shi and L.O. Chua, “Resistive grid image filtering: input/output analysis via the cnn framework,” Circuits and Systems I: Fundamental Theory and Applications, IEEE Transactions on, vol. 39, no. 7, pp. 531 – 548, jul 1992.

[8] Helga Kolb, “How the retina works,” American Scientist, vol. 91, no. 1, pp. 28–35, Jan-Feb 2003.

[9] Richard H. Masland, “Neuronal cell types,” Current Biology, vol. 14, no. 13, pp. 497–500.

[10] D.A. Baylor, M.G.F. Fuortes, and P.M. O’Bryan, “Receptive fields of the cones in the retina of the turtle,” Journal of Physiology, vol. 214, pp. 265–294, 1971.

[11] Silke H., Ulrike G., and Heinz W., “The cone pedicle, a complex synapse in the retina,” Neuron, vol. 27, no. 1, pp. 85 – 95, 2000.

[12] Helga Kolb, “Organization of the outer plexiform layer of the primate retina: Electron microscopy of golgi impregnated cells,” Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, vol 258. no. 823, pp. 261-283, May 1970.

[13] John Canny, “A computational approach to edge detection,” Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. PAMI-8, no. 6, pp. 679–698, Nov 1986.

[14] P.C. Yu, S.J. Decker, H.-S. Lee, C.G. Sodini, and Jr.Wyatt, J.L., “Cmos resistive fuses for image smoothing and segmentation,” Solid-State Circuits, IEEE Journal of, vol. 27, no. 4, pp. 545–553, Apr 1992.

[15] L. Chua, “Memristor-the missing circuit element,” Circuit Theory, IEEE Transactions on, vol. 18, no. 5, pp. 507–519, Sep 1971.

[16] Dmitri B. Strukov, Gregory S. Snider, Duncan R. Stewart, and R. Stanley Williams, “The missing memristor found,” Nature, vol. 453, no. 7191, pp. 80–83, May 2008.

[17] Bernabe L., Teresa S., Luis A. C., Jose A. P., Carlos Z., and Timothee M., “On spike-timing-dependent-plasticity, memristive devices, and building a self-learning visual cortex,” Frontiers in Neuroscience, vol. 5, no. 0, 2011.

[18] Feijun J. and B.E. Shi, “The memristive grid outperforms the resistive grid for edge preserving smoothing,” in Circuit Theory and Design, 2009. ECCTD 2009. European Conference on, Aug 2009, pp. 181 –184.

[19] Christine A. Curcio, Kenneth R. Sloan, Robert E. Kalina, and Anita E. Hendrickson, “Human photoreceptor topography,” The Journal of Comparative Neurology, vol. 292, no. 4, pp. 497–523, Feb 1990.

[20] Thomas C. Hales, “The honeycomb conjecture.,” Discrete & Computational Geometry, pp. 1–22, 2001. vol. 258, no. 823, pp. 261–283, May 1970 .

Brain activity measurement with implantable microchip

Zoltán Kárász
(Supervisor: Dr. Péter Földesy)
karzo@digitus.itk.ppke.hu

***Abstract*— This paper presents a low power neural signal amplifier with tunable cut-off frequencies. The presented compact amplifier used for sensing different type of neural signals reduces the size and the power consumption of the whole circuit. The distinguishing features of this solution are the large time constant, linearity, and small achievable area, which are realized with a configurable series of pseudo resistances. The proof of concept has been manufactured on TSMC 90nm technologies.**

***Index Terms*— Brain-computer interface, large time constant, neuro-amplifier, pseudo-resistor.**

I. Introduction

The biomedical field is one of the most dynamically developing research areas in the analog IC design, especially those concerning low-power implementation including implantable without battery. The examination procedures need more time for the functional result than available using other observing techniques as the FMRI or using simple EEG [1,2]. Even though the portability of the measuring instrument is not an important issue for the animal studies, it is in the human experiments.

Our interest concerts indeed the implantable cortical micro sensor arrays, which causes minimal structural damages in the analyzed region. From the engineer's aspect measuring the brain activity could be simplified to an electrical connection between the brain tissue and the electrode. The implantable neural recording devices have to achieve strict specifications, including the power consumption, noise and distortion requirements, defined maximal thermal dissipation and specified input frequency range. The presented architecture is an Operational Trans-conductance Amplifier (OTA) based capacitive feedback single input differential output amplifier including pseudo-resistors chains to achieve programmable large time constant with significantly reduced distortion and robustness.

II. Backgorund of Neural Recording

A. Neural Signal

Even thought a neuron can produce 100 μV internal voltage changes relative to the extracellular fluid, this can be recorded directly only with patch-clamp electrodes, but the in-vivo chronic recording using multi-electrode arrays which able to utilize the smaller extracellular potentials from several micrometers from the cell [3].

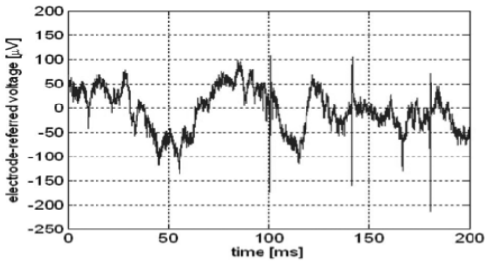


Fig. 1 Typical neural signal

The amplitude of the signal is on the order of 100 μV (*Fig. 1)*[4]. The neural action potentials are called „spikes". Neurons rarely fire more rapidly than 100 spikes per second, although rapid bursts of several spikes are possible. Neurons produce spikes of nearly identical amplitude and duration and information is encoded in the timing of spikes.

The low-frequency (under 200 Hz) oscillations are known as Local Field Potentials (LFPs). They arise from the synchronous activity of many neurons in one region of the brain. These neurons are far away from the electrode for their individual action potentials to be detected, but the many neighboring cells create a large signal that is easily detected [3]. In some experiments the electrode arrays scared the tissue around microelectrode tips. This scar tissue tends to attenuate spike signals from nearby neurons, but LFP signals are less affected.

In many applications, it is desirable to separate LFP from spike signals so they may be analyzed separately. This is easily accomplished by linear filtering since LFPs occupy frequencies from approximately 0.5 – 200 Hz, while spikes have energy concentrated in the 300 Hz – 7 kHz range.

B. Amplifier Requirements

The realization of large time constants is fundamental for design filters with very low cut-off frequencies especially in implantable biomedical sensors. The filters are required to be tunable. In addition, realizations with low power dissipation and small size are also critical. Several approaches for the design of integrators with very large time constants have been reported [5-7]. The trivial solution to employ on-chip physical

resistor and capacitor requires large chip area and it would not be tunable. The possible solutions can be categorized into pseudo-resistor implementations [3,5,9,10], switched-capacitor (SC) methods [13-15] and operational trans-conductance amplifier capacitor (OTA-C) techniques with very small trans-conductance's [13-15] to allow the on-chip capacitance to be kept manageable low.

C. MOS Pseudo-Resistor

This work is based on pseudo-resistors, as they outperform other solutions in term of power and area efficiency to reach large time constant. The pseudo-resistance has good size and parasitic values (in the range of fF), but it also has some serious non-ideal behavior, which means poor robustness and bad distortion in the LFP range.

To able to handle the pseudo element it is necessary to modeling the resistance of the MOS transistor.

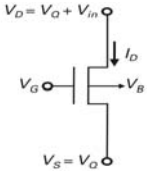


Fig. 2 Transistor model

A descriptive linear model bases on the following components [6]: the source diffusion; the channel resistance; accumulation resistance; component resistance; drift region resistance; substrate resistance. For more appropriate result it is needed a nonlinear approximation.

$$\left.\frac{1}{R}\right|_{V_{DS}=0} = \left.\frac{dI_D}{dU_D}\right|_{V_{DS}=0} = -\left.\frac{dI_D}{dU_S}\right|_{V_{DS}=0} = g_m = \frac{2I_S}{\phi_t}\left(\sqrt{1+i_f}-1\right) \quad (1)$$

In the strong inversion (2) and weak inversion (3) region it is possible to explain the resistance as following:

$$\left.\frac{1}{R}\right|_{V_{DS}=0} = g_m = \mu C_{ox}\frac{W}{L}\left(V_G - V_{T_0} - nV_Q\right) \quad (2)$$

$$g_m = \frac{2I_S}{\phi_t}\left(\frac{1}{2}i_f\right) = \frac{2I_S}{\phi_t}\exp\left(\frac{V_G - V_{T_0} - nV_Q + n\phi_t}{n\phi_t}\right) \quad (3)$$

where n is the slope parameter.

The most prevalent utilization of the MOS transistor as a resistor is the pseudo-resistor. That is construing the features of this solution, like the minimal size, simplicity and the outstanding effective resistance [9].

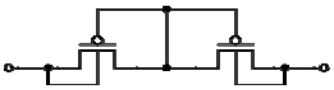


Fig. 3 Schematic of the pseudo-resistor element

The basic symmetric element contains two transistors that are connected as a MOS diode and a parasitic source-bulk diode connected in anti-parallel. If the voltage across the device is small enough, then neither diode will conduct strongly, and the effective resistance is very large (> 10 GΩ).

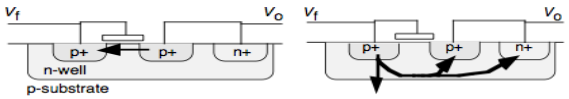


Fig. 4 Diode-connected and PN junction is forward-biased MOS transistor cross-section image

For voltage polarity $V_o > V_f$ across the element (*Fig. 4*) the side of V_o in the MOS case acts as the source of the transistor. For the opposite polarity, the driven side is a forward-biased source-gate junction.

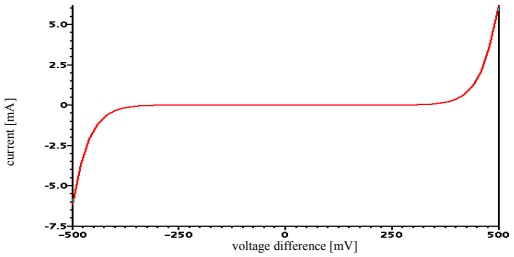


Fig 5. Current voltage relation on a pseudo transistor

The current-voltage relationship (*Fig. 5*) [5, 10] of the expansive element means that the effective resistance of the element is large for small signals and small for large signals. Therefore the adaptation is slow for small signals and fast for large signals.

The nonlinear variation of the resistance in the feedback loop means the transfer-function would not be permanent at the whole working period. If the cut of frequency is altered the whole distortion increases. This effect impairs significant in the lower frequency range (under 100 Hz).

Another relevant problem to address with this solution is the large impact of the technological parameters and the operational conditions. The biomedical applications have strict operating requirement about the temperature (30-44 C°) that actually reduce the variation, but still remain large manufactured uncertainty (which depends on technology node).

III. Basic Neuro-Amplifier Topology

The amplifier is based around an operational transconductance amplifier (OTA) that produces a current applied to its input (Fig. 6) [4,5,10-12]. A capacitive feedback network consisting of C_1 and C_2 capacitors sets the mid-band gain of the amplifier. The input is capacitively coupled through C_1, so any dc offset from the electrode-tissue interface is removed. C_1 should be made much smaller than the electrode impedance to minimize signal attenuation. The R_2 elements shown in the feedback loop set the low-frequency amplifier cut-off.

The approximate transfer function is given by

$$\frac{v_{out+}-v_{out-}}{v_{in}} = \frac{C_1}{C_2}\frac{1-sC_2/G_m}{\left(\frac{1}{sR_2C_2+1}\right)\left(s\frac{C_LC_1}{G_mC_2}+1\right)} \quad (4)$$

The midband gain A_M is set by the capacitance ratio C_1/C_2, and the gain is flat between the lower and upper cutoff

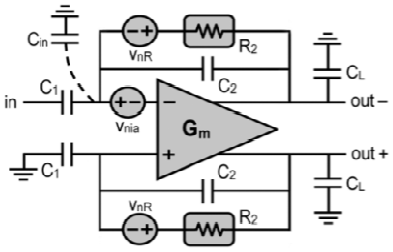


Fig. 6 Schematic of the capacitive feedback amplifier

frequencies f_L and f_H. The product of R_2 and C_2 determines the lower cutoff frequency, while the upper cutoff is determined by the load capacitance C_L, the OTA trans-conductance G_m, and the mid-band gain. Capacitive feed introduces a right-half-plane zero at f_z, but this zero can be very at high frequency by setting

$$C_2 \ll \sqrt{C_1 C_L} \qquad (5)$$

so that it has little practical effect on amplifier operation. The OTA contributes noise primarily between f_L and f_H. Below a particular frequency called f_{corner}, the noise contribution from v_{nR} will dominate. If R_2 is implemented as a real resistor so that its noise spectral density is

$$v_{nR}^2(f) = 4kTR_2 \qquad (6)$$

and $C_1 >> C_2$, C_{in}, then f_{corner} is approximately

$$f_{corner} \approx \sqrt{\frac{3C_L}{2C_1} f_L f_H} \qquad (7)$$

A similar result is obtained for pseudo resistor element element used as R_2 in. To minimize the noise contribution from the R_2 elements, we should ensure that $f_{corner} << f_H$.

If the noise contribution from R_2 is negligible and $C_1 >> C_2$, C_{in}, then the output rms noise voltage of the neural amplifier is dominated by the noise from the OTA.

$$v_{nia}^2 = \frac{16kT}{3g_{m1}}\left(1 + 2\frac{g_{m3}}{g_{m1}} + \frac{g_{m7}}{g_{m1}}\right) \qquad (8)$$

where g_{m1} is the trans-conductance of the input devices M_1 and M_2. The noise of the cascode transistors is negligible.

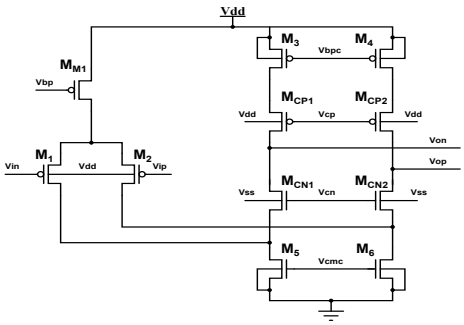


Fig. 7 Differential input cascoded OTA

In that case the load capacitance is determined by

$$C_L = \frac{4kT}{V_{ni}^2 3A_M} \qquad (9)$$

In practical implantable multi electrode systems, the size of the capacitances is very limited, because of the minimal fabrication size for the C_2 and the available space for the C_1, which is why we have to choose them deliberately.

IV. Proposed Series-Connected Digitally Controllable Pseudo-Resistor

There is a possible tradeoff between the noise and distortion. Using more pseudo resistor element in series helps decreasing the nonlinearity effect at the price of increasing noise figure. In this section this tradeoff is analyzed on resistor-chains, which contain different number of pseudo resistor element.

The series of pseudo-resistors results in decreasing distortion approximately linearly with the number of elements, due to the voltage different would be smaller between the two sides of each element (Fig. 8).

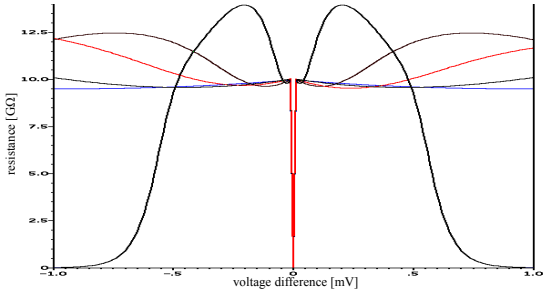


Fig. 8 Resistance variation at different number pseudo-resistor in series (curves PS2, PS8, PS16, PS32, PS64 respectively) [GΩ / mV]

In order to fulfill the accuracy requirements in the whole system, we need satisfy the total harmonic distortion (THD) on the every frequency as well. For a typical 8-bit accuracy case, we would need to keep at least the 60 dB level for the frequency range of interest (Fig. 9).

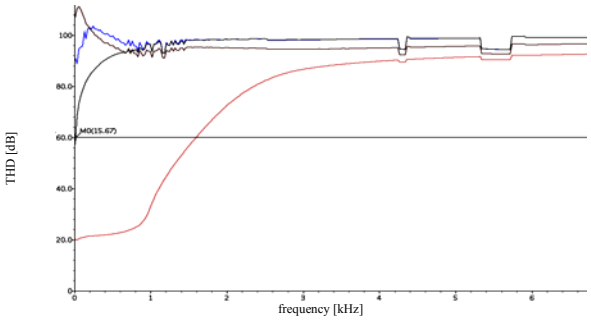


Fig. 9 THD of the Harrison-topology with different resistor implementation showing the constraint for a typical 60 dB system

A. Gated Pseudo-Resistance

Because of the high corner deviations and the frequency tuneability, another important aspect in the design is the resistance control.

It possible to give controllability to the resistance if we use switches to shortcut the remaining part of the chain (Fig. 10). This gated structure needs to be designed at least the required resistance plus the corner variations. Note that the large

number of the series connected pseudo resistor still does not have area large overhead neither the parasitic.

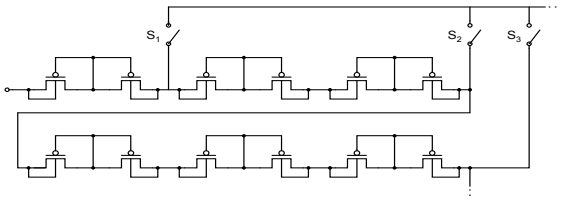


Fig. 10 Gated pseudo-resistance

The switch implementation needs careful design as well. Large open state impedance are required so that they could be commensurable to the pseudo-resistances, otherwise the leakage will reduce the overall resistance; hence they must be optimized to the OFF resistance oppositely the general usage.

Another issue is the sizing of the different segments. It is not effective to use identical resistors if we like to tuning and compensating with the same chain (Fig. 11). The exact choice of distribution (linear, exponential, or binary weighted) depends on the required cut-off frequencies and the degree of the corner deviations.

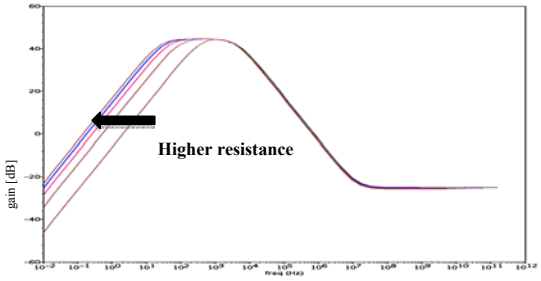


Fig. 11.a Transfer function at different f_L

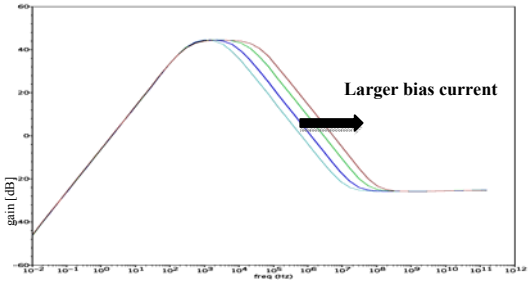


Fig. 11.b Transfer function at different f_H

Finally, we got a programmable solution that helps us to increase the robustness against the technology parameter variation, to reduce the significant distortion and gives us the possibility to choose the cut-off frequency.

V. Measurements

For proof of concept we designed and sent for manufacturing this architecture with 32 pseudo-resistor and a low power LNA. Targeted technology is the TSMC 90 nm LP-RF. By the time of the revision process and camera-ready paper deadline the ASIC measurements will be available and presented.

Current [A]	Noise [μV_{RMS}]
50n	7,8
100n	7,6
500n	7,1
1u	6,9

Table 1. Different bias current influence on noise, (f_H=1kHz)

Frequency [kHz]	Noise [μV_{RMS}]
1	7,6
4	8,3
7	8,6
11	8,6
18	8,1

Table 2. Different high cut-off frequency, (I_{bias}=100nA)

VI. Conclusions

The main contribution in this paper, that we presented an integrated low noise amplifier circuit for the battery less implantable neural recording, and reviewed the most important design considerations. The MOS pseudo resistor chain is genuine innovation which is not used any other solutions on this area. The comparison between the switched-capacitance, the pseudo resistance and the modified OTA topologies as generally are not definite. As long as the current cancellation and division generate a continuously current consumption and not gives any chance for tuning the transfer-function [8], till then the switched capacitor provide a fine tuning method but generates high distortion. The basic MOS pseudo resistance not able the handle the low frequency input, because the bad distortion and sensitivity for the corner variation as a SC resistances. The gated chain could be the optimal solution. It gives the tuning range to decreasing the corner effect and to be able handle the local field potential range. In summary in this paper, an integrated tunable low noise amplifier circuit is presented for implantable neural recording, and introduced a MOS pseudo resistor chain outperforms existing solutions in terms of area and linearity.

References

[1] S. Chen, B. Mulgrew, and P. M. Grant, "A clustering technique for digital communications channel equalization using radial basis function networks," *IEEE Trans. Neural Networks*, vol. 4, pp. 570–578, Jul. 1993.

[2] Avestruz et al. A 5 µW/Channel Spectral Analysis IC for Chronic Bidirectional Brain-Machine Interfaces. Solid-State Circuits, IEEE Journal of (2008) vol. 43 (12) pp. 3006-3024

[3] Rieger et al. A 230-nW 10-s time constant CMOS integrator for an adaptive nerve signal amplifier. Solid-State Circuits, IEEE Journal of (2004) vol. 39 (11) pp. 1968- 1975

[4] Harrison and Charles. A low-power low-noise CMOS amplifier for neural recording applications. Solid-State Circuits, IEEE Journal of (2003) vol. 38 (6) pp. 958- 965

[5] Gozzini et al. Linear transconductor with rail-to-rail input swing for very large time constant applications. Electronics Letters AB - ER - (2006) vol. 42 (19) pp. 1069- 1070

[6] Wattanapanitch et al. An Energy-Efficient Micropower Neural Recording Amplifier. Biomedical Circuits and Systems, IEEE Transactions on (2007) vol. 1 (2) pp. 136-147

[7] Rieger et al. Design of a low-noise preamplifier for nerve cuff electrode recording. Solid-State Circuits, IEEE Journal of (2003) vol. 38 (8) pp. 1373- 1379

[8] Triantis and Demosthenous. An improved, very long time-constant CMOS integrator for use in implantable neuroprosthetic devices. Circuit Theory and Design, 2005. vol. 3 pp. III/15- III/18 vol. 3

[9] Zou et al. A 1-V 1.1-µW sensor interface IC for wearable biomedical devices. Circuits and Systems, 2008. ISCAS 2008. IEEE International Symposium on (2008) pp. 2725-2728

[10] Yin and Ghovanloo. A Low-Noise Preamplifier with Adjustable Gain and Bandwidth for Biopotential Recording Applications. Circuits and Systems, 2007. pp. 321-324

Terahertz imaging

Image acquisition in the terahertz frequency domain

Domonkos Gergelyi
(Supervisor: Dr. Péter Földesy)
gerdo@digitus.itk.ppke.hu

***Abstract*—Image capturing in the terahertz frequency domain is a great challenge. Depending on the realized measurement configuration it has various application areas. Non-invasive analysis of different tissues is one of its most important applications. Our new CMOS based detector can be a good basis for such an employment. In the followings I present our terahertz measurement setup and summarize my work on implementing a terahertz imager.**

Keywords-componen, terahertz, imaging, sensor fusion

I. INTRODUCTION

Terahertz imaging is not an old research area, see [1] for the first promising imaging techniques, and commercial applications of terahertz imaging. Speaking about the frequency spectrum this domain between the millimeter and infrared range is mentioned as the "terahertz gap". This motivates the most rapid development among the frequency domains. If terahertz imaging became cheaper and faster, numerous applications would be built on it regarding biology, material engineering and medicine. For instance: the thorough observation of cell cultures and thick excisions, which involves the qualitative analysis of the matter; the examination of different surface structures; exploring the roots of tooth and diagnosing skin cancers (especially basal cell carcinoma).

II. THE CHALLENGES OF SENSING

Due to many physical limitations, active image acquisition was realized, which utilizes dedicated illumination source. However, the power of the incident radiation is still quite low, hence the SNR of the detector becomes critical. To alleviate this fact, we attempted to cancel out a certain portion of the noise by utilizing compressed sensing, cf. [2]. Thorough analysis and experimentation showed that it is indispensible to boost up the SNR by oversampling in contrast to the CS literature. In parallel, we have to rather exploit the advances of greater sensor arrays. This is made possible by the utilization of purely CMOS based detectors.

III. OUR THERMOPILE SETUP

At the very beginning, we used thermopile based sensing elements. These had high theoretical sensitivity (100V/W) but we can achieve low SNRs during their practical applications. The other problems with these detectors were their immature c manufacturing process as THz sensors. In addition it was difficult to integrate them into a complete system.

IV. OUR ACTUAL SETUP

In our actual setup we are using the first generation CMOS detector which was manufactured at 180 nm technology. This provides relatively low self noise.

Different antenna and detector configurations were realized to help determining the best arrangement. Figure 1. shows the optical setup.

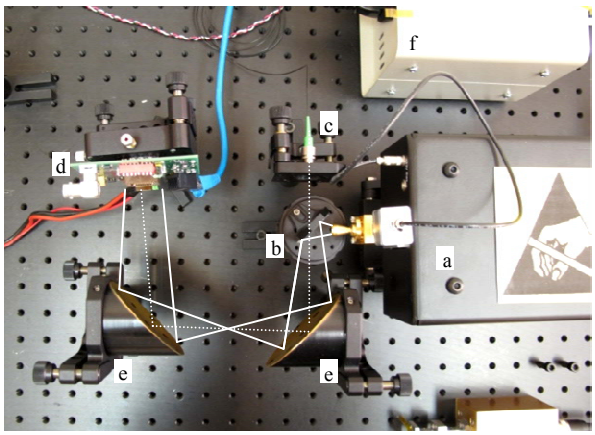


Figure 1. Our actual optical setup. a) WR 9.0 AMC (amplifier multiplier chain) b) beam splitter c) laser for optical positioning d) detector bord e) parabolic mirrors f) common power supply (the 9-14GHz base oscillator cannot be seen).

V. FUTURE PLANS

The second generation CMOS detectors embedded into a system on a chip is going to arrive from the factory in the near future, which has been designed by our group. These are designed at 90 nm TSMC CMOS technology, which means several advances and new challenges within the process of sensing.

The self noise of the sensor is significantly increasing, but higher sensitivity and bigger responses are also expected.

Higher sensor density and more efficient area utilization could be achieved by smart detector array arrangement and careful placing of the auxiliary electronics, cf. [3], [4].

As mentioned before by utilizing the detector arrays higher accuracy and significantly faster acquisition can be achieved.

IV. THE PRINCILE OF SENSING

Figure 2. indicates the basic principle of the sensing. The detector is a MOS transistor, which is biased to a near threshold gate potential. This results in a low DC current on the transistor at small source to drain voltage. The gate of the detector is attached to the antenna through a matched feed line. This current is modulated by the terahertz signal on the gate of detector. The process does not include any rectification, only the macroscopic changes of current are detected during a longer sampling period. The transistor's channel behaves as active, nonlinear media, resulting in measureable change in the device current.

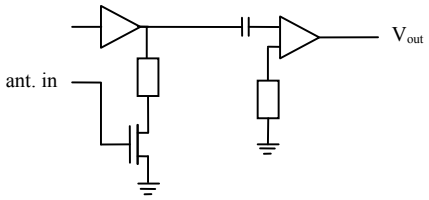


Figure 2. Schematic of a detector element without the antenna and the auxiliary processing elements.

V. CREATING TERAHERTZ IMAGE

To start building a practically useful image with a few sensors it is necessary to utilize a scanning device. That way the terahertz image is made by pixel by pixel sampling. For that purpose we designed a high precision mover to operate the sample to be observed. In this case one can accurately position the sample relative to the individual sensor elements on the chip.

A. The mechanics

The sample mover consists of X, Y, Z components. To perform 2D scanning the samples are moved in the X, Y plane. The Z dimension is to bring the sample into focus and to make possible other corrections regarding the scanning plane.

B. Motor control

To automate the operation of the scanner two phase hybrid stepper motors are used. Compact complete solutions are available for this problem, but to have full control over the dynamics custom solutions are needed. For that two other approaches were investigated as well. A fully PC based control provides a low cost solution for testing purposes. The other approach involves additional dedicated DAQ cards and makes possible to integrate the control of the whole measurement setup into a single controller interface.

In our application relatively high torque is needed (0.5 Nm). Thus we have to accelerate the mechanical parts smoothly. However stepper motors driven in full step mode can vibrate especially at lower speeds. To cope with the problem we obtain full control of the dynamics by directly driving the DMOS full bridges. We produce signals, which make possible seamless motions at minimal losses of torque.

VI. SENSOR FUSION

A. Overview

My ultimate goal is to produce detailed multi-spectral images about 1-3 mm thick samples which contain the precious information provided by the terahertz wave penetrated through the material.

If the samples are such tissues what are transparent to light, we can make infrared images as well and register the results with the terahertz measurements easily in the setup. This condition holds in general as the considered samples are thin enough to be transparent for a certain degree. Therefore it was obvious to apply sensor fusion and to synthesize such an image that flexibly joins the data gathered from different sources.

B. Image segmentation

The infrared image is segmented first by applying a Canny edge detector. This is for an optional enhancing of the borders of different structures. After that I utilize region growing to obtain full segmentation. The two results can be combined. To explain the selected region growing method in more details I want to give a more comprehensive view of my motivations. Different imaging techniques have to be developed which are built upon each other:

First, I utilize one of our 180nm CMOS detectors to make a scanned image. Here the optical setup is simple as only one pixel and constant optical path (moving sample) is used. This indicates that local artifacts can appear only at the borders of the sample holder. Therefore by image processing simple global operators can be used to segment the image and define significant regions.

As a second step, the capabilities of the detector array have to be exploited. It makes possible to speed up the scanning process or enhance the spatial precision of the sensing. But in this case due to the inhomogeneity of the sensor array and the different optical setup I have to concern significant local disturbances. For this reason transitivity is needed by the segmentation.

Therefore an enhanced centroid region growing can be used to classify the pixels that are related to each other. Thus a cumulative distance measure is created, which is based on the distance from the neighboring pixels, the distance from the central element and the distance from the statistics of the given region. By the weight of the three components we can balance the different properties of the classifier for instance the transitivity or its sensitivity to drift.

C. The speed criteria

It is important to keep the time needed for the scanning as low as possible. In one dimensional scanning high resolution can be achieved when the sample travels at a constant speed. To ensure this constant accurately tuned speed proper controlled acceleration phases should be realized. To keep the scanning time low different spatial resolution has to be used among the X and Y directions to enhance the speed of the acquisition.

By changing the direction of the “constant speed mode” from X to Y one can achieve better spatial sampling distribution.

CONCLUSION

Terahertz imaging is possible with the novel CMOS detectors which can reduce the price of this technology. It is worth to focus on designing greater detector arrays which can depress the needed time for image acquisition. Such results can advance terahertz imaging to become more general and further its application in new areas.

ACKNOWLEDGMENT

The work is supported by the Hungarian Scientific Research Fund - National Office for Research and Technology OTKA-NTKH CNK-77564 project.

REFERENCES

[1] D. M. Mittleman; R. H. Jacobsen; M.C. Nuss, “T-ray imaging”, IEEE Journal of Selected Topics in Quantum Electronics, Vol. 2, No. 3. pp: 679-692.

[2] G. Hosein Mohimani, M. Babaie-Zadeh, C. Jutten, “A fast approach for overcomplete sparse decomposition based on smoothed l_0 norm”, IEEE Transactions on Signal Processing, Vol. 57, No. 1, Jan. 2009, pp. 289-301.

[3] A. Rodríguez-Vázquez, R. Domínguez-Castro, F. Jiménez-Garrido, S. Morillas, A. García, C. Utrera, M. Dolores Pardo, J. Listan, and R. Romay, “A CMOS Vision System On-Chip with Multi-Core, Cellular Sensory-Processing Front-End”, in Cellular Nanoscale Sensory Wave Computing, edited by C. Baatar, W. Porod and T. Roska, ISBN: 978-1-4419-1010-3, 2009

[4] P. Földesy, Á. Zarándy, Cs. Rekeczky, and T. Roska „Configurable 3D integrated focal-plane sensor-processor array architecture”, Int. J. Circuit Theory and Applications (CTA), pp: 573-588, 2008.

Development of Thermopile Type THz detector Measurements of Performance

Endre László
(Supervisors: Dr. Péter Földesy and Dr. Gábor Battistig)
laszlo@mfa.kfki.hu

***Abstract*—A new type of thermopile structure is developed for THz radiation detection. A regular thermopile is based on the Seebeck effect. In the present article the infrared operation principle of the micromachined thermopile is demonstrated. The idea of the linearly arranged dipole antennas is detailed, namely that the polarity sensitive antenna array is capable of detecting THz radiation. The fabrication process of such devices is discussed. Two type of responsivity measurements are performed. One for measuring the infrared responsivity and another measurement for THz responsivity. These measurements show that the THz responsivity is about 96 V/W. The infrared responsivity in vacuum chamber measurement has increased significantly. This implies that by using a proper vacuum casing the performance of the detector for THz radiation detection could be improved and the device could reach as high as 1176 V/W.**

Keywords: thermopile, THz, MEMS

I. INTRODUCTION

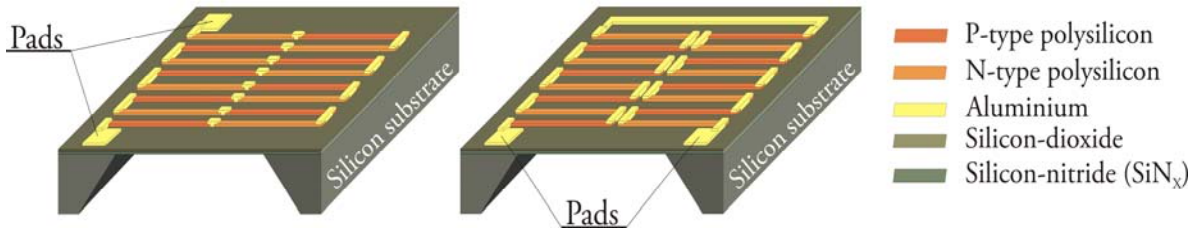
The operation principle of the device is based on the Seebeck effect which is detailed in [1],[3]. This type of thermoelectric effect is extensively used in thermopair temperature sensors. These sensors allow temperature measurements. A thermopile is constructed of similar thermopairs connected in series in order to multiply the output signal to a well conditioned level. By using MEMS technology these devices could be miniaturized. A simplified sketch shows such a device in Fig. 1.b). The strong difference in the thermal resistance along the loop is crucial in the operation principle. Primarily it is used for sensing the infrared radiation [3]. Due to the good heat isolation of the membrane the absorbed power heats up the inner ends of the thermopair loops, while the outer ends are thermalized by the substrate to the temperature of the surrounding. The great asymmetry of the heat conduction is the key point of the device. The device can be fabricated by an electric heater on the membrane and then it can sense the gas flow as it results in the cooling of the heater [4]; in addition this construction is an electronic device realizing Quadratic Transfer Characteristics [5]. Regarding the THz and mid-infrared radiation the thermopiles were used only for read-out the temperature increment caused by absorption of the radiation in the feed-point resistance of the metallic antenna [6], or in the thin-film absorber [7].

Another construction for sensing the mm wave and THz radiation was suggested in [2]. In this case the thermopairs are arranged linearly instead of loops, as it is schetched in Fig. 1. a). Here the thermocouple lines act as short circuited dipole antennas. The high-frequency electric field parallel to these lines induce currents and Joule heating in them.

II. FABRICATION PROCESS

The description of the fabrication process can be found in [1]. The conventional poly-silicon thermopile technology was combined and improved by double side bulk silicon micromachining [3],[8],[9]. For reducing the residual stress in the suspended membrane a stacked layer structure was adopted containing a double layer of non-stoichiometric silicon-nitride (SiN_x) and silicon-oxide (SiO_2) with the adequate thickness ratio [10]. The main steps of the process are shown in Fig. 2.

A low stress non-stochiometric silicon-nitride (SiN_x) is deposited on the substrate using low pressure chemical vapor deposition (LPCVD) at 830 ^oC temperature. A SiH_2Cl_2:NH_3 = 4:1 gas mixture is employed within this process. To reduce the stress in the 700nm thick silicon-nitride membrane, a 200nm thick SiO_2 is deposited on top of it. CVD process with SiH_4 precursor is used for this purpose at 450 °C. Using LPCVD at 630 °C and SiH_4 as precursor, functional poly-silicon is deposited and etched for the given geometry (10μm wide poly-silicon strips) defined by the lithography. Ion implantation of boron and phosphorus ions at 40keV is used for setting the p and n doping of the poly-silicon stripes. It is followed by the two step annealing process (600 °C and 1050 °C) which results in 23.7Ω/□ and 37Ω/□ sheet resistance for n and p type poly-silicon respectively. The contact pads and the wiring was structured from evaporated Al. KOH backside anisotropic alkaline etching at 78 °C temperature is deployed for removing the substrate underneath the membrane. Finally the chips were diced and mounted on the prototype panels.



a) b)

Figure 1. Thermopile structures: a) linear arrangement of thermopairs – act like antennas; b) looped arrangement – conventional micromachined thermopile.

III. Infrared Responsivity Measurements

Conventional thermopile detectors originally detect the incident infrared radiation that heats the membrane region. From the point of characterizing of the THz decection properties of the modified thermopile the infrared responsivity of the device also delivers useful information. The sensitivity is defined as the output voltage of the device for a given power that heats the membrane [V/W].

In order to measure the responsivity for infrared radiation a measurement setup has been built that is capable of measuring the performance of the device in atmospheric conditions and in vacuum as well. The vacuum chamber is used for this purpose. Within this chamber a black body with $50\ \text{mm} \times 50\ \text{mm}$ painted with black paint (emissivity $\varepsilon \approx 1$) is placed 20 mm from the detector. The black body is heated with temperature controlled Peltier elements ($\sim$150 W). By setting the temperature of the black body to a specific intensity, the emitted radiation power can be estimated by the Stefan-Boltzmann law:

$$P_e = A \cdot \varepsilon \cdot \sigma \cdot T^4$$

The emitted power varies between 1.36W and 2.74 W for 40°C and 100°C respectively. At this point it has to be noted that the detector senses a temperature gradient, thus the room temperature influences the output voltage.

As mentioned earlier the detector is placed 20 mm from the black body. The two surfaces are parallel to each other, i.e. both surface lie on the x-y plane but with a 20 mm shift on the z axis. The view factor [12] between the two surfaces can be calculated using analytic or numeric methods. The detector has a $1.6\ \text{mm} \times 0.745\ \text{mm}$ effective area. Using the geometries this gives the $F_{12} = 0.000314231$ view factor. Having these all the necessary parameters the incident power on the detector can be calculated:

$$P_i = F_{12} \cdot P_e$$

Figure 3. shows the results of the measurements. The detector in hand has been measured atmospheric and vacuum conditions as well. The responsivity on atmospheric pressure is 12V/W. By creating vacuum in the chamber the responsivity of the device rose to 147 V/W. This phenomena occurs because of the thermal conductivity of the air. The heat that build up on the membrane region is conducted to the environment thru the conductivity of the air. Thus the temperature gradient along the membrane drops significantly which result in the drop of the output thermovoltage. the vacuum in the chamber reaches the 10^{-2} Pa pressure. A proper vacuum casing thus is necessary for the efficient operation. In the mean time it has to be noted that the accumulated heat in the structure falls more slowly in vacuum as the conductivity of the air doesn't help cooling it down.

IV. THz Responsivity Measurements

For a first attempt to prove that the induced current heats up dipole antennas was demonstrated in the K_u band at 13 GHz. The first near THz measurements were performed by P. Basa and G. Károlyi at the Universität Duisburg-Essen at 100 GHz using a microwave source. The measurement proved that the device (not the one that is the subject of the present article) is capable of detecting microwave radiation with a responsivity of 5.6 V/W, where the power was estimated as the product of the chip area and the radiation intensity. I.e. the reflected and transmitted powers are not known. It should be noted here that this situation regards to all radiation measurements, even to the IR and to the broad-band THz.

The first terahertz measurement of the device were performed in the DESY laboratory in Hamburg using a broadband pulsed THz laser source. The source consists of an impulse fs laser source, a Lithium-niobate ($L_iN_bO_3$) non-linear crystal, an attenuator and the necessary equipment. Fig. 4. shows the measurement setup set for calibration. The fs laser hits the $L_iN_bO_3$ crystal that emits a ps pulse that has most of its power in the THz region. The theory behind pulse generation is detailed in [11]. The source provided ps impulses with 1.6μJ energy at about 1kHz frequency, which in sum add up 1.6mJ for the broadband frequency region. The timeseries and the spectra of the impulse can been seen on Fig. 5.

In our measurements the path of the fs laser has been blocked between the waveplate and the mirror. This way only the THz radiation reached the device. The thermopile sample has been placed in the focus beam. The spot size at the focus was about 2 mm. The obtained sensitivity for the THz radiation can be seen on Fig. 3. The measurements have been done in atmospheric conditions. The detector has a 96 V/W responsivity for such a broadband radiation. Such a detector has a 10-100 ms time constant that is order of magnitudes

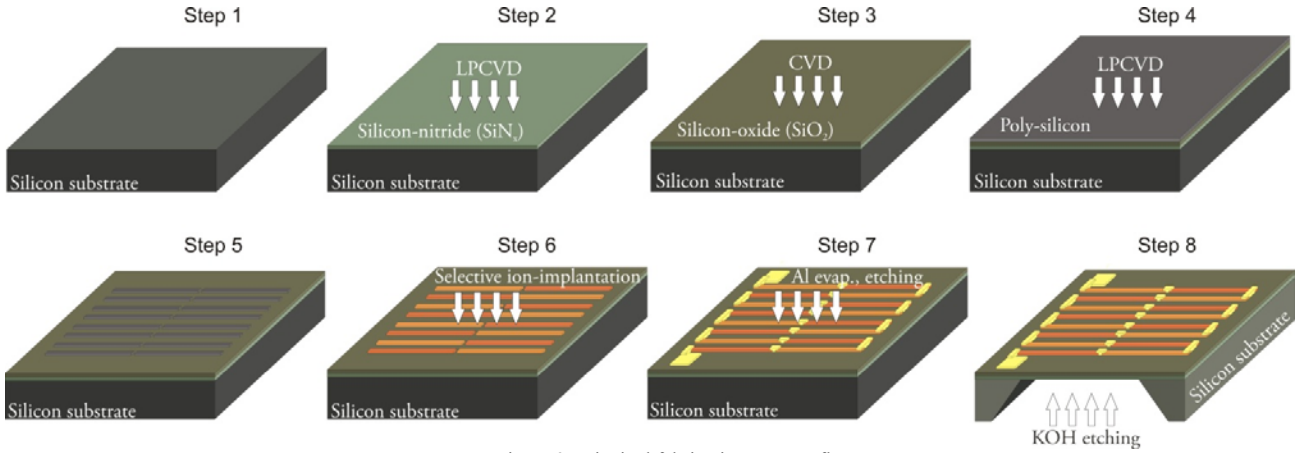


Figure 2. Principal fabrication process flow.

higher than the what the impulses are generated (1ms). Thus the detector senses the sum of the impulse energy in a time interval. It has to be noted that by a proper vacuum casing the responsivity of the detector could be increased by the same magnitude as the vacuum conditions increased the performance for the infrared radiation. This is so, because the conductivity of the air is only affecting the performance of the detector. Using such a casing the responsivity could reach the 1176 V/W value.

CONCULSIONS

First the infrared operation principle of the micromachined thermopile has been demonstrated. Then the idea of the linearly arranged dipole antennas has been detailed, namely that the polarity sensitive antenna array is capable of detecting THz radiation. The fabrication process of such devices is discussed. Two type of responsivity measurements have been performed. One for measuring the infrared responsivity. It has been done by using a temperature controlled black body in atmospheric and vacuum conditions. Another measurement for THz responsivity has been done using an impulse laser source. As a conclusion it can be stated that by using a proper vacuum casing the responsivity of the device could reach as high as 1176 V/W.

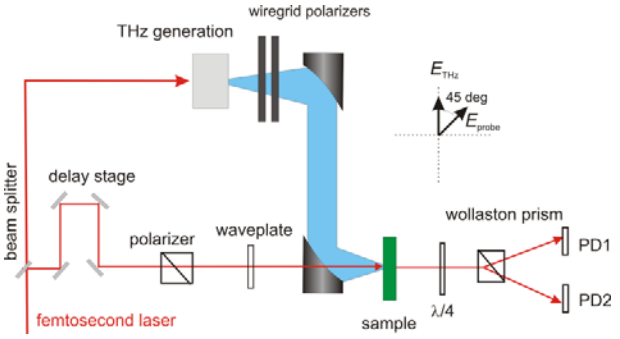


Figure 4. Pulsed laser THz measurement setup.

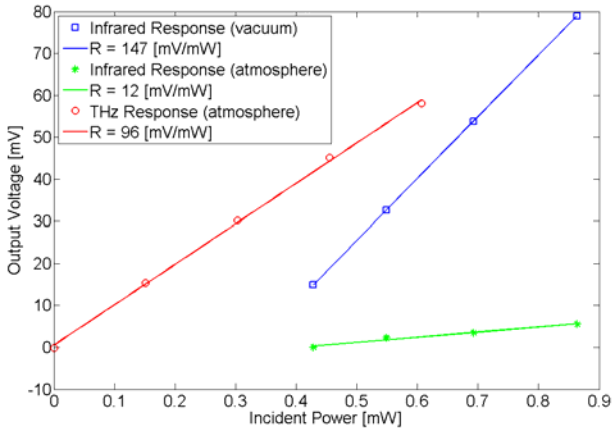


Figure 3. Measured responses for infrared and THz radiations in atmospheric and vacuum conditions. The broad band THz source was used (see below).

ACKNOWLEDGMENT

The measurements with the broadband impulse THz source were performed at DESY, Hamburg. The kind help and contribution of Matthias Hoffman and János Hebling is acknowledged.

REFERENCES

[1] B. Szentpáli, G. Matyi, P. Fürjes, E. László, G. Battistig, I. Bársony, G. Károlyi, T. Berceli, "THz detection by modified thermopile," poster at SPIE Microtechnologies, Prague, Czech Republic, April 2011.

[2] B. Szentpáli, P. Basa, P. Fürjes, G. Battistig, I. Bársony, G. Károlyi, T. Berceli, V. Rymanov and A. Stöhr, "Detection of millimeter waves by thermopile antennas," Appl. Phys. Lett. , vol. 96, 2010.

[3] A. Graf, M. Arndt, M. Sauer and G. Gerlach, "Review of micromachined thermopiles for infrared detection," Meas. Sci. Technol., vol. 18, is. 7, pp. R59-R75, 2007

[4] M. Dijkstra, T. S. J. Lammerink, M. J. de Boer, J. W. Berenschot, R. J. Wiegerink, M. C. Elwenspoek, "Low-Drift U-Shaped Thermopile Flow Sensor," IEEE Sensors 2008, pp. 66-69, 26-29 Oct. 2008.

[5] P.G. Szabó, V. Székely, “Characterization and modeling of electro-thermal MEMS structures,” Microsyst. Technol. vol. 15, pp. 1293-1301, 2009.

[6] N. Chong, H. Ahmed, “Antenna-coupled polycrystalline silicon air-bridge thermal detector for mid-infrared radiation,” Appl. Phys. Lett. vol. 71, pp. 1607-1609, 1997.

[7] I. KasaIynas, A. J. L. Adamta, T. Klaassena, N. J. Hovenid, G. Pandraudc, V. P. Iordanovtc, P. M. Sarro, “Some properties of a room temperature THz detection array,” Proc. of SPIE, vol. 6596, 65960J, doi: 10.1117/12.726404, 2007.

[8] H. Seidel, L. Csepregi, A. Heuberger, H. Baumgärtel, “Anisotropic etching of crystalline silicon in alkaline solutions,” J. Elchem. Soc. vol. 137, pp. 3612-3625, 1990.

[9] É. Vázsonyi, Z. Vértesy, A. Tóth, J. Szlufcik, “Anisotropic etching of silicon in a two-component alkaline solution,” J. Micromech. Microeng. vol. 13, pp. 165-169, 2003.

[10] D. Resnik, U. Aljančič, D. Vrtačnik, M. Možek and S. Amon, “Mechanical stress in thin film microstructures on silicon substrate,” Vacuum, vol. 73, pp. 623-628, 2004.

[11] K. Polgár, L. Kovács, I. Földvári, I. Cravero, “Spectroscopic and eletrical conductivity investigation of Mg doped LiNbO3 single crystals,” Solid State Comm., vol. 59, pp. 375, 1986.

[12] Y. A. Cengel, A. J. Ghajar, Heat and Mass Transfer: Fundamentals and Applications, 4th ed., 2010.

Efficient Mapping of Mathematical Expressions to FPGAs: Placement Problem

Csaba Nemes
(Supervisors: Dr. Péter Szolgay and Dr. Tamás Roska)
nemes.csaba@itk.ppke.hu

***Abstract*—Computationally intensive problems can be represented with data-flow graphs and automatically transformed to locally controlled floating-point units via partitioning. In theory the lack of global control signals enables high performance implementation however placing and routing of the partitioned circuits are not trivial. In practice to create a high performance implementation the clusters should be placed efficiently on the surface of an FPGA using the physical constraining feature of CAD tools. In the paper a new partitioning strategy is presented which not only minimizes the number of cut nets but produces partition which can be mapped without long interconnections between the clusters. The new strategy is demonstrated during the automatic circuit generation from a complex mathematical expression. The proposed partitioning method produces more cut nets than common strategies however the resulting partition can be easily mapped and operate on significantly higher frequency.**

I. Introduction

Computational problems defined on a mesh can be efficiently accelerated by FPGAs. In this type of problems a mathematical expression has to be evaluated continuously and many times over different points of the mesh. The implemented circuit can be decomposed to a memory interface and an arithmetic unit. Our primary aim is to automatically map the given mathematical expression to the FPGA. Automation can speed up the development process and graph optimization techniques can produces even better solutions than manual designs. In the paper a mathematical expression related to a Computational Fluid Dynamics (CFD) [1] problem is used as a test case.

To reach high operating frequency the arithmetic unit shall be partitioned and a local control unit should be assigned to every cluster. In a previous work [2] it was demonstrated that the ideal partitioning minimize the number of the cut nets and the number of I/O connections of the clusters is less than roughly 10. The latter constraint will guarantee that the signals of the control unit will have tolerable fanout and will not decrease the operating frequency of the rest of the circuit. FIFO buffers are used to synchronize the data-flow between the clusters therefore minimization of the cut nets will minimize the number of extra FIFOs and the area requirements of the circuit. There is a trade-off between the speed of the control unit and the size of the circuit, however the operating frequency of the system is determined by the slowest arithmetic unit. In [2] a simple greedy algorithm was proposed which finds a partition with valid I/O constraint however other existing methods can be modified for this task too.

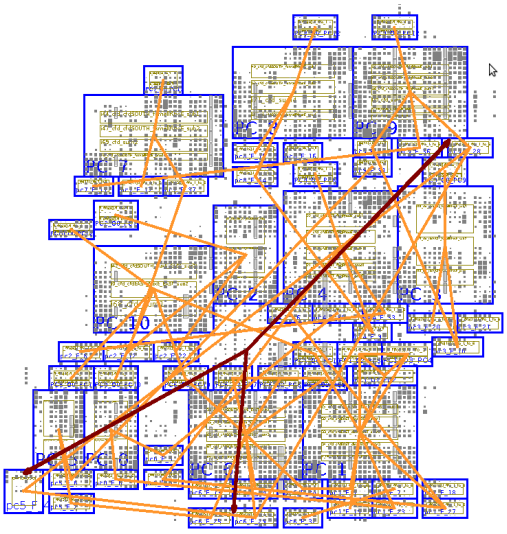


Fig. 1. Placement constraints and the placed instances of a circuit generated from a partition which was created by the greedy partitioner [2] used as a reference. The connectivity of the clusters are indicated by orange lines. The edge cut of the clusters was minimized but the clusters cannot be mapped to FPGAs without long connections. Red lines indicate a critical net (also shown on Figure 2.) associated to an input variable of the original mathematical expression. The input variable is used in three different operations in the expression and the associated three arithmetic units belong to three different and distant clusters. The FIFOs shall be placed close to their clusters and also close to each other but this cannot be solved at the same time.

Two famous graph partitioning algorithm (hMetis [3], spectral technique [4]) were modified and circuits were generated from the partitions produced by all three algorithm. Placement and routing of the circuits in all three cases gave slower operating frequency than expected because distant clusters were connected by long nets. These slow interconnects cannot be eliminated even by using placement constraints because the topology of the clusters is not suitable to keep all the connected clusters close to each other on the surface of the FPGA. In the paper a new partitioning strategy is proposed to create a partition which can be drawn into the plane without long interconnections to overcome the limitations of the common partitioning strategies.

II. Drawbacks of min-cut partitioning algorithms

The number of cut edges can be minimized by common graph partitioning techniques and the size and I/O connections of the clusters also can be balanced or constrained.

The number of cut nets shall be minimized to reduce the extra area requirements of the circuit while constraining the number of I/O connections of the clusters provides high-speed local control units. Implementation of the partitioned circuits without physical constraints gave very poor timing results. CAD tools give the designer the ability to constrain the physical position of parts of the circuit. Constraining the placement of the partitions and the connecting FIFOs gave better timing results in all three cases but it was still far from the theoretical maximum. An example circuit partitioned by the greedy algorithm with placement constraints is shown in Figure 1. A critical net with a fanout of three which limits the operating frequency is colored by red. The net corresponds to an input variable (also indicated by red in Figure 2.) of the mathematical expression which is used in three different operations in the expression. If the partitioning algorithm puts the arithmetic units associated with the three operations to different clusters three FIFOs are generated to solve the synchronization problem. The input of the FIFOs are connected to the same source and to the corresponding clusters. Slow nets cannot be avoided if the clusters cannot be placed close to each other. This phenomena also exists at lower layers (indicated by green in Figure 2.) if a vertex provides input to several vertices they should belong to the same or neighboring clusters otherwise the operating frequency will be limited by the long interconnection. The other two reference algorithms have the same weakness and create partitions which cannot be placed without nets connecting distant clusters. This phenomena gave the idea to design a new algorithm which creates slightly more cut edges but the resulting partition can be placed without long connections between the clusters.

III. Proposed algorithm

The main idea of the algorithm is to draw the graph into the plain before the partitioning starts. If a representation of the graph which minimizes the distance between the connected edges is given a simple greedy clustering algorithm can provide a partitioning without long interconnections between the clusters. Furthermore even the placement becomes trivial and can be easily automated.

One of the initial step of the algorithm is to create a bipartite graph from the original data-flow graph. Every floating-point unit is represented by a vertex of the graph and has a delay associated with it. A bipartite graph can be easily created via a breadth-first search which visits every vertex of the graph and computes the level of the vertex and the time required for the input to reach the given vertex based on the levels and delays of its ancestors. If the levels of its ancestors are different the algorithm can inserts the proper number of extra vertices (delays) after the problematic ancestor. In physical implementation these delays will be shift registers which hold the data for the proper number of clocks. The computed levels will determine the vertical coordinates of the vertices and the layer in which they are.

Vertices get horizontal coordinates randomly then the number of edge crossings is minimized to create a good initial solution. Minimal edge crossing objective does not guarantee good placement but it was found to be a good initial solution for our vertex swapping iterative algorithm. In physical implementation an edge crossing is not a limiting factor if the edges goes to the same cluster. However the global operating frequency of the system can be affected by edges going to different or non neighboring clusters.

A. Objective

The objective is to minimize the distance between the connected vertices. As the partition clusters are determined based on the position of the vertices this objective automatically avoid long interconnections between the clusters. The distance between two vertices are determined according to their horizontal coordinates:

$$distance(A,B) := \begin{cases} (x_A - x_B)^2 & \text{if A and B are connected} \\ 0 & \text{otherwise} \end{cases}$$

where x_A and x_B are the horizontal coordinates of vertex A and B respectively. Both horizontal and vertical coordinates are integer numbers. The physical size of the floating-point units are not considered in this representation and set to 1. Vertical coordinates can be neglected because the distance is always one as the graph is a bipartite graph.

Manhattan distance cannot be applied: if vertex A is connected to two other vertices which are relatively far from each other it only guarantees that vertex A will be somewhere between the two vertices but not at the middle of the interval.

B. Proposed algorithm

The proposed algorithm consists of two greedy phases. The first phase collects global information about the graph by positioning the vertices, while clusters are created in the second phase using the information encoded in the spatial position of the vertices.

1) Barycenter and iterative movement: The first greedy phase tries to minimize the distance between the connected vertices which is the objective function. A simple swap-based algorithm like Kernighan-Linn [5] have been designed which minimize the distance between vertices of all layers together. This algorithm can be easily trapped in the local minimum therefore the initial placement of the vertices is critical. Barycenter heuristic [6] is a fast and simple algorithm to create an initial solution for our purposes however various min-crossing algorithms can be chosen to gain even better initial solution [7]. Barycenter heuristic tries to minimize the edge crossing in a layered digraph. The minimization of the edge crossing is NP-complete, even if there are only two layers [8] however barycenter heuristic is one of the best heuristic available [9]. Barycenter method is basically a layer-by-layer sweep method: in every iteration one layer of the digraph is fixed and the vertices of the next layer is arranged. The horizontal coordinate (x_A) of each vertex(A) is chosen to the barycenter of its neighborhood in the fixed layer:

$$x_A := \frac{1}{deg(A)} \sum_{v \in N_A} x_v$$

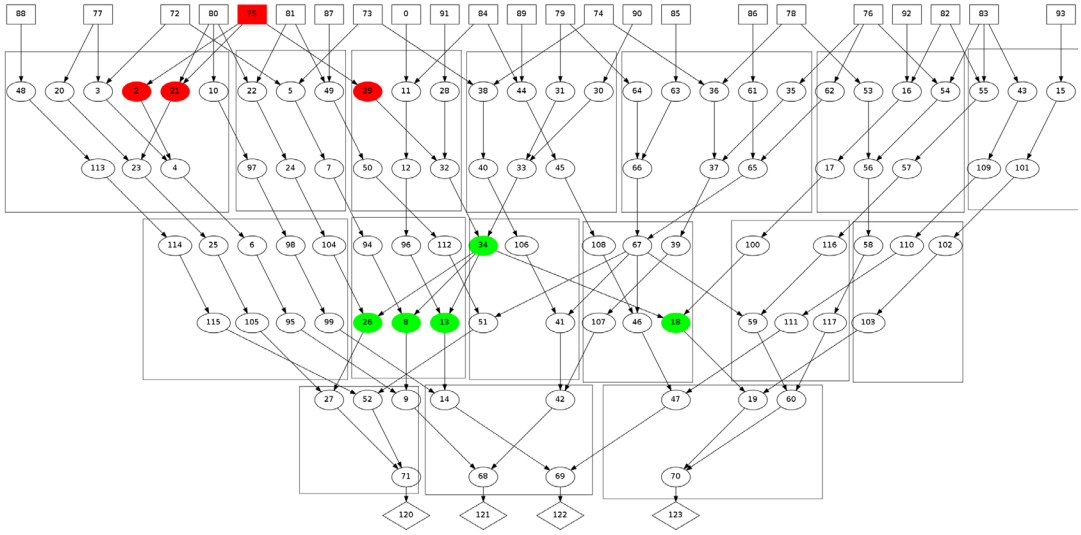


Fig. 2. Partition of the CFD graph created by the proposed algorithm. Inputs and outputs of the mathematical expression are represented by small rectangles and diamonds while floating-point units are represented by circles respectively. The clusters of the partitioning are indicated by big rectangles. Vertical and horizontal position of the vertices were determined in the first phase of the algorithm to minimize the distance between the vertices. Clusters can be easily mapped to FPGAs keeping the same relative positions as shown in Figure 3. Vertices colored by red and green are typical examples for vertices which should be placed close to each other because they will have common nets in the implementation.

where N_A denotes the set of the neighbors of vertex A and x_v denotes the horizontal coordinate of vertex v.

Result of the first phase applied to the examined graph shown on Figure 2.

2) Rectangular clustering based on position: The second phase is another simple greedy algorithm which search the results produced by the first phase for rectangular clusters. Height of the rectangular domains can be chosen arbitrary however in our examples it is set to two. The algorithm is started from the top left corner and the largest possible rectangular cluster is searched which still meets the I/O constraint. Next the algorithm moves left and search for the largest possible rectangular cluster of the rest of the unclustered vertices. If there are no more unclustered vertices on the selected layers the algorithm moves down and continues with the lower layers.

In the second phase decisions can be made on global information even with a simple greedy mechanism. Better digraph or planar partitioning algorithms are published in the literature and will be considered in our future works however even this algorithm makes the circuit generation significantly better than the previously used reference algorithm [2]. The resulting partition clusters of the second phase is shown in Figure 2.

3) Main steps of the algorithm: The main steps of the algorithm are summarized in Algorithm 1:

Algorithm 1 Main steps of the proposed algorithm.

1: Create delay vertices to make the graph bipartite and associate every vertex with a level according to how many levels its ancestors has.
2: Place every vertex into a layer according to the associated level. (The order of vertices on a given layer is random.)
3: Create an initial solution for the iterative swap-based algorithm by barycenter heuristic.
4: Find the vertical position of the vertices and a local minimum of the objective function using a simple iterative swap-based algorithm.
5: Determine the rectangular clusters based on the vertical and horizontal positions of the clusters using a greedy algorithm.

IV. Implementation results

The circuits were implemented on a Xilinx Virtex-6 SXT FPGA (XC6VSX315T) with speed grade -1. Position of specific parts of the circuit is constrained by using the Xilinx PlanAhead [10] software. It enables the user to create rectangular placement constraints also called pblocks. FIFOs and floating-point units were generated by the Xilinx Core Generator [10].

The arithmetic unit of the CFD problem consists of nearly 50 floating-point units therefore placement of smaller partitioned graphs with 5-10 vertices had been investigated before its implementation. The Xilinx P&R tools are likely to disperse the registers of the FIFOs which can limit the operating frequency of the circuit therefore separate pblocks should be created for the FIFO buffers. On the other hand the Xilinx P&R tools were able to place and route the floating-point units inside the clusters if one pblock is defined for each cluster. Therefore this method has been applied during the implementation of the partitioned CFD graph and one pblock is generated for every FIFO and every cluster.

Currently the pblocks was placed manually with the help

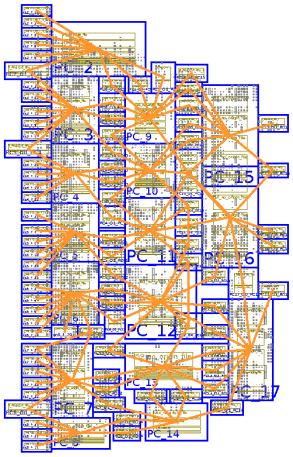


Fig. 3. Placement constraints and the placed instances of the arithmetic unit of the CFD problem generated by the proposed algorithm. However it has more clusters and cut edges compared to the result of the reference algorithm (see Figure 1.) mapping of the clusters to the FPGA is trivial and can be automated based on the spatial position of the vertices (see Figure 2.)

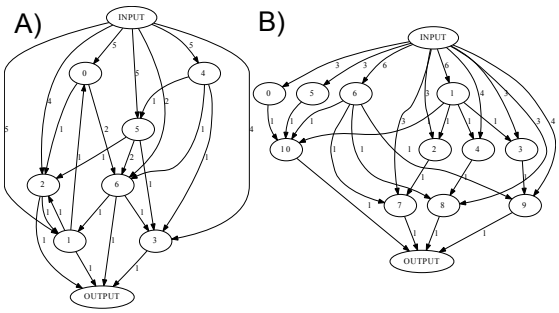


Fig. 4. Connections of the clusters plotted by graphviz [11] in case of hMetis (A) and greedy (B) partitioner.

of a graphviz [11] plot depicting the connections between the clusters (see Figure 4). In case of the proposed algorithm placement was very straightforward and the resulting layout is shown in Figure 2.

In case of the reference algorithms the placement of the clusters to minimize the length of their connection was very challenging. Placement constraints and the placed instances are shown in Figure 1. in the case of the greedy reference algorithm. Red lines indicate a critical net with a large propagation delay which cannot be avoided because the endpoints are placed into distant clusters. Placement results of the proposed algorithm is displayed in Figure 3. and no long connections can be observed between the clusters. Partitioning results, resource utilization and operating frequency are compared on Table I. Partitioning results of the proposed algorithm is slightly worse and the area requirements are greater than the reference algorithms however operating frequency is significantly increased.

V. Conclusion

Graph partitioning strategies can create partitions which can be converted to circuits with small extra area requirements and high operating frequency. However operating frequency cannot reach the theoretical maximum if the clusters cannot be mapped to the surface of the FPGA efficiently. An algorithm was successfully designed in which the vertices of the input graph are placed to minimize the distance of the connected vertices then the positioned vertices are partitioned. The resulting clusters can be mapped to the FPGA without long interconnections and the operating frequency can be further improved in the price of a slight area increase.

TABLE I
Implementation results of various partitioning strategies using the CFD graph.

	without partitioning*	modified hMetis	Greedy [2]	Proposed algorithm
Number of clusters	-	7	11	16
Number of extra FIFOs	23	49	55	89
Number of Slice Registers	15,534	18,866	19,148	21,998
Number of Slice LUTs	12,084	14,275	14,614	16,883
Number of occupied Slices	5351	4,284	4,801	5,751
Clock frequency (MHz)	133.97	325.627	275.482	369.959

∗Implemented on Virtex-5 FPGA.

References

[1] S. Kocsárdi, Z. Nagy, A. Csík, and P. Szolgay, "Simulation of 2D inviscid, adiabatic, compressible flows on emulated digital CNN-UM," *International Journal on Circuit Theory and Applications*, vol. 37, no. 4, pp. 569–585, 2009.

[2] C. Nemes, Z. Nagy, M. Ruszinkó, A. Kiss, and P. Szolgay, "Mapping of high performance data-flow graphs into programmable logic devices." *NOLTA 2010. International symposium on nonlinear theory and its applications.*, pp. 99 –102, Sept. 2010.

[3] G. Karypis and V. Kumar, "hmetis 1.5: A hypergraph partitioning package," http://www.cs.umn.edu/ metis, 1998.

[4] C. J. Alpert and S.-Z. Yao, "Spectral partitioning: The more eigenvectors, the better," in *Proc. ACM/IEEE Design Automation Conf*, 1994, pp. 195–200.

[5] B. W. Kemighan and S. Lin, "An efficient heuristic procedure for partitioning graphs," 1970.

[6] K. Sugiyama, S. Tagawa, and M. Toda, "Methods for visual understanding of hierarchical system structures," *Systems, Man and Cybernetics, IEEE Transactions on*, vol. 11, no. 2, pp. 109 –125, feb. 1981.

[7] G. D. Battista, P. Eades, R. Tamassia, and I. G. Tollis, "Graph drawing: Algorithms for the visualization of graphs," 1999.

[8] P. Eades and N. C. Wormald, "Edge crossings in drawings of bipartite graphs," *Algorithmica*, vol. 11, pp. 379–403, 1994.

[9] M. Jünger and P. Mutzel, "2-layer straightline crossing minimization: Performance of exact and heuristic algorithms," 1996, this article was published in 1997 by Brown University in the journal "Journal of Graph Algorithms and Applications", volume 1, pages 1-25. [Online]. Available: http://gdea.informatik.uni-koeln.de/22/

[10] "Xilinx product homepage," http://www.xilinx.com, 2010.

[11] J. Ellson, E. R. Gansner, E. Koutsofios, S. C. North, and G. Woodhull, "Graphviz - Open Source Graph Drawing Tools," *Graph Drawing*, pp. 483–484, 2001.

Cellular Stochastic Optimization for Integral Calculation

András Horváth

(Supervisors: Dr. Miklós Rásonyi and Dr. Tamás Roska)
horan@digitus.itk.ppke.hu

***Abstract*—During the first four semesters We have developed, simulated and tested new topographic, cellular variants of stochastic optimization methods: genetic algorithms and particle filter algorithms. We also created a method in which particle filters can be used for numerical integration.**

***Keywords*-Particle filter, Processor array**

I. Introduction

Sequential Monte Carlo methods arose for the computation of optimal estimates in nonlinear and non-Gaussian state-space models where analytic solutions are not available. They found applications in diverse areas such as localization, navigation, tracking, robotics and signal processing, see e.g. [1] for a representative sample. More recently they have been applied in financial mathematics (to stochastic volatility models and to the computation of credit losses, see [2] and [4]). For the mathematical theory, consult [3].

In this short summary I will briefly describe our variant of the Genetic Algorithm, that suits ideally on a coarse-grain architecture.

The new variants of these algorithms could be implemented on an array of processors and, using parallelism and local communication, could greatly enhance computational speed without substantial loss in precision.

It is already clear that the basic idea of the new algorithm may be applied to a much more general class of stochastic optimization. This is subject of current research.

II. Genetic algorithms

Genetic algorithms (GAs) and their utility in practical problems have been introduced by John H Holland in 1975 [5]. Since then these methods (with various alterations) have been applied to a large class of problems. These heuristic search algorithms, inspired by natural selection and evolutionary mechanisms, can provide solutions faster than exhaustive search, and give better results than greedy algorithms.

It can be observed in biological networks and among organisms that mate selection and genetic inheritance between generations happens locally, according to a topographic rule. The genomes that are able to overcome the challenges of nature in a well-defined environment (in the territory of the individual) can be inherited to the next generation, and spread out in a diffuse-topographic way in the population.

Examining evolution and natural selection we can see that a parallel and topographic approach is more realistic and similar to the original, motivating idea than the commonly used "global", non-topographic selection rules. This also refers to GAs, and a parallel topographic implementation can outperform its "regular" single-core ancestors.

Genetic algorithms are often used in complex tasks with strict time limits (or with high memory requirements). A parallel, fast, effective implementation can be useful for solving a wide range of problems.

The global selection and information gathering makes general genetic algorithms unfeasible to be implemented on a CNN in an effective way. Cellular architectures encapsulates the previously mentioned similarities and advantages that our living environment has in case of the natural selection. An altered version, the cellular genetic algorithm with local selection rule can easily exploit all the advantages of cellular architectures.

The algorithm was tested on two typical benchmark problems the N-queen problem (see [8], [7],[6]) and the knapsack problem ([9],[11]). These two problems are the most frequently used tasks to gauge the efficiency of genetic algorithms.

1) Description of the cellular genetic algorithm: The regular genetic algorithm can not be effectively implemented on a parallel architecture because for the selection step we need to collect fitness values from all the genomes. All the other steps could be easily implemented on a fine-grained (single instruction multiple data) architecture, where every processing unit represents one genome.

Mutation, as well as fitness calculation effects the state of one element only, thus their calculation does not need information from any other genomes. Hence these operations can be carried out easily in a parallel way on every processing unit.

We will use an altered version of GA, a two dimensional variant of the so-called cellular genetic algorithm. The theory of these algorithms was introduced in [14], [13]. Here we will implement a two-dimensional variant on an nxn array, where the radius of the communicating neighbourhoods can be set arbitrarily, by repeating the parent selection step with one neighbourhood radius. We can set the exploration/exploitation ratio with this parameter, instead of using an nxm grid and varying the n/m ratio, as in [12]. In this method we

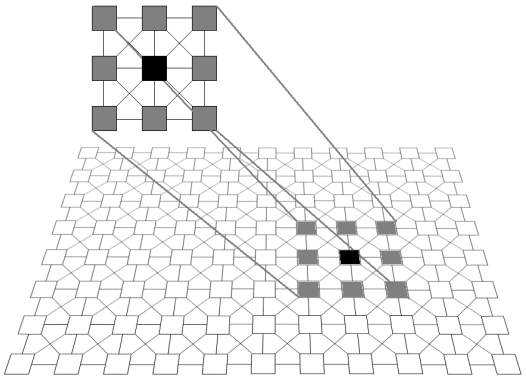


Fig. 1. The small set, the neighbourhood of a given processor (representing a particle during the algorithm), that determines the calculations. The grid structure represents how the information spreads out from processor to processor, from neighbourhood to neighbourhood.

will determine the parents locally, the selection of the fittest genomes and the recombination is done in a topographic way, just like in the case of real organisms.

In an iteration every genome will select the fittest genomes in a neighborhood of radius $NegihSize$, and these parents will create the gene pool of the respective genome for the next generation.

We will make one crucial alteration with respect to the original GA, so as to avoid the global ordering of genomes according to their fitness. We will rather find and use the local extrema (minima or maxima) of the fitness values. With this we can ensure parallel execution, and a prefect mapping to a cellular architecture.

In the present paper we will use only the simplest, but also most general variation of cGA. This contains all the important features from the implementation point of view. This implementation can be used for solving other types of problems, and can be adapted to improved versions of GAs, too.

Because heuristic improvements (like [15] or [16]) nearly always intervene at the stage of mutation, fitness calculation or at other genome-dependent stages they can also be realized in the topographically distributed, multi-parallel implementation on CNN architecture.

The pseudo code of a cGA using deterministic sampling is the following: Algorithm 1.

The parameters of the algorithm are the following:

$PopsSize$: the size of the population, i.e. the number of genomes used in an iteration.

$MutFact$: mutation factor.The probability that a randomly selected gene will change its value.

$NegihSize$: Neighborhood size. This parameter determines the size of neighborhood in which the parents are searched. The larger value means more possible parent candidates.

A. Performance analysis of cellular genetic algorithms

Various versions of cellular genetic algorithms have been investigated and proved to have a faster convergence than

Algorithm 1 Cellular Genetic Algorithm

Require: $MutFact\ NegihSize\ PopsSize\ MaxIter$
Ensure: $gmin$
  $gmin \Leftarrow 1$
  $Iter \Leftarrow 0$
  while $Iter < MaxIter$ AND $gmin \neq 0$ **do**
    {1- initialization};
    for $i = 0 toPopSize$ **do**
      for every $gene$ in g_i **do**
        $gene \Leftarrow randomgene()$
      end for
    end for
    {2- selection recombination};
    for $i = 0 toPopSize$ **do**
      for each $neighbour$ in $LocalNeighbourhood(g_i, NegihSize)$ **do**
        $Parent1$, $Parent2 \Leftarrow SelectParents(neighbour)$
      end for
      {recombination};
      $gnew_{i+j} \Leftarrow recombine(Parent1, Parent2)$
    end for
    $g \Leftarrow gnew$
    {mutation};
    for $i = 0$ to $PopSize$ **do**
      for every $gene$ in g_i **do**
        $a \Leftarrow randomnumber(0, 1)$
        if $a < MutFact$ **then**
          $gene \Leftarrow randomgene()$
        end if
      end for
    end for
    $Iter \Leftarrow Iter + 1$
  end while

their 'regular' counterparts [17]. However, a two-dimensional mapping on the algorithm, with varying neighbourhood size (repeated steps on neighbourhoods in one iteration) has never been implemented on a cellular architecture. Only one dimensional cellular arrays [14] or grids with fix negihbourhood radius (usually one) but varrying grid shapes have been used [10].

Because GAs (especially cellular versions) are too complex to investigate their convergence speed analytically, we decided to simulate them on a virtual cellular architecture before the actual implementation.

Simulations were carried out with the following parameters: 1096 genomes for the N-queen problem and 512 genomes for the knapsack problems (bounded and unbounded versions) were generated with uniform distribution in the state space. 1000 repeated tests were executed and averaged for every single parameter set. According to the compiler results for the N-queen problem for a population with 1096 elements one iteration takes 15434 clock cycles (the cores on the Xenon chip are operating with 100 Mhz, this means 154 usec per iteration). With this performance we can execute 6479 iterations in one

second. For this problem $4-5$ iteration is usually enough to find the optimal solution. With 6479 iterations we expect that one could solve much more difficult problems in less than a second.

The total running time was distributed amongst the operations as follows:

Calculation of the fitness values: 6698 clock cycles (67 usec).

Selection of two parents (with neighborhood radius 1, i.e. from nine elements): 5700 clock cycles (57 usec). We remark that with radius k it would take $k \times 57$ usec, because we have to iterate the selection k times. The execution with larger neighborhood radius has also been tested on the real architecture.

Crossover: 2850 clock cycles (28 usec).

Mutation: 96 clock cycles (0.9 usec).

The bounded knapsack problem can be solved with the cGA in an average 7.12 iterations while the unbounded one in average 13.47 iterations. With the GA we will need 11.43 and 17.92 average iterations for the bounded and the unbounded knapsack problems.

III. Integral calculation

In the previous semesters I have examined and implemented a cellular version of the Particle Filter algorithm, that suits ideally on kilo-processor architectures.

It was shown that Particle filters (PF) can be used for state estimations and probability estimations. They work well with in case of difficult, non-linear, Markovian models, even in those cases where the Kalman filters can not be used.

Particle filters without resampling are not efficient enough to be used in complex, practical problems.

On the other hand Particle filters with resampling have good and much higher performance but they will lose the ability to be used for probability estimations.

With a small alteration PF with resampling can be used for probability estimation, this change requires some extra computation, but the number of particles can be decreased drastically.

A. PFs with resampling

However even with resampling one can calculate the probabilities between state transitions with the following 'trick'.

Lets assume our process is in step j (no resampling have occurred in the previous steps), before resampling and state estimation (here i can make assumption about state x_j).

We can calculate the weights for every particle, and divide them into two groups: particles with matching and differing observations. Here we can estimate, that until point j, the probability that our trajectory will be identical with the given observation (o_t) is:

$$P(y_j = o_j) = \frac{G_j}{N} \tag{1}$$

Where G is the number of matching trajectories and N is the total number of particles ($N = W + G$). After the resampling step all the particles will be considered as 'matching particles (G), so $G = N$ if we would make a calculation after the resmapling. Because this is still iteration j, the probability has to be equal to the previous expression. To calculate the same probability after the resampling step we can introduce a 'correcting factor': Before the resampling step we had G matching particles from N particles, and now we have N matching particles. We can add virtual particles (they will not effect our computation) and consider that in the beginning we have started the algorithm with $N * (N/G)$. This way we can conserve the calculated probability. (we can note the number of virtual and real particles at iteration j with $N_j = N * (N/G)$)

After the k-th iteration the probability of a trajectory will be:

$$P(y_k = o_k) = \frac{G_k}{N_k} \tag{2}$$

because the probability was conserved during every resampling.

This way I can calculate probabilities with resampling particle filters, which makes the number of particles to be used exponentially less. for this model and for a trajectory with 100 samples: the Pf with resampling could calculate the probability with 300 particles with the same accuracy. With the old method (without resmapling) I needed 10000000 particles.

The main point is, that this way the number of used particles will not depend on the length of the trajectory.

IV. Estimation of conditional probability of a trajectory

In many application our aim is not to determine the probability of the observations, but to determine the probability of a trajectory of the hidden states.

from a more practical point of view: -first we have an observation according to our model -from this observation we can approximate the trajectory of the unknown hidden states, usually there are more possible sequence of hidden states to results the given observations (especially if our observation is nonlinear) - as the last step we should determine the probability of one selected sequence of hidden states that matches our observation.

In a more mathematical form:

$$\begin{aligned} p(x_t = s_t, x_{t-1} = s_{t-1} ... x_1 = \\ s_1 | y_t = o_t, y_{t-1} = o_{t-1} ... y_1 = o_1) \end{aligned} \tag{3}$$

where s_t is a given series of hidden states and o_t is a sequence of observation.

A. Results

With this type of calculation one iteration of θ' with 1000 particles and over a 30 steps long trajectory could be calculated in 5.6 second on a single core architecture and θ' can be calculated with average error 0.02.

In case of the normal method (without resampling)θ' can not be estimated with 1000 particles, because in most of the cases there are no trajectories that will match our observations.

This way for a longer trajectory the computation time will increase linearly (instead of the exponential growth of the normal method).

With 10000 particles θ' can be approximated with 0.006 average error.

With 10000 particles the 'regular method' could approximate theta with 0.4 average error.

B. Estimation of the probability of a trajectory

Estimation of a trajectory based on a model is relatively simple. The problem and its solution can be explained through the following simple example:

Let us have a first order autoregressive model:

$$x_t = x_{t-1} * \theta + N(0,1) \tag{4}$$

Where x_t is a system state at time t, the state transition rule is known, but we will not have the x values during the simulation.

From the system we will have the following observation:

$$y_t = \begin{cases} \lfloor x_t \rfloor, if : x_t - \lfloor x_t \rfloor < 0.5 \\ \lfloor x_t + 1 \rfloor, if : x_t - \lfloor x_t \rfloor \geq 0.5 \end{cases} \tag{5}$$

We call y the observed state. This is a regular quantized signal based on nearest neighbor quantization, these observations can be seen in many practical problems.

This is a function of x_t for simpler notation we will use it as:

$$y_t = \phi(x_t) \tag{6}$$

In practical problems we usually have one trajectory generated by the previous model: r_t which contains a given number e.g.: 100 iterations $t = 1, 2...100$.

We have a previously defined series of the observations: o_t and our aim is to estimate the probability that from the given model we will get the given trajectory of observations:

$$p(y_t = o_t, y_{t-1} = o_{t-1} ... y_1 = o_1) \tag{7}$$

or with a simpler notation: $P(y_t = o_t)$ for every $t = 1, 2...100$

In case of models with infinite state space this calculation is hard (can not be done) analytically, hence the non-linear observations.

Particle filters are used in case of this difficult stochastic models, and non-linear observations. They are useful in state estimation, because the distribution of the particles follows the distribution of the model. From these distributions we can make probability estimation according to our trajectory.

V. Conclusion

In the previous semesters I have implemented a cellular version of the genetic algorithm and tested it not only in simulations but also on the Xenonv3 architecture. This algorithm uses the advantages of topographic algorithms and the structure of cellular neural networks, see [18]. The present study shows, that this algorithm – implemented on cellular architecture like the Xenon-v3 – could provide a solution to many problems where optimization task is complex and an optimal or suboptimal solution has to be reached with a strict time limit. I have also introduced a brief description how particle filters with resampling can be used for the estimation of the probability of different trajectories. This method allows numerical integration of certain autoregressive moving average processes, however the validation, simulation and application of this method is in progress.

References

[1] Djurić, P.M. and Godsill, S. J., guest editors). Special issue on Monte Carlo Methods for Statistical Signal Processing *IEEE Transactions on Signal Processing*, Vol. 50, Feb. 2002.

[2] Doucet, A. and Johansen, A. M. A tutorial on particle filtering and smoothing: fifteen years later. *To appear in: Oxford Handbook of Nonlinear Filtering*, Oxford University Press, 2010.

[3] Del Moral, P. Genealogical and Interacting Particle Systems with Applications *Feynman-Kac Formulae* . Springer, 2004.

[4] Carmona, R., Fouque, J.-P. and Vestal, D. Interacting particle systems for the computation of rare credit portfolio losses. *Finance and Stochastics*, Vol. 13, pp. 613–633, 2009.

[5] J. H. Holland, *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*. MIT Press, 1975

[6] C. Letavec, and J. Ruggiero, "The n-Queens Problem", *INFORMS Transactions on Education*, vol. 2, no. 3, pp. 101-103, (May. 2002)

[7] A. Bruen and R. Dixon, "The n-queens problem" *Discrete Mathematics*, vol. 12, pp. 393-395, (1975)

[8] W. Ahrens, *Mathematische Unterhaltungen und Spiele*. Leipzig: B.G. Teubner, 1910

[9] K. Kogan, "Unbounded knapsack problem with controllable rates: the case of a random demand for items," *Journal of the Operational Research Society*, vol. 54, no. 6, pp. 594-604, (2003)

[10] M. Tomassini: "The parallel genetic cellular automata: Application to global function optimization," *Proceedings of the International Conference on Artificial Neural Networks and Genetic Algorithms*, Innsbruck, Austria, pp. 385391, (1993)

[11] J. Puchinger, G. R. Raidl, and U. Pferschy, "The Multidimensional Knapsack Problem: Structure and Algorithms" *INFORMS Journal on Computing*, vol. 22, no. 2, pp. 250-265, (2010)

[12] E. Alba and M. Tomassini, "Parallelism and Evolutionary Algorithms" *IEEE Transactions on Evolutionary Computation*, vol. 6, no. 5, pp. 443-462, (October 2002)

[13] M. Nowostawski and R. Poli, "Parallel genetic algorithm taxonomy" *Third International Conference Knowledge-Based Intelligent Information Engineering Systems* Adelaide, South Australia, pp: 88-92, (1999)

[14] H. H. Hoos, and T. Sttzle: *Stochastic Local Search Foundations and Applications*. Morgan Kaufmann - Elsevier, 2004

[15] G. Rudolph and J. Sprave, "A Cellular Genetic Algorithm with Self-Adjusting Acceptance Threshold", *First International Conference on Genetic Algorithms in Engineering Systems: Innovations and Applications*, Sheffield, UK, pp: 365 - 372, (1995)

[16] L. Loukil, M. Mehdi, N. Melab, E. G. Talbi, and P. Bouvry, "A parallel hybrid genetic algorithm-simulated annealing for solving Q3AP on computational grid" *IEEE International Symposium on Parallel and Distributed Processing*, Chengdu, China, pp: 1 - 8, (2009)

[17] E. Alba and B. Dorronsoro, "The Exploration/Exploitation Tradeoff in Dynamic Cellular Genetic Algorithms" *IEEE Transactions on Evolutionary Computation*, vol. 9, no. 2, pp. 126-142, (April 2005)

[18] Roska T, Chua LO, The CNN universal machine: an analogic array computer, 3rd ed. *IEEE Transactions on Circuits and Systems II-Analog and Digital Processing* 40:(3) pp. 163-173, 1993.

The Amoeba Constructive Metaheuristic and Its Derivatives for the Sequential Ordering Problem(SOP)

Antal Hiba
(Supervisors: Dr. Peter Szolgay and Dr. Miklos Ruszinko)
hiban@digitus.itk.ppke.hu

***Abstract*—Novel heuristic methods for complex optimization problems are based on high level algorithmic models, called metaheuristics. These models provide an ability to create generalized algorithms, and also open the way for defining hybrid algorithms. Hybrid metaheuristics are often parallel, and use multiple heuristics, which are communicating with each other, during the optimization process.**
In this paper a new constructive metaheuristic method will be introduced. It grows particular solutions parallel, which are fused step-by-step to a single general solution of the optimization problem. It is similar to the behavior of amoebas. The Amoeba method uses subordinate heuristics, which can be chosen optionally. We show a derivative heuristic of Amoeba for the Sequential Ordering Problem(SOP), called SOP-Amoeba1, which can keep the local error summing down, and provides similar quality solutions to the naive greedy method.
SOP-A1 algorithm is parallel, and this method has more additional benefits than growing a single particular solution.

I. Introduction

A. Metaheuristics and the search landscape

Metaheuristics typically define generalized ways of scouting the search landscape. Search landscape in combinatorial optimization, is a discrete, high dimensional solution space, with a neighborhood function, which depends on the allowed elementary modifications on solutions.[1]
Basic metaheuristics are Local Search, Simulated Annealing(Kirkpatric 1983), TABU Search(Glover 1989), Variable Neighbourhood Search(VNS), Guided Local Search(GLS), Ant-Colony Optimization(ACO), Particle Swarm Optimization(PSO) and Evolutional Optimization(Holland 1975, Fogel 1994), brief summary and further references can be found in [2]. All methods have to deal with the diversification-intensification problem: on one hand the convergence which leads into a local optimum have to be restrained, on the other hand it is necessary to reuse the knowledge from best known solutions. All of the above mentioned heuristics are iterative, and define how to move in the search landscape, and how to handle the local optimum problem. Iterative methods use constructive heuristics to build initial solutions, and improve them till it is possible, so iterative methods provide better solutions, but it takes more time. Constructive heuristics always take part in iterative and population based methods, and be used alone in special cases, where the execution time is critical, so the examination of these methods are important.

[1]we get all of the neighbors of a solution with one of the elementary modifications

B. Hybrid metaheuristics

The goal of hybridization is combining the advantages of the participating methods. The integration of a constructive heuristic into an iterative one, is a trivial example. There are two basic forms of hybridization: integrative combination, and collaborative combination. Integration means that a metaheuristic has integrated other methods, which leads better solutions. In collaborative combinations, metaheuristics run parallel and exchange information to each other. This is an emerging research direction, summaries in [1], [3] show promising examples. A hybrid metaheuristic method was suggested for handling SOP in [4], where an Ant-Colony System hybridized with a special 3-exchange local search. In that work, the authors used TSPlib reference problem set, which included both real world SOP instances, and random problems for testing optimality. In this paper we use the same problem set for measuring performance. We hope that the metaheuristic, which will be presented here, can be applied in the future hybrid methods, but in this work we show only the method itself, and its derivatives for the SOP.

II. SOP-Amoeba

A. Amoeba metaheuristic

The main concept of Amoeba method is building many particular solutions parallel. This method is a kind of greedy heuristic, in each step a solution element is chosen. There are many solution parts, and all of them want to join to an other solution part. In finite steps we have only one particular solution, which will be the output of the method.

Amoeba Metaheuristic:
1 create a partition of the task
2 initialize the elementary particular solutions
while num of independent particular solutions $\neq 1$
3 update particular solutions
4 choose a part
5 the chosen part incorporates its candidate

Amoeba has two subordinate heuristics: one for choosing a part(part-choosing heuristic), and one for selecting candidates

for each part(candidate-choosing heuristic). When the number of independent solutions reaches 1, we have a solution for the primary problem. It is trivial that the naive greedy method is an Amoeba, which always chooses the same part during the process.
We can create Amoeba methods for different optimization problems by defining the above mentioned points of the Amoeba metaheuristic. In graph-based problems, the particular solutions are subgraphs(for example: paths), which create connections to each other, when a subgraph incorporates its candidate subgraph.

B. Sequential Ordering Problem(SOP)

The sequential ordering problem(SOP) is a combinatorial optimization task, which was defined by Escudero(1988). SOP can be derived from the traveling salesman problem(TSP), by adding precedence constraints to the cities. Every city can have a list of other nodes, which have to be before them in the solution. Because of all TSP instances are special SOP ones, SOP is also NP-hard. Many real-world tasks leads to SOP: pick-up and delivery tasks, planning robot actions(to build multilevel buildings with minimal cost), ordering problems with cost function minimization. Formal definition:
Given: A directed complete graph DK_n and a weight-function $c : E(DK_n) \rightarrow R_+$ and a poset P on set $V(DK_n)$ with relation R(P).
Task: Give a permutation $V(DK_n) = \{v_1, v_2, ..., v_n\}$, where $\sum_{i=1}^{n-1} c(e(v_i, v_{i+1}))$ is minimal on the linear extensions of P.

R(P) is a set of constraints, $(v_i, v_j) \in R(P)$ means that v_i have to be before v_j. Linear extension of P is a total order, where all of the relations in R(P) are satisfied. In practice, SOP is given by a matrix A, where $A_{i,j} = -1$ if $(v_j, v_i) \in R(P)$, $A_{i,j} = c(e(v_i, v_j))$ otherwise.

C. SOP-Amoeba

Amoeba metaheuristic is a generalized optimization model. Here we show a specialized Amoeba method which constructs solutions for the SOP. If the following points are inserted into the Amoeba metaheuristic, we get the SOP-Amoeba method.

1. CREATE PARTITION OF THE TASK:

We divide the primary task into n decision tasks. If all nodes decide which are the next nodes in the route, we get a candidate solution. Elementary particular solutions are sections(paths), which have only one node. These sections joins to each other till we get a Hamiltonian path, where the constraints defined in R(P) are satisfied. At this point, each node v_i is a section and has these data structures:
$CAN(v_i)$: list of candidate sections for section v_i
$Route(v_i)$: Inner route, the list of nodes in section v_i
$is\_independent(v_i)$: v_i can be chosen
In this method the first step is only theoretical, but it is possible, that the partition depends on the problem instance.

2. INITIALIZE THE ELEMENTARY PARTICULAR SOLUTIONS:

```
1  for all section v_i
2      Route(v_i) = {v_i} // inner route
3      CAN(v_i) = {v_1, v_2, ..., v_{i-1}, v_{i+1}, ..., v_n}
4      // where c(e(v_i, v_j)) ≤ c(e(v_i, v_{j+1})) ∀j ≠ i
5      is_independent(v_i) = TRUE
6  // Add transitive relations to R(P):
7  if (a, b) ∈ R(P) and (b, c) ∈ R(P) and (a, c) ∉ R(P)
8      (a,c)→R(P)
9  Add E to V(DK_n)
10 // with c(e(v_i, E)) = inf and c(e(E, v_i)) = inf
11 Route(E)={}
12 CAN(E)={}
13 is_independent(E) = FALSE
14 for all section v_i
15     add (v_i, E) to R(P)
16     add as last element E to CAN(v_i)
```

In the second step we can find the candidate-choosing heuristic. In the $CAN(v_i)$ lists the order of the candidates depending on that method. In this work the candidate-choosing heuristic is a naive greedy one, which chooses the closest possible node first.

3. UPDATE PARTICULAR SOLUTIONS:

```
1  for all independent v_i
2      num = 0
3      for each v_j ∈ CAN(v_i) // from j=1
4          if (v_j, v_i) ∈ R(P)
5              delete v_j from CAN(v_i)
6              continue with next iteration
7          if ∃v_k (v_i, v_k) ∈ R(P) and (v_k, v_j) ∈ R(P)
8              delete v_j from CAN(v_i)
9              continue with next iteration
10         if ∃v_k (v_k, v_i) ∈ R(P) and (v_j, v_k) ∈ R(P)
11             delete v_j from CAN(v_i)
12             continue with next iteration
13         num = num + 1
14         if num == 2
15             break loop
```

Before every iteration in the main loop of Amoeba, we update the particular solutions. In the pseudo code, we show the case, where are enough to provide validity of the first and the second candidates. Each candidate, which leads to deadlock in the later steps are deleted. To avoid deadlock, we need to check three conditions for every v_i and $CAN(v_i)$, if one of them not meets, we delete the candidate definitely.

4. CHOOSE A PART:

We can choose parts optionally, for example randomly, but we can use heuristics to select, which part can choose its first candidate. The variants defined here, are made to increase optimality of the solution. All of these methods suppose the usage of naive greedy candidate-choosing heuristic[2].

[2]every section chooses the closest valid section

Variations of part-choosing heuristic

1) Amoeba-Greedy: This is the naive variant of part-choosing. Simply chooses the same v_i during the execution. Only one particular solution grows to a candidate solution.

2) Amoeba-1: We choose a part with MAXIMAL possible loss.

$$p\_loss(a) = c(e(a, CAN(a).sec)) - c(e(a, CAN(a).first))$$

This loss occurs when an other section makes connection, and **a** can not choose its first candidate.

3) Amoeba-2: If a part incorporates its candidate, other parts are not able to connect to that candidate anymore, so they realized their possible losses. We choose the part, which causes MINIMAL possible loss.

$$c\_loss_2(a) = \sum_{i^*} p\_loss(v_{i^*})$$

$$i^* : CAN(v_{i^*}).first = CAN(a).first \quad v_{i^*} \neq a$$

4) Amoeba-3: This part-choosing method is a combination of the previous two heuristics. With α, β parameters we weighed the two goals: choose parts with high possible loss, and minimize the caused loss on other parts.

$$c\_loss_3(a) = \alpha \cdot c\_loss_2(a) - \beta \cdot p\_loss(a)$$

5) Amoeba-4: A modification of SOP-Amoeba3. This method takes the inner circle into consideration in the calculation of caused loss. If the candidate wants to choose **a**, it can not, because this would cause inner circle.

$$c\_loss_4(a) = \alpha \cdot (c\_loss_2(a) + C) - \beta \cdot p\_loss(a)$$

$$C = \begin{cases} p\_loss(CAN(a).f) & if \quad CAN(a).f = a \\ 0 & otherwise \end{cases}$$

6) Amoeba-5: We can calculate the accurate value of caused loss for a candidate choose. We select the part, which causes the minimal loss. This method summing the possible losses for all parts, which won't be able to choose their first candidate at the next iteration of Amoeba algorithm.

$$c\_loss_5(a) = \alpha \cdot \sum_{i^*} p\_loss(v_{i^*}) - \beta \cdot p\_loss(a)$$

$$i^* : CAN(v_{i^*}).f = CAN(a).f \; v_{i^*} \neq a$$

$$\bigcup v_{i^*} = CAN(a).f \; and \; CAN(v_{i^*}).f = a$$

$$\bigcup (v_{i^*}, a) \in R(P) \; and \; (CAN(a).f, CAN(v_{i^*}).f) \in R(P)$$

$$\bigcup (a, v_{i^*}) \in R(P) \; and \; (CAN(v_{i^*}).f, CAN(a).f) \in R(P)$$

This loss with $\alpha = 1, \beta = 0$ surely appears in the resulting solution, so the sum of these losses, with the sum of minimal cost candidates at first step, as offset[3], is a lower bound on the cost of the resulting solution. We can find a perfect greedy candidate by calculating the loss for all possible part-candidate pairs[4]. If we use candidate $v_j \neq CAN(a).f$, the arising loss on **a** have to be added.

When we have a part and a candidate for it, we are ready to create a new particular solution by combining them. If section **A** joins to section **B**, we get section **A'** = {**A,B**}. **B** will be not independent. We have to refresh R(P), because R(P) contains relations on the independent sections only. All of the relations with **B** given to **A'**, and **A'** inherits the candidate list of **B**. Step 5. holds the transitive closed property of R(P).

5. CHOOSEN PART **A** INCORPORATES ITS CANDIDATE **B**:

```
1  // All constraints for B given to A
2  for all v_i : (B, v_i) ∈ R(P)
3       (A, v_i)→R(P)
4       delete (B, v_i) from R(P)
5  for all v_i : (v_i, B) ∈ R(P)
6       (v_i, A)→R(P)
7       delete (v_i, B) from R(P)
8  // Refresh A's candidate list
9  delete A from CAN(B) list // inner circle
10 CAN(A) = CAN(B)
11 // Create connection
12 Route(A) = {Route(A), Route(B)}
13 // Refresh candidate lists points to B
14 for all independent section v_i
15      delete B from CAN(v_i)
16 // Independence
17 if B not independent // iff B contains E
18      is_independent(A) = FALSE
19 else
20      is_independent(B) = FALSE
21 // Clear B
22 Route(B)={} CAN(B)={}
23 // Hold the transitive closed property of R(P)
24 for ∀v_j : (A, v_j) ∈ R(P)
25      for ∀v_i : (v_i, A) ∈ R(P)
26           if (v_i, v_j) ∉ R(P)
27                (v_i, v_j)→R(P)
```

In each iteration, the number of independent solutions decreased by 1, so in finite steps we get the state, where we have only one independent section. This section has the solution in its $ROUTE()$ structure. At the end of the execution, R(P) is an empty constraint set, because in each iteration, the base-set of R(P) is reduced by 1.

III. RESULTS

A. Characteristics of SOP-Amoeba

SOP-Amoeba is a deterministic[5] constructive heuristic method, which builds a single solution from many particular solutions, connecting two parts in each step. There are no back-track, if two nodes are connected in a part, they will be connected in the result too, so we can start the execution of a

[3]this is the trivial lower-bound on the solution's cost

[4]not only for part-first_candidate pairs

[5]with deterministic subordinate heuristics

solution, before the execution of the optimization ends.
In SOP-Amoeba, the candidate-choosing and part-choosing methods can be selected optionally. After step 3 in the main loop, the valid candidates for all independent sections are known, this information can be used by other heuristics in hybridizations. The option of choosing provides several advantages: we can support the growing of predecessor-independent parts, for multi agent planning; we can find successor for the more important nodes first; we can handle time-window problems[6].
The method declared by the pseudo codes, with Amoeba-1 part-choosing, is the SOP-Amoeba1 algorithm. The size of the input is n^2, where n is the number of nodes included in a problem. SOP-Amoeba1 has $O(c \cdot n^2 \cdot log(n))$ computational complexity, this comes from shorting lists with n elements for all nodes in step 2. The memory usage is mainly depends on the $CAN()$ lists, which need $O(c \cdot n^2)$ memory space. The algorithm is complete, so if the input is contradiction free, we get a linear extension of P. If the constraints are unsatisfiable, one of the independent sections will have no valid candidates. SOP-Amoeba is a parallel method: step 2-4 can be executed parallel for all sections, one synchronization needed in each iteration, when the chosen section incorporates its candidate in step 5.

B. Measurement results

We use TSPlib SOP instances for testing optimality: there are real-life problems for example (rbgxxxa), which derived from a stacker crane application, and random generated instances (prob.x). A comparison between different part-choosing methods and best known solutions is shown in Figure 1. We found that Amoeba1 is the most effective on the test instances, with 26% average distance from best known solutions. All best known solution comes from iterative methods, we mentioned that iterative methods provide superior solutions, but they need more time. The leading heuristics[4] have random components, these methods can have 1-10% dispersion between solutions generated for the same instance. In Figure 1 we show results for the Amoeba-Greedy(SA-G), Amoeba-1(SA-1) and Amoeba-5(SA-5) part-choosing methods only, because these are the extreme methods. The SOP-Amoeba1 construction method is better than the naive greedy by 7,4% average. It is interesting that Amoeba-1 beats Amoeba-5 by 5% too, so the calculation of caused_loss less effective, than choose the parts, with highest possible_loss. This result showed that we can reach better solutions by growing multiple particular solutions, the local error summing is restrained by the Amoeba-1 part-choosing method.

PROBLEM	N	|R|	SA-1	SA-5	SA-G	BEST 2007	SA-1 % from BEST	SA-1 % from SA-G	SA-1 % from SA-5
ESC07	9	7	2200	**2125**	2700	2125	3,529	-18,519	3,529
ESC12	14	11	**1829**	2115	2034	1675	9,194	-10,079	-13,522
ESC25	27	11	**2410**	3460	3360	1681	43,367	-28,274	-30,347
ESC47	49	32	**1645**	3556	3843	1288	27,717	-57,195	-53,740
ESC63	65	233	**66**	74	76	62	6,452	-13,158	-10,811
ESC78	80	283	22310	**22075**	22600	18230	22,381	-1,283	1,065
ft53.1	54	12	**8807**	12195	10404	7531	16,943	-15,350	-27,782
ft53.2	54	30	**9581**	14237	12656	8335	14,949	-24,297	-32,704
ft53.3	54	217	**12162**	16216	15127	10935	11,221	-19,601	-25,000
ft53.4	54	759	**17176**	18914	18549	14425	19,071	-7,402	-9,189
ft70.1	71	17	**41808**	46375	46060	39313	6,347	-9,231	-9,848
ft70.2	71	48	**43883**	48731	48359	40422	8,562	-9,256	-9,948
ft70.3	71	215	**47772**	51159	52067	42535	12,312	-8,249	-6,621
ft70.4	71	1325	**57148**	59898	62534	53562	6,695	-8,613	-4,591
kro124p.1	101	33	**51235**	55232	52575	40186	27,495	-2,549	-7,237
kro124p.2	101	68	60699	66488	**57723**	41677	45,641	5,156	-8,707
kro124p.3	101	266	79250	**71872**	77266	50876	55,771	2,568	10,265
kro124p.4	101	2305	108199	**93484**	98427	76103	42,174	9,928	15,741
p43.1	44	11	29020	**28995**	29630	27990	3,680	-2,059	0,086
p43.2	44	34	56020	56825	**29725**	28330	97,741	88,461	-1,417
p43.3	44	96	**30110**	84525	31340	28680	4,986	-3,925	-64,377
p43.4	44	496	**84360**	84490	85250	82960	1,688	-1,044	-0,154
prob.100	100	41	3455	**2815**	3311	1385	149,458	4,349	22,735
prob.42	42	19	422	**393**	458	243	73,663	-7,860	7,379
rbg048a	50	447	411	**405**	506	351	17,094	-18,775	1,481
rbg050c	52	508	**503**	534	568	467	7,709	-11,444	-5,805
rbg109a	111	5329	**1131**	1155	1443	1038	8,960	-21,622	-2,078
rbg150a	152	10334	1862	**1843**	2168	1750	6,400	-14,114	1,031
rbg174a	176	13955	2168	**2137**	2444	2033	6,640	-11,293	1,451
rbg253a	255	30181	3125	**3122**	3558	2987	4,620	-12,170	0,096
rbg323a	325	48202	**3333**	3344	4032	3157	5,575	-17,336	-0,329
rbg341a	343	54303	2984	**2875**	3786	2597	14,902	-21,183	3,791
rbg358a	360	56536	3152	**2956**	4110	2599	21,277	-23,309	6,631
rbg378a	380	63585	3287	**3277**	4109	2833	16,025	-20,005	0,305
ry48p.1	49	12	22452	**19352**	22493	15805	42,056	-0,182	16,019
ry48p.2	49	26	26491	**20701**	20911	16666	58,952	26,685	27,970
ry48p.3	49	132	31011	**26970**	27342	19894	55,881	13,419	14,983
ry48p.4	49	596	**35796**	36684	41176	31446	13,833	-13,066	-2,421
						Avg:	26,078	-7,418	-5,054
						Med:	14,925	-9,667	-0,873
						Max:	149,458	88,461	27,970
						Min:	1,688	-57,195	-64,377

Fig. 1. Measurement results: Problem instances comes from TSPlib. N is the number of nodes, $||R||$ the number of constraints. The results shows the cost of the solutions given by different methods.

IV. CONCLUSION

In this paper we showed a generalized constructive metaheuristic method called Amoeba, and a derivative algorithm SOP-Amoeba for handling the sequential ordering problem. We define many part-choosing methods for SOP-Amoeba, and found that the Amoeba1 is the most effective one. In SOP-Amoeba we can define optional subordinate heuristics. If we do the test defined in point 3 of the algorithm, we know the valid candidates for all particular solutions, this gives opportunities to create SOP-Amoebas for specific tasks, and this information can be used by other heuristics in hybrid methods.

[6]in this case we may need back-tack

REFERENCES

[1] Christian Blum, Maria Jose Blesa Aguilera, Andrea Roli, Michael Sampels(Eds.)
Hybrid Metaheuristics: An Emerging Approach to Optimization
Springer, 2008.

[2] C. Blum and A. Roli. Metaheuristics in combinatorial optimization: Overview and conceptual comparison. ACM Computing Surveys, 35(3):268308, 2003.

[3] G. R. Raidl. A unified view on hybrid metaheuristics. In F. Almeida, M. Blesa, C.Blum, J. M. Moreno, M. Perez, A. Roli, and M. Sampels, editors, Proceedings of HM 2006 3rd International Workshop on Hybrid Metaheuristics, volume 4030 of Lecture Notes in Computer Science, pages 112. Springer-Verlag, Berlin, Germany, 2006.

[4] Luca Maria Gambardella, Marco Dorigo: An Ant Colony System Hybridized With A New Local Search For The Sequential Ordering Problem Informs Journal On Computing, 2000.

Hardware acceleration of 3D HSCN-TLM Method

László Füredi
(Supervisor: Dr. Péter Szolgay)
furedi.laszlo@itk.ppke.hu

***Abstract*—Transmission line analysis in high frequencies needs use of full wave numerical approaches like Transmission Line Matrix method (TLM) which is one of the most powerful and useful approach in electromagnetism. It can be used for analysis of geometrical or non-geometrical, isotropic or anisotropic, homogeneous or inhomogeneous structures with or without electrical or magnetic losses. In this paper a general discretization for the 3D symmetrical condensed node (SCN) with Stubs TLM is given. Acceleration of the a 3D SCN TLM computation on a high performance field programmable gate array (FPGA) is described as well. The key points of the implementation are the minimization of the memory transfers and maximization of the operating frequency.**

***Index Terms*—Transmission-Line Matrix Method, Time Domain Electromagnetism, FPGA, models of computation, High-performance computing**

I. Introduction

A new paradigm of computer programming is required when algorithms has to be implemented on parallel processors. To obtain more performance out of numerical models running on traditional serial computers higher clock frequency is needed which will produce exponentially more heat. Another way to get more performance out of these models is to adapt them to the new multi- [1], [2] and many-core [3], [4] CPUs or GPUs or array computers where small pieces of the problem is processed simultaneously on all cores; the heat increase is now linear. This type of problems has a good mapping potential to parallel architectures, because several equations should be calculated using only a small local data set from a large array of data.

The challenge of this work is to find an optimal (maximum operating frequency, minimum power usage, minimum area, minimum memory latency) implementation of TLM method for FPGAs.

II. 3D Transmission-Line Modeling Method Node Types

A. Transmission-Line Modeling Method

The transmission-line modeling method was first published in 1971 [5], since then it has being referred as TLM. Sometimes is called transmission-line matrix method. The TLM discrete models of electromagnetic fields have being applied for modeling 1- 2- and 3-dimensional environments with time varying and frequency applications. This is an explicit numerical method which is based on the solution of waves on electric transmission line.

A TLM model is built from a set of transmission line segments, lumped resistors, and sources. An inductor and a capacitor can be replaced by a stub or a series transmission lines. A resistor is modeled by an infinite transmission line stub. A cubic element is modeled by a set of uniform transmission lines connected to the six adjacent cubic elements of the analyzed environment. Losses of the medium, the external sources, the mutual effects and the non-uniformity are also can be added to the TLM circuit to represent a grid node. Each grid node is formed by twelve, fifteen or more transmission lines. All the series transmission lines have characteristic impedance

$$Z_0 = \sqrt{\frac{L}{C}} = \frac{1}{Y_0}[\Omega] \tag{1}$$

while the environment non uniformity is represented by capacitors C_s and inductors L_t in each 3D grid node (p,q,r) [6]. As the intrinsic impedance of the medium is modeled by the characteristic impedance of a transmission line, the following equations relate all inductances and capacitances involved at each node:

$$c = \frac{1}{\sqrt{LC}} = \frac{1}{\sqrt{\varepsilon_0 \mu_0}}[m/s] \tag{2}$$

The inductors and capacitors are modeled as stub or link lines. The inductors are represented as short-circuit stubs, and the capacitors are open-circuit stubs. The impedances Z_{C_s} and Z_{L_t} are the stub transmission lines characteristic impedances for C_s and L_t, respectively.

B. Symmetrical Condensed Node

There are several types of 3D TLM nodes in the class of symmetrical condensed node (SCN) models, the simplest one is the symmetrical condensed node [7] and the most complex is the general symmetrical condensed node (GSCN) [8].

The symmetrical condensed node is a dodecaport machine which mimics the electromagnetic field on a piece of space. A general symmetrical condensed node can model variations of the medium as electric, magnetic and loss non uniformity [8].

The complexity of the hybrid symmetrical condensed node [9] model is between the symmetrical condensed node and the general symmetrical condensed node. The hybrid symmetrical condensed node considers only the variation of the electrical permittivity and the variation of the losses along the medium being modeled. This method is ideal for parallel implementation because it provide a reasonable compromise between model accuracy and implementation complexity.

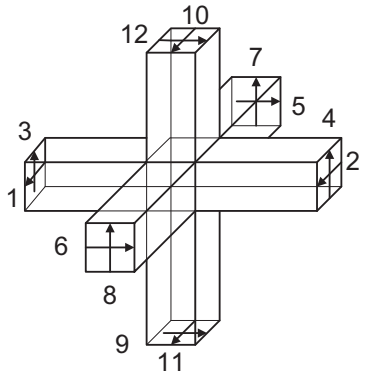


Fig. 1. 3D TLM SCN input voltage polarization

C. Node Port Numbering

Numbering of the neighborhood of the actual cell is organized in a logical way; it is easy to keep in mind and to analyze the direction of propagation. For the three directions x, y, z, the first four voltages propagates in the x direction, the middle four voltages propagates in the y direction and the last four voltages propagates in the z direction. Even numbers are at the lower scale and odd numbers are at the higher scale. The numbering are also organized in a sequence of pairs: [(1,2), (3,4); (5,6), (7,8); (9,10), (11,12)].

Based on this numbering the relation between the incident voltage at time step n+1 to the reflected voltage at the adjacent node at time step n is

$$V_{\varphi\pm1}^{i,n+1} = V_{\varphi}^{r,n}|_{AdjacentNode} \tag{3}$$

where i is the incident voltage; r is the reflected voltage; n is the discrete time step; φ is the node port, $\varphi\epsilon\{1,2,3,4,5,6,7,8,9,10,11,12\}$;

D. Scattering Matrix

Relation between incident and reflected voltages of one cell at time step n are described by the TLM scattering matrix. For the hybrid symmetrical condensed node the relation of the incident voltage to the reflected voltages are shown in Table I., where columns 13, 14 and 15 shows the coefficients for the shunt capacitances, while columns 16, 17 and 18 takes the current sources for the x, y and z directions into consideration. Columns from 1 to 12 are for the node ports, the table lines are grouped in three parts, one for each direction. In this table the port numbers are followed by the respective direction just to ease interpretation. Blank spaces represent a zero element. Therefore, fifteen linear equations should be computed where each equation requires eight coefficients, seven input voltages and one current source.

$$a_{pp} = -\frac{Y_0 + G_s + 2(Y_l - Y_t)}{2[Y_0 + G_s + 2(Y_l + Y_t)]} \tag{4}$$

$$b_{pq} = -\frac{2Y_t}{Y_0 + G_s + 2(Y_l + Y_t)} \tag{5}$$

$$c_{pp} = -[\frac{Y_0 + G_s + 2(Y_l - Y_t)}{2[Y_0 + G_s + 2(Y_l + Y_t)]} + \frac{R_t Y_t}{2(R_t Y_t + 4)}] \tag{6}$$

$$d_{pq} = \frac{2}{R_t Y_t + 4} \tag{7}$$

$$g_{pq} = \frac{2Y_0}{Y_0 + G_s + 2(Y_l + Y_t)} \tag{8}$$

$$h_{pq} = \frac{Y_0 - G_s - 2(Y_l - Y_t)}{Y_0 + G_s + 2(Y_l + Y_t)} = g_{ij} - 1 \tag{9}$$

$$k_{pq} = \frac{1}{Y_0 + G_s + 2(Y_l + Y_t)} \tag{10}$$

$$V_{16} = j_x Z_0 \Delta z \Delta y \quad V_{17} = j_y Z_0 \Delta x \Delta z \quad V_{18} = j_z Z_0 \Delta x \Delta y \tag{11}$$

$$G_{S_p} = \sigma_{ep} \frac{\Delta q \Delta r}{\Delta p}[S] \tag{12}$$

$$R_{mp} = \sigma_{mp} \frac{\Delta q \Delta r}{\Delta p}[\Omega] \tag{13}$$

$$Y_{pq} = \frac{1}{Z_{pq}} = \sqrt{\frac{C_{pq}}{L_{pq}}}[S] \qquad Y_{O_q} = \frac{2C_O^i}{\Delta t}[S] \tag{14}$$

$$\Delta t = \Delta p \sqrt{C_{pq} L_{pq}}[m/s] \tag{15}$$

$$C_{rp}\Delta r + C_{qp}\Delta q + C_O^p = \varepsilon_0 \varepsilon_p \frac{\Delta q \Delta r}{\Delta p}[F] \tag{16}$$

$$L_{rp}\Delta r + L_{qp}\Delta q = \mu_0 \mu_p \frac{\Delta q \Delta r}{\Delta p}[H] \tag{17}$$

where R_t magnetic losses [Ω] at direction t = x, y or z;
G_s electric losses [S] at direction s = x, y or z;
C_{pq} Capacitance [F/m] of the six lines;
L_{pq} Inductance [H/m] of the six lines;
C_{O_p} Capacitance [F] of the three open lines stubs;
p,q,r are replaced by x, y, z.

TABLE I
HSCN SCATTERING MATRIX

		Y_l	x	x	x	x	y	y	y	y	z	z	z	z	Capacitors			Sources		
		Y_t	z	z	y	y	z	z	x	x	y	y	x	x						
		Y_s	y	y	z	z	x	x	z	z	x	x	y	y	x	y	z	x	y	z
R_t	G_s		1y	2y	3z	4z	5x	6x	7z	8z	9x	10x	11y	12y	13	14	15	16	17	18
y	x	9x			d_{zx}	$-d_{zx}$	b_{zx}	b_{zx}			a_{zx}	c_{zx}			g			k		
y	x	10x			$-d_{zx}$	d_{zx}	b_{zx}	b_{zx}			c_{zx}	a_{zx}			g			k		
	x	13x					b	b			b	b			h			k		
z	x	5x	d_{yx}	$-d_{yx}$			a_{yx}	c_{yx}			b_{yx}	b_{yx}			g			k		
z	x	6x	$-d_{yx}$	d_{yx}			c_{yx}	a_{yx}			b_{yx}	b_{yx}			g			k		
z	y	1y	a_{xz}	c_{xy}			d_{xy}	$-d_{xy}$					b_{xy}	b_{xy}		g			k	
z	y	2y	c_{xy}	a_{xy}			$-d_{xy}$	d_{xy}					b_{xy}	b_{xy}		g			k	
	y	14y	b	b									b	b		h			k	
x	y	11y	b_{zy}	b_{zy}					d_{zy}	$-d_{zy}$			a_{zy}	c_{zy}		g			k	
x	y	12y	b_{zy}	b_{zy}					$-d_{zy}$	d_{zy}			c_{zy}	a_{zy}		g			k	
y	z	3z			a_{xz}	c_{xz}			b_{xz}	b_{xz}	d_{xz}	$-d_{xz}$					g			k
y	z	4z			c_{xz}	a_{xz}			b_{xz}	b_{xz}	$-d_{xz}$	d_{xz}					g			k
	z	15z			b	b			b	b							h			k
x	z	7z			b_{yz}	b_{yz}			a_{yz}	c_{yz}			d_{yz}	$-d_{yz}$			g			k
x	z	8z			b_{yz}	b_{yz}			c_{yz}	a_{yz}			$-d_{yz}$	d_{yz}			g			k

Voltage numbering is followed by the associate polarized direction just to ease the understanding. Sub-indexes l, t and s are replaced by x, y and z. $Y_t - Ypq; Y_l - Y_{O_q}$.

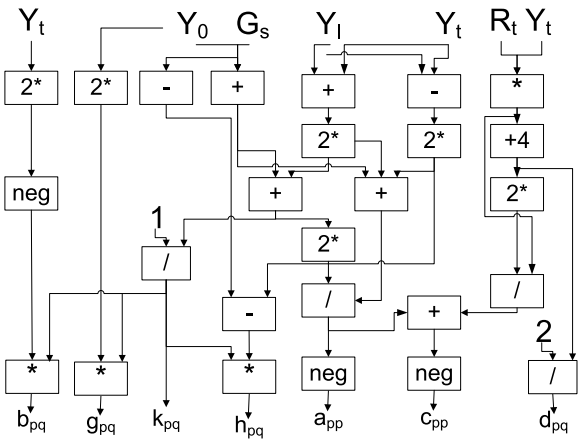


Fig. 2. Scattering matrix calculation

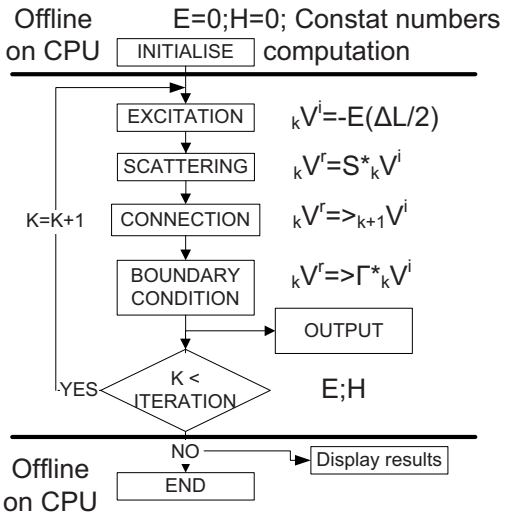


Fig. 3. The implemented architecture

III. Implementation on FPGA

The structure of the TLM model simulation is shown in Fig 3. First the Electronic (E) and Magnetic (H) field values are initialized to 0, and several constants, based on the properties of the different materials used in the model, are computed. This calculation should be done only once therefore it is implemented on the host PC.

Electronic and Magnetic field values are stored in the on-board memory of the FPGA card. Off chip memory bandwidth of the FPGA is significantly smaller than the GPGPU systems. Therefore efficient buffering of the loaded values is critical for high performance implementation. The models usually contains few different materials (chopper, dielectric material, air) therefore the precomputed constant values can be stored in the on-chip memories to save bandwidth. Instead of loading all of the six neighbors which used during the computation, a simple buffering scheme is used to further reduce memory bandwidth. Two plates are cut from the computational domain and these state values are stored in the on-chip block memories.

During the excitation step 12 voltage values are calculated according to Figure 1. For this step only 3 multipliers are required because changing the sign of a floating point-number can be carried out by an inverter and multiplication by integer power of two can be computed using an adder/subtractor to modify the exponent of the number. In addition the computed values can be reused during the computation of the neighboring cells.

The values of the scattering matrix is calculated in the next step. Most of the computing resources are required during this step. Elements of the scattering matrix is computed according to Table I using equations 4-10. Fortunately several partial results can be shared between the equations as shown in Figure 2. and the area requirements can be decreased significantly.

The reflected voltages are computed by a matrix vector multiplication using the previously computed scattering matrix and the incident voltages. Due to the symmetry properties of the scattering matrix the computation can be broken down to several 2×2 matrix vector multiplications and additions. For this step 81 adder and 48 multiplier is required (Table II.).

TABLE II
Required resource of the operations

	Multiplier	Adder	Divisor	Subtractor	Change sign	Division with 2
Input	3	0	0	0	0	0
S matrix	12	75	21	9	21	48
Matrix multip.	48	81	0	0	0	0
Output	21	24	0	9	3	6
Sum	84	180	21	18	24	54

In the final step E and H values are calculated (see 18,19) using the reflected voltages and saved to the off-chip memory.

$$E_x = -\frac{V_1^i + V_2^i + V_9^i + V_{10}^i}{2\Delta l} \tag{18}$$

$$H_x = \frac{V_3^i - V_{11}^i + V_{12}^i - V_4^i}{2Z_0\Delta l} \tag{19}$$

The memory requirement of each cell is 49 byte when double precision is used. This contains 6 time dependent state values and 1 byte to select the appropriate boundary conditions, signal sources and materials. Part of this array data is stored in the on-chip BRAM memories of the FPGA. On the largest Virtex-6 FPGA 781714 cell values can be stored on-chip therefore the maximum size of the grid is $625 \times 625 \times 625$ cells. There are not enough logic and DSP resources on Virtex-5 FPGA to implement the full architecture without resource sharing (Table III.).

IV. Performance

Performance of the software simulation depends on the size of the simulated structure if the size is larger than $128 \times 128 \times 128$ the performance drops to a lower level, due to the memory bottleneck and L3-cache memory occupancy in Intel CPUs. The optimal grid size for GPUs is around $160 \times 160 \times 160$, while an FPGA can efficiently compute

TABLE III
REQUIRED RESOURCES ON FPGA

Sum (D)	Multiplier	Adder	Divisor	Subtractor	Change sign	Division with 2	SUM	Virtex-5	Virtex-6
LUT	23436	146340	5859	14634	360	1296	191925	149760	295200
FF	34692	129420	8673	12942	0	3456	189183	149760	393600
DSP	924	540	231	54	0	0	1749	1056	2016
BRAM	0	0	0	0	0	0	49	11664000	38304000
					Number of cached cells			238040	781714

TABLE IV
PERFORMANCE COMPARISON OF DIFFERENT IMPLEMENTATIONS

	Implementations			
	Intel Q9300[10]	Intel Xeon X5550[10]	NVIDIA GTX480[10]	XC6VSX475T
Implementation type	Software	Software	Software (Cuda)	FPGA
Technology (nm)	45	45	65	40
Clock Frequency (MHz)	2500	2666 (3060)	1400	300
Number of Processing Elements	4 Cores	4 Cores	480 Cuda Cores	1 PE
Power Consumption(W)	95	95	450	120
Million cell iteration/s	10,5	18,5	180	472
Speedup	1	1.76	17.11	44.95

on $625 \times 625 \times 625$ sized grid. Performance of the different implementations are summarized on Tabel XY. For easier comparison to related work a $128 \times 128 \times 128$ sized grid is used during the computations. Even implementing a single processing element on an FPGA is 44.95 times faster than an Intel Q9300 processor when all four CPU cores are used. The Virtex-6 FPGA based solution is 2.6 times faster compared to a high performance General-purpose graphics processing unit (GPGPU), where the GPU is using all of the stream processors during the computation.

V. CONCLUSION, FUTURE WORK

An optimized architecture to calculation the HSCN-TLM method was successful, using a Virtex-6 SX475T FPGA. The solution is optimized to reduce the memory transfer between the FPGA and the off-chip memory. The main parameters of the architecture are described and compared to the parameters of the software simulation of the HSCN-TLM model running on high performance processors such as Intel Q9300 and Intel Xeon 5550 and running on GPGPU such as NVIDIA GTX480. The proposed architecture is significantly faster than a multicore processor or GPGPU.

The current implementation is optimized for a single FPGA in the future the architecture can be extended to use multiple FPGAs for even faster simulation. On the other hand the bandwidth limitation of the current architecture can be solved by using the new Virtex-7 generation FPGAs where more logic and I/O resources are available. Finally the accelerator architecture should be integrated to an existing software environment to utilize the fast computation of the TLM method in a high speed PCB design scenario.

REFERENCES

[1] J. Su, W. Guo, J. Wei, S. Shi, and B. Jiang, "Transaction level modeling tradeoff on accuracy and speed: A case study," in *Computer Design and Applications (ICCDA), 2010 International Conference on*, vol. 5, june 2010, pp. V5–14 –V5–18.

[2] O. Aghzout and F. Medina, "Accelerated computation of the propagation constants of multiconductor planar lines," *Microwave and Guided Wave Letters, IEEE*, vol. 10, no. 5, pp. 165 –167, may 2000.

[3] F. Rossi and P. So, "Accelerated symmetrical condensed node tlm algorithms for nvidia cuda enabled graphics processing units," in *Electromagnetics in Advanced Applications, 2009. ICEAA '09. International Conference on*, sept. 2009, pp. 170 –173.

[4] P. So, "Time-domain computational electromagnetics algorithms for gpu based computers," in *EUROCON, 2007. The International Conference on Computer as a Tool*, sept. 2007, pp. 1 –4.

[5] P. B. Johns and R. L. Beurle, "Numerical solution of two-dimensional scattering problems using a transmission-line matrix," *Iet Software/iee Proceedings - Software*, 1971.

[6] C. Christopoulos, *The transmission-line modeling (TLM) method in electromagnetics*, ser. Synthesis lectures on computational electromagnetics. Morgan & Claypool Publishers, 2006. [Online]. Available: http://books.google.com/books?id=pPD481tv6TgC

[7] P. B. Johns, "A symmetrical condensed node for the tlm method," *Microwave Theory and Techniques IEEE Transactions on*, vol. 35, no. 4, p. 370377, 1987.

[8] V. Trenkic, C. Christopoulos, and T. Benson, "Theory of the symmetrical super-condensed node for the tlm method," *Microwave Theory and Techniques, IEEE Transactions on*, vol. 43, no. 6, pp. 1342 –1348, jun 1995.

[9] R. Scaramuzza and A. Lowery, "Hybrid symmetrical condensed node for the tlm method," *Electronics Letters*, vol. 26, no. 23, pp. 1947 –1949, nov. 1990.

[10] M. A. F. Mattos, "A 3d electromagnetic field model with transmission-line modeling for parallel processing," *National Laboratory of Scientific Computation (2010)*. [Online]. Available: http://www.lncc.br

Dynamic Feature and Signature Analysis for Multiple Target Tracking

Vilmos Szabo
(Supervisor: Dr. Csaba Rekeczky)
szavi@digitus.itk.ppke.hu

***Abstract*— In this paper, we address the problem of multi-target tracking (MTT) with dynamically changing objects. It preforms robust multiple object tracking in a noisy, cluttered environment with closely spaced targets. Our method uses the Interacting Multiple Model (IMM) estimation framework to estimate and predict the state space of the maneuvering targets. In the first part, we demonstrate a method to simultaneously estimate the state-space and the tracking index by adaptive learning to ensure smooth spatio-temporal tracking. The second part deals with dynamical selection of parallely extracted features based on a spatio-temporal consistency metrics to maximize the robustness of data association and reduce the overall complexity of the algorithm. The proposed model is very general and can be applied for a solution of a variety problems. Experiments on simulated computer generated videos and also real-world videos validate the proposed method.**

***Index Terms*—multiple-target tracking, object detection, data association, feature extraction, dynamic feature selection, motion filters**

I. Introduction

Multiple target or object tracking (MTT) is a very important task in many computer vision applications [1]. However, it can become a challenging problem, especially if the object is in a dynamically changing environment. A number of computer vision applications could be characterized by two complex stages of processing. The first stage is the topographic image acquisition, which may include pre-processing, background subtraction [2], image segmentation [3], and post-processing. The second stage is a non-topographic sensing which includes the feature-signature extraction [4] and the target tracking back-end.

A video scene usually contains an unknown number of objects, and multiple observations for each object which needs to paired in a spatial and temporal consistent manner; this is called data association [5]. The data association step is followed by state-estimation of each track. At last, predictive-corrective filtering is applied to estimate the next state of each object.

The overall numerical complexity of the tracking algorithm is crucial in order to meet the systems real-time demand. Application areas may include traffic monitoring, vehicle navigation, automated surveillance and biological applications.

II. Dynamic Tracking

A. Algorithm Overview

Multiple target tracking can be defined, as estimating the trajectory of objects in the image plane as they move around in the scene. Generally, an object segmentation algorithm runs on each frame of the video flow in order to detect objects.

The detected objects are then assigned to consistent labels, called *tracks*. The temporal analysis of tracks can be used to identify and select features that best represent each object. The final goal of target tracking is to determine the position of an object or a bounding box on each frame of the video sequence.

B. Hierarchical Feature Extraction

The input image is highly redundant. The transformation, to reduce the dimensionality of input data while keeping relevant information content, is called *feature extraction*. For each object, a number of features are calculated from the image. A set of seven statistically independent features groups are determined that can identify and describe each object in a given frame.

The result of the segmentation algorithm is a binary mask, where black pixels correspond to the background pixels and white pixels correspond to the foreground pixels. The x and y coordinates are the centroids of each connected foreground pixel on the mask image. The shape of the objects is described by the eccentricity metric, which measures how much the object deviates from a circle. In order to extract the color information, the original color input image is converted to YCbCr color space. For each chromatic channel (luminance, blue and red) the average intensity value is extracted. All features are is normalized between 0 and 1 values in order to make comparable measurements between each frame of the video sequence.

C. Data Assignment

The assignment of measurements to consistent tracks is accomplished using a combinatorial optimization algorithm called the *Hungarian* method. Only features that are selected contribute to the calculation of the distance matrix. The assignment algorithm is used to match the current and predicted states together with minimal cost. The time complexity of the assignment algorithm is low order polynomial. More complex data association models can be applied, such as Monte Carlo data association.

In a real-time application, the number of features should be minimal to increase the speed of the system, but all the relevant information must be kept. This can be done by creating a hierarchy in the features based on their confidence or robustness. The noisy feature channels should be filtered out. The *tracking system* consists of feature selection, data assignment, state space estimation, prediction and error correction.

D. Feature Selection

The feature selection is done by analyzing the spatial and temporal property of each feature channel. The "good" features are selected based on a spatio–temporal consistency metric. Let $\mathbf{x}_k^i$ and be the feature state space vectors at frame k for the i-th object ($i = 1..m$). Let $\mathbf{D}_k(n)$ denote a vector which represents each features discriminative power (eq. 1). This can be defined as the minimum of pair wise l_1 distance of the current state space vector n-th component between the i and j-th objects. This will give high values if the inter object distance is high in the feature space.

$$\mathbf{D}_k(n) = \min_{\substack{i,j \\ i<j}} \left\{ d_1(\mathbf{x}_k^i(n), \mathbf{x}_k^j(n)) \right\} \tag{1}$$

The second term of the consistency metric is the inverse of the residual gradient magnitude of the previous state space estimation (eq. 2). This penalties feature that change considerably in time.

$$\mathbf{R}_k(n) = \frac{1}{\frac{1}{2}\left(1 + \frac{1}{m}\sum_{i=1}^{m} |\mathbf{x}_{k-1}^i - \mathbf{x}_k^i|\right)} - 1 \tag{2}$$

The final consistency metric (eq. 3) is defined by a linear μ parameter homotopy of the first part (D) and the second parts (R).

$$\mathbf{C}_k = (1-\mu)\mathbf{Q}_k + \mu\mathbf{R}_k \tag{3}$$

Note that since all the state variables are normalized between 0 and 1 the resulting components of the consistency metric $\mathbf{C}$ (eq. 3) will also be between 0 and 1. The final feature selection for frame k will be denoted by $\mathbf{S}_k$ (eq. 4). These features are the ones that are well separated from each other and do not change considerably in time.

$$\mathbf{S}_{k,1\times s} = \arg\max_{m=1..s} \mathbf{C}_k(m) \tag{4}$$

E. Event Detection

The above tracking system works very well if there are no events happening in the scene. In this section we define a number of events which needs to handled in order to increase the performance of the tracking system.

1) Track initialization when an object enters the scene.
2) Track intersection when two objects collide.
3) Track persistence when an object is being partially occluded by the background.
4) Track deletion when an object exits the scene.

Entering and exiting the scene can usually be detected if a rectangle is drawn around the edge of the scene. We can predict the intersection of two or more object by checking weather their predicted bounding boxes will overlap. If intersection occurs then we reconfigure the tracking indexes these object to enter a mode when we rely more on the model prediction than on the measurements, since the measurement of the intersecting object will be corrupted.

III. Single Object State Estimation

A. Deterministic Linear Dynamic System

The general equation of deterministic linear dynamic system can be given by

$$x(k+1) = \Phi(k)x(k) + \Gamma v(k) \tag{5a}$$
$$z(k) = H(k)x(k) + w(k) \tag{5b}$$

In 5a the x(k) is the target state vector at time k, Φ is the state transition matrix, v(k) is the unknow target maneuver, Φ is the state transition matrix, and Γ is the noise gain. The equation 5b is the output equation, where z(k) is the output and H(k) is the measurement matrix. The process noise v(k) is assumed to be zero-mean, white Gaussian noise with covariance

$$E[v(k)v(k)'] = Q(k) \tag{6}$$

w(k) is the measurement noise which is also assumed to be zero-mean, white Gaussian noise with covariance

$$E[w(k)w(k)'] = R(k) \tag{7}$$

The process noise, the measurement noise and the initial state are assumed mutually independent.

B. Kinematic Models

Multiple models can be used to describe the path of a moving object. Two well know models are the:

1) discrete White noise acceleration (DWNA)
2) discrete Wiener process acceleration (DWPA)

DWNA and DWPA are also called second and third order polynomial models. The corresponding state transition matrix Φ eq. 8, the gain vector Γ eq. 9 and the output vector H eq. 10 are given by following equations.

$$\Phi = \begin{bmatrix} 1 & T & \frac{1}{2}T^2 \\ 0 & 1 & T \\ 0 & 0 & 1 \end{bmatrix} \tag{8}$$

$$\Gamma = \begin{bmatrix} \frac{1}{2}T^2 \\ T \\ 1 \end{bmatrix} \tag{9}$$

$$H = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} \tag{10}$$

C. The Tracking Index

The tracking index (Λ) is the only design parameter of out tracking system, however it needs to be estimated beforehand to ensure high tracking performance. The derived steady-state filters are the $\alpha-\beta, \alpha-\beta-\gamma$ filters and the coefficients α,β and γ are the steady state Kalman gain components.

$$\lim_{k\to\infty} W(k) \triangleq [g_1 \; g_2 \; g_3]' \triangleq \left[\alpha \; \frac{\beta}{T} \; \frac{\gamma}{2T^2}\right] \tag{11}$$

The tracking index is a dimensionless, proportional to the ratio of position uncertainty due to the target maneuverability and the sensor measurement eq. 12. The tracking index is discussed more rigourously in [1] or in [6].

$$\Lambda \triangleq \frac{\sigma_v T^2}{\sigma_w} \tag{12}$$

Equations 13, 14 and 15 summarizes the exact formulae of the steady state Kalman gains. The position gain coefficient:

$$\alpha = -\frac{1}{8}\left(\Lambda^2 + 8\Lambda - (\Lambda+4)\sqrt{\lambda^2+8\Lambda}\right) \tag{13}$$

The velocity gain coefficient:

$$\beta = \frac{1}{4}\left(\Lambda^2 + 4\Lambda - \Lambda\sqrt{\Lambda^2+8\Lambda}\right) \tag{14}$$

The acceleration gain coefficient:

$$\gamma = \frac{\beta^2}{\alpha} \tag{15}$$

D. Estimating the Tracking Index

In this section we give a adaptive algorithm to estimate the tracking index (Λ) with a simple recursive adaptive filter. Let $\tau(k)$ be a track consisting of the position states of an object at time k, then $\tau^*_\epsilon(k)$ is an exponential decaying moving average process with ϵ parameter (eq. 16).

$$\tau^*_\epsilon(k) \quad = \epsilon\tau(k) + (1-\epsilon)\tau(k-1) \tag{16}$$

The adaptation of the tracking index Λ is performed by eq. 17. The adaptation will force Λ in such a way that the predictions of state estimation will fall within a gating-range (G_r). For high noise measurement it will decrease Λ meaning we will rely more on the model estimation than the measurement. For low noise measurements it will increase the tracking, and rely more on the measurements and it enables for agile maneuvers in the future.

$$\widetilde{\Lambda}(k+1) = \begin{cases} (1+\delta)\Lambda(k) & |\tau^*_\epsilon(k) - \tau(k)| \le G_r \\ (1-\delta)\Lambda(k) & |\tau^*_\epsilon(k) - \tau(k)| > G_r \end{cases} \tag{17}$$

Note that δ is chooser such that it results in a slower adaptation then the kinematic parameter estimation. An other solution can be if we periodically alternate the state parameter adaptation and tracking index adaptation cycles. The final tracking index value Λ is truncate between a minimal (Λ_{min}) and maximal (Λ_{max}) values (eq. 18).

$$\Lambda(k+1) = \min\left(\max\left(\widetilde{\Lambda}(k+1), \Lambda_{min}\right), \Lambda_{max}\right) \tag{18}$$

In our experiment we used $\Lambda_{min} = 0.001$ and $\Lambda_{max} = 5$.

IV. Reference Videos

The algorithm was evaluated on six video flows. The first three videos are computer generated video flows for which very accurate references can be synthesized.

A. Simulation Videos

1) Scene 1 (Shapes): (350 frames, 320 × 240 pixels, 5 objects, G=25). The first video *Scene 1 (Shapes)* contains five dynamically changing objects. Each object is able to change its location, visibility, orientation, color, shape, noise, and inner structure according to the following list:

- Location: [0–1]
- Visibility: [0–1]
- Orientation: [0–360°]
- Color: [red, green, cyan, blue]
- Shape: [circle, triangle, square, pentagon]
- Noise: [on off]
- Inner structure: [dots, lines, concentric circles]

2) Scene 2 (Bipeds): (550 frames, 490 × 270 pixels, 6 objects, G=70). The second video flow is called *Scene 2 (Bipeds)*. This scene contains walking humans with crossing and overlapping paths; they are in partial and full occlusion, entering and exiting the scene.

3) Scene 3 (Cars): (310 frames, 490 × 270 pixels, 6 objects, G=30). The third video flow is called *Scene 3 (Cars)*. The first two scenes contain non-rigid objects, while the third scene contains only rigid objects.

Figures 1, 2 contain the simulation videos and their corresponding references.

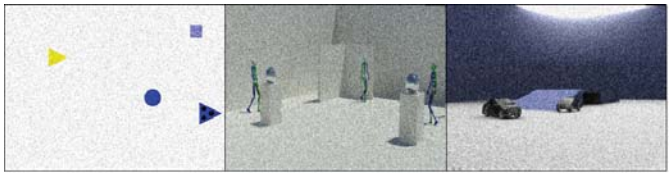


Fig. 1. Computer generated video flows with added Gaussian noise. From left to right: Scene 1 (Shapes), Scene 2 (Bipeds), Scene 3 (Cars)

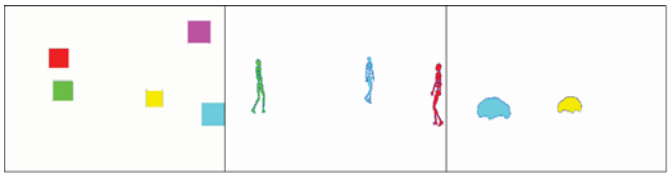


Fig. 2. Object ID map flows for the simulation videos. Each object is assigned a unique color. From left to right: Scene 1 (Shapes), Scene 2 (Bipeds), Scene 3 (Cars). The images have been modified for printing.

B. Real-World Videos

1) Scene 4 (Intersection 1): (315 frames, 420 × 270 pixels, 6 objects, G=5). The first real video is an intersection containing one car, and three pedestrians. Since our segmentation algorithm currently is unable to separate jointly initialized groups, we will track them as a single group

until separation. However in the corresponding references we included both the group bounding box and the single person bounding boxes.

2) Scene 5 (Intersection 2): (360 frames, 420 × 270 pixels, 5 objects, G=5). The second real video is the same intersection, but it contains four cars and a single pedestrian.

3) Scene 6 (PETS: (430 frames, 320 × 240 pixels, 7 objects, G=5). The last scene is the PETS 2001 video.

V. Performance Evaluation

The algorithm has been evaluated using multiple metrics. The dynamic feature selection can be characterized by the spatio-temporal consistency metric. Figure 3 summarizes the distribution of the consistency metric for each feature. For each feature the minimum, maximum and the mean value is shown in descending order. Larger values represent more discriminative features.

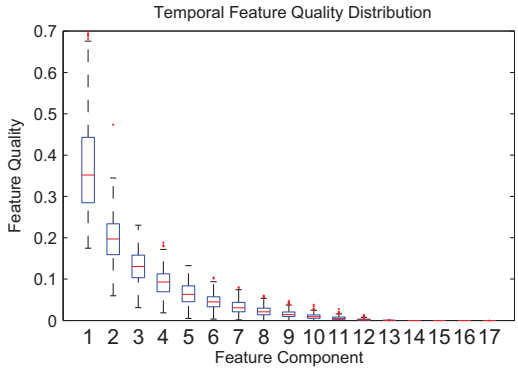


Fig. 3. Summary of feature quality distribution on the six video sequences.

Figure 4 shows the probability a feature being selected at a given time. The more discrimnative features are selected with higher probability for tracking the objects.

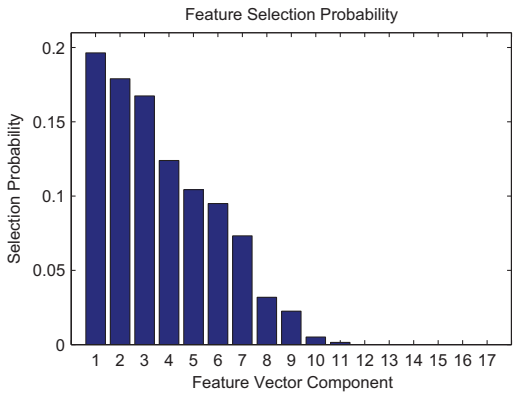


Fig. 4. This figure summarizes the probability of selection of a single feature.

The last figure 5 show the results of dynamic tracking on all of the video sequences used for the evaluation.

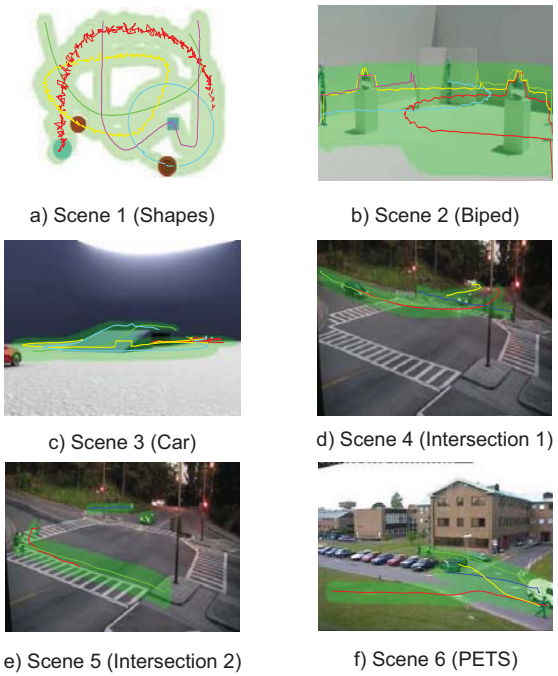


Fig. 5. Summary of tracking videos showing track information. The first three videos are the computer generated videos and the last three videos are the real-life videos used in the evaluations.

VI. Conclusion

The tracking framework that was evaluated uses adaptive tracking index estimation with dynamic feature and signature selection. This algorithm can be used to track objects in a changing environment after a topographic CNN-like segmentation and feature extraction. The algorithm arranges the parallelly extracted features into a hierarchy, based on their consistency measurement. The overall complexity is reduced by keeping only the relevant features for tracking the objects in the scene. This saves considerable processing time. Based on tests on synthesized videos and real video sequences, it has been confirmed that selecting 3-4 features dynamically could result in as good tracking as using all the features.

References

[1] Y. Bar-Shalom, X. R. Li, and T. Kirubarajan, *Estimation with Applications to Tracking and Navigation*, 1st ed. Wiley-Interscience, Jun. 2001.

[2] D.-M. Tsai and S.-C. Lai, "Independent component analysis-based background subtraction for indoor surveillance." *IEEE Transactions on Image Processing*, vol. 18, no. 1, pp. 158–67, Jan. 2009.

[3] Y. Pan, J. D. Birdwell, and S. M. Djouadi, "Preferential image segmentation using trees of shapes." *IEEE Transactions on Image Processing*, vol. 18, no. 4, pp. 854–66, Apr. 2009.

[4] J. Wang and Y. Yagi, "Integrating color and shape-texture features for adaptive real-time object tracking." *IEEE Transactions on Image Processing*, vol. 17, no. 2, pp. 235–40, Feb. 2008.

[5] Q. Yu and G. Medioni, "Multiple-target tracking by spatiotemporal Monte Carlo Markov chain data association." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 12, pp. 2196–210, Dec. 2009.

[6] P. Kalata, "The Tracking Index: A Generalized Parameter for alpha-beta and alpha-beta-gamma Target Trackers," *Aerospace and Electronic Systems, IEEE Transactions on*, vol. AES-20, no. 2, pp. 174–182, Mar. 1984.

Investigating Hungarian part-of-speech tagging methods

György Orosz
(Supervisor: Dr. Gábor Prószéky)
oroszgy@itk.ppke.hu

***Abstract*—In this paper while we are presenting part-of-speech tagging we are also describing the most common methods: Hidden Markov models and Maximum entropy models. After reviewing the widely used freely available taggers we introduce the basics of a newly developed hybrid system. We are showing that this system is competitive with the state-of-the-art Hungarian one. Finally we introduce some techniques with which we think our results is going to be improved.**

***Keywords*-part-of-speech tagging; natural language processing; stemming; maximum entropy**

I. Introduction

In corpus linguistic, part-of-speech (POS) tagging, also called grammatical tagging or word-category disambiguation, is the process of marking up the words in a text (corpus) as corresponding to a particular part of speech, based on both its definition, as well as its context. In another contexts the task can be interpreted to involve word stemming as well. This approach make sense especially for those languages when stemming is not an easy truncating task and depends heavily on the grammatical label.

Several natural language task requires the accurate assignment of POS tags to previously unseen text. Due to the availability of large corpora which have been manually annotated with POS information, many taggers use annotated texts to learn either probability distributions or rules and use them to automatically assign POS tags to unseen text. The stochastic approaches are the most widely used in nowadays since its robustness, henceforth we are not investigating other methods.

In the second section we are presenting the most widely used methods for the task and are also introducing the Hungarian morphological coding systems. After that we are reviewing the freely available taggers that are commonly used for research purposes and also presenting the current Hungarian results in the field. In the fourth section we are introducing the basics of a new hybrid POS tagger. Finally we are presenting current results compared with the state-of-the-art and introducing ideas how it shall been improved.

II. Background

In the field of natural language processing many task, such as POS tagging or named entity recognition (NER), can be interpreted as a labelling or classifying task. For solving these kind of classification tasks the most popular approaches are based on one of the following stochastic learning technique: Hidden Markov Models (HMM), Maximum Entropy Modelling (also called Maxent modelling or MEMM) and Conditional Random Fields (CRF). Hereinafter we are describing the two most widely use ones.

A. Hidden Markov models

A HMM – as described by Thede and Harper [1] – is a statistical construct that can be used to solve classification problems that have an inherit state sequence representation. The model can be visualised as an interlocking set of states. These states are connected by a set of *transition probabilities*, which indicate the probabilities of travelling from one state to another. A process begins in some state then at discrete time intervals it moves to a new one ruled by the transition probabilities. In a Hidden Markov model the exact path sequence of the states that the process generates is unknown (*hidden*). In each state one of a set output symbols is emitted by the process. The emitting is determined by a probability distribution that is specific for each state. The output of the process is a sequence of *output symbol*s. In practice for labelling problems, especially for part-of-speech tagging, second order models are used. In this case the state transitions and the *omitting probabilities* depend on the previous two states.

The following five element describes formally a HMM:

1) N, the number of distinct states in the model. In part-of-speech tagging it is the number of all the possible syntactical labels, where each state represents a tag.
2) M, the number of distinct output symbols in the alphabet of the HMM. In grammatical tagging this is the size of the lexicon.
3) $A = \{a_{ij}k\}$, the state transition probability distribution, where $a_{ijk} = P(\tau_p = t_k | \tau_{p-1} = t_j, \tau_{p-2} = t_i)$ so it is the probability to move to the next state depending on the previous two states.
4) $B = \{b_{ij}(k)\}$, the observation symbol probability distribution. $b_{ij}(k)$ is the probability that the k-th symbol is emitted when the model is in the state j and was in i. In part-of-speech tagging it is the probability of emitting the word w_k when the model is at the state labelled by the tag t_j and was in t_i.
5) $\pi = \{\pi_i\}$, the initial state distribution. π_i is the probability that the model will start in state i. In our case it is the probability of a sentence starting with t_i.

Baum-Welch algorithm is used for estimating the 3), 4) and 5) distributions from the training corpus. In the other hand, to tag a W sentence using the given model, the most probable T tag sequence is supposed to be found that has the maximal value according to the probability $P(W|T)$. Viterbi algorithm is a common method for calculating the globally most likely sequence according to a HMM.

This model has the problem of handling word forms that are not seen in the training data. A common solution is to incorporate contextual features for unknown words (e.g. affixes) that are gathered in the distribution $c_{ij}(k)$ which is also utilised for tagging. Smoothing methods are also commonly utilised handling data spareness.

B. Maximum Entropy models

Maxent models for natural language labelling tasks were firstly successfully introduced by Ratnaparkhi [2] in 1996. These models are extensively used in several fields of natural language processing (NLP) with state-of-the art results , such as statistical parsing , NER and POS tagging [3].

The principle of Maximum Entropy – as described by Le-Hong et al. [4] – states that given a set of *observations*, the most likely underlying probability distribution is that which has minimal bias – that is maximal entropy – while verifying the statistical properties measured on the observation set. In POS tagging a maximum entropy tagger learns a log-linear conditional probability model from the tagged text. One of its most appreciable strengths compared with other methods is that is potentially allows a word to be tagged with a label it has never been observed before. In the maximum entropy framework it is also possible to define and incorporate complex properties, not restricted to n-gram as it is in a Markov model.

The constraints of the model are the expectations of $f_i(i \in [1,k])$ *feature functions* according to the joint distribution p are equal to the *empirical expectations* of the feature functions in the training data distribution $\widehat{p}$:

$$\mathbb{E}_p f_i(h,t) = \mathbb{E}_{\widehat{p}} f_i(h,t) \tag{1}$$

In order to make the model easily computable $p(h,t) \approx \widehat{p}(h) \cdot p(t|h)$ estimation is used. Applying this to (1) it can be written as:

$$\sum_{h \in \mathcal{H}, t \in \mathcal{T}} \widehat{p}(h) p(t|h) f_i(h,t) = \sum_{h \in \mathcal{H}, t \in \mathcal{T}} \widehat{p}(h,t) f_i(h,t) \tag{2}$$

($\mathcal{H}$ is the space of the possible contexts and $\mathcal{T}$ is the set of tags.) The model that is the solution for this constrained optimisation task is a log-linear model with the parametric form:

$$p(t|h;\lambda) = \frac{1}{Z(h,\lambda)} \exp \sum_{i=1}^{k} \lambda_i f_i(h,t) \tag{3}$$

Where $Z(h,\lambda)$ is the normalising term that equals to $\sum_{h \in \mathcal{H}} \exp \sum_{i=1..k} \lambda_i f_i(h,t)$. There exist efficient algorithms for estimating the $\lambda = (\lambda_1, \ldots, \lambda_k)$ parameters of the model such as Generative Iterative Scaling (GIS), Improved Iterative Scaling (IIS), L-BFGS[1] and conjugate gradient methods.

[1]A limited-memory quasi-Newton method for unconstrained optimisation.

In the other hand for classifying according to a model usually the Beam search or the Viterbi algorithm is used. Given a word w and its context h, the model assigns probability for each possible tag t. The probability of a tag sequence $t_1 \ldots t_n$ according the the sentence $w_1 \ldots w_n$ can be estimated as:

$$p(t_1 \ldots t_n | w_1 \ldots w_n) \approx \prod_{i=1}^{n} p(t_i|h_i) \tag{4}$$

Morphological coding systems

For Hungarian language – as described by Farkas et al. [5] – morphological coding systems – that are in use – are the MSD, KR and the HUMOR ones. The *MSD* morphological coding system was developed for a bunch of languages including Hungarian. Within the codes the first position determines the part-of-speech while other positions offer other types of linguistic information (e.g. in the case of verbs, the type, mood, tense, number and person are provided). The *KR* coding system was developed with respect to the morphology of the Hungarian language, however, its basic syntax is language-independent. Linguistic information is encoded in hierarchical attribute value matrices: there are default values (e.g. singular or 3rd person) and only those that differ from these manifest in the code. The *HUMOR* coding system is used by the unification-based morphological analyser (MA) HUMOR [6]. The code is an aggregation of a set of labels, which are generated for morphemes by their capability of fusing with others. These labels are describing the fusing property or even the forbiddance. A word can only be built upon morphemes that are not forbidden to fuse with their neighbours.

III. Part-of-speech tagger implementations

There are several freely available taggers in the web [7], now we introduce the most popular ones.

- **TnT** – It [8] is based on a Hidden Markov model that also incorporates a combination of unigram, bigram, trigram model and several lexical properties for handling unknown words. For many languages it is successfully used because of its simplicity and fastness.
- **HunPOS** – An open source OCaml re-implementation [9] of the TnT. Widely used tagger which also able to use a POS lexicon to restrict the tag set for unknown words.
- **MXPOST** – Maximum Entropy based part-of-speech tagger [10], which simultaneously uses many contextual features to predict the POS tag.
- **OpenNLP** – Open source JAVA implementation of the Maximum Entropy machine learning method, with many application (such as POS tagger). It is implemented following the guidelines by Ratnaparkhi [10].
- **Stanford tagger** – A log-linear part-of-speech tagger [3] that uses a cyclic dependency network for incorporating preceding and following contextual predicates as well.

The first Hungarian stochastic POS tagger was firstly introduced by Oravecz and Dienes in 2002 [11]. It is built upon the TnT, and incorporates ambiguity class features with which

it was reported to achieve 98.11% accuracy on the Hungarian National Corpus [12]. This corpus is labelled with MSD codes, so the tagger also uses these annotations.

Halácsy et al. [13] showed that using a Maxent model with proper contextual features combining with the output of the TnT it is possible to make a more accurate[2] system. In 2007 his team created the HunPOS, which is currently reported to be the state-of-the-art for Hungarian POS-tagging with 98.24% accuracy on the Szeged Corpus. These tools are also using the MSD annotation system.

Two other results should be mentioned as well related to the previous works:

1) Kuba et al.(2005) [14] investigated other machine learning methods for the task, such as C4.5, RGLearn, TBL and the combination of these, but the result achieved by Oravecz [11] still remained the most accurate.
2) Zsibrita et al. [15], [16] created a basic language processing toolkit (magyarlanc) for Hungarian which includes a part-of-speech tagger as well. This part of the system is based on the Stanford's one and uses both of the MSD and KR coding systems. Investigations were only made on the whole toolkit showing an 91.02% F-value.

IV. Towards to a hybrid system

Most of the previously seen taggers hit their limits in Hungarian texts when the document has a lot of unknown words[3]. In these cases incorporating a morphological analyser could help with dealing them. This idea was also used in the magyarlanc, but not in the others. There is also a lack of such a software which utilises the HUMOR[4] and is able to do stemming simultaneously. Because of this we decided to create a software which: incorporates a morphological analyser and is able to find the proper stems for the all of the words (even for OOV[5] ones). The new POS tagger (HumPOS) uses the HUMOR as its analyser. We decided to build or system upon a Maximum Entropy model since its model fits well for the Hungarian morphology. We decided to start the development with the OpenlNLP [17] toolkit, since it is applied for many languages and several tasks.

We needed a HUMOR labelled training corpus but non of the previous ones were like this so Attila Novák[6] created one from the Szeged Corpus. There are two other integral part of our system which is build upon these lexical resources.

- The *token frequency table* which holds data about each token's frequency in the corpuses above merged together with the Morphdb.hu [18].
- The *morphological guesser* is a table generated by HUMOR. It has transformation-suffix-analysis triples where an analysis is permitted for a suffix if there exists a word with the suffix having the proper analysis. The transformation is description about getting the word's stem.

POS tagger at work

The HumPOS uses the Maxent model for building its own model of the part-of-speech tags. During the training we use the features suggested by Ratnaparkhi [10].

For tagging the Beam search method is utilised with a threshold 3[7]. In each step of the search the morphological analyser validates the found elements, ensuring that the analysises are valid ones. If no valid ones are found, then the tagger gets all the possible outputs from the HUMOR and ranks them according their lemma's frequency value, finally keeps the top three analysises.

Stemming is done by the HUMOR: if there are more than one stem for the given tag, then the system selects the most frequent one. HumPOS tries to guess a valid stem for the OOV words as well. After tagging an OOV word the HumPOS looks up and applies proper triples and selects a lemma as seen before.

V. Evaluation

For the evaluation the training corpus[8] was divided into three parts. 80% is used for training, 10% is for development and other 10% for testing. Since the research is still in progress we are only presenting results which are made on the development set.

In this linguistic task the performance is measured with (token) accuracy that is calculated: R/A where R the right tagged words count, and A is the total tokens amount. Sometimes it is also calculated for the unknown and ambiguous tokens as well. Measuring sentence accuracy is done similarly: a sentence is well labelled if all the tokens are well tagged. It is important to notice that performances of systems which does not use the same morphological coding system[9] should not be compared, since the sizes of the coding systems may vary between ten to thousands.

The first part-of-speech tagger that we are investigating is a baseline (BLT) tagger. It assigns to a known word its most frequent POS tag. For unknown words the tag `[FN][NOM]` is applied, since this is the most frequent amongst all. The HunPOS can utilise a morphological lexicon, we are evaluating this with (HP2) and without (HP1) the lexicon. Finally HMP1 stands for the tagger that is built upon the OpenNLP toolkit without the HUMOR and HMP2 is developed with it.

Previously seen [9], [11], [13] that a baseline tagger usually has a relative good performance in this task. Depending on the corpus it's token accuracy is usually between 85-90%. Investigating our baseline result – 92.30% that is significantly

[2]The accuracy is 98.17% which is measured on the Szeged Corpus.

[3]In part-of-speech tagging unknown words are those which are not seen previously during the training.

[4]Since its generative behaviour it covers the widest scale of Hungarian word forms.

[5]OOV words are those which are not recognised by the analyser.

[6]Morphologic Ltd.

[7]Ratnaparkhi [10] and Toutanova [19] showed previously that beam sizes above 3 do not influence the performance of a Maximum Entropy model.

[8]It has a size with 70084 sentences

[9]This is even more true for different languages.

TABLE I
POS TAGGERS COMPARISON

	BLT	HP1	HP2	HMP1	HMP2
Token accuracy	92.30%	98.04%	98.97%	96.41%	98.87%
Ambiguous ta.	92.11%	96.89%	97.37%	95.85%	97.08%
Unknown ta.	14.54%	84.23%	96.90%	72.91%	96.13%
Sentence accuracy	35.75%	74.73%	84.81%	59.61%	83.48%

higher then presented by Halácsy and Oravecz – we can take the cognisance that the corpus on which our evaluation is run is more consistent than the Szeged Corpus and the Hungarian National Corpus. We can also say – according to Halácsy et al. [13] – that using a MA only for filtering makes a stochastic tagger more robust in this way increasing the overall accuracy. What is remarkable that using this simple idea we could reduce the amount of the maximum entropy method's token errors with 68.48% and the amount of the sentence errors as well with 59.05%. Finally it can be concluded that we managed to develop a system, which is in all of the measured attributes close to the state-of-the-art.

VI. Conclusion

In this article we introduced the most common used stochastic methods that are used in statistical natural language processing. We also reviewed the most popular part-of-speech taggers paying particular attention to the Hungarian ones showing their current results. With a combination of currently available methods we introduced a POS tagger which is capable of determining the proper lemmas as well. Finally it was shown that using a morphological analyser in a Maximum Entropy model the state-of-the-art results are approachable.

Future plans

Our goal is to create the state-of-the-art system for Hungarian. For this we think there is still place for improvement in HumPOS. We are planning to incorporate more contextual features in the model such as the ambiguity class of tokens. Current researches [3] showed that using cyclic dependency network for English POS tagging significantly increases the performance of such a system. We deeply believe that it is adaptable for Hungarian as well. Another way to improve the performance is to adapt rules during the Beam search to filter out those cases which is known to be linguistically impossible. The error table[10] of HumPOS can be very useful for determining these rules. Since the lack of a proper lemmatized testing corpus our stemming method performance is still waiting to be evaluated.

References

[1] S. M. Thede and M. P. Harper, "A second-order Hidden Markov Model for part-of-speech tagging," in *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics -*. Morristown, NJ, USA: Association for Computational Linguistics, Jun. 1999, pp. 175–182.

[2] A. Ratnaparkhi, "Maximum Entropy Models for Natural Language Ambiguity Resolution," Ph.D. dissertation, University of Pennsylvania, 1998.

[3] K. Toutanova, D. Klein, C. Manning, and Y. Singer, "Feature-rich part-of-speech tagging with a cyclic dependency network," in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, M. Hearst and M. Ostendorf, Eds., no. June. Edmonton, Canada: Association for Computational Linguistics, 2003, pp. 173–180.

[4] P. Le-Hong, A. Roussanaly, T. M. H. Nguyen, and M. Rossignol, "An empirical study of maximum entropy approach for part-of-speech tagging of Vietnamese texts," in *Traitement Automatique des Langues Naturelles - TALN 2010*, Aug. 2010.

[5] R. Farkas, D. Szeredi, D. Varga, and V. Vincze, "MSD-KR harmonizáció a Szeged Treebank 2.5-ben," in *VII. Magyar Számítógépes Nyelvészeti Konferencia*, A. Tanács and V. Vincze, Eds. Szeged: Szegedi Tudományegyetem, 2010, pp. 349–353.

[6] G. Prószéky and B. Kis, "A Unification-based Approach to Morpho-syntactic Parsing of Agglutinative and Other (Highly) Infelctional Languages," in *27th Annual Meeting of the Association for Computational Linguistics*. Maryland, USA: Associtation for Computer Linguistics, 1999, pp. 261–268.

[7] C. Manning, "Statistical natural language processing and corpus-based computational linguistics: An annotated list of resources," 2010. [Online]. Available: http://nlp.stanford.edu/links/statnlp.html

[8] T. Brants, "TnT - A Statistical Part-of-Speech Tagger," in *Proceedings of the sixth conference on Applied natural language processing*, no. i, Universit{ä}t des Saarlandes, Computational Linguistics. Association for Computational Linguistics, 2000.

[9] P. Halácsy, A. Kornai, and C. Oravecz, "HunPos: an open source trigram tagger," in *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*. Prague, Czech Republic: Association for Computational Linguistics, Jun. 2007, pp. 209–212.

[10] A. Ratnaparkhi, "A maximum entropy model for part-of-speech tagging," in *Proceedings of the conference on empirical methods in natural language processing*, vol. 1, 1996, pp. 133–142.

[11] C. Oravecz and P. Dienes, "Efficient stochastic part-of-speech tagging for Hungarian," in *Proceedings of LREC*, 2002, pp. 710–717.

[12] T. Váradi, "The Hungarian National Corpus," in *In Proceedings of the Second International Conference on Language Resources and Evaluation*, Las Palmas, 2002, pp. 385–389.

[13] P. Halácsy, A. Kornai, C. Oravecz, V. Trón, and D. Varga, "Using a morphological analyzer in high precision POS tagging of Hungarian," in *Proceedings of LREC*. Citeseer, 2006, pp. 2245–2248.

[14] A. Kuba, L. Felföldi, and A. Kocsor, "POS tagger combinations on Hungarian text," in *2nd International Joint Conference on Natural Language Processing*. Jeju Island, Republic of Korea: Association for Computational Linguistics, 2005.

[15] J. Zsibrita, V. Vincze, and R. Farkas, "Ismeretlen kifejezések és a szófaji egyértelműsítés," in *VII. Magyar Számítógépes Nyelvészeti Konferencia*, A. Tanács and V. Vincze, Eds. Szeged: Szegedi Tudományegyetem, 2010, pp. 275–283.

[16] J. Zsibrita, I. Nagy, and R. Farkas, "Magyar nyelvi elemző modulok az UIMA keretrendszerhez," in *VI. Magyar Számítógépes Nyelvészeti Konferencia*, A. Tanács, D. Szauter, and V. Vincze, Eds. Szeged: Szegedi Tudományegyetem, 2009, pp. 394–395.

[17] J. Baldridge, T. Morton, and G. Bierner, "The opennlp maximum entropy package," Tech. Rep., 2002.

[18] V. Trón, P. Halácsy, P. Rebrus, A. Rung, P. Vajda, and E. Simon, "Morphdb.hu: Hungarian lexical database and morphological grammar," in *Proceedings of LREC*, Genoa, 2006, pp. 1670–1673.

[19] K. Toutanova and C. Manning, "Enriching the knowledge sources used in a maximum entropy part-of-speech tagger," in *Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics*. Hong Kong: Association for Computational Linguistics, 2000, pp. 63–70.

[10] An error table of a part-of-speech tagger shows the most frequent cases where the system made a mistake.

Investigating the possibilities using SMT for text annotation

László János Laki
(Supervisor: Dr. Gábor Prószéky)
laki.laszlo@itk.ppke.hu

***Abstract*—This paper deals with a new sight of view of POS tagging and lemmatization. Nowadays there are several methods to solve these tasks separately, but this is unusual to make this process in one step. If we define text annotation problem as a translation from a simple text to an annotated and lemmatized one statistical machine translation could be used to solve these task simultaneously. Secondly opposite to the current machine translation systems SMT finds the rules without defining feature sets. The goal of this study was to investigate the possibilities using SMT system for POS tagging and lemmatization.**

Keywords—Statistical Machine Translation (SMT), text annotation, lemmatization, Szeged Korpusz

I. Introduction

Wide spectrum of opportunities was opened because of the fast development of information technology in almost all disciplines. This evolution could be detected on the field of computational linguistics as well. Processing of huge text materials has become easier, even the efficiency of these systems is increasing. It is really difficult to find out the regularities in the grammar of different languages with rule-based methods, although statistical methods could provide obvious solutions for these tasks. The aim of this paper is to analyze and implement a Hungarian POS tagger and lemmatizer system based on statistical machine translation methods (SMT) and estimates the possibilities in this field.

There are several methods for text annotation; the two most widely-used are the rule-based technique and the method based on machine learning. Both methods have many advantages, but they have similar problems as rule-based SMT systems. It is very difficult to set up the proper rules for the system that covers all possibilities and the process needs serious attention. It is really hard to define the rules because of the exceptions and ambiguities.

Therefore statistical methods are used more than rule-based ones. In case of Hungarian language there are almost only machine learning (ML) methods for POS (Part-Of-Speech) tagging. The main advantage of such techniques is that the system finds the rules itself. For this feature sets has to be defined. To find all the features it could be a difficult task, and the complexity of the system is terribly increasing with the increase of the size of the feature set, which made the process much slower.

Since the task can be considered as a translation between two languages; an SMT system could be used to make the translation, if the simple text and the annotated one are considered as bilingual corpora. The benefits of this method are that the system finds the rules without defining feature sets and it could do the POS tagging and the lemmatization simultaneously. Secondly this is a language independent method, where the performance and the language of the system mainly depends on the quality of the corpus.

In the second section I present the background of the SMT system, the task of the text annotation and the lemmatization. After that I represent the evaluation and results of my system. Finally some improvements are presented.

II. Statistical Machine Translation

Machine translation is a basic branch of statistical language processing. Statistical Machine Translation (SMT) has a great advantage over rule-based translation; namely a bilingual corpus is needed to set up the system training set, the knowledge of the language grammatics is not required to create the architecture of SMT system. This corpus is the input of the system, from which statistical observations and rules are determined. The idea of the SMT method comes from speech recognition systems. [1]

The phrase, which we want to translate – i.e. the source sentence, is the only certain thing we know in the beginning of the translation. Therefore, the system can be defined as a noisy channel. A set of target sentences is passed through this channel. The output of the channel is compared with the source sentence. The result of the translation is the phrase, which provides the most appropriate match with the source sentence. This process can be formulated by Bayes' theorem as the product of two stochastic variables called language model and translation model. The next figure represents this method.

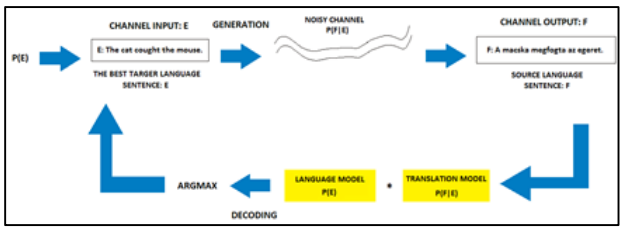


Figure 1. The idea of the SMT process

A. Frameworks

Now I describe the frameworks that implements SMT methods and I used.

1) MOSES

Several methods were studied, which are able to obtain information from parallel corpora. Finally, we decided to use IBM models, which are relatively accurate, and the used algorithm was adaptable to our task. Based on these findings we decided to use the MOSES framework [2, 3], which implements the above mentioned IBM models. This system includes algorithms for the preprocessing of the parallel corpus, the setup of translation and language models, the decoding and the optimization to the BLEU score.

2) JOSHUA

The second SMT system framework is JOSHUA [4], which not only applies word or phrase level statistics, but takes into account the morphological characteristics of the language. Chomsky's generative grammars are able to help us to solve this task. The languages, which could be described with grammatical rules, belong to the class of regular languages and context-free grammars (CFG). The great advantage of the JOSHUA system is that it is able to translate between these CFG rules in such a way, that rules can be specified for both source and target languages, furthermore the probability of the transformation into each other.

B. Evaluation

To evaluate the systems evaluation an automatic method was used. The BiLingual Evaluation Understudy [5] – BLEU score – is a frequently used algorithm to measure the quality of an SMT system. The essence of this method is that the SMT translations are compared with the reference sentences. The system calculates the result score in the interval from 0 to 1. For a better readability it is presented in percent form.

Beside the automatic method the precision of the system was calculated as well. This evaluation was used in sentence and in token level as well.

C. Text annotation

During this work in the field of machine translation the POS tagging of the corpus had to be used. That gave the idea to use SMT system as a POS tagger. POS tagging is the process of marking the words with a corresponding part-of-speech tags based on both its definition, as well as its context. This task is a really important in the field of computational language technology because this is the base of numerous methods and complex algorithms (e.g.: Information Retrieval, machine translation, word sense disambiguation, parsing, etc.) Part-of-speech tagging is harder than just having a list of words and their part of speech, because some words can represent more than one part of speech at different times, and because some part of speech are complex or unspoken. For example the Hungarian word *'vár'* has two different meanings. These two meanings – *'wait'* and *'castle'* - have different part of speech; i.e. verb and noun.

1) Coding systems for POS tags

There are three types of coding systems used for morphological coding in Hungarian language; namely KR, HUMOR and MSD-systems. The morphology of Hungarian language was taken into consideration, when KR-coding system was developed. Its basic syntax is language-independent.

The HUMOR [6] morphological coding system is based on unification. Different labels are used as tags. These labels are generated based on the capability of fusing of morphemes with others. The different labels could allow or contradict each other. One word can be built up from morphemes, for which the labels do not exclude each other.

The MSD-coding system was used to analyze the corpus from morpho-syntactical point of view. The MSD-coding system is applied for coding the different attributes of words – mainly morphologically-, and could be used for most of European languages. The morpho-syntactical attributes of words – for example type, mood, tense, number, person etc. – are represented as a character set. In the place of attributes, which are missing or not interpreted in natural languages, character '-' is used. In first position the main POS categories of words are available.[7]

D. Existing implementation of POS tagging

Oravecz and Dienes made the first stochastic Hungarian POS tagger based on MSD codes in 2002. It was reported to achieve 98.11% accuracy. [8]

Halácsy et al. used a Maxent model with proper textual features. His team created the HunPOS which is currently reported to be the state-of-the-art for Hungarian POS-tagging with 98.24% in 2007. It based on MSD code as well. [9]

E. Lemmatization

In computational linguistics, lemmatization is the algorithmic process of determining the lemma for a given word. The lemma is the dictionary form of a word. Since the process may involve complex tasks such as understanding context and determining the part of speech of a word in a sentence (requiring, for example, knowledge of the grammar of a language),it is a hard task to implement a lemmatizer for a language like Hungarian, because words appear in several inflected forms.

F. Corpus

The Szeged Corpus 2 [10] was used as parallel corpus, which was made by the Language Technology Group of the University of Szeged. This XML-based database contains both plain texts and their part-of-speech clarified version using the MSD-coding system. The advantage of the corpus is that it was manually controlled, therefore it is a very accurate data set. Further benefit is that it is general and not topic specific. The only main disadvantage is that we get a relatively small corpus. It contains 1.2 million words, which covers 155 500 different word forms and cover further 250 thousand punctuation marks. Nevertheless it seemed to be suitable for the task, because the morphological tags are of a limited number in a language, so they can occur frequently enough even in a smaller corpus. For testing the system the subpart of the corpus was used which contained 1500 randomly selected sentences.

III. RESULTS AND EVALUATION

It is easy to see that POS tagging can be considered as a transformation problem between simple text and annotated one, therefore the question arises how SMT systems could be used as POS tagger. In the first test the system was trained with the unmodified text of the corpus. The foreign language was the simple text, the target language was the lemmatized and annotated sentence. Table 1 shows the results of the system.

TABLE 1.

The system's performance I.			
System	***BLEU score***	***Correct***	***Incorrect***
MOSES	90.97%	90.29%	9.71%
JOSHUA	90.96%	90.79%	9.21%

Some mistakes of the system were realized during the evaluation. The first and maybe the main problem came from the structure of the corpus. In the annotated corpus there are the lemmas of the words connected to the morphological tags, but in the case of multi-word phrases (for example: multi-word proper names, verb phrases) the tag joins only the last word of each phrase. The lack of the marks of the words contained a unique part of speech sequence that makes false translation model. Because of that mistake the system had joint a plus tag (*')_[PUNCT]'*) to the proper names which resulted in worse results. To solve this problem all independent tags were joint to the previous word. The results can be seen in Table 2:

TABLE 2.

The system's performance II.			
System	***BLEU score***	***Correct***	***Incorrect***
MOSES	90.97%	90.80%	9.20%
JOSHUA	90.96%	90.72%	9.28%

From the table we can see that besides unchanged BLEU score the precision of the system is increased with 0,5-0,6 percent. This is because unnecessary elements were not put to the translation, but the translation of multi-word phrases still did not work. To solve this issue, it is indispensable to join these phrases to find named entities. Multi-word phrases were joint in the corpus to improve the system .The implementation of a named entity recognizer was not included in this study. After the training we got these results.

TABLE 3.

The system's performance III.			
System	***BLEU score***	***Correct***	***Incorrect***
MOSES	90.96%	91.05%	8.95%
JOSHUA	90.77%	91.07%	8.93%

Numerically from the 1500 sentence of the test set there were 691 absolute correct and 1309 sentences with any mistake. At first sight this is a quite strange rate, but if we see the results of the token level (32031 correct and 3135 incorrect) we got much better numbers. Table 3 shows that the join of related words increased the precision of the system, however the BLEU score was lower than the previous system's result. The evaluation revealed that the wrongly annotated sentences could be divided into two categories. The first category is when the system does not do the translation, but gives back the original word (2302 pieces). In most cases these words are not included in the corpus, so they could not be in the translation model. If the SMT does not find an entry in the corpus it puts the original form into the translation . The other type of error is the set of incorrect annotations (833 pieces). Two subcategories can be recognized in this case. The first is when the system can find correctly the main POS tag of the word but it makes a fail in the further analysis; secondly when it fails to find the correct main POS tags.

It is easily reasonable that the quality of the system can increase with post processing. If the system marks the unknown words with the most probable tag (i.e. noun) there is the biggest chance to get the correct POS tag. From the incorrect tagged words of the test set 1330 pieces are noun. Secondly the bigger parts of the multi-word sentences are proper names. So after the post processing the precision of the system is 95.35%. Table 4 shows an example from the results of the system.

TABLE 4.

Example from the text annotation result	
System name	***System's translation***
Simple text:	mindenképp kötelességtudó szeretnék lenni , de azért nem olyan fanatikus szinten , mint egyes felnőttek , hogy még a családot is feláldozza a kötelesség miatt .
Reference annotation	mindenképp_[Rg] kötelességtudó_[Afp-sn] szeret_[Vmcp1s---n] lesz_[Vmn] ,_[Punct] de_[Ccsp] azért_[Rd] nem_[Rm] olyan_[Pd3-sn] fanatikus_[Afp-sn] szint_[Nc-sp] ,_[Punct] mint_[Cssp] egyes_[Afp-sn] felnőtt_[Nc-pn] ,_[Punct] hogy_[Cssp] még_[Rx] a_[Tf] család_[Nc-sa] is_[Ccsp] feláldoz_[Vmip3s---y] a_[Tf] kötelesség_[Nc-sn] miatt_[St] ._[Punct]
SMT annotation:	mindenképp_[Rg] kötelességtudó_[Afp-sn] szeret_[Vmcp1s---n] lesz_[Vmn] ,_[Punct] de_[Ccsp] azért_[Rd] nem_[Rm] olyan_[Pd3-sn] fanatikus_[Afp-sn] szint_[Nc-sp] ,_[Punct] mint_[Cssp] egyes_[Afp-sn] felnőtt_[Nc-pn] ,_[Punct] hogy_[Cssp] még_[Rx] a_[Tf] család_[Nc-sa] is_[Ccsp] feláldoz_[Vmip3s---y] a_[Tf] kötelesség_[Nc-sn] miatt_[St] . _[Punct]
SMT only POS tagger:	[Rg] [Afp-sn] [Vmcp1s---n] [Vmn] [Punct] [Ccsp] [Rd] [Rm] [Pd3-sn] [Afp-sn] [Nc-sp] [Punct] [Cssp] [Afp-sn] [Nc-pn] [Punct] [Cssp] [Rx] [Tf] [Nc-sa] [Ccsp] [Vmip3s---y] [Tf] [Nc-sn] [St] [Punct]

A question could arise; namely whether if the system is able to distinguish the disambiguation in the text, and give a proper POS-tag? The answer was given by the properties of the SMT system. For those phenomena that have enough examples in the corpus, the SMT system performs really well. If all polysemous phrases are in the training corpus in the fine frequent, the system could decide correctly between the meanings of the phrases. Unfortunately such kind of corpus has not been produced yet and probably it will not be created in the near future. Since the Szeged Korpusz is relatively small, probably it does not contain all meanings of a multiple-meaning phrase. In such cases the system will choose the most

probable entry of the models; so it could not solve the question of the disambiguation with that resource.

It follows from the above that any rules could be learnt from an appropriate corpus. Since the size of our corpus is fix, improvement of the quality of the system could be gained with the decrease of the complexity of the annotation task. In our case it could be achieved if we translated the simple text to the "language" of the POS tags. Therefore the lemmas were not written before the tags, because the number of these tags are much fewer than Hungarian words, a relatively accurate system could be built from a smaller corpus. On the other hand, if the lemmas are left from the annotation and we translate only to the set of tags, the order of the morphemes in the sentence will be much more weighted in the translation and the language model as well.

TABLE 5.

The system's performance IV.			
System	***BLEU score***	***Correct***	***Incorrect***
MOSES	88.65%	91.22%	8.77%
JOSHUA	88.57%	91.09%	8.91%

The results of the system are shown in Table 5. We can again observe the decrease of the BLEU score compared to the baseline system's one, but the precision was the best. Numerically 703 correct sentence and 1297 incorrect one, 32081 correct tokens and 3085 incorrect ones. That system did not annotate 2303 pieces of words from which 1330 pieces are nouns, so after the post processing the precision of the system was 95.01%. Despite the fact that there was no lemmatization the algorithm could be used, because another system would be trained for the lemmatization task and we would combine the results of the systems, thus the quality of the baseline system would be increased.

In the previous case the task of lemmatization has been lost in order to set focus on the POS tagging task. With the decrease of the number of the words of a language the quality of the SMT system has been increased. It follows that further decreasing this number, the results should be better. Therefore in the next experiment I examined the results of the system if the annotation is done on a previously lemmatized corpus. The results can be seen in TABLE 6:

TABLE 6.

The system's performance V.			
System	***BLEU score***	***Correct***	***Incorrect***
MOSES	80.05%	76.35%	23.65%

Against the expectations a huge decrease of the accuracy can be observed. Against the previous results when 6323 pieces of tokens were incorrect from which only 716 pieces were not annotated by the system. So because of the decrease of the complexity of the foreign language the system could annotate the most part of the text, in parallel the number of the incorrect annotations increased. That is due to the ambiguity problem, because a noun and a verb have different affixes, but have the same lemma. In that case the system can only calculate with the environment of the word. To find the correct meaning of the word a corpus is needed where both meanings are represented in a convenient number.

IV. Conclusion

In this article I investigated the possibilities using SMT system for the task of POS tagging and lemmatization. I find that these tasks can be considered as a translation from the language of simple texts to the language of its annotations. I presented that the quality of the system is quite good, the best result has 95.35% accuracy. Altough that quality is far from the state-of-the-art system's performance, but I gave an absolutely automated system which finds itself the rules without previously setting any feature sets. Secondly against the proper POS taggers this system made the annotation and the lemmatization in the same time. Even then the results warned us that the statistical methods alone are not enough to solve this task either, it needs some kind of hybridization. The given results were encouraging and they pointed out that this way of research contains further possibilities.

References

[1] L. J. Laki and G. Prószéky, "Statisztikai és hibrid módszerek párhuzamos korpuszok feldolgozására," in *VII. Magyar Számítógépes Nyelvészeti Konferencia*, (Szeged), pp. 69–79, Szegedi Egyetem, 12 2010.

[2] P. Koehn, *Moses - A Beam-Search Decoder for Factored Phrase-Based Statistical Machine Translation Models.*, 2009.

[3] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, "Moses: Open Source Toolkit for Statistical Machine Translation," in *Proceedings of the ACL 2007 Demo and Poster Sessions*, (Prague), pp. 177–180, Association for Computational Linguistics, 2007.

[4] Z. Li, C. Callison-Burch, C. Dyer, J. Ganitkevitch, S. Khudanpur, L. Schwartz, W. N. G. Thornton, J. Weese, and O. F. Zaidan, "Joshua: an open source toolkit for parsing-based machine translation," in *Proceedings of the Fourth Workshop on Statistical Machine Translation*, StatMT '09, (Stroudsburg, PA, USA), pp. 135–139, Association for Computational Linguistics, 2009.

[5] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, (Stroudsburg, PA, USA), pp. 311–318, Association for Computational Linguistics, 2002.

[6] G. Prószéky and B. Kis, "A unification-based approach to morpho-syntactic parsing of agglutinative and other (highly) inflectional languages," in *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, (Stroudsburg, PA, USA), pp. 261–268, Association for Computational Linguistics, 1999.

[7] R. Farkas, D. Szeredi, D. Varga, and V. Vincze, "Msd-kr harmonizáció a szeged treebank 2.5-ben," in *VII. Magyar Számítógépes Nyelvészeti Konferencia*, (Szeged), pp. 349–353, Szegedi Egyetem, 12 2010.

[8] C. Oravecz, P. Dienes, and P. Dienes, "Efficient stochastic part-of-speech tagging for hungarian," in *In Proc. of the Third LREC, pages 710–717, Las Palmas, Espanha*, p. ELRA., 2002.

[9] P. Halácsy, A. Kornai, C. Oravecz, V. Trón, and D. Varga, "D.: Using a morphological analyzer in high precision pos tagging of hungarian," in *In: Proceedings of LREC 2006*, pp. 2245–2248, 2006.

[10] D. Csendes, C. Hatvani, Z. Alexin, J. Csirik, T. Gyimóthy, G. Prószéky, and T. Váradi, "Kézzel annotált magyar nyelvi korpusz: a Szeged Korpusz.," in *I. Magyar Számítógépes Nyelvészeti Konferencia*, pp. 238–247, Szegedi Egyetem, 2003.

Information-retrieval from medical diagnoses and anamneses with text mining algorithms

Ferenc Ott
(Supervisor: Dr. Gábor Prószéky)
ferenc.ott@gmail.com

Abstract – **In this presentation I introduce a concept of a system and its modules that can be helpful in Hungarian medical texts, mostly diagnoses and anamneses. The medical corpus is from SOTE, and it is a big help for us. Text mining can be very helpful for doctors and researchers, but the texts to be used should guarantee anonymity. For this reason some pre-processing steps was needed to substitute real proper names of the documents. New information found with the help of medical text mining can help in identification of diseases in question.**

Index Terms: medical diagnoses; anamneses; text mining in medical texts; Latin morphological parsing; medical ontology

I. INTRODUCTION

The base of the text mining methods is the rich representation of the texts. There are different approaches: rule-based and statistical methods. The representation of the knowledge is the gate between processing structured information and unstructured information.

Four main approaches of text mining methods can be distinguished:

- **Data-oriented definition:** text mining is an operation on symbols of the texts; it tries to find associations with statistical, data-mining methods.
- **Structure-oriented definition:** this method uses morphological approaches.

Intelligence-oriented definition: this method uses ontologies and taxonomies as representations.

- **Business-oriented definition:** this concept is much more pragmatic; it is a system built from modules that work together to grab human readable information from databases, diagnoses, anamneses.

Medical text mining is a special area of general text processing, because medical texts contain a lot of Latin terms or Medical Latin expressions that can be handled by automatically methods. The Medical Latin language is a special language differing from the Classical Latin language. To analyze a Medical Latin diagnosis or anamnesis, we have to have special methods and pre-processing steps, which I've sold with automatically methods. My aim is to build a system, which can be helpful – by analyzing diagnoses and anamneses – for doctors and for researchers to get more information about diseases, and to be supported in their decision-making. Some typical Medical Latin expressions I've retrieved from existing medical dictionaries.

II. INFORMATION RETRIEVAL FROM MEDICAL DICTIONARIES

I have worked with the following dictionaries:

1. Hungarian-Latin morphological dictionary
   M-80500 M 00 805-808 LAPHÁM TUMOROK
   M-80502 L 01 Carcinoma papillare in situ
   M-80502 V 01 Papillaris in-situ carcinoma

2. Hungarian-Latin topographical dictionary:
   T-00400 M 05 nyálkahártyai
   T-00400 L 01 tunica mucosa
   T-00400 M 01 nyálkahártya
3. Hungarian-Latin procedure dictionary:
   P1-95348 M 01 háromosztatúideg-kimetszés
   P1-95348 V 01 trigeminálisideg-excízió

4. Hungarian-Latin disease dictionary:

$D4-00A41$	*L 01*	*multiplicata*	*epiphyseal*	*dysplasia*
$D4-00A41$	*V 01*	*multiplex*	*epifízeális*	*diszplázia*
$D4-00A41$	*M 01*	*többszörös csövescsonti porcvégi rendellenes fejlődésű*		

5. Multilingual anatomical dictionary:

LA	*cervicalis [-e]*
EN	*cervical; C. (pertaining to the neck)*
DE	*zervikal; Zervix-; Hals- (zum Hals gehörend)*
HU	*cervikális; nyakhoz tartozó*

III. THE MEDICAL LATIN PARSER

The parser is partially written by me. I added special medical affixes to it, like *-isis* or *-aris*. I have also developed a taxonomy-building ability: an XML database has been from the above dictionaries. The parser's input and output are simple text files (see Figure 1.)

```
Latin jelentesek:
Elso latin jelentes:                abdomen
Masodik latin jelentes:             alvus
Harmadik latin jelentes:
Negyedik latin jelentes:


Magyar jelentesek:
Elso magyar jelentes:               has
Masodik magyar jelentes:
Harmadik magyar jelentes:
Negyedik magyar jelentes:
------------------------------------


A(z) filamentum
 elemzese es forditasa:
----------------------------
Forditasi eredmeny:

Szoto:          filament
Szo:            filamentum
Kepzo:
Kepzofajta:
Toldalek:       um
```

Figure 1: parsing of word "*abdomen*"

IV. BUILDING A MEDICAL LATIN CONCEPT DATABASE

To make the program understand the Medical Latin expressions, I have built an XML-based expressions database containing the expressions of the above medical (anatomical and pathological) dictionaries. The result is human readable (Figure 2).

```
- <reszSzo>
    <szoto>intervertebral</szoto>
    <szo>intervertebralis</szo>
    <eset>Acc</eset>
    <nem>M</nem>
    <szofaj>M</szofaj>
    <toldalek>es</toldalek>
    <szam>Tobbes</szam>
```

Figure 2: parsing of word "*abdomen*"

In this XML file every expression or a simple word is a <kifejezes>; every part of which is a <reszkifejezes>. If <reszkifejezes> contains more than one word, that is, the expression contains more subexpressions with more than one word, it is a <reszszo> (Figure 2).

V. THE HUNGARIAN MEDICAL CORPUS

The Hungarian medical corpus is from the SOTE, and it is containing 465 MB plain text, diagnoses, anamnesis, medical expressions.

```
#### 1S3 2004.05.11
   ERRU129

#### 1SA 2005.02.03
      Semmelweis Egyetem Szemészeti Klinika Tömő u.
      1083 Budapest Tömő u. 25-29.
      Általános Ambulancia
      Intézetvezető: Prof. Németh János
      Tel.: (1) 210-0280/51710

#### 1SA 2005.03.04
      Semmelweis Egyetem Szemészeti Klinika Tömő u.
      1083 Budapest Tömő u. 25-29.
      Általános Ambulancia
      Intézetvezető: Prof. Németh János
      Tel.: (1) 210-0280/51710

                              A M B U L Á N S   K E Z E L Ő L A P


      Diagnózis
        DIAGNÓZISOK megnevezése                    Kód   Dátum      Év  K V T
        Sárgafolt és hátsó pólus sorvadás          H3530 2005.03.04       1
        Sárgafolt és hátsó pólus sorvadás          H3530 2005.03.04       3

      Beavatkozások
         Kód     Megnevezés                                  Menny.  Pont
         12098   Fénytörés meghatározása komputerrel             1    185
         12111   Accomodatio vizsgálata                          1     93
         12113   Fénytörés szubjektív meghatározása              1    112
         12130   Üvegtest biomikroszkópos vizsgálata             1    274
         12204   Ophthalmoscopia, binocularis, indirekt módsze   1    274
         12210   Réslámpa vizsgálat                              3    357
         12231   Kancsalsági szög, szemmozgások vizsgálata       1    164
         12240   Szaruhártya-görbület mérése                     1    168
         12291   Cataracta átvilágítás, entoptikai vizsgálat     1    206

      2010.11.16


#### 1SA 2005.06.27
      Semmelweis Egyetem Szemészeti Klinika Tömő u.
      1083 Budapest Tömő u. 25-29.
      Általános Ambulancia
      Intézetvezető: Prof. Németh János
      Tel.: (1) 210-0280/51710
```

VI. PROLOG

Prolog is a logic programming language associated with artificial intelligence and computational linguistics. Prolog has its roots in formal logic, and unlike many other programming languages, Prolog is declarative. The program logic is expressed in terms of relations, represented as facts and rules.

A computation is initiated by running a *query* over these relations. Prolog can be used to parse natural language texts (Figure 5).

```
<sentence>   ::= <stat_part>
<stat_part>  ::= <statement> | <stat_part> <statement><statement>  ::=
<id> = <expression>
<expression> ::= <operand> | <expression> <operator> <operand>
<operand>    ::= <id> | <digit>
<id>         ::= a | b
<digit>      ::= 0..9
<operator>   ::= + | - | *
```

Figure 5: Language description in Prolog

Some Prolog implementations, notably SWI-Prolog, support server-side web programming with support for web protocols, HTML and XML. There are also extensions to support semantic web formats such as RDF and OWL. I use Prolog for managing the output of the Stanford parser. Because the output is in tree-format, the words and the grammatical information are in brackets, it can easily be handled with this programming language.

VII. WORKING WITH SWI-PROLOG IN TEXT-MINING

SWI-Prolog is a freely usable Prolog implementation. I have tried it on a Hungarian corpus, with a Hungarian morphological parser [9.]. The parser's output was similar, like the Stanford parsers output, so both software result can be handled with the Prolog's methodic. An example of the Prolog's input file (that is the morphological parsers output file) is shown by Figure 6.

```
csomo('COMPL','(null)','(null)',['num'-'SG','pers'-'P3','case'-'ACC'],[
  csomo('NP-FULL','érték','(null)',['num'-'SG','pers'-'P3','case'-'ACC'],[
    csomo('NP','érték','(null)',['num'-'SG','pers'-'P3','case'-'ACC'],[
      csomo('NP-FULL','portfolió','(null)',['num'-'SG','pers'-'P3','case'-'DAT'],[
        csomo('NP','portfolió','(null)',['num'-'SG','pers'-'P3','case'-'DAT'],[
          csomo('ADJP','tőzsde','(null)',['num'-'nil','case'-'NOM'],[
            csomo('ADJX','tőzsde','(null)',['num'-'nil','pers'-'nil','case'-'NOM'],[
              csomo('ONADJ','tőzsde','tőzsdei',['case'-'NOM','pers'-'P3','num'-'SG']
          csomo('NX','portfolió','(null)',['num'-'SG','pers'-'P3','case'-'DAT'],[
            csomo('N','portfolió','portfoliójának',['case'-'DAT','pers'-'P3','num'-'
      csomo('NP','érték','(null)',['num'-'SG','pers'-'P3','case'-'ACC'],[
        csomo('NX','érték','(null)',['num'-'SG','pers'-'P3','case'-'ACC'],[
          csomo('N','érték','értékét.',['case'-'ACC','pers'-'P3','num'-'SG'],[])])])
```

Figure 6: Morphological output

The Prolog program grabs out the relevant words from the text and writes them out to a text file (part):

```
mondat_fonev([A|L],Gyujto,MondatFonevk):-
        A = csomo(Tagok,Szotari,_,_,Gyerekek),
        (Tagok == 'N',
        hozzaad(Szotari,Gyujto,Gyujto1),
        !,
        mondat_fonev(Gyerekek,Gyujto1,Gyujto2),
        mondat_fonev(L,Gyujto2,MondatFonevk)
        ;
        !,
        mondat_fonev(Gyerekek,Gyujto,Gyujto1),
        mondat_fonev(L,Gyujto1,MondatFonevk)
        ).
```

VIII. CONCLUSION

My goal is to develop an efficient algorithm and system to automatically process medical texts. The system will consist of the following modules:

- a Medical Latin parser (which I have finished),
- an XML-based expressions-database, which I have also finished)
- the MySQL-based BioLexDB
- the Stanford-parser (what I have tested this year)
- a medical corpus (I have started to build partly from internet sites and partly from still ready corpora)

The downloader script and the text-cleaning script are ready. The parser is also a good tool to analyze the data. My Latin parser and my medical XML database from medical Latin dictionaries serve as a basis of the system. The BioLexDB can be a good add-in, the Prolog NLP functionality makes easier to deal with the output of the morphological analyzer.

REFERENCES

[1] Sholom M. Weiss, Nitin Indorkhya, Thong Zang, Fred J. Damerau: "Text Mining, Predictive Methods for Analyzing Unstructured Information," Springer 2005.

[2] Brian Roark and Richard Sproat: "Computational Approaches to Morphology and Syntax" 2008 Massachusetts Institute of Technology, September 2008, Vol. 34, No. 3, Pages 453-457.

[3] Tikk Domonkos, Biró György, Szidarovszky Ferenc P., Kardkovács Zsolt T., Héder Mihály és Lemák Gábor. "Magyar internetes gazdasági tematikájú tartalmak keresése,". in IV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY-06) pp. 3–14, Szeged, 2006.

[4] György Szarvas, Richárd Farkas, Róbert Busa-Fekete: „State-of-the-art anonymisation of medical records using an iterative machine learning framework," in Journal of the American Medical Informatics Association, 2007 Vol. 14. pp 574-580.

[5] Nagy József: Orvosi latin nyelvi alapismeretek, Medicina, 2009.

[6] D. Tikk, Zs. T. Kardkovács, Z. Andriska, G. Magyar, A. Babarczy, and

web searcher," in Proc. of ICCC-04, 2nd IEEE Int. Conf. on Computational Cybernetics, pp. 303–309, Vienna, Austria, 2004.

[7] J. P. Pestian, C. Brew, P. Matykiewicz, D. J. Hovermale, N. Johnson, K. B. Cohen, W. Duch; "A shared task involving multi-label classification of clinical free text," Proceedings ACL:BioNLP, Prague, June 2007

[8] http://nlp.stanford.edu/software/lex-parser.shtml

[9] Morphologic HumorESK C library

[10] http://www.computationalmedicine.org/

Design of a test setup for the flexor-extensor mechanism of a biomechatronic-hand

Norbert Sárkány
(Supervisors: Dr. Péter Szolgay and Dr. György Cserey)
sano@digitus.itk.ppke.hu

***Abstract*—This paper presents a design of an anthropomorphic biomechatronic hand(bmt-h), focusing on the design of the fingers and its bio-inspired flexor-extensor like low-level control. The kinematic description, the detailed explanation and presentation of the 3D CAD design are included. The description of the applied 3D tactile and magnetic sensors are also detailed in the article. Matlab simulation results and also the first experiments of the hardware prototype gave promising results and show that the approach can be an effective solution for the need of a hand-like actuator in robotics.**

I. Introduction

In the last twenty years there was an extensive research about robotic hands, which goal was to design and develop an anthropomorphic dexterous hand [1], [2], [3], [4], [5], [6]. There are two main concept of designs, one with a local control where the actuators are in the hand [1], [3], [4], its reduces the amount of space which it requires, and the weight. The small weight is always an important aspect but in many cases this reduces the DOF. The second design is where the actuated structure and the actuator mechanism are separated and connected with artificial tendons, such a hand is capable to do manipulation tasks like a human hand can do, here every joint has an independent control, and there are no passively controlled joints.

The commercially available prosthetics are similar to the first type but are limited in their movement capability and they have a lack of sensory information and a not so sophisticated control.

In this paper a design of an anthropomorphic biomechatronic hand will be presented. Primarily the design of the finger, the concept of a full hand and the experimental tests.

The main goal of the research is to have an artificial hand which can be used in robotic applications and which could give a basis of new prosthetics design too.

In Section II. the basic design concepts of a BMT-H is presented, in Section III. shows a detailed mechanical structure of a finger. Section IV. discuss the type of sensors in the human hand and our selections which correspond to them. In Section V. it will be shown a tests setup for the first prototype and the experimental results. Finally, conclusions and future work are discussed in Section. VI.

II. Biomechatronic Hand Design

"The human hand is a precise tool". It has 27 bones and it can be divided into three main parts [7]:

- Carpus – wrist
- Metacarpus - mid hand
- Phalanx – fingers

The exact value of the DOF in the human hand is determined by the number of independent muscle groups. If we consider a situation where each DOF has three positions (two end positions and a middle one) then there would be 3^{27} available positions, which is more then ~ 7 billion states.

One of the major functions of the hand is to grip and manipulate objects. Gripping objects generally involves flexing the fingers against the thumb. Depending on the type of grip (Fig. 1), muscles in the hand act to modify the actions of long tendons that emerge from the forearm and insert into the digits of the hand. It produces combinations of joint movements within each digit that cannot be generated by the long flexor and extensor tendons alone coming from the forearm.

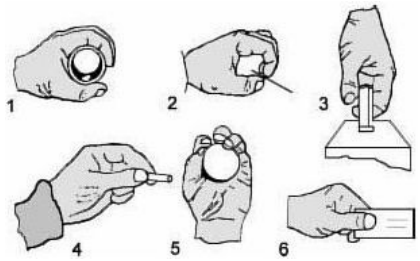


Fig. 1. The main gripping groups [8].

The main requirements of biomechatronic design to be considered from the beginning where the following: biomimetic structure, control. To meet these conditions two separated systems has to be implemented:

- actuation system: includes the actuators, force transfers (cables, artificial tendons) , the sensors, and the control electronic
- actuated system: the mechanical structure of the BMT-H

The actuation system bases on the extensor-flexor mechanism of the human hand. This mechanism contains the muscle and tendon complex, one for the extensor and one for the flexor.

Muscle force is proportional to physiologic cross-sectional area, and muscle velocity is proportional to muscle fiber length. The strength of a joint, however, is determined by a number of biomechanical parameters, including the distance between muscle insertions and pivot points and muscle size.

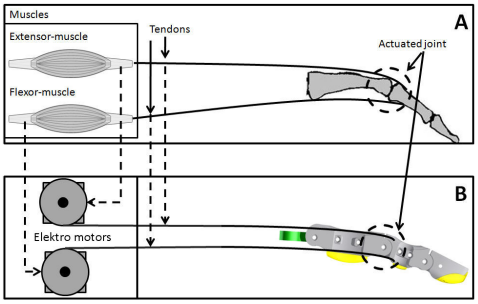


Fig. 2. A.) The extensor-flexor structure of the human finger. B.) The actuation system of the BMT-H

We represent the muscles in our design with servo actuators in the configuration that can be seen in Fig. 2.

In the human body the muscle power is transferred by the tendons. The tendons can be modeled as elastic fibers, force can be applied lengthwise on them until the maximum length is reached. After the stress ceases, the stored mechanical energy releases.

Two main parameters stiffness and hysteresis describe the tendons behavior in the actuation system .

Tendon stiffness is a mechanical property in the biological structure which based on the force applied on the muscle end tendon. When a greater force must be applied to achieve a given extension then the tendon is stiffer, otherwise it is more compliant.

The second property is hysteresis which is the amount of lost energy during the recoil from the extension.

Tendons have viscous and elastic properties too, which means it is a so called viscoelastic material, this property is observed during deformation. In our concept we use Kevlar fibers as tendons. It has the following properties we demanded for artificial tendons. High tensile strength at low weight, low elongation to break, high modulus (structural rigidity), high toughness (work-to-break), excellent dimensional stability, high cut resistance and self-extinguishing.

The structure of the BMT-H mimics that of the human hand it has 21 DOF and can be divided into four main parts, to the fingers, the thumb, palm and wrist, the size of the hand was based on average adult male hand size. The maximum length of the hand fully stretched is 212 mm, and the width is 85 mm shown in Fig. 3.

The thumb has to be designed to preform grasping tasks by thumb oposation.

The actuating system is located in the forearm just like in the human structure.

The first prototype was made using 3D prototype technology. The technology allows to produce three-dimensional objects from a different type of materials. The BMT-H is made up of two types of prototyping material a hard rigid one from which the fingers are and a soft compliant one used for the pads of the finger and in the palm.

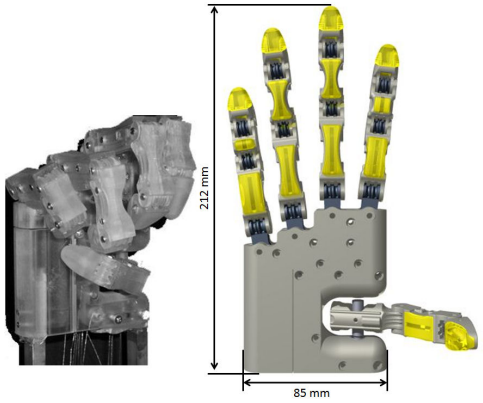


Fig. 3. The first full BMT-H prototype, CAD design.

III. Finger prototype design

The BMT- finger prototype is designed by reproducing, as close as possible to the size and kinematics of the human finger.

Each finger has two interphalangeal joints (IP's), distal (DIP) and proximal (PIP). Between the proximal phalanges and the metacarpals are the knuckles or metacarpophalangeal (MCP) joints. The IP and MCP joints are capable of flexion (bending) and extension (straightening). In addition, the MCP joints are capable of abduction (spreading of the fingers) and adduction (bringing the fingers together) [9] [10].

For the kinematic description of the finger we used a three segment anthropomorphic model, showed in Fig. 4.

Where q_i is the angel between x_{i-1} and x_i, C_{q_i} is the Cos of the q_i angle, S_{q_i} is the Sin of the q_i angle, l_i is the length of the given phalanges, α_i is the angle about common normal, from z_{i-1} zo z_i and d_i is the offset along z_{i-1} to the common normal. We can use simplifications where α_i and d_i are zero.

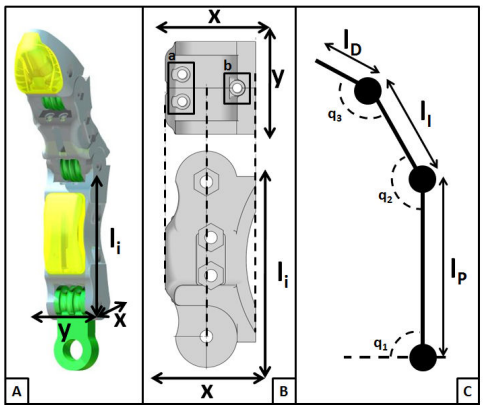


Fig. 4. The CAD design of the finger (A). Intermediate phalanges top and side view (B), the guiding of the flexor tendons(a) and for the extensor (b). The kinematic description of the finger, l_i is the length of the given phalanges, x is the width, y hight of the phalanges are , q_i is the angle between the i and $i+1$ phalanges

The thumb has been obtained by simply removing the distal phalanx from the standard finger concept.

IV. Sensors

The human hand is not just a mechanical but a sensing tool as well. The hand can distinguish between different grasped objects based on tactile information (size, shape, material, and surface properties). The five main *tactile* sensors in the hands are: Meisner, Merkel, Ruffini, Vater-Pacini for actuali touch sensing and free nerv endings for pain.

Besides the sense of touch the other important sens is *proprioception*. Proprioception is the sense of the relative position of neighboring parts of the body. Proprioception provides slow feedback on the status of the body internally. It is the sense that indicates whether the body is moving with the required effort, as well as where the various parts of the body are located in relation to each other.

In order to achive such a big modality of sensing in the BMT- H it would be needed a lot of different types of sensors which would required more space, computation power and draining. We selected the following sensors for the different senses.

A. Touch sensing

1) 3D Force sensor: The sensor (Model: TactoPoint, Tactologic, HUN) is a small and sensitive contact-force measuring unit, consisting of a single three-axial taxel. The three-axial signal provides the possibility for basic dynamic, spatial-temporal tactile measurements, while the extremely tiny size (1×2 mm sensor chip size) allows to be placed even in the phalanges under the viscoelastic pads [11].

2) Pressure sensor: A $44x44$ pressure array (Model: 5051 , TekScan, USA), consists of two thin, flexible polyester sheets. The electrically conductive electrodes, wThis sensor is designed in to the palm.

B. Proprioception

The designed sensor (Model:AS5045 , Austriamicrosystems, A) which is a magnetic rotary encoder for accurate angular measurement over a full turn. It is a system-on-chip, combining integrated Hall elements, analog front end and a digital signal processing in a single device. To measure the angle, only a simple two-pole magnet, rotating over the center of the chip, is required. It is a contact less, 12 bit high resolution, absolute type high speed digital readout (10 KHz) sensor. The sensor requires a magnet which has a magnetic field between 45–70 mT and it is made of *Neodymium Iron Boron* (NdFeB) or *Samarium Cobalt* (SmCo).

V. Test results

A. Test setup

The test setup which was used is shown in Fig. 5. A digital camera was mounted in order to obtain a stable position perpendicular to the plane of the movement of the finger prototype at this configuration the MCP joint was fixed and the PIP and DIP joints were controlled. The finger was actuated with servo motors (Model: SG90, TowerPro, TW). The test setup was controlled by a microcontroller (Model: PIC18f2321, Microchip, USA). The test setup was based on position control. The tendon in the test setup was a kevlar cable.

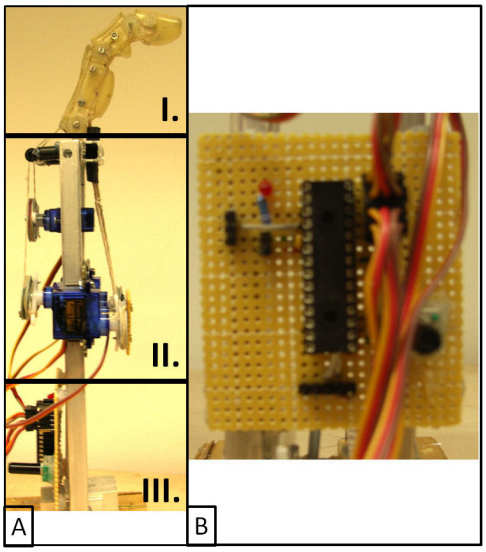


Fig. 5. Finger test setup (A) the finger prototype (I.), actuators (II.) and its control unit (III.), (B).

Before the hardware test computer simulations were made on the position control. The simulation results are shown in Fig. 6 and Fig. 7. The primary goal of the simulation was to determine the length change of the tendons at a given joint angle. The changing of tendons length were calculated in two ways the first one was based on the model of the first prototype showed in Fig. 6, where the actually angle change can be calculated with the law of cosine because it is a line between the emergent (B/K) point and the insertion (B/J) point. This makes the change nonlinear, and it causes a changing lever arm. At the next simulation (Fig. 7) the tendon was constrained on a arc between the emergent (B/K) point and the insertion (B/J) point to archive a linear change of the tendon length and a constant lever arm.

The phalanges angels were measured by a given PWM duty cycle shown in Fig. 8. The operation angle of the servo is $0^o - 170^o$, to control the servo actuators it needs a 50Hz PWM with the up-time of : $0.45ms - 1.95ms$ (this parameters are the rounded up average of the used actuators) that means to step one degree we needed a $\approx 9\mu s$ stepping in the duty cycle. In Fig. 8 four measurments can be seen: the (a) which is the initial state where every joint angle is 90^o in the (b) the DIP joints expected position was 45^o and the achieved was $\approx 35^o$, at (c) the PIP joint was controlled. It has to be noted that in this test the DIP joint was also controlled to hold the position with respect to the PIP, the expected angle was 25^o and the achieved was $\approx 23^o$. The (d) shows the end state of the finger where the expected angle was 180^o and the achieved was $\approx 168^o$, because of the maximal angel of the servos.

VI. Future Works & Conclusions

The functional test showed promising results, but there is still room for improvement. First of all the investigation of

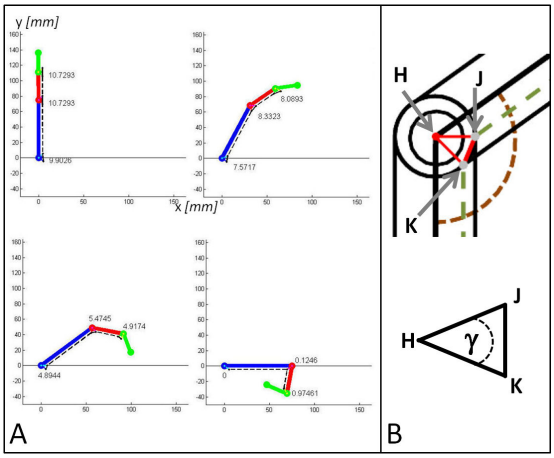


Fig. 6. Finger motion simulation, the shown values in (A) are the length of the tendons at the joints,(B) $JK = \sqrt{HK^2 + HJ^2 - 2 * HK * JK * cos(\gamma)}$ where HK: distance between pivot and emergent, HJ: distance between pivot and the insertion and HK = HJ, γ is the angle between HK and HJ

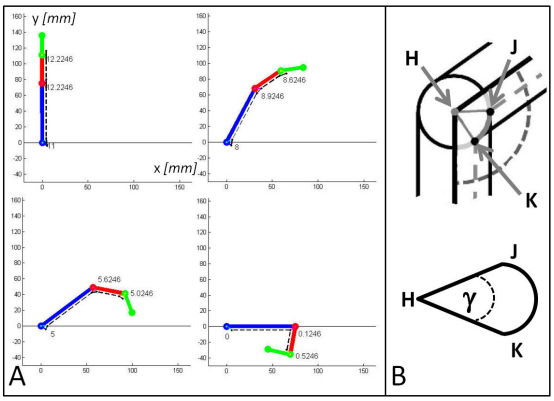


Fig. 7. Finger motion simulation,the shown values in (A) are the length of the tendons at the joints,(B) $JK = \gamma * HK$ where HK: distance between pivot and emergent, HJ: distance between pivot and the insertion HK = HJ, γ is the angle between HK and HJ

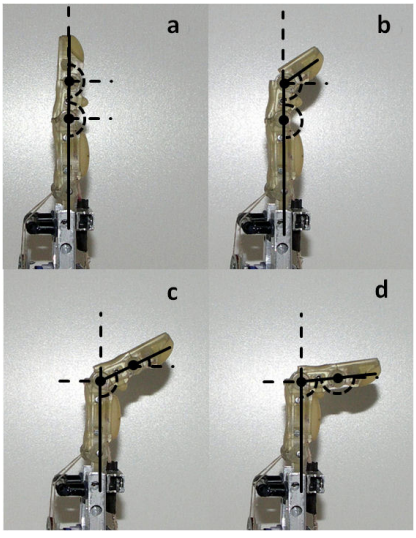


Fig. 8. The angles of the finger joints are measured using the pivots axes as mark points.

the movement of the human hand to achieve "human-like" behavior, the muscle structure of the forearm to improve a better optimal strategy for the actuation system to reduce the amount of them. A low-level control which works as reflexes. A second important objective that we are pursuing is to implement a global–control, based on the functional, biological structure of the cerebellum. This very challenging goal could ultimately lead to the development of a novel biomechatronic hand.

Acknowledgment

The Office of Naval Research (ONR) and the Operational Program for Economic Competitiveness (GVOP KMA) that supports the multidisciplinary doctoral school at the Faculty of Information Technology of the Pázmány Péter Catholic University, the Bolyai János Research Scholarship of the Hungarian Academy of Sciences and the NTP-OKA-I. for they support are gratefully acknowledged. The authors are also grateful to Professor Tamás Roska, and the members of the Robotics lab for the discussions and their suggestions. And special thanks go to Varinex Zrt. for the prototyping.

References

[1] H. Kawasaki, T. Komatsu, and K. Uchiyama, "Dexterous anthropomorphic robot hand with distributed tactile sensor: Gifu hand II," *Mechatronics, IEEE/ASME Transactions on*, vol. 7, no. 3, pp. 296–303, 2002.

[2] V. Weghe, M. Rogers, M. Weissert, and Y. Matsuoka, "The ACT hand: design of the skeletal structure," in *Robotics and Automation, 2004. Proceedings. ICRA'04. 2004 IEEE International Conference on*, vol. 4, pp. 3375–3379, IEEE, 2004.

[3] J. Butterfass, M. Grebenstein, H. Liu, and G. Hirzinger, "DLR-Hand II: Next generation of a dextrous robot hand," in *Robotics and Automation, 2001. Proceedings 2001 ICRA. IEEE International Conference on*, vol. 1, pp. 109–114, IEEE, 2006.

[4] M. Carrozza, B. Massa, S. Micera, R. Lazzarini, M. Zecca, and P. Dario, "The development of a novel prosthetic hand-ongoing research and preliminary results," *Mechatronics, IEEE/ASME Transactions on*, vol. 7, no. 2, pp. 108–114, 2002.

[5] S. Jacobsen, E. Iversen, D. Knutti, R. Johnson, and K. Biggers, "Design of the Utah/MIT dextrous hand," in *Robotics and Automation. Proceedings. 1986 IEEE International Conference on*, vol. 3, pp. 1520–1532, IEEE, 2002.

[6] C. Lovchik and M. Diftler, "The robonaut hand: A dexterous robot hand for space," in *Robotics and Automation, 1999. Proceedings. 1999 IEEE International Conference on*, vol. 2, pp. 907–912, IEEE, 2002.

[7] R. Tubiana, "Architecture and functions of the hand," *The hand*, vol. 1, 1981.

[8] G. Monkman, S. Hesse, and R. Steinmann, *Robot grippers*. John Wiley and Sons, 2007.

[9] R. Drake, W. Vogl, and A. Mitchell, *Gray's anatomy for students*. Elsevier/Churchill Livingstone Philadelphia, 2005.

[10] J. Doyle and M. Botte, *Surgical anatomy of the hand and upper extremity*. Lippincott Williams & Wilkins, 2003.

[11] G. Vasarhelyi, M. Adam, E. Vazsonyi, Z. Vizvary, A. Kis, I. Barsony, and C. Ducso, "Characterization of an integrable single-crystalline 3-d tactile sensor," *Sensors Journal, IEEE*, vol. 6, no. 4, pp. 928–934, 2006.

[12] J. Criag, "Introduction to robotics: mechanics and control," 2005.

Integrating and exploiting many-core architecture capabilities in MIMO communication systems

Csaba Máté Józsa
(Supervisor: Dr. Géza Kolumbán)
jozcsma@itk.ppke.hu

***Abstract* —Current single-input single-output (SISO) technologies data rate cannot be significantly improved because of the severe bandwidth regulations. A promising solution to significant increase of bandwidth efficiency, transmission capacity and system robustness is the exploitation of the spatial dimension. Multiple–input multiple-output (MIMO) systems can turn multipath propagation, scattering from a limiting factor to a very powerful feature. The implementation of wideband MIMO system posts a major challenge to hardware designers due to the huge computing power required for MIMO detection. With the help of the General Purpose Graphical Processing Unit (GP-GPU) the computing power is not a limiting factor anymore. In this paper after introducing the basic concepts of the MIMO system, the widely used system models, and detection algorithms are presented, finally the importance of many-core architectures are shown in the field of digital signal processing.**

Index Terms – multiple-input-multiple-output (MIMO), wireless systems, channel capacity, detection algorithms, graphical processing unit (GPU)

I. INTRODUCTION

Nowadays, the challenges in wireless communications systems are: increasing the link throughput (i.e., bit rate) and the network capacity[5][6][7]. The limiting factors of such systems are usually, the equipment cost and the radio propagation, and the frequency spectrum. However to fulfill the above goals, future systems should be characterized by improved spectral efficiency.

Research in the information theory, has revealed that important improvements in information rate can be achieved when multiple antennas are applied at both the receiver and transmitter side. The key feature of multiple-transmit multiple-receive antenna, i.e., Multiple-Input Multiple-Output (MIMO), systems is the ability to turn multipath propagation, traditionally a pitfall of wireless transmission, into a benefit for the user. MIMO effectively takes advantage of random fading and when available, multipath delay spread, for multiplying transfer rates.

The success of MIMO lies in the fact that the performance of wireless systems are many orders of magnitude improved at no cost of extra spectrum only complexity is added to the different algorithms and hardware.

The MIMO techniques can basically be split into two categories: Space-Time Coding (STC)[10] and Space Division Multiplexing (SDM). STC increases the robustness of the wireless communication system by transmitting different representations of the same data stream (by means of coding) on the different transmit branches, while SDM achieves a higher throughput by transmitting independent data streams on the different transmit branches simultaneously and at the same carrier frequency.

However the complexity of the detector algorithms used in the receiver structures is defined by many factors (coding, channel, antenna mapping) there are some cases, when the detection cannot be done by a simple hardware components, instead it's worth using many-core architectures such as graphical processing units (GPUs), or field programmable gate arrays (FPGAs). Nowadays the multi-core architectures are playing a prominent role in computing theory. This is because their price is getting cheaper and the computational possibilities are far beyond that of the general purpose processors. This is very important because the computational power of the supercomputers can be reached by scientists with this kind of systems at a reasonable cost. Despite the fact that the number of this new architectures are permanently increasing, the change is not trivial. The successful utilization of the facilities provided by these parallel architectures necessitates rethinking the programming paradigms, to redesign and implement the old algorithms and softwares.

II. SYSTEM MODEL

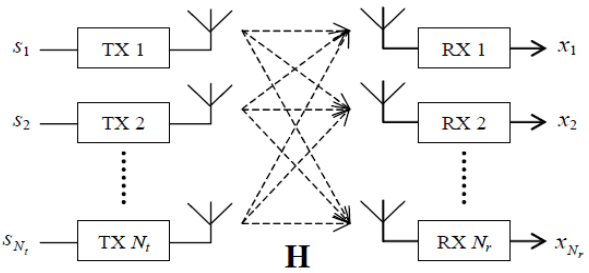


Figure 1: Model of MIMO communication system

As shown in Figure 1, a MIMO wireless communication system has N_t transmit and N_r receive antennas. In the case of SDM [8], the different transmit antennas radiates different streams on the same frequency. This can be denoted by the following equation:

$$\mathrm{s}(t)=\begin{pmatrix} s_1(t) \\ s_2(t) \\ \vdots \\ s_{N_t}(t) \end{pmatrix},\ \mathrm{x}(t)=\begin{pmatrix} x_1(t) \\ x_2(t) \\ \vdots \\ x_{N_r}(t) \end{pmatrix} \text{ and } \mathrm{H}(t)=\begin{pmatrix} h_{11}(t) & h_{12}(t) & \cdots & h_{1N_t}(t) \\ h_{21}(t) & h_{22}(t) & \cdots & h_{2N_t}(t) \\ \vdots & \vdots & \ddots & \vdots \\ h_{N_r1}(t) & h_{N_r2}(t) & \cdots & H_{N_rN_t}(t) \end{pmatrix}$$

$$\mathrm{x}(t)=\int_{-\infty}^{\infty} \mathrm{H}(t,\tau)\mathrm{s}(t-\tau)\mathrm{d}\tau$$

In the case of narrowband systems, where the system bandwidth is smaller than the coherence bandwidth the system equation can be rewritten as:

$$\mathbf{x}(t) = \mathbf{H}\mathbf{s}(t)$$

Usually this model is further simplified, as a result the continuous time waveforms will be replaced by symbols and assuming quasi-stationary channel during a packet transmission, the channel matrix will contain complex numbers, representing the gain and the phase deviation of the channel [1]. In this case the system model can be defined as follows:

$$\mathbf{x} = \mathbf{H}\mathbf{s}$$

Where **s** represents the transmitted MIMO symbol vector, **x** represents the received MIMO symbol vector, and **H** represents the channel matrix. In this case the vector and matrix elements are complex numbers.

III. Channel Capacity

One way to express the gain of a MIMO system over a SISO system is by means of the capacity. In general, the capacity is defined by information theory as an upperbound on the information rate for error-free communication. The capacity of a MIMO communication link depends not only on the fading statistics, as for a SISO link, but also on the spatial correlation of the channel. This results in a random capacity whose instant value depends on the corresponding instantaneous **H** matrix. When the instant capacity is less than the chosen rate, a channel outage occurs.

For a memoryless 1x1 (SISO) system the capacity is given by

$$C = \log_2(1+\rho|h|^2) \text{ b/s/Hz}$$

Where h is the normalized complex gain of a fixed wireless channel or that of a particular realization of a random channel, and ρ is the SNR at any RX antenna. With the increasing number of receive and transmit antennas the capacity of a MIMO system, for N TX and M RX antennas, is given by [11]

$$C = \log_2\left[\det(\mathrm{I}_M + \frac{\rho}{N}\mathrm{H}\mathrm{H}^*)\right] \text{ b/s/Hz}$$

Where $\mathbf{H}^*$ means the transpose-conjugate of the **H**, which is the M x N channel matrix.

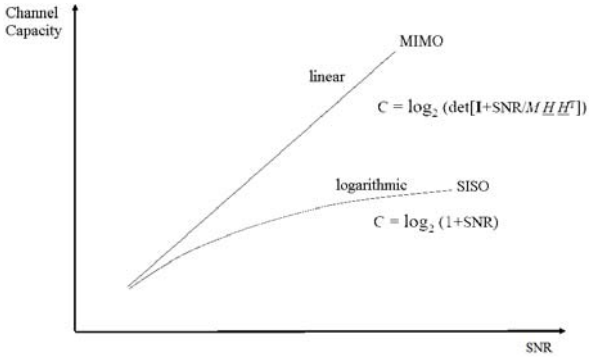


Figure 2. MIMO vs SISO channel capacity

Foschini [11] and Telatar [12] both demonstrated that the capacity in MIMO systems grows linearly with rather than logarithmically as in the SISO case. This result can be intuited as follows: the determinant operator yields a product of nonzero eigenvalues of its (channel-dependent) matrix argument, each eigenvalue characterizing the SNR over a so-called channel eigenmode.

Since the rank equals the number of nonzero eigenvalues, it represents the number of available spatial subchannels. The number of spatial subchannels (or eigenmodes) indicates the number of parallel symbol streams that can be transmitted through the MIMO channel, using the same frequency bandwidth, and is hence a measure of the capacity of the MIMO channel.

With spatial multiplexing, the virtual subchannels of a MIMO channel are exploited by sending independent data streams on multiple transmit antennas to improve data rates. Symmetric orthogonal channels are desirable since they do not have null modes, and hence do not lose transmitted information. Moreover, these channels can be inverted in the receiver without noise amplification, leading to a good system performance. Two concrete measures for the orthogonality of a MIMO channel are the condition number and Effective Degrees Of Freedom (EDOF).

IV. Detection Algorithms

A. Linear MIMO Detection[13]

A straightforward approach to recover **s** from **x** is to use an $Nt \times Nr$ weight matrix **W** to linearly combine the elements of **x** to estimate **s**, $\hat{\mathbf{s}} = \mathbf{W}\mathbf{x}$.

1) Zero forcing (ZF)

Zero Forcing is a linear MIMO technique. The processing takes place at the receiver where, under the assumption that the channel transfer matrix is invertible, is inverted and the transmitted MIMO vector is estimated.

This principle is based on a conventional adaptive antenna array technique, namely, linear combinatorial nulling. In this technique, each substream in turn is considered to be the desired signal, and the remaining data streams are considered as "interferers". Nulling of the interferers is performed by linearly weighting the received signals such that all interfering terms are cancelled

2) Minimum Mean Squared Error (MMSE)

A drawback of the ZF is that nulling out the interference without considering the noise could boost up the noise power significantly, which in turn results in performance degradation. To solve this, MMSE minimizes the mean squared-error.

B. Nonlinear MIMO Detection

1) V-Blast

A popular nonlinear combining approach is the vertical Bell Labs layered space time algorithm (VBLAST). It uses the detect-and-cancel strategy similar to that of decision-feedback equalizer. Either ZF or MMSE can be used for detecting the strongest signal component used for interference cancellation. The performance of this procedure is generally better than ZF and MMSE.

2) Maximum Likelihood Detection (MLD)[8]

In MLD, **s** is estimated according to the Maximum Likelihood principle. The idea is to find a vector $\mathbf{s}_j$ for which the probability $p(s_j \mid x)$ is maximized (with $1 \le j \le K$), where *K* denotes all possible transmitted vectors:

$$K=M^{N_t}$$

with *M* representing the number of constellation points. Using Bayes' rule, this probability may be expressed as:

$$p(s_j \mid x) = \frac{p(x \mid s_j)p(s_j)}{p(x)}$$

where $p(x \mid s_j)$ is the conditional probability density function (pdf) of the observed vector, given that $\mathbf{s}_j$ is transmitted. Probability $p(s_j)$ is the probability of the *j*-th vector being transmitted. If the *K* vectors are equally probable to be transmitted, then $p(s_j) = 1/K$. Furthermore, the denominator is independent of $\mathbf{s}_j$. Consequently, finding the vector that maximizes $p(s_j \mid x)$ is equivalent to finding the vector that maximizes $p(x \mid s_j)$.

The (conditional) pdf $p(x \mid s_j)$ is a complex multivariate normal distribution. The general formula of a complex multivariate normal distribution **x**, with mean and covariance matrix **Q**, can be shown to be:

$$p(x) = \det(\pi \mathrm{Q})^{-1} e^{(-(x-\mu)^* \mathrm{Q}^{-1}(x-\mu))}$$

For a specific channel matrix **H** and given $\mathbf{s}_j$, this leads to:

$$p(x \mid \mathrm{H}, s_j) = \det(\pi \mathrm{Q})^{-1} e^{(-(x-\mathrm{H}s_j)^* \mathrm{Q}^{-1}(x-\mathrm{H}s_j))}$$

Consequently, finding the maximum of the conditional probability $p(s_j \mid x)$ leads to:

$$\hat{\mathrm{s}} = \arg\max_{s_j \in \{s_1,\ldots,s_K\}} p(\mathrm{x}|\mathrm{H},\mathrm{s}_j) = \arg\min_{s_j \in \{s_1,\ldots,s_K\}} \left\|\mathrm{x}-\mathrm{Hs}_j\right\|^2$$

The last formula is the MLD solution. Note that MLD is optimal in performance, because finding the maximum of the conditional probability $p(s_j \mid x)$ leads to the minimization of the symbol error probability. Note that the MLD solution requires an exhaustive search through all possible transmitted vectors *K*. So, the complexity is proportional to *K*, which is the main disadvantage of this method. For a small number of transmit antennas (Nt<5), however, the complexity seems reasonable.

C. Linear Adaptive MIMO Detection [13]

Instead of assuming known channel matrix **H**, which usually requires channel testing before each transmission and then calculating **W** in a bursty manner, adaptive algorithms estimate **W** directly through iteration via the use of a known training sequence at the beginning of each transmission.

1) Least Mean-Square (LMS)

LMS is an estimate of the steepest descent algorithm and updates **W** according to

$$\mathrm{W}_i = \mathrm{W}_{i-1} + \mu\left[\mathrm{s}_i - \mathrm{W}_{i-1}\mathrm{x}_i\right]\mathrm{x}_i^*,$$

where μ is the update step size.

2) Recursive Least-Squares (RLS)

RLS is the recursive solution to the exponentially weighted least-squares (LS) problem.

D. Computational Complexity

The comparison of different MIMO detection algorithms implementation complexity is shown in Table 1. Since the hardware cost of each algorithm is highly implementation-specific, the comparison gives a rough estimation of the required multiplications for each algorithm.

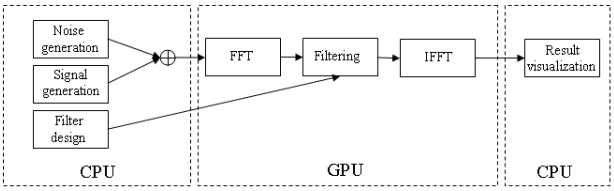
Algorithm	Real Multiplications	Giga Op. per Second 4×4	8×8	16×16
LMS	$8N_tN_r + 2N_t$	3	13	52
RLS	$14N_r^2 + 8N_tN_r + 6N_r$	9	36	143
ZF/MMSE	$4N_t^3 + 8N_t^2N_r$	19	154	1,229
ZF/MMSE -VBLAST	$N_t^4 + \frac{8}{3}N_t^3N_r$ $+2N_t^3 + 4N_t^2N_r$	33	452	6,622
ML	$4N_tN_rM^{N_t}$	410	4×10^5	10^{11}

Table I Computational complexity of different detection algorithms

The GOPS figure is calculated for 25MHz bandwidth and 4-QAM modulation. Due to the over-simplified nature of these assumptions, the estimation in Table I is only meaningful in the order of magnitude sense. The estimations for LMS and RLS are per iteration and those for the nonadaptive algorithms are for estimating **W** and do not take into account the cost of estimating the channel matrix **H**.

V. IMPLEMENTATION

Before redesigning some of the well known serial detector algorithms, we wanted to test the capabilities of a GTX 460 NVidia GPU in a digital filtering problem, just for a proof of concept. The following block diagram shows the problem flow:



The key concept is that the FFT [3], inverse FFT and filtering algorithms can be parallelized, thus giving this computationally intensive parts to the GPU we reach significant speedups. The signal and noise generation, the filter design and the visualization of the filtered signal is done by CPU.

As shown in Figure 3. we can see that increasing the points number of the FFT, the CPU's execution time is exponentially growing while the GPU's execution time follows a linear growth. The results show that if we want to evaluate a 10^6 points FFT, the GPU will perform 10 times faster than the CPU.

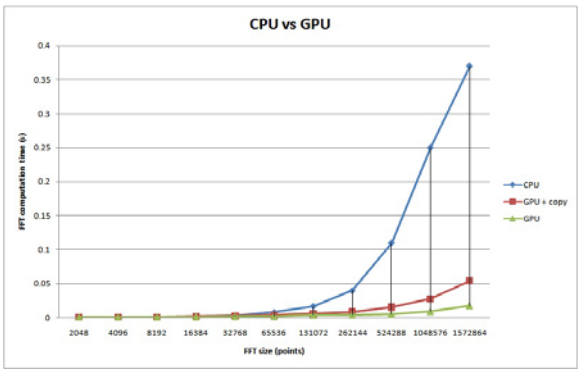


Figure 3. Signal frequency domain filtering time (CPU vs. GPU)

VI. CONCLUSIONS

In this paper we showed the importance of MIMO communication systems, presented the basic models, derived the channel capacity, presented some of the detection algorithms, and proved that GPU's can significantly improve the computational time of digital signal processing algorithms. In the future work we would like the examine the possibilities of parallelizing the presented detection algorithms and to reach high data rates while keeping the bit error rate as low as possible with these very powerful many-core architectures.

REFERENCES

[1] J.G. Proakis, "Digital Communications", 4th edition, 2001

[2] Simon Haykin, "Communication Systems", 4th edition, 2001

[3] http://developer.download.nvidia.com/compute/cuda/3_1/toolkit/docs/CUFFT_Library_3.1.pdf

[4] Albert van Zelst, "MIMO OFDM for Wireless LANs", Technische Universiteit Eindhoven, 2004

[5] E. Dahlman, S. Parkwall, and J. Sköld, "4G LTE/LTE-Advanced for Mobile Broadband"

[6] V. Tarokh, "New Directions in Wireless Communication Research", Harvard University, 2009

[7] A. Sibille, C Oestges, A. Zanellas, "MIMO From Theory to Implementation", 2011

[8] A. van Zelst, R. van Nee, G.A. Awater, "Space Division Multiplexing(SDM) for OFDM systems", IEEE Conf. Proce. Veh. Technol., vol. 2., 2000

[9] H. Xu, D. Chizhik, H. Huang, R. Valenzuela, "A Generalized Space-Time Multiple-Input Multiple-Outpu (MIMO) Channel Model", IEEE Trans. Commun., vol. 3., no. 3, May 2004

[10] D. Gesbert, M Shafi, D. Shiu, P. Smith, A. Naguib, "From Theory to Practice: An Overview of MIMO Space-Time Coded Wireless Systems", IEEE Journal Commun., vol. 21., no. 3, April 2003

[11] G. J. Foschini and M. J. Gans, "On limits of wireless communications in a fading environment when using multiple antennas," Wireless Pers. Commun., vol. 6, pp. 311–335, Mar. 1998.

[12] E. Telatar, "Capacity of multiantenna Gaussian channels," AT&T Bell Laboratories, Tech. Memo., June 1995.

[13] J. Wang and B. Daneshrad, "A Comparative Study of MIMO Detection Algorithms for Wideband Spatial Multiplexing Systems", IEEE Wireless Commun., vol. 1.,May 2005

In-line color digital holographic microscope for biological water quality measurement

Márton Kiss
(Supervisor: Dr. Szabolcs Tőkés)
kisma1@digitus.itk.ppke.hu

***Abstract*—We introduce a color digital holographic microscope for measuring the biological content of water samples. Our approach uses single shot RGB exposure in an in-line holographic setup to obtain color images. With the application of appropriate numerical algorithms we can fulfill color crosstalk compensation, segmentation, and twin image removal tasks, and we obtain good quality color image reconstructions with 1µm resolution from a 1 mm^3 volume. We briefly compare the conventional color CCD/CMOS and the Foveon X3 sensor for color digital holographic applications. The in-line holographic setup and reconstruction algorithms are presented with demonstrative simulations, experimentally captured and numerically reconstructed images.**

Keywords— Color Digital Holography, Digital Holographic Microscope, Color crosstalk compensation

I. Introduction

The possibility to obtain volumetric information from a single image using wave field propagation algorithms makes digital holography a promising method to investigate freely moving biological organisms in water samples. As in ordinary holography, digital holography captures the interference pattern of two beams, namely the object and the reference beam, but instead of using a high resolution holographic recording medium, the digital holographic system uses an image sensor to capture the holograms. The reconstruction of this digitally recorded hologram is done numerically by the simulation of light propagation. Digital holography is frequently used to capture 3D-4D information of objects within a volume.[1], [2] In most cases the recorded holograms are monochromatic, however, since the color of the object can carry relevant information several multi-wavelength optical setups emerged. These approaches usually use time multiplexed recordings, where the objects are illuminated by only a single wavelength at a given time, the recording is done by a monochromatic sensor and the captured image is reassembled algorithmically. [3] There are experiments where objects were simultaneously illuminated by several wavelengths [4] and the refractive index of phase samples was determined using multi wavelength illumination. [5] To our knowledge there are methods that use only monochromatic, [6] two-color, [7] or sequentially exposed three color cases, [8] where digital holography is used in microscope. In the case of single shot three-color exposure, color-crosstalk can cause noise by false reconstructions. Our previous in-line approach used a conventional color CCD with a Bayer filter array,[9] whereas here we investigated the behavior of the Foveon sensor with our color compensation method, which is capable to greatly reduce the color crosstalk caused by the overlapping transmission curves. By utilizing the Foveon sensor the sampling issues of the conventional color CCD can be avoided. In the case of in-line holography, for multiple object recognition on additional segmentation and twin image removal is required. Our experimental results are presented in the paper.

II. Optical setup

A. In-line holographic setup

We use an in-line holographic setup where the illumination beam also serves as the reference beam. The other frequently used method is the off-axis holographic setup, where an additional reference beam is used which intersects with the object beam at a none zero angle on the hologram plane. Off-axis setups can easily provide better quality images since the twin image and zero order terms are spatially separated from the object during the reconstruction process, and in multi color cases wavelength separation can also be done. [4] However, the off-axis architectures are more complex and costly systems, they can retrieve only approximately 1/3 of the resolution of the image sensor as the object, and thus it inherently sacrifices the majority of the achievable resolution. Although in-line architectures use the full resolution of the image sensor, but due to the overlapping noise terms (twin image, color-crosstalk) additional data processing steps are needed. An other very interesting feature of in-line systems is the capability to work with light sources that have a short coherence length. [10] Our approach uses the in-line setup, and can be seen in Fig. 1. A flow cell of 0.8 mm thickness containing fresh water algae samples is illuminated by a multi wavelength laser beam. We use an Olympus LUCPLFLN20X microscope objective with an achromatic tube lens of 150mm focal length. Previously we used a conventional Nikon D60 single-lens reflex camera as a detector, but now we investigated the behavior of the SIGMA SD14 which utilizes a FOVEON X3 image sensor.

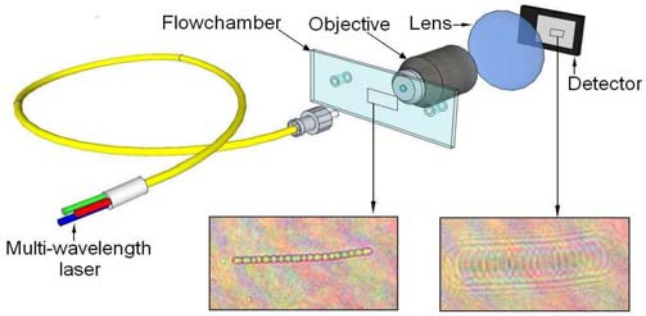


Figure 1. The used in-line holographic setup

B. Light Source

Digital holography assumes that the reference beam is known, therefore usually a distortion free single mode Gaussian beam is applied, where the Rayleigh range of the beam can be either treated as a plane wave or a spherical wave at the detector plane. Traditionally, a nearly perfect spherical wavefront is created by low-pass filtering of a focused laser beam with a pinhole of a few μm in diameter. This pinhole can be treated as a point like light source. Unfortunately, it tends to get blocked by dust or other particles easily. Furthermore, the proper adjustment of such a pinhole to the focused laser beam can be troublesome. Consequently, we implement a point like light source with an alternative method by using the fiber end of a single mode optical cable, which has 4.5μm mode field diameter. We have built an optical system which simultaneously illuminates the sample with red, green, and blue lights, thus it becomes possible to capture holograms of the object in the RGB color scheme at the same time, thus allowing us examine to moving samples. Using a proper detector even video rate color holographic recording can be achieved. We use pigtailed laser diodes as the red (650nm) and violet (406nm) light source, while a frequency doubled Nd:YAG laser is coupled into a fiber to get the green (532nm) one. To achieve simultaneous, parallel illuminating wavefronts fiber couplers are applied to guide all the lasers into a single fiber. We use Thorlabs FC632-50B-FC fiber couplers, which was designed for the 632nm wavelength, thus their performance is not ideal for the other two wavelengths. Nevertheless, cascading them as shown in Fig. 2 with properly set intensities provided us a coherent "white" laser source with nearly perfect spherical wavefronts.

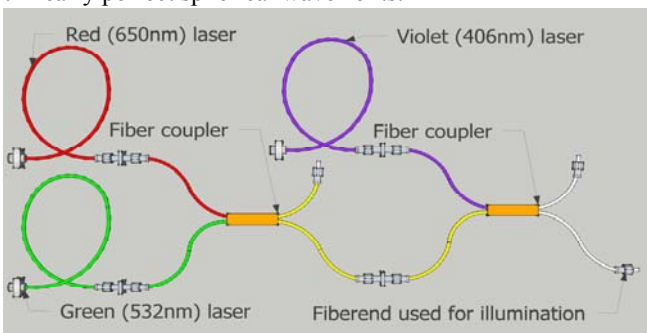


Figure 2. The cascaded fiber couplers used to create a multi-colored light source used for illuminating the object.

C. Detector

The image sensor is the most crucial part in a digital holographic optical system. The pixel pitch, the effective area, and the image quality of the sensor is one of the main limiting factors for the observable volume size and for the resolution of the whole microscope. Most single chip color cameras use a Bayer filter to capture color information. This filter is a 2D array of color filters with RGB colors shown in Fig. 3a. The sensors photosensitive elements have similar absorption spectra, but the transmission characteristics of the superimposed Bayer filter makes them color sensitive. However, the transmission spectrum (see Fig. 3b) usually overlaps and thus the obtained image contains considerable color crosstalk. In our previous setup we used a Nikon D60 camera which uses a Bayer filter array. Since the transmission spectrum of the Nikon D60's Bayer filter is not available we measured the transmission of each of the wavelengths we used. Table 1 shows the results for reference. Using a detector with a

TABLE I
NORMALIZED TRANSMISSION OF THE BAYERN PATTERN'S COLOR FILTERS AT THE NIKON D60.

Illumination wavelength	Red Filter	Green Filter	Blue Filter
650 NM	1	0.042838	0.005157
532 NM	0.066383	1	0.032673
406 NM	0.178884	0.115504	1

Bayer filter also causes sampling problems of the wavefront, and thus various color artifacts. The color crosstalk of a Bayer filter array is relatively small, but the sampling method of these sensors makes color crosstalk compensation problematic. These difficulties can also be overcome by using a direct image sensor such as the Foveon X3. Unlike the Bayer filter array, the Foveon image sensor takes advantage of the fact that red, green, and blue light penetrate silicon to different depths, thus it is capable to capture the full range of colors at each pixel location. (See Fig. 3c) The main weakness of this sensor

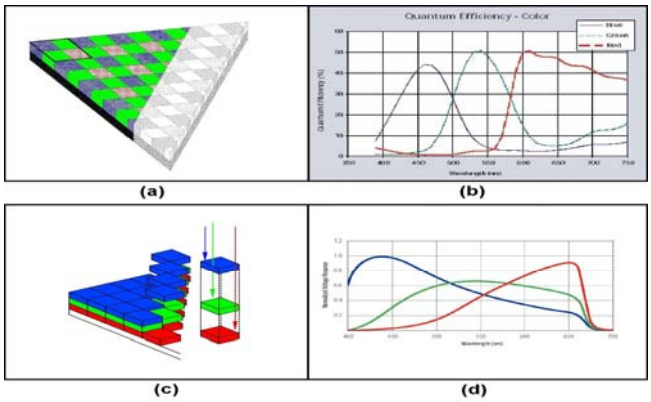


Figure 3. a) Bayer filter array on the detector surface. b) Usual color sensitivity spectrum of a color CMOS sensor with Bayer filter array. c) The color sensitive pixel structure of the Foveon X3 sensor d.) Color sensitivity spectrum of the Foveon X3 image sensor.

is the considerable higher color crosstalk which can be seen in Fig. 3d. We also measured the color crosstalk of the Foveon image sensor (see Table 2) at the used wavelengths.

TABLE II
NORMALIZED TRANSMISSION OF THE FOVEON X3 SENSORS COLOR CHANELS AT THE SIGMA SD14.

Illumination wavelength	Red Filter	Green Filter	Blue Filter
650 NM	1	0.585268	0.301762
532 NM	0.503208	1	0.852527
406 NM	0.062338	0.076821	1

III. LIGHT PROAGATION ALGORITHM

There are three main methods for digital emulation of propagating wave fields between parallel planes.[11] These methods are the single Fourier transform based Fresnel method, the convolution based Fresnel method, and the angular spectrum method. They commonly use fast Fourier transform to calculate the propagated electric field distribution. During our measurements we used the angular spectrum method because it works even for small propagation distances, as does not use paraxial approximations. An extension of this method was recently found by Matsushima, [12] which improves the accuracy of the angular spectrum method for larger distances by proper sampling.

A. The multi-wavelength algorithm

A quick overview of the algorithm can be seen in Fig.(4). We start with the raw sensor data of the captured hologram. Due to the overlapping transmission curves, a color crosstalk compensation has to be made on the red, green, and blue components, to acquire the three holograms of the object at the different wavelengths. By using the angular spectrum method on one of the holograms the whole volume is reconstructed layer by layer to find the position of the objects. If there are two objects close to each other laterally, but at different depths the diffraction pattern of the first object is usually so spread out that it overlaps with the second object at the plane where the second object is in focus. To eliminate this problem we use a segmentation process where the objects in the examined volume are removed one after an other, so only the examined object remains. The next step is a twin image removal process to achieve the best possible image quality. The three holograms are then recombined to the color image which is then processed by an object classification algorithm to recognize the different types of algae in the water sample.

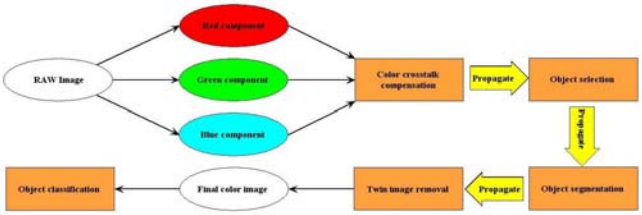


Figure 4. The schematics of the algorithm

IV. ERROR COMPENSATION

A. Color crosstalk compensation

As mentioned before, to record the color digital hologram the sample is illuminated with the combination of three laser beams with the wavelength of 406nm, 532nm, and 635nm. Each of these waves creates their own separate hologram, but the Sigma SD14's Foveon X3 image sensor captures it at the same time. Due to the considerable color crosstalk shown in Table 2 the three holograms are mixed at the color channels of the sensory data. Because of the wavelength difference each hologram has to be processed individually, thus the first task is to separate them. We found a very simple method to overcome this problem. Using the values in Table 2 we can calculate the effect of the overlapping spectrum. Using

$$\overline{\overline{T}} = \begin{bmatrix} 1 & 0.503208 & 0.062338 \\ 0.585268 & 1 & 0.076821 \\ 0.301762 & 0.852527 & 1 \end{bmatrix}$$

effect of the overlapping sensitivity curves can be calculated as:

$$\overline{M} = \overline{\overline{T}} \cdot \overline{R}$$

$$\begin{pmatrix} M_{red} \\ M_{green} \\ M_{blue} \end{pmatrix} = \begin{bmatrix} 1 & 0.503208 & 0.062338 \\ 0.585268 & 1 & 0.076821 \\ 0.301762 & 0.852527 & 1 \end{bmatrix} \cdot \begin{pmatrix} R_{red} \\ R_{green} \\ R_{blue} \end{pmatrix}$$

where $\overline{M}$ is the measured intensity vector and $\overline{R}$ is the real intensity vector. Here by multiplying the previous equation with $\left(\overline{\overline{T}}\right)^{-1}$ from the left we get:

$$\overline{R} = \overline{\overline{T}}^{-1} \cdot \overline{M}$$

$$\begin{pmatrix} R_{red} \\ R_{green} \\ R_{blue} \end{pmatrix} = \begin{bmatrix} 1.407488 & 0.677852 & -0.035667 \\ -0.846575 & 1.477796 & -0.060752 \\ 0.297002 & -1.055312 & 1.062555 \end{bmatrix} \cdot \begin{pmatrix} R_{red} \\ R_{green} \\ R_{blue} \end{pmatrix}$$

Using this simple matrix multiplication we can compensate for the color crosstalk of the sensor.

B. Segmentation

Reconstruction of the in-line holograms is a hard problem not only due to the twin image (zero order terms) caused noises, [14] but the neighboring objects diffraction can also distort the reconstructed image quality. Furthermore, the common phase retrieval algorithms do not tolerate this type of biases.[15] Therefore, first we segment the recorded hologram according to the occurring objects sub-holograms. This segmentation is based on the objects reconstruction distance and their corresponding spatial supports, which can be determined by an algorithm that use an appropriate focus measure.[16] The algorithm is based on the special inner structure of the in-line holograms. The support of the segmented object decreases the aperture of the other objects and thus decreases the achievable resolution, but since we can achieve almost perfect segmentation, all the diffraction caused by the other objects becomes negligible. Diffraction pattern of

the other objects are abolished almost perfectly, and this way enhancing the quality of the reconstruction.

C. Twin image removal

The hologram of the segmented object is still an in-line hologram and so it is contaminated by twin image and zero order noises. As the estimated object support and reconstruction distance is known we can apply the well known Gerchberg-Saxton-Fienup phase retrieval algorithm. [17] Our solution extends this method to the recorded multi-wavelength holograms. This way we can remove the twin image diffraction pattern from the reconstructed object.

D. Abberation compensation

Although the main advantage of digital holography is the ability to reconstruct sharp images at different layers of a volume algorithmically, thus virtually increasing the depth of field of a microscope system, there are other benefits of knowing the whole wavefront. For example, after obtaining the recorded holograms, we can correct the aberrations of the optical system numerically. Using color holography the most common aberrations are the lateral and the transversal chromatic aberration. [3] Since we used a high end microscope objective, these chromatic aberrations were not observable, but we would like to note that using a less sophisticated microscope objective they cause considerable distortions. In these cases measuring the transfer function of the system for the whole volume, one can design an inverse transformation method to compensate for chromatic and some other type of aberrations as well.

V. Conclusion

Our main goal is to develop a Color Holographic Video Microscope to measure the biological content of water samples. The introduced color crosstalk compensation makes it possible to use simultaneous three-color illumination, and this way to record a hologram of the examined volume with three colors at video frame rate, since the illuminating colors are not time multiplexed. The color crosstalk usually causes smaller noise then the twin image. However, as twin image elimination is a deconvolution task, its success and quality considerably depends on the noise of the input in-line hologram. This way, without color crosstalk compensation, there is little hope of good quality twin image removal. Our earlier sampling method [9] can cause aliasing, which can be avoided by using the Foveon X3 sensor. The introduced color compensation method can also be used in the case of off axis architectures, where the effect of color crosstalk seems more crucial as the twin image noise does not exist. An in-line three color (RGB) digital holographic microscope setup using a Sigma SD14 camera was presented. We detailed the color compensation method, which is capable to greatly reduce, the color crosstalk caused by the overlapping color sensitivity spectrum of the image sensor. We gave experimental results of our segmentation and twin image removal algorithm. The performance of our color DHM is demonstrated by reconstructed images from a volume of flowing water containing algae.

References

[1] 1. Xu, WB Jericho, M. M. I. K. H., “Digital in-line holography for biological applications,” PNAS 98(20), 11301–11305 (2001).

[2] Garcia-Sucerquia, J., Xu, W., Jericho, S., Klages, P., Jericho, M., and Kreuzer, H., “Digital in-line holographic microscopy,” Applied optics 45(5), 836–850 (2006).

[3] Zhao, J., Jiang, H., and Di, J., “Recording and reconstruction of a color holographic image by using digital lensless Fourier transform holography,” Optics Express 16(4), 2514–2519 (2008).

[4] Khmaladze, A., Kim, M., and Lo, C.-M., “Phase imaging of cells by simultaneous dual-wavelengthreflection digital holography,” Opt. Express 16(15), 10900–10911 (2008).

[5] Desse, J.-M., Picart, P., and Tankam, P., “Digital three-color holographic interferometry for flow analysis,” Optics Express 16(8), 5471 – 5480 (2008).

[6] Marquet, P., Rappaz, B., Magistretti, P., Cuche, E., Emery, Y., Colomb, T., and Depeursinge, C., “Digital holographic microscopy: a noninvasive contrast imaging technique allowing quantitative visualization of living cells with subwavelength axial accuracy,” Optics letters 30(5), 468–470 (2005).

[7] Kuhn, J., Colomb, T., Montfort, F., Charrire, F., Emery, Y., Cuche, E., Marquet, P., and Depeursinge, C., “Real-time dual-wavelength digital holographic microscopy with a single hologram acquisition.,” Optics Express 15(12), 7231 – 7242 (2007).

[8] Mo, X., Kemper, B., Langehanenberg, P., Vollmer, A., Xie, J., and von Bally, G., “Application of Color Digital Holographic Microscopy,” in [DGaO Proceedings], (2009).

[9] Göröcs, Z., Kiss, M., Tóth, V., Orzó, L., and Tőkés, S., “Multicolor digital holographic microscope (DHM) for biological purposes,” in [Proceedings of SPIE], 7568, 75681P (2010).

[10] Oh, C., Isikman, S., Khademhosseinieh, B., and Ozcan, A., “On-chip differential interference contrast microscopy using lensless digital holography,” Opt. Express 18, 4717–4726 (2010).

[11] Kreis, T., Adams, M., and Jueptner, W., “Methods of digital holography: a comparison,” Proceedings of SPIE 3098, 224 (1997).

[12] Matsushima, K. and Shimobaba, T., “Band-limited angular spectrum method for numerical simulation of free-space propagation in far and near fields,” Optics Express 17(22), 19662–19673 (2009).

[13] Shen, F. and Wang, A., “Fast-Fourier-transform based numerical integration method for the Rayleigh- Sommerfeld diffraction formula,” Applied optics 45(6), 1102–1110 (2006).

[14] Koren, G., Polack, F., and Joyeux, D., “Iterative algorithms for twin-image elimination in in-line holography using finite-support constraints,” Journal of the Optical Society of America A 10(3), 423–433 (1993).

[15] Denis, L., Fournier, C., Fournel, T., and Ducottet, C., “Numerical suppression of the twin image in inline holography of a volume of micro-objects,” Measurement Science and Technology 19(7), 074004 (10pp) (2008).

[16] Bergoend, I., Colomb, T., Pavillon, N., Emery, Y., and Depeursinge, C., “Depth-of-field extension and 3D reconstruction in digital holographic microscopy,” in [Proc. of SPIE], 7390, 73901C–1 (2009).

[17] Fienup, J. R., “Phase retrieval algorithms: a comparison,” Appl. Opt. 21(15), 2758–2769 (1982).

Evaluation of movement parameters in constrained tracking arm movements

Bence Borbély
(Supervisor: Dr. József Laczkó)
borbely.bence@itk.ppke.hu

***Abstract*—A three-dimensional (3D) measurement method of constrained tracking arm movements is presented. Healthy controls and post-stroke patients performed predefined tracking arm movements repeatedly. During the movement task the subject had to follow a moving disk (target) on a computer screen with a mouse pointer. The pointer was moved by means of mouse using a digitizer tablet. Each task consisted of four trial sets with fixed path shape (circle or square) and movement speed (normal or fast) conditions of the target. Spatial coordinates of anatomical landmarks of the subject's arm and the surface electromyogram (EMG) of the main arm muscles were recorded with an ultrasound based movement analyzer system. Seven joint angles defining the spatial configuration of the arm were calculated from recorded landmark coordinates. Kinematical methods based on this seven-degrees-of-freedom model and EMG processing techniques are presented.**

Keywords-tracking; post-stroke; rehabilitation;

I. INTRODUCTION

Stroke is one of the major causes of motor disability. One third of the patients suffered from stroke will live with decreased functional capacity. About half of all hemiplegic survivors will be left with a non-functional arm. Thus, post-stroke rehabilitation is the key of (partly) recovering an individual's lost motor capabilities. To obtain objective information concerning arm function, kinematical measurement methods have been developed to assist the clinical evaluation process. These approaches include joint torque analysis [1], spatial target reaching movements [2,3] and planar circle drawing methods [4-6]. In most of these studies robotic exoskeleton systems were applied to measure the kinematical properties of the subjects' movements, i.e. the end-effector path, the joint angles of the arm or the active range of motion. These methods showed a good correlation with the Fugl-Meyer scale applied in clinical studies [7], however using such a robotic device can induce some restrictions to the area of motion of the subjects' arm compared to other commercially available optical or ultrasound-based movement analyzer systems.

Computer-aided diagnostic methods have been developed to examine the effects of movement disorders affecting movement patterns while the subjects were using a computer mouse [8,9]. These studies showed that endpoint variability of repeatedly executed drawing arm movements carries important features in the diagnostic process. Variances of movement patterns at different levels (endpoint, arm-configuration and muscle activity) have been analyzed through both clinical measurements and simulation methods in the last few years [10-12]. The results of these studies show that variability on different movement levels indicates important features of specific arm movements thus can be used for clinical diagnostic purposes.

In this study we investigated the influence of hemiparetic stroke on constrained tracking movement patterns. During a specific visuo-motor task the subjects had to follow a target moving in a plane (2D) with constant speed and path, while the spatial (3D) kinematics of the arm was recorded with an ultrasound-based movement analyzer system. The goal of this study is to evaluate whether a constrained movement task examined with an extended measurement apparatus can produce useful information assisting the diagnostic process of hemiplegic patients.

II. MATERIALS AND METHODS

A. Subjects

Nine healthy right handed subjects and nine right-handed patients with hemiparetic stroke affecting the right (dominant) upper limb participated in these measurements after giving informed consent. None of them presented any visual deficit. The measurements were performed at the National Institute for Medical Rehabilitation in Budapest (NIMR). The procedures of the measurement were approved by the Ethics Committee of the NIMR.

B. Apparatus and procedure

The subjects sat on a chair in front of a computer desk with monitor stand. The subject's position was adjusted so that she or he was able to reach any part of the tablet's active surface without totally stretching the arm or bending the trunk. The subjects were instructed not to bend their trunks during the movement. The measurement setup is shown in Fig. 1.

In each trial the subject saw a moving disk on the screen. The instruction was to follow the moving target with the mouse pointer as accurately as possible. The mouse pointer could be moved with a specific mouse device moving on the digitizer tablet.

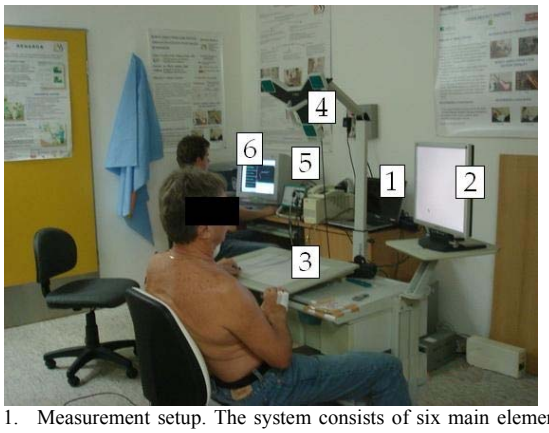


Figure 1. Measurement setup. The system consists of six main elements. 1. Measurement control PC, 2. Stimulus monitor, 3. Digitizer tablet, 4. Movement analyzer ultrasound microphone, 5. Synchronizing hardware element, 6. Movement analyzer control PC

During the experiment two paths (circle with diameter 23cm and square with edge length 23 cm) and two speeds (normal and fast) were applied for the moving target. The normal speed was determined by a speed calibration method for each subject. Each measurement consisted of four trial sets according to the possible combinations of conditions. Each trial set consisted of ten trials with fixed conditions. No learning phase was allowed prior to the measurement. The effect of fatigue was not investigated.

C. Instrumentation

The presentation of the described visual stimulus, the related speed calibration method and handling of the digitizer tablet were implemented in MATLAB environment on a laptop computer. The position of the mouse pointer was recorded with the digitizer tablet. The task screen and the tablet were connected to the laptop. Arm movements were recorded by a ZEBRIS CMS 70P ultrasonic movement analyzer system equipped with six ultrasound-emitting markers and four bipolar surface electromyogram (EMG) electrodes. The marker and electrode positions are listed in Table 1. Muscle activities of four main arm muscles (deltoid anterior, deltoid posterior, biceps and triceps) were recorded simultaneously with 3D marker coordinates. The synchronized operation of the described system was assured by a self-developed hardware element.

TABLE **1**. ULTRASOUND MARKER POSITIONS AND EMG RECORDING LOCATIONS

Channel number	Marker position	EMG position
1	Proximal head of the clavicle	Deltoid anterior
2	Shoulder	Deltoid posterior
3	Elbow	Biceps
4	Wrist1 (ulna distal head)	Triceps
5	Wrist2 (radius distal head)	---
6	Endpoint (tip of index finger)	---

III. DATA ANALYSIS

A. Kinematic data

1) Seven-degrees-of-freedom arm model

The recorded position data were sorted into four groups according to the different movement conditions. After temporal normalization (100 frames) seven joint angles were computed at each time percent for the whole arm (a similar angle definition method can be found in [13]). The defined joint angles are listed in Table 2. The instantaneous arm-configuration was defined by these seven joint angles.

TABLE 2. DEFINED JOINT ANGLES

Joint	Specific angle
Shoulder	Shoulder frontal
	Shoulder sagittal
	Upper arm rotation
Elbow	Elbow flexion
	Lower arm rotation
Wrist	Wrist elevation
	Wrist azimuth

For the calculation of the individual angles an iterative rotation algorithm was used incorporating Rodrigues' general rotation formula:

$$\mathbf{v}_{rot} = \mathbf{v}\cos(\theta) + (\mathbf{k}\times\mathbf{v})\sin(\theta) + \mathbf{k}(\mathbf{k}\cdot\mathbf{v})(1-\cos(\theta)). \quad (1)$$

Using this method a given vector in the three-dimensional space ($\mathbf{v}\in R^3$) can be rotated about any arbitrary rotation axis of unit length ($\mathbf{k}\in R^3$) by a given angle of rotation (θ). During the calculation the joint angles are computed in proximal to distal order. In the first step of the algorithm shoulder frontal angle is calculated as the inverse cosine function of the dot product of a unit reference vector and the normalized upper arm vector. Following this, the upper arm vector is rotated using (1) by a negative value of the computed shoulder frontal angle about the angle-specific rotation axis (which is the cross product of the reference vector and the normalized upper arm vector). This sets the arm to a virtual position in which the shoulder frontal angle is zero, thus it is a new reference position for the next angle calculation using the described dot product method. This step is repeated iteratively for every defined joint angle in the distal way. In each step an angle-specific reference vector is used for angle calculation, and the corresponding rotation axis for the negative rotation is applied.

This algorithm performs the computation of seven joint angles that are minimally required to determine the exact position of the human arm in the three-dimensional space. Any further kinematic computation can be performed on the basis of this seven-degrees-of-freedom representation.

2) Kinematic stability

When a task is executed repeatedly, variability on different movement levels is an important descriptor of movement stability (i.e. less variance across trials means more stable

movement execution). The total variance of repeatedly executed movement patterns can be computed as:

$$V(t) = \frac{\sum_{k=1}^{N} \left(\left\| \mathbf{x}_k(t) - \bar{\mathbf{x}}(t) \right\|_2 \right)^2}{N-1}, \qquad (2)$$

where $V(t)$ denotes variance with respect to normalized time, N is the number of trials, $\mathbf{x}_k(t)$ denotes the actual movement parameter (i.e. endpoint coordinates ($\mathbf{P}_k(t) \in R^3$) or arm-configuration vector ($\mathbf{AC}_k(t) \in R^7$)), and $\bar{\mathbf{x}}(t)$ contains the average value across trials. $\mathbf{x}(t)$ and $\bar{\mathbf{x}}(t)$ are vectors in both cases, so the L^2 norm is used to assure that the equation applies the Euclidean distance (scalar) of the two quantities.

3) Structure of kinematic variance – the uncontrolled manifold approach

Total joint angle variance given by (2) can further be decomposed into two vectors of orthogonal subspaces defined by the selected task variable [10]. In our case, endpoint position planned to be chosen as task variable. This variable defines an uncontrolled manifold (UCM), and a subspace orthogonal to the UCM (ORT) in the corres-ponding joint space. The method calculates the projection of the total variance into these subspaces using the null space of the Jacobian matrix in the mean arm-configuration, and the orthogonal subspace to this null space.

Comparing these projected variances can give us deeper understanding of a movement disorder and the effectiveness of the applied rehabilitation techniques on kinematic level.

B. EMG data

1) Amplitude estimation by windowed RMS smoothing

The recorded EMG data were sorted into four groups according to the different movement conditions just like in the case of kinematic data. For each trial EMG processing consisted of three stages:

1) The data was filtered with a 4th order Butterworth band-pass filter (50-400 Hz)

2) The filtered signal was smoothed with the following windowed RMS algorithm:

$$RMS(t) = \sqrt{\frac{\sum_{k=t-a}^{t+a} EMG(k)^2}{2a+1}}, \qquad (3)$$

where $RMS(t)$ is the root mean square value of the interval $[t-a, t+a]$, $EMG(k)$ is the filtered electromyogram at time instant k. The window size is $(2a+1)$.

3) Temporal normalization (100 frames) of the filtered and smoothed signal was applied.

With this method, a good estimation of the EMG amplitude can be obtained with respect to time, thus the control of the specific movement can be observed at muscle activation level.

2) Spectral analysis – a wavelet-based approach

While the amplitude estimation method can give a good insight into timing of muscles and force-EMG relationship, it has nothing to show about the spectral properties of the recording. The frequency content of an EMG signal can be analyzed using the Fourier transform. The main drawback of the Fourier transform is that it requires recording of the EMG signal over a substantial period of time, thus the temporal aspect of the signal disappears. Shortening the time period leads to the Short-time Fourier transform with its disadvantages that are not discussed here.

There are other approaches to resolve events in the EMG signal by time-frequency analysis. One promising method applies a nonlinearly scaled filter-bank defined by specific wavelets to extract the power of the EMG in discrete frequency bands but continuously over time [14]. Using this method, specific event-related features of the muscle activity can be observed without modifying the original signal's power properties. Based on this concept a specific method was developed to analyze the fatigue effect during cycling movements [15]. Principal component analysis (PCA) was applied to define a time-intensity pattern-space in which the effect of mild fatigue can be numerically expressed. Another important application of the referred wavelet analysis is the removal of the electrocardiogram (ECG) signal from the surface EMG recording [16]. This technique involves independent component analysis (ICA) to separate the ECG signal from the EMG.

Applying these methods to our measured data will give us deeper insight into motor control processes in the examined movement patterns. In addition, these techniques can be used for tracking of motor unit recruitment in patients during the rehabilitation period.

IV. Results and Discussion

In this work a measurement system capable of recording important features of the human upper extremity during constrained tracking arm movements was constructed. Based on this system a measurement protocol was developed to examine the described movement patterns of healthy control subjects and post-stroke hemiplegic patients.

During data analysis a model with seven degrees of freedom was developed to describe the spatial kinematics of the measured arm. Based on this model, endpoint and arm-configuration variances of both control and patient groups were calculated. A representative example of the difference between the two groups in variability is shown in Fig. 2. (only one subject is considered from each group).

Statistical analysis was performed on kinematic variances. We computed a three-level repeated measure ANOVA on mean endpoint and arm-configuration variances with movement path shape, movement speed and subject group being the repeated factors. Significant effects were found between control and patient groups. These results are being prepared for further publication.

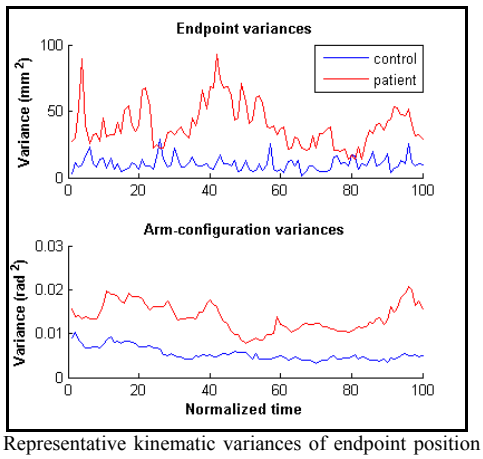


Figure 2. Representative kinematic variances of endpoint position and arm-configuration with respect to time of a control and a patient subject. The patient has larger kinematic variability in both movement parameters than the control subject, showing that computed variance can characterize the actual movement control state of the subject.

Other methods, like the uncontrolled manifold concept has been employed to analyze the kinematics of constrained arm movements in more detail. The UCM technique can reveal the hidden structure of kinematic variance enabling us to have a closer look on the stability of movement control [17].

Concerning EMG analysis, two methods were presented to show the potential of biosignal processing. Using the amplitude estimation technique we are able to analyze movement control processes at muscle activation level, while the wavelet-based approach can give us deeper insight into event-related muscle control and validation of rehabilitation processes. Furthermore, it can provide a new processing method in myoelectric prosthesis control.

V. Further Plans

Applying the measured kinematic data we are planning to perform the UCM analysis for both control and patient data to show the structural differences between healthy and damaged movement control systems.

Besides describing the current state of motor control stability we would like to monitor the effectiveness of currently used movement rehabilitation techniques.

In addition, our goal is to analyze the recorded EMG data with the described processing methods. We would like to use the wavelet-based decomposition to discover the spectral content of muscles during specific constrained arm movements. We are planning to use these results to reveal the relation between the spatial movement of the arm and muscle activities. This would be very beneficial for EMG based methods in neuroprosthesis control.

Acknowledgment

We are grateful to Dr. Gábor Fazekas for the support of this study from clinical aspects and to Györgyi Stefanik for her help during the measurements. We express our thanks to Attila Tihanyi for his help in engineering solutions and to Dr. József Takács for his help in measurement protocol design.

References

[1] J.P. Dewald and R.F. Beer, “Abnormal joint torque patterns in the paretic upper limb of subjects with hemiparesis.,” *Muscle & nerve*, vol. 24, Feb. 2001, pp. 273-83.

[2] M.C. Cirstea and M.F. Levin, “Compensatory strategies for reaching in stroke.,” *Brain : a journal of neurology*, vol. 123 (Pt 5, May. 2000, pp. 940-53.

[3] M.D. Ellis, T. Sukal, T. Demott, and J.P. A, “Augmenting Clinical Evaluation of Hemiparetic Arm Movement With a Laboratory-Based Quantitative Measurement of Kinematick as a Function of Limb Loading,” *Biomedical Engineering*, vol. 22, 2010, pp. 321-329.

[4] S. Vuillermot, A. Pescatore, L. Holper, D.C. Kiper, and K. Eng, “An extended drawing test for the assessment of arm and hand function with a performance invariant for healthy subjects.,” *Journal of neuroscience methods*, vol. 177, Mar. 2009, pp. 452-60.

[5] L. Dipietro, H.I. Krebs, S.E. Fasoli, B.T. Volpe, and N. Hogan, “Submovement changes characterize generalization of motor recovery after stroke.,” *Cortex; a journal devoted to the study of the nervous system and behavior*, vol. 45, Mar. 2009, pp. 318-24.

[6] T. Krabben, B.I. Molier, A. Houwink, J.S. Rietman, J.H. Buurke, and G.B. Prange, “Circle drawing as evaluative movement task in stroke rehabilitation: an explorative study.,” *Journal of neuroengineering and rehabilitation*, vol. 8, Jan. 2011, p. 15.

[7] A.R. Fugl-Meyer, L. Jääskö, I. Leyman, S. Olsson, and S. Steglind, “The post-stroke hemiplegic patient. 1. a method for evaluation of physical performance.,” *Scandinavian journal of rehabilitation medicine*, vol. 7, Jan. 1975, pp. 13-31.

[8] J. Laczkó, ML. Latash, “Components ofthe End-Effector Jerk during Voluntary Arm Movements,” *Journal of Applied Biomechanics*, 2000, pp. 14-26.

[9] Z. Kereszténvi, P. Cesari, G. Fazekas, and J. Laczkó, “The relation of hand and arm configuration variances while tracking geometric figures in Parkinson’s disease: aspects for rehabilitation.,” *International journal of rehabilitation research.*, vol. 32, Mar. 2009, pp. 53-63.

[10] D. Domkin, J. Laczko, M. Djupsjöbacka, S. Jaric, and M.L. Latash, “Joint angle variability in 3D bimanual pointing: uncontrolled manifold analysis.,” *Experimental brain research.*, vol. 163, May. 2005, pp. 44-57.

[11] J. Laczkó, Z. Kereszténvi, “Variances of hand positions and arm configurations during arm movements under external load and without external load,” *Motor control*, 2007, p. 127.

[12] R. Tibold and J. Laczkó, “Three-dimensional model to predict muscle forces and their relation to motor variances in reaching arm movements,” *Journal of Applied Biomechanics*, In Press.

[13] M. Desmurget, C. Prablanc, Y. Rossetti, M. Arzi, Y. Paulignan, C. Urquizar, and J.C. Mignot, “Postural and synergic control for three-dimensional movements of reaching and grasping.,” *Journal of neurophysiology*, vol. 74, Aug. 1995, pp. 905-10.

[14] V. von Tscharner, “Intensity analysis in time-frequency space of surface myoelectric signals by wavelets of specified resolution.,” *Journal of electromyography and kinesiology : official journal of the International Society of Electrophysiological Kinesiology*, vol. 10, Dec. 2000, pp. 433-45.

[15] V. von Tscharner, “Time-frequency and principal-component methods for the analysis of EMGs recorded during a mildly fatiguing exercise on a cycle ergometer.,” *Journal of electromyography and kinesiology : official journal of the International Society of Electrophysiological Kinesiology*, vol. 12, Dec. 2002, pp. 479-92.

[16] V. von Tscharner, B. Eskofier, and P. Federolf, “Removal of the electrocardiogram signal from surface EMG recordings using non-linearly scaled wavelets.,” *Journal of electromyography and kinesiology : official journal of the International Society of Electrophysiological Kinesiology*, Apr. 2011.

[17] J. Laczkó, S. Jaric, D. Domkin, H. Johansson, and M. Latash, “Stabilization of Kinematik Variables in the Control of Bimanual Pointing Movements,” *Intnl. Joint Conference on Neural Networks. Washington DC.*, 2001, pp. 1256-1260.

Characterizing problems and hardware for massively parallel implementation

István Zoltán Reguly
(Supervisors: Dr. András Oláh and Dr. Tamás Roska)
reguly.istvan@itk.ppke.hu

Abstract— **As Moore's law in single processor cores nearly came to a halt, and parallel architectures are present in all modern computers (namely the GPUs) and many embedded systems (FPGAs), there is a growing demand to map problems to these parallel hardware. The main purpose of my research is to develop a methodology that can assist parallelization by describing the problem space and by defining a "walk" of this space obeying the hardware limitations. This process is illustrated by three problems: CNN simulation on FPGA, global optimization with genetic algorithms on GPU and solution of PDEs with the finite-difference time-domain method on GPU.**

Keywords—gpgpu, fpga, parallel programming, cnn, genetic algorithms, partial differential equations

I. INTRODUCTION

For decades the main goal of microprocessor vendors was to produce higher and higher clock rates. However at around 3 GHz, the power requirements became unreasonably high, and performance had reached a limit. The long-standing practice of serial programming had become unsustainable: single thread performance no longer increases exponentially with time. GPGPU (general purpose computing on graphics processing units) and FPGA (field-programmable gate array) based software and hardware co-design has become popular means to assist general-purpose processors in performing complex, data and/or computation intensive tasks. Today most computers ship with at least one such accelerator (usually a GPU), and already the CPU and GPU have appeared on the same silicone [1].

Different applications and algorithms however place unique and distinct demands on computing resources, so implementations that work well with one accelerator will not necessarily map to another. Programming methodologies range from direct hardware designs for FPGAs [2], through assembly and domain specific languages, to high level languages supported by GPUs [3].

There is little research on how to use these special purpose processors as accelerators for general-purpose computations, how accelerators and tasks map. Also a challenge facing developers is to understand application behavior on different accelerators, to learn how to map, partition and execute parallel algorithms on these new hardware. As a first step to understand these issues, this work studies three different applications on different hardware: genetic algorithms, CNN simulation and solving a PDE with the FDTD method.

The structure of the paper if the following: in section 2 the problem is described in greater detail with additional pointers to related work in this field. In section 3 the approach to parallelizing problems is described. In section 4 three case studies are presented. Section 5 summarizes this works and points out possible future directions.

II. PROBLEM FORMULATION AND RELATED WORK

Most work was focused on specific problems and applications; systematic approaches like autoparallelization have yielded few results, and usually focus on specific pieces of code that allow trivial parallelization [4]. Some propose a structured approach to parallelizing individual problems: using structural, computational and implementation patterns to find and exploit inherent parallelism [5].

Modeling of algorithms on parallel architectures has also proven difficult; most performance models [6,7] require low-level source code to make estimates, which obviously requires implementing the problem first. These models offer good quantitative estimates but require much effort.

The availability of cheap, general purpose programmable Graphical Processing Units has made an impact on high performance computing, modern supercomputers utilize GPUs [8] because of their good performance to price/power/effort ratio. GPUs require power in the range from Watts to hundreds of Watts to be effective. On the other hand, very low power FPGAs are available, which are very energy efficient, enabling their use in small scale, mobile and embedded applications. Another problem is the price - few hundred dollar, high power GPU and a many thousand dollar, low power FPGA – not a realistic comparison. The other differentiating factor is the programming style, which usually requires very different algorithms on these two platforms to solve a given problem.

Because of these, current attempts at comparing the two platforms [9, 10, 11] seem unrealistic and unfair.

For these reasons I think it would be important to classify problems of different kind and scale by observing the requirements of applications to decide whether it makes sense to solve the problem on a given platform. For example in case of a large-scale molecular dynamics simulation the primary concerns are computational speed and cost, so the GPU is a clear winner, but in case of embedded signal processing the power constraints disqualify the use of GPUs.

III. THE PROPOSED APPROACH

The goal of this research if to create a methodology, that given a problem one could analyze and decompose it, then map it onto a mathematical structure that exposes possible ways of parallelization. After this, specific methods of implementation can be explored, given the constraints of the hardware, and finally a decision can be made which one to use. To do this, different metrics are required that describe the different aspects of the hardware and the problem; like speed, power, area, bandwidth, computational and data complexity etc.

How can such a problem be attacked? The approach has two sides: the hardware has to be abstracted, the problem more specific. In most cases this can be done using problem/implementation classes or patterns. An abstraction of the hardware is its computational and memory hierarchy, and a specific algorithm can be described with patterns of computation, and flow of data.

The first classification possibility is the constraints of the application, like expected speed performance or power limitations. In many cases this can completely determine the choice of hardware class, but sometimes parallelization is not worth the effort (e.g. High power gpu vs. cpu or low power fpga vs. microcontroller).

The problem can usually be decomposed into overall structural patterns that describe the high level computational and communicational steps that later will be parallelized. Such patterns are pipe and filter, agent and repository, iterative refinement, map reduce, process control etc. The high level computational steps can be classified using the 13 motifs (or dwarves) [5], these classes describe many computational and communication structures that offer different possibilities of parallelization. These first two steps of classifications are hardware independent and ideally they allow for different implementations.

The two lowest levels of classification are execution strategy and implementation strategy. Different implementation strategies are allowed by different hardware, we can identify different data and program structures. More elaborate ones like fork-join and shared hash table for CPUs, SPMD and distributed array on GPUs, some cases of strict-data-par and memory level parallelism on FPGAs. The lowest level is execution strategy, which is usually a hardware's own, like SIMD, thread pools, or digital circuits.

The connections between the higher and lower levels, are the parallel algorithm strategies that describe a way of exploiting the parallelism within the upper level computational patterns, and take into account lower level patterns. On each level of classification the program can be described by a graph of computations and communications, that are restricted at the lowest level by hardware processing and memory hierarchies. Possible transformations of these graphs can be entry points into lower level classifications.

The concept is to use a graph to describe all calculations in a given (sub-)task, marking dependencies. Memory accesses and dependencies are embedded within this graph. During an actual implementation a parallel walk is defined on this graph, which depends on both the parallel algorithm strategy class and the hardware: graph partitions and other properties may require different strategies to utilize or circumvent hardware limitations (like synchronization, communication restrictions, bandwidth).

This model can support different metrics like computational time, computational and memory access efficiency. Based on the properties of the graph, several complexity metrics can be introduced that describe the computational and data dependencies: these could point out how "long" these dependencies last, and how "far" they reach, they are closely related to the implementation requirements like communication and synchronization.

The structured decomposition of most problems into structural and computational patterns, their description using computational and data flow graphs can yield qualitative models and estimates that could both help to determine the best fitting platform and to guide implementation.

IV. CASE STUDIES

Three problems have been selected for further examination on two platforms: on an FPGA and a GPU. First I describe the hardware and programming concepts of both platforms, then the implementation of the algorithms.

A. Graphical Processing Units (GPUs)

GPUs are inexpensive, commodity devices, that were developed for acceleration of video games. The main advantage of a GPU is a very high memory bandwidth, up to 200 GB/s on latest Fermi architectures [12]. On the GPU chip, there are hundreds of programmable cores, with possibly thousands of active threads that are executed in a single program multiple data (SPMD) fashion. GPUs are flexible and easy to program using high level languages and APIs that abstract away hardware details, like Nvidia's own C extension, CUDA [3]. In CUDA the GPU is treated as a co-processor that executes data-parallel kernels with thousands of threads. Threads are grouped into *thread blocks*. Threads within a block can share data using fast on-chip shared memory and synchronize using hardware-supported barriers. Each thread block is assigned to a *streaming multiprocessor* (SM). Coordination between thread blocks is only possible through a much slower global memory (two orders of magnitude). Note that the programming model for a CUDA kernel is scalar not vector. The current Tesla architecture combines 32 scalar threads into SIMD groups called *warps*, that are executed in lockstep.

Experiments were made on different GPU hardware: GeForce 320M with 6 SMs, a Tesla S1070 with four Tesla C1060 inside, each equipped with 30 SMs. Each SM has 8 *streaming processors* (SPs), with each group of 8 SPs sharing 16kB of per-block shared memory [13]. The global memory space of the 320M is shared from system memory, so it is much slower than dedicated RAM, and cannot utilize coalesced memory accesses. There is 6GBs of global memory on each Tesla unit, in total 24GBs for the whole rack, with global memory bandwidth of 100 (400) GB/s.

B. Field Programmable Gate Arrays (FPGAs)

Compared to the fixed hardware architecture of the GPU, FPGAs are essentially high density arrays of uncommitted

logic and are very flexible in that developers have direct control over hardware infrastructure and trade-off resources and performance by selecting the appropriate level of parallelism to implement an algorithm. In the FPGA paradigm, the hardware structure is used to approximate a custom chip. This eliminates the inefficiencies caused by the traditional Neumann execution model and can achieve vastly improved performance and power efficiency. Though vendors provide IP cores that offer the most common processing functions, programming in VHDL and creating the entire design from scratch is a costly and labor intensive task.

VHDL is one of the most widely used hardware description languages. It supports the description of circuits at a range of abstraction levels varying from gate level netlists up to purely algorithmic behavior [14]. Very efficient hardware can be developed in VHDL but it requires a great deal of programming effort. FPGAs consist of hundreds of thousands of programmable logic blocks and programmable interconnects that can be used to create custom logic functions, and many GPFA products also include some hardwired functionality for common functions. Fast programmable FPGAs cost orders of magnitude more than GPUs, however their power consumption is much lower. Fixed logic circuits can be manufactured for a fraction of the price of a programmable board.

C. Genetic algorithm on a GPU

Genetic Algorithms (GAs) were introduced in the 1960s [15], since then it has become a popular tool for researchers: they are effective in handling a wide range of difficult real-world problems, mainly global optimization [16]. In general Gas use selection, mutation and crossover to generate new points in the search space. The set of points forms the population of the algorithm, where the initial population is usually generated at random. In each iteration the fitness of each individual is evaluated, and the best are selected to form the next population, introducing random mutations and crossover in the meanwhile. The iteration goes on until a stopping condition is reached.

Genetic algorithms offer straightforward parallelization, as the evaluation of individuals is independent of each other. Several studies have shown it is possible to offload part or all of the algorithm to the GPU [17,18]. The structural constraints of the GPU do not enable global synchronization on the GPU that would be necessary to implement single population iterations, the following methods of coordination have been explored:

- Offload the evaluation of the fitness function to the GPU, the rest is done on the CPU. Involves heavy use of CPU-GPU bandwidth.
- The whole process is done on the GPU, this requires global synchronization every iteration, which implies starting a new kernel every iteration.
- The population is divided into subpopulations (islands), communication between the islands occur only every few iterations (migration). Theory of distributed GAs is described in [19].

The implementation optimized a distributed routing table for Wireless Sensor Networks (WSNs). We used a two-tiered GA approach: one for individual paths and one for routing tables with all paths. The network map is uploaded to a fast *constant memory*, and in the first stage each thread block is assigned one routing table to optimize, with individual fitness functions for each thread based on the length of the path. Afterwards the fitness of the table is calculated per thread block based on the total length of all paths and the load of individual nodes (to avoid bottlenecks). The second stage is a GA on routing tables, when individual paths are handled as atomic parts of an individual. The algorithm is depicted on Fig 1.

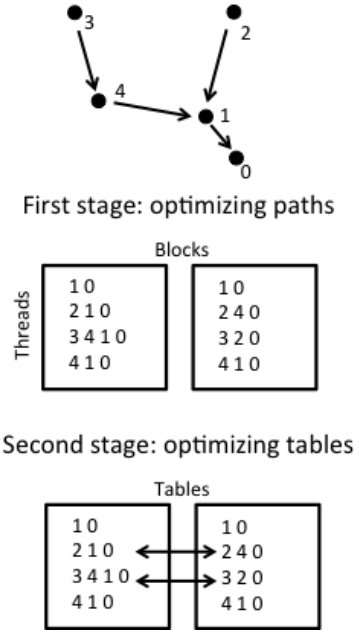


Figure 1. Optimizing routing tables in a WSN

D. Digital CNN simulation on an FPGA

The CNN-UM [20] is a powerful analogic machine, providing massive parallelism with its cellular structure. The digital simulation is done by converting the continuous time wave equation to discrete time.

$$\dot{x_{ij}} = -x_{ij} + \sum_{C(k,l)\in S(i,j)} A(i,j;k,l)y_{kl} + \sum_{C(k,l)\in S(i,j)} B(i,j;k,l)u_{kl} + z_{ij}$$

This results in a structured grid computational pattern, where each grid point is a CNN cell, connected to its eight neighbors in case of radius 1. As the target platform was an FPGA, the chosen implementation pattern was the pipeline.

We used a Spartan 3 FPGA for the implementation, clocked at 50 MHz. A 32*32 BRAM was created to store the state of the CNN, data was uploaded via USB, and displayed through a VGA connection. Because the update of an individual cell requires data from its neighborhood, the use of a buffer was

necessary which contained the data from neighboring cells before they were updated. As the algorithm processes each cell row-by-row, a shift register of size 32*2+1 was used. The pipeline is seven stages deep: thus seven cells were being processed at the same time.

E. FDTD solution of PDEs on a GPU

Finite Difference Time Domain methods have been widely used for the solution of PDEs, several works demonstrate the solution of e.g. Navier-Stokes equations [21]. This method is also a case of the structured grid implementation pattern. The algorithm discretizes the equation in both space and time domain, and updates individual cells that depend on their neighbors. The test problem was the simple heat equation:

$$u_t' = u_{xx}' + u_{yy}'$$

The implementation separates the space domain into 64*63 grid point sized chunks and assigns them to thread blocks, that store them in local shared memory for fast access. This kind of separation however requires the exchange of ghost cells between chunks, and this makes global synchronization necessary. Global synchronization by running kernels back-to-back introduces a significant control overhead, and the contents of the shared memory are also lost between kernels.

A novel synchronization method has been developed making use of atomic memory accesses: by using only two counters in global memory and increasing them atomically make any number of global synchronizations possible. In my experience when the work between the thread blocks is well distributed, the overhead is only 1-2 memory accesses (nanoseconds) per synchronization, while starting kernels requires microseconds. In this case the shared memory persists, so whole chunks need not be written out to global memory, only ghost cells, which reduces memory bandwidth use by 93%. Experiments show 5000% speed increase over Matlab on a single Tesla GPU, and 60% increase over back-to-back kernel runs.

V. Conclusion and Future work

This paper shows the necessity of parallel mapping of problems, and develops a concept how individual problems should be described to enable possible parallelization for different hardware.

Three case studies are presented, that show how different problems can be implemented in very different ways onto different hardware: Genetic Algorithms, CNN simulation and FDTD methods, on GPUs and FPGAs. A novel global synchronization method is developed for the GPU, and tested in the FDTD algorithm.

In the future I would like to focus on structured and unstructured grid algorithms for PDEs, the use of different algorithms like the Finite Element Method, the Pseudo-Spectral Method and others, as they provide different approaches to solve the same problem. The question of parallelizability of these algorithms for different hardware would be an interesting task.

References

[1] AMD Fusion. http://fusion.amd.com

[2] P. J. Ashenden, “The VHDL Cookbook”, 1990

[3] NVIDIA CUDA Compute Unified Device Architecture Reference Manual, NVIDIA Corporation, 2008.

[4] H. Kasim, V. March, R. Zhang, S. See, “Survey on Parallel Programming Model” in Network and Parallel Computing, Springer / Heidelberg p. 266, 2008

[5] K. Keutzer, B. Massingill, T. Mattson, B. Sanders, “A Design Pattern Language for Engineering (Parallel) Software: Merging the PLPP and OPL projects”, 2nd Annual Conference on Parallel Programming Patterns (ParaPLoP'10), Carefree, AZ, March 30, 2010

[6] K. Kothapalli, R. Mukherjee, M. S. Rehman, "A performance prediction model for the CUDA GPGPU platform", In Proceedings of 2009 International Conference on High Performance Computing (HiPC) (December 2009), pp. 463-472. doi:10.1109/HIPC.2009.5433179

[7] S. Hong, H. Kim, „An analytical model for a GPU architecture with memory-level and thread-level parallelism awareness“ In the Proceedings of SIGARCH Comput. Archit. News, June 2009

[8] Top500 Supercomputers http://www.top500.org

[9] Y. Zhang, Y. Shalabi, R. Jain, K. Nagar, J. Bakos, "FPGA vs. GPU for sparse matrix vector multiply", International Conference on Field-Programmable Technology, 2009.

[10] Asano, S., Maruyama, T., & Yamaguchi, Y. (2009). Performance comparison of FPGA, GPU and CPU in image processing. In Field Programmable Logic and Applications, 2009. International Conference on (p. 126-131). doi:10.1109/FPL.2009.5272532

[11] Papadonikolakis, M., Bouganis, C., & Constantinides, G. (2009). Performance comparison of GPU and FPGA architectures for the SVM training problem. In International Conference on Field-Programmable Technology, 2009. (p. 388-391). doi:10.1109/FPT.2009.5377653

[12] Nvidia, „Fermi Compute Architecture Whitepaper“, http://www.nvidia.com/content/PDF/fermi_white_papers/NVIDIA_Fermi_Compute_Architecture_Whitepaper.pdf

[13] E. Lindholm, J. Nickolls, S. Oberman, and J. Montrym. NVIDIA Tesla: A unified graphics and computing architecture. IEEE Micro, 28(2):39–55, 2008.

[14] N. Calazans, E. Moreno, F. Hessel, V. Rosa, F. Moraes, and E. Carara. From VHDL register transfer level to SystemC transaction level modeling: A comparative case study. In Proceedings of the 16th Symposium on Integrated Circuits and Systems Design, page 355, 2003.

[15] Holland, J.H.: Adaptation in Natural and Artificial Systems. University of Michigan Press (1975)

[16] Goldberg, D.E.: Genetic Algorithms in Search, Optimization, and Machine Learning. Addison-Wesley, Reading (1989)

[17] Fok, K.L., Wong, T.T., Wong, M.L.: Evolutionary Computing on Consumer-Level Graphics Hardware. IEEE Intelligent Systems 22(2), 69–78 (2007)

[18] Wong, M.L., Wong, T.T., Fok, K.L.: Parallel Evolutionary Algorithms on Graphics Processing Unit. In: Proceedings of IEEE Congress on Evolutionary Computation 2005 (CEC 2005), pp. 2286–2293 (2005)

[19] Eric Cantú-Paz. Efficient and Accurate Parallel Genetic Algorithms, volume 1 of Genetic Algorithms and Evolutionary Computation. Kluwer Academic Publishers, 1st edition, 2000.

[20] Chua LO, Roska T, Venetianer P, The CNN is universal as the turing machine, IEEE Transactions on Circuits and Systems I - Fundamental Theory and Applications 40:(4) pp. 289-291. (1993)

[21] Julien C. Thibault and Inanc Senocak. CUDA implementation of a Navier-Stokes solver on Multi-GPU desktop platforms for incompressible flows. Orlando, FL, January 2009. 47th AIAA Aerospace Sciences Meeting.

Packet scheduling algorithm for WSN

Kálmán Tornai
(Supervisor: Dr. János Levendovszky)
tornai.kalman@itk.ppke.hu

***Abstract*—In this paper, we develop optimal scheduling mechanisms for packet forwarding in Wireless Sensor Network (WSN), where clusterheads are gathering information with a predefined Quality of Service (QoS). The objective is to ensure balanced energy consumption and to minimize the probability of packet loss, subject to time constraints (i.e. different nodes must send all their packets within a given time interval). Novel solutions of scheduling are developed by combinatorial optimization, and by quadratic programming methods. In our approach, the scheduling of packet forwarding is broken down to a discrete quadratic optimization problem and the optimum is sought by a Hopfield Network (HNN) yielding a solution in polynomial time. The scheduling provided by the Hopfield Network indeed guarantees uniform packet loss probabilities for all the nodes and saves the energy of the clusterheads. In this way, the longevity of the network can also be increased.**

I. Introduction

Data gathering from a set of sensor nodes to a Base Station (BS) by using a cluster-based routing topology is commonly used in wireless sensor networks (WSNs) [1], [2]. In this kind of networks tiny sensor nodes communicate in short distances and collaboratively work to fulfill the application specific objectives of WSN. Many of the envisioned applications involve the collection of bursty data traffic generated by events which are be delivered to the BS as quickly and as reliably as possible in order to recognize emergency situation. In these applications packet delay and packet loss probability are of crucial importance [3], [4].

In this paper an optimal packet scheduling scheme is proposed, which minimizes the associated packet loss probability under time constraints. Scheduling has been intensively researched in the telecommunication literature [5], [6], [7], however the main focus was on buffered architectures. In this paper, we will present the problem as a matrix optimization. Since clusterhead based routing is a commonly used solution in WSN (e.g. LEACH protocol [8] or other hierarchical solutions proposed in [9], [10]), we assume that each node can only send one packet to a selected clusterhead at each time instant. The overall number of packets node i wants to send to the clusterhead is denoted by X_i . All of the packets must be sent within a time interval $K_i : K_i > X_i$. The scheduling of packet transmissions by node i can be expressed by a binary vector $\mathbf{c}^{(i)}$ of length K_i with weight X_i, where component $c_l^{(i)} \in \{0, 1\}$ indicates whether packet is transmitted or not to the clusterhead at time instances l. Since the clusterhead has a finite energy, thus at each time instant it selects randomly at most Cap packets to re-transmit them to the BS. In this paper, we optimize the scheduling vectors in order to balance the packet loss probabilities the nodes suffer from if the aggregated load at a given time instant exceeds Cap. Composing a transmission matrix $\mathbf{C}$ of these vectors, one may seek the optimal matrix, which enforces uniform packet loss probability for each user under the time constraints.
To find the optimal matrix, we present a novel approach based on quadratic optimization which can then be tackled by the Hopfield Neural Network (HNN).

II. The model

Let us assume that there are J number of nodes transmitting packets to the cluster head. The capacity of the cluster head is denoted by Cap. The amount of packets to be sent by node j is denoted by X_j, while the time constraint in which the transmission is to be finished is denoted by K_j. The time is measured in discrete units thus $K_j, j = 1, \ldots, J$ are assumed to be integers. The scheduling of node j is represented by a binary vector $\mathbf{c}(j) \in \{0, 1\}^{K_j}$ where if $c_l(j) = 1$ then a packet is sent to the cluster head at time instant l. The scheduling matrix $\mathbf{C}$ can be constructed from vectors $\mathbf{c}(j), j = 1, \ldots, J$ which form the row vectors of $\mathbf{C}$ and the number of columns is taken as $L = \arg\max_j K_j$. For example in the case of $J = 3$, $X_1 = 2$, $X_2 = 5$, $X_3 = 3$, $K_1 = 8$, $K_2 = 10$ and $K_3 = 4$ one specific scheduling matrix looks as follows:

$$\mathbf{C} = \begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}.$$

The aggregated number of incoming packets in the cluster head at time instant l is given as $\sum_{j=1}^{J} c_l(j)$ being the weight of the lth column vector in matrix $\mathbf{C}$. The cost of reception (which needs the power of clusterheads) is proportional to the number of received packets at time instant l expressed as $a_l := \sum_{j=1}^{J} c_l(j)$. Let us assume that the number of arriving packets at the cluster head at time instant l exceeds the capacity of the cluster head. Formally:

$$\sum_{j=1}^{J} C_{jl} > Cap. \tag{1}$$

In this situation the cluster head decides in uniform random fashion which packet will be discarded. In this case the probability of packet loss suffered by a given node is

$$P\,(\text{packet loss at node i}) = \frac{\sum_{j=1}^{J} C_{jk} - Cap}{\sum_{j=1}^{J} C_{jk}}, \forall i. \tag{2}$$

There are two different goals when optimizing matrix $\mathbf{C}$.

- In the first case, we seek the optimal matrix $\mathbf{C}_{opt}$ for which the overall cost is minimal. This can be formulated as follows:

$$\mathbf{C}_{opt} : \min_{\mathbf{C}} \sum_{j=1}^{J} \sum_{l=1}^{L} \sum_{i=1}^{J} C_{jl} C_{il}. \quad (3)$$

  The constraint can be expressed as $\sum_{l=1}^{L} C_{jl} = X_j, j = 1, \ldots, J$ and if the last nonzero component of row j is at location M_j, such that $M_j : \sum_{l=1}^{L} C_{jl} = X_j, j = 1, \ldots, J$ then $M_j \leq K_j, j = 1, \ldots, J$. We fill refer to this problem as optimizing the Overall Cost (OC).
- In the second case, we seek an optimal matrix $\mathbf{C}_{opt}$ where the aggregated number of incoming packets are balanced with respect to time. More precisely, we try to achieve that nodes schedule their packet transmission in such a way that that each time instant the culsterheads receive more or less the same number of packets. This objective can be formulated as follows:

$$\mathbf{C}_{opt} : \min_{\mathbf{C}} \sum_{l=1}^{L} \sum_{k=1}^{L} \left(\sum_{j=1}^{J} C_{jl} - \sum_{j=1}^{J} C_{jk} \right)^2. \quad (4)$$

  The constraints are the same as before.
  When we seek the balanced solution by optimizing (4) we provide approximatively the same Quality of Service in terms of packet loss probability to all nodes. We will refer to this problem as Balanced Cost (BC).

III. Solution by quadratic programming and Hopfield Network

The Hopfield Network is a recurrent neural network [11] the dynamics and optimization capabilities of which have been intensively studied [12]. The state transition rule is described as follows: [13]

$$y_i(k+1) = -\text{sgn}\left(\sum_{j=1}^{N} \tilde{W}_{ij} y(k)_j - b_i \right), \; i = \text{mod}_{\; N} k, \quad (5)$$

The convergence to a steady state has been proven [14] by using a quadratic Lyapunov function given as

$$\mathcal{L}(\mathbf{y}) := \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} W_{ij} y_i y_j - \sum_{i=1}^{N} y_i b_i = \frac{1}{2} \mathbf{y}^T \mathbf{W} \mathbf{y} - \mathbf{b}^T \mathbf{y}. \quad (6)$$

Matrix $\tilde{\mathbf{W}}$ has zero diagonal elements. Thus, a class of combinatorial optimization problems, which can be mapped into a quadratic objective function where the optima are sought over binary arrays can be efficiently solved by Hopfield Network.

A. Optimizing the Overall Cost by Hopfield Network

We now map the optimization problem defined by expression (3) into a quadratic objective function which can then be minimized by a Hopfield Network in polynomial time [15], [16], [17]. The constraints are included as additive terms in the goal function. There are three parameters: α, β, δ. They are found heuristically in the course of the optimization.
The endeavor is to map the objective function into (6). In the first step we encode $\mathbf{C}$ row-vise onto vector $\mathbf{y}$ (which is the state of vector of the corresponding Hopfield Network).

$$\mathbf{C} = \begin{pmatrix} C_{11} & C_{12} & \cdots & C_{1L} \\ \vdots & \vdots & \ddots & \vdots \\ C_{J1} & C_{J2} & \cdots & C_{JL} \end{pmatrix} \rightarrow$$

$$\mathbf{c} = (C_{11}, C_{12}, \cdots, C_{1L}, C_{21}, \cdots, C_{2L}, C_{J1}, \cdots, C_{JL})^T.$$

To carry out the mapping, one has to solve the following equations (7) (8) (9):

$$\alpha \sum_{j=1}^{J} \sum_{l=1}^{L} \sum_{i=1}^{J} C_{jl} C_{il} = -\frac{1}{2} \mathbf{y}^T \mathbf{W}_A \mathbf{y} - \mathbf{b}_A^T \mathbf{c}, \quad (7)$$

$$\beta \sum_{j=1}^{J} \left(\sum_{l=1}^{K_j} C_{jl} - X_j \right)^2 = -\frac{1}{2} \mathbf{y}^T \mathbf{W}_B \mathbf{y} - \mathbf{b}_B^T \mathbf{c}, \quad (8)$$

$$\delta \sum_{j=1}^{J} \sum_{l=K_j}^{L} (C_{jl})^2 = -\frac{1}{2} \mathbf{y}^T \mathbf{W}_C \mathbf{y} - \mathbf{b}_C^T \mathbf{c}. \quad (9)$$

The negative sign on the right hand side is due to the minimization. Calculating matrices $\mathbf{W}_A$, $\mathbf{W}_B$, $\mathbf{W}_C$, $\mathbf{b}_A$, $\mathbf{b}_B$ and $\mathbf{b}_C$, respectively, the parameters of the quadratic form which is minimized by the Hopfield Network is as follows:

$$\mathbf{W} = \mathbf{W}_A + \mathbf{W}_B + \mathbf{W}_C \in \mathbb{R}^{JL \times JL},$$

$$\mathbf{b} = \mathbf{b}_A + \mathbf{b}_B + \mathbf{b}_C \in \mathbb{R}^{JL \times 1}.$$

Solving (7) yields the following weight matrix and bias vector:

$$\mathbf{b}_A = \mathbf{0}_{JL \times 1}, \quad (10)$$

$$\mathbf{W}_A = 2\alpha \begin{pmatrix} \mathbf{I}_{L \times L} & \mathbf{I}_{L \times L} & \cdots & \mathbf{I}_{L \times L} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{I}_{L \times L} & \mathbf{I}_{L \times L} & \cdots & \mathbf{I}_{L \times L} \end{pmatrix}. \quad (11)$$

Processing the second equation (8) and taking into account that $\sum_{j=1}^{J} X_j^2 = const$ which does not have any effect on the extremal point, we obtain the following equation:

$$-\frac{1}{2} \mathbf{c}^T \mathbf{W}_B \mathbf{c} - \mathbf{b}_B^T \mathbf{c} = \beta \sum_{j=1}^{J} \left(\sum_{l=1}^{K_j} C_{jl} \right)^2 - 2\beta \sum_{j=1}^{J} \left(\sum_{l=1}^{K_j} C_{jl} \right) X_j$$

The solution is (12) and (13), where $\mathbf{1}_{K_j, K_j}$ denotes a matrix all elements of which are ones and where $\mathbf{b}_{Bj} = X_j \cdot \begin{pmatrix} \mathbf{1}_{1 \times K_j} & \mathbf{0}_{1 \times L - K_j} \end{pmatrix} \in \mathbb{R}^{1 \times L}$:

$$\mathbf{b}_B = 2 \begin{pmatrix} \mathbf{b}_{B1} & \mathbf{b}_{B2} & \cdots & \mathbf{b}_{BJ} \end{pmatrix}. \quad (12)$$

The solution of the third equation (9) is

$$\mathbf{b}_C = \mathbf{0}_{JL \times 1}, \quad (14)$$

$$\mathbf{W}_C = -2\delta \begin{pmatrix} \mathbf{C}_1 & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{C}_J \end{pmatrix}. \quad (15)$$

$$\mathbf{W}_B = -2\beta \left(\begin{array}{cc|cc|c|cc} \mathbf{1}_{K_1\times K_1} & \mathbf{0}_{K_1\times L-K_1} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0}_{L-K_1\times K_1} & \mathbf{0}_{L-K_1\times L-K_1} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \hline \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \hline \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{1}_{K_J\times K_J} & \mathbf{0}_{K_J\times L-K_J} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0}_{L-K_J\times K_J} & \mathbf{0}_{L-K_J\times L-K_J} \end{array} \right) \tag{13}$$

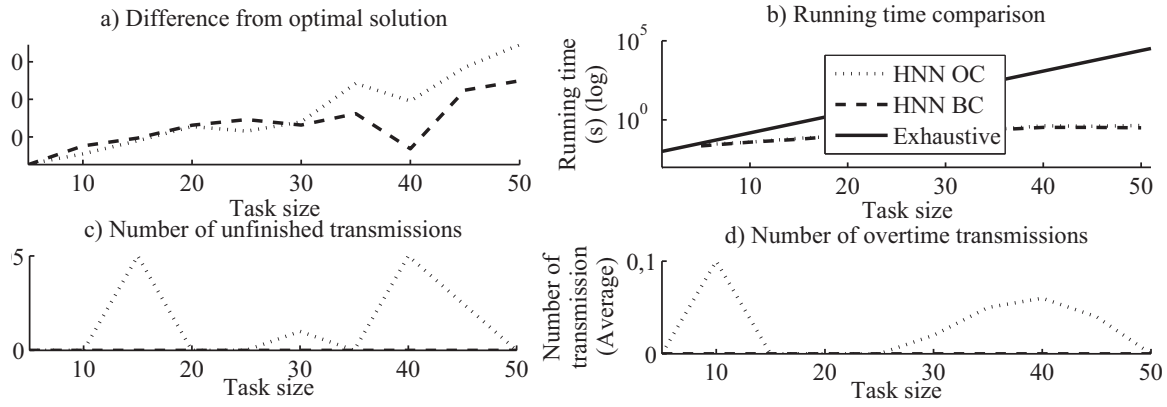


Fig. 1. a) The cost difference between the optimal solution and the solutions achieved by the different methods b) Running time of methods c) The number of incomplete transmissions d) Number of packets transmitted overtime

Here $\mathbf{C}_j = \begin{pmatrix} \mathbf{0}_{K_j\times K_j} & \mathbf{0}_{K_j\times L-K_j} \\ \mathbf{0}_{L-K_j\times K_j} & \mathbf{I}_{L-K_j+1\times L-K_j+1} \end{pmatrix} \in \mathbb{R}^{L\times L}$ which is special $(L\times L)$ "identity" type of matrix where the first K_j diagonal is zero. [18]

B. Optimizing the Balanced Cost by Hopfield Network

Previously the solution of the Overall Cost (3) was introduced. For the case of Balanced Cost (4) the goal function has to be be rewritten. The three parameters are to be found heuristically in the course of the optimization.

The first term of objective function have to be extracted. Performing mathematical transformations we obtain the final form. Since the parts corresponding to (8) and (9) are the same as in the first problem (the constraints are not changed) the following new equation has to be solved:

$$2L\alpha_1 \sum_{l=1}^{L} \left(\sum_{j=1}^{J} c_{jl} \right)^2 = -\frac{1}{2}\mathbf{y}^T\mathbf{W}_{A_1}\mathbf{y} - \mathbf{b}_{A_1}^T\mathbf{c}, \tag{16}$$

$$-2\alpha_2 \left(\sum_{l=1}^{L}\sum_{j=1}^{J} c_{jl} \right)^2 = -\frac{1}{2}\mathbf{y}^T\mathbf{W}_{A_2}\mathbf{y} - \mathbf{b}_{A_2}^T\mathbf{c}. \tag{17}$$

The solution for equation (16) is the following

$$\mathbf{b}_{A1} = \mathbf{0}_{JL\times 1}, \tag{18}$$

$$\mathbf{W}_{A1} = 2\alpha_1 L \begin{pmatrix} \mathbf{I}_{L\times L} & \mathbf{I}_{L\times L} & \cdots & \mathbf{I}_{L\times L} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{I}_{L\times L} & \mathbf{I}_{L\times L} & \cdots & \mathbf{I}_{L\times L} \end{pmatrix}. \tag{19}$$

Then solving (17) yields the following weight matrix and bias vector

$$\mathbf{b}_{A2} = \mathbf{0}_{JL\times 1}, \tag{20}$$

and

$$\mathbf{W}_{A2} = 2\alpha_2 \mathbf{1}_{JL\times JL}. \tag{21}$$

IV. Performance analysis

In this section we compare the optimum obtained by exhaustive search to the one achieved by the Hopfield Network. The simulations were carried out for $J = 20$ nodes and the corresponding $X_j, j = 1, ..., J$ and $K_j, j = 1, ..., J$ constraints have been chosen randomly in the range of $X_j \in [5, 100]$ and $K_j \in [5, 100]$. The results have been evaluated after selecting several constraints randomly, then running the simulations and taking the average error of the solutions achieved by the Hopfield Network and by exhaustive search.

The obtained results are depicted by the 1. figure. Analyzing the result, one may note that the solutions provided by Hopfield Network does not always yields the optimal solution as it is demonstrated by "Chart a)" of the figure. Furthermore the results for the original objective function (3) and the results for the modified, second objective function (4) are similar.

On the other hand, the solutions of the HNN only slightly differs from the optimum, and its great advantage is that this method can easily be reconfigured when the parameters of the problem are changing. It must also be noted, that a good quality solution is achieved by the Hopfield Network in polynomial time as opposed to the exponential complexity of the exhaustive search method. These running times are compared to each other in the "Chart b)" of the figure.

In the other charts the amount of relative error is indicated, which error is due to the Hopfield Network being not always capable of providing a valid solution by fulfilling all the constraints. "Chart c)" exhibits the cases when only a portion of the packets are sent, whereas on "Chart d)" the packets are

sent after the deadline has expired. It is noteworthy, that these errors can be further reduced by the fine tuning of parameters α, β, δ as can be seen in [19] and is demonstrated by running the modified Hopfield recursion.

Figure 1 shows that the result of the two goal functions does not differ significantly. The main difference is demonstrated by Table I and Figure 2, respectively. The near optimal solution provided by the Hopfield Network proves to be satisfying.

In order to measure how balanced a solution is, we introduce the entropy of the weight distribution of the columns in matrix $\mathbf{C}$ as follows

$$H(\mathbf{p}) = \sum_{k=1}^{L} -p_k \ln p_k, \tag{22}$$

where $p_k = \dfrac{\sum_{j=1}^{J} C_{jk}}{\sum_{k=1}^{L}\sum_{j=1}^{J} C_{jk}}$. Table I demonstrates that the solution provided by HNN for the Balanced Cost problem has the highest weight entropy (i.e the most uniform weight distribution of the columns), hence it fulfills the constraint related to weight balancing.

Method	Entropy
Original HNN	5.4369
Modified HNN	5.6883
Exhaustive Search	5.7066

TABLE I
ENTROPY OF WEIGHT DISTRIBUTION

Figure 2 shows the advantages of the Modified HNN solution, since packet loss probability has significantly been decreased. One can see that the probability of packet loss is under 0.1 in every time slots, while the Original HNN solution yield higher packet loss probabilities.

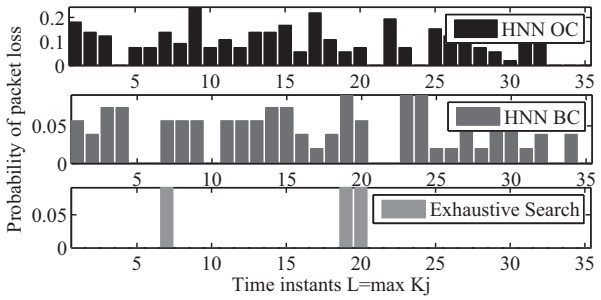


Fig. 2. Packet loss probabilities

V. CONCLUSION

In this paper new methods were proposed to provide optimal scheduling for packet-transmissions in WSN. The objective of optimal scheduling was on the one hand defined by the Overall Cost and by the Balanced Cost. In the latter case the new method provides the same QoS for the nodes in terms uniform packet loss probabilities. Both of the corresponding objective functions have been mapped into quadratic optimization and then solved by the Hopfield Network. In this way, good quality solutions have been obtained in real-time. Using these optimized scheduling matrices in WSN, the clusterhead nodes can save energies which will increase the longevity of the network. As the simulation results have demonstrated the solutions provided by the Hopfield Network are very close to the ones achieved by exhaustive search.

REFERENCES

[1] A. Rogers, D. D. Corkill, and N. R. Jennings, "Agent technologies for sensor networks," *IEEE Intelligent Systems*, vol. 24, pp. 13–17, 2009.

[2] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, "Wireless sensor networks: a survey." *Computer Networks*, vol. 38, no. 4, pp. 393–422, 2002.

[3] C.-Y. Chong and S. P. Kumar, "Sensor networks: evolution, opportunities, and challenges," *Proceedings of the IEEE*, vol. 91, no. 8, pp. 1247–1256, 2003.

[4] P. Inverardi and L. Mostarda, "A distributed monitoring system for enhancing security and dependability at architectural level," pp. 210–236, 2007.

[5] O. Aumagie, E. Brunet, N. Furmento, and R. Namyst, "NewMadeleine: a Fast Communication Scheduling Engine for High Performance Networks," RR-1421, Tech. Rep., 2007.

[6] H. Choi, J. Wang, and E. A. Hughes, "Scheduling for information gathering on sensor network," *Wirel. Netw.*, vol. 15, no. 1, pp. 127–140, 2009.

[7] H. Li, P. Shenoy, and K. Ramamritham, "Scheduling messages with deadlines in multi-hop real-time sensor networks," in *RTAS '05: Proceedings of the 11th IEEE Real Time on Embedded Technology and Applications Symposium*. Washington, DC, USA: IEEE Computer Society, 2005, pp. 415–425.

[8] W. R. Heinzelman, A. Chandrakasan, and H. Balakrishnan, "Energy-efficient communication protocol for wireless microsensor networks," in *HICSS '00: Proceedings of the 33rd Hawaii International Conference on System Sciences-Volume 8*. Washington, DC, USA: IEEE Computer Society, 2000.

[9] S.-H. Cha and M. Jo, "An energy-efficient clustering algorithm for large-scale wireless sensor networks," in *GPC'07: Proceedings of the 2nd international conference on Advances in grid and pervasive computing*. Berlin, Heidelberg: Springer-Verlag, 2007, pp. 436–446.

[10] K. Wu, C. Liu, Y. Xiao, and J. Liu, "Delay-constrained optimal data aggregation in hierarchical wireless sensor networks," *Mob. Netw. Appl.*, vol. 14, no. 5, pp. 571–589, 2009.

[11] S. Haykin, *Neural Networks, A Comprehensive Foundation*, 2nd ed. Pearson, Prentince Hall, 2005.

[12] J. Sima, P. Orponen, and T. Poika, *Some Afterthoughts on Hopfield Networks*. Springer Berlin / Heidelberg, 2010.

[13] M. Satoshi, "Stability of solutions in Hopfield neural network," *IEEE Transaction on Neural Networks*, vol. 9, no. 6, pp. 1319–1330, 1998.

[14] J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities," *Proc. NatL Acad. Sci.*, vol. 79, pp. 2554–2558, 1982.

[15] J. J. Hopfield and D. W. Tank, "Neural computation of decisions in optimization problems," *Biological Cybernetics*, vol. 55, pp. 141–146, 1985.

[16] J. Mandziuk, "Solving the Travelling Salesman Problem with a Hopfield - type neural network," *Demonstratio Mathematica*, vol. 29, no. 1, pp. 219–231, 1996.

[17] J. Mandziuk and B. Macukow, "A neural network designed to solve the N-Queens Problem," *Biological Cybernetics*, vol. 66, no. 4, pp. 375–379, 1992.

[18] J. Levendovszky, E. Laszlo, K. Tornai, and G. Treplan, "Optimal pricing based resource management," in *International Conference on Operations Research Munich 2010*, September 2010, p. 169.

[19] C. Douligeris and G. Feng, "Using Hopfield Networks to Solve Assignment Problem and N-Queen Problem: An Application of Guided Trial and Error Technique," *Springer-Verlag, Heidelberg*, vol. 2308, p. 325, 2002.

Parameter Extraction of Phonocardiographic Signals with Murmur

Ádám T. Balogh
(Supervisors: Dr. Ferenc Kovács and Dr. Tamás Roska)
balogh.adam@itk.ppke.hu

***Abstract*—The continuous monitoring of clinically relevant parameters of the patent ductus arteriosus (PDA) in case of preterm neonates is not solved. Murmur is one of the important symptoms of the PDA, which can be sensitively detected with phonocardiography based methods. This has been shown in a previous study. In this paper two approaches for parameter extraction of the murmur is presented. In the first one time and frequency analysis is applied, but the extracted parameters show only a moderate correlation with the clinically important paramters (NRMSE $>$ 20 %). In the second approach joint time-frequency analysis is performed by applying the Matching Pursuit decomposition. Using this method promising results have been achieved by comparing the time-frequency signatures of murmurs related to different graded PDA-s, where the grade corresponds to the LA/Ao ratio. Further measurements are needed for the quantitative verification of this assumption.**

***Keywords*-phonocardiography; patent ductus arteriosus; murmur; characteristic heart sound; Matching Pursuit decomposition**

I. Introduction

The ductus arteriosus is an essential fetal vascular structure. This means that its closure during pregnancy may lead even to right heart failure. After birth, in case of normal neonates, the ductus arteriosus starts closing with the first intake of breath allowing the development of the normal human circulation [1].

The persistence of the ductal patency after birth is abnormal and has several consequences, such as respiratory problems and hypertrophy of the left atrium and ventricle. Nevertheless the physiological impact and clinical significance of a PDA depends above all on its size and the state of the underlying cardiovascular system.

The closure of the PDA may occur spontaneously or due to a surgical or transcatheter intervention. In case of preterm infants pharmacological closure is also possible [2]. The risk of PDA is clearly much greater in case of preterms, which is due to physiological factors related to prematurity [3]. The main diagnosis of PDA in case of preterms is done with echocardiography, which needs expertise, and sophisticated and expensive equipment.

These aspects show the need for simple tools for helping the diagnosis and the monitoring of the PDA in case of preterm infants. Phonocardiography comes into view based on the observation that one of the fundamental symptoms is murmur. Although earlier studies investigated the murmurs related to PDA in preterm infants [4], non of them tried to find a relationship between various parameters of the heart sound and of the PDA. In this study this problem is investigated.

II. Materials and Methods

A. Database

In this study 25 preterm newborns have been examined, with an average of 3 measurements per infant. Hemodynamically significant PDA was verified by echocardiography in case of 15 infants but only 8 of those were examined over several days because the others had either also some other malformation or some other circumstances made further measurements not possible. Preterms without PDA were measured as a control group. The diagnostic parameters of the PDA acquired with echocardiography were all collected for later comparison with phonocardiographic parameters. In case of the 8 newborns mentioned above, the PDA was closed by means of pharmacological treatment (4 infants) or surgical intervention (4 infants).

These infants, except one, all weighed less than 2300 g at birth, with an average weight of 1400 g. Except one, all of them were less than 33 weeks of gestation, with an average of 29. They were examined on average on their 6th day after birth and those with PDA then every day until the closure of the PDA, which was verified by echocardiography.

The measurements were made with a self-made electronic stethoscope. Each measurement consisted of about three 30 seconds long phonocardiographic records which were recorded at 48 kHz, with a resolution of 16 bits. According to our observation the main components of the heart sounds lie in the low frequency range, thus after prefiltering, the data was resampled at 3000 Hz and only the useful part of the record (at least 10 secs) was kept for further analysis.

B. Preprocessing

For the identification of the murmur a characteristic heart sound was calculated for each record [5]. Using this technique murmur was identified in case 5 out of the 8 preterm neonates. Their records were used for further analysis. According to our measurements, murmur in case of preterms with PDA is usually late systolic. This murmur was detected by thresholding the average envelope calculated using the Hilbert transform [6]. Due to the continuous noise coming from the medical ventilator the envelope had a positive baseline shift. This was corrected by estimating the baseline of the average envelope based on the histogram of the envelope values. Those envelope values were set to zero which were smaller than two times the baseline value. In case of records with murmur the average

envelope had still non-zero values in the systolic segment after the baseline correction. This was used to determine the length of a time window for observing the murmur in the given record.

C. Parameter extraction based on time and frequency analysis

Using the time window mentioned above the murmur was analyzed in all heart cycles from which the characteristic heart sound was calculated. During the analysis the following parameters were extracted: the length of the murmur, average maximal murmur amplitude, average maximal S2 amplitude, the ratio of the previous two, average mean, maximal and minimal instantaneous frequency, the average jitter of the instantaneous frequency and the frequency limits of the murmur. The instantaneous frequency for a given time instance was estimated by calculating the first moment of the Fourier transform in a 20 ms time window as follows:

$$IF[n] = \sum_{f=F_1}^{F_2} (S[f])^2 \cdot f \;\; / \;\; \sum_{f=F_1}^{F_2} (S[f])^2 \qquad (1)$$

where $IF[n]$ is the instantaneous frequency at the time instance n, $S[f]$ is the Fourier transform of the 20 ms long heart sound signal segment at the frequency f, and F_1 and F_2 are the limits of the investigated frequency interval.

Frequency limits are estimated by finding the maximal and minimal frequencies of the thresholded spectrogram, in this study a threshold of -50 dB was employed. The spectrogram was calculated in the time window defined for the murmur with a 20 ms Hamming window.

D. Matching Pursuit decomposition of heart cycles

Matching Pursuit (MP) is an efficient method for optimal decomposition of an $x[n]$ signal into time-frequency atoms regarding the l_2 norm. This time-frequency representation is usually of better resolution than the conventional spectrogram, but is also free of the cross terms typical for the Wigner-Ville distribution. I applied the MP implementation of Leung et al. [7] and extended it with a feature enhancement approach of Tang et al. [8].

The decomposed signal takes the following form:

$$x[n] = \sum_{i=1}^{N} A_i \cdot e^{-((n-n_i)\Delta t)^2/(2\sigma_i)^2} \cdot cos(2\pi f_i n \Delta t + \theta_i) \quad (2)$$

where $x[n]$ is made up of N gaussian modulated cosine time-frequency atoms with A_i maximal amplitude, f_i frequency , θ_i phase, respectively. The gaussian modulation function has a maximum at $t = n_i \Delta t$ and has a standard deviation of σ_i.

The decomposition is performed based on an iterative way using the spectrogram of the investigated segment. The applied algorithm consists of following steps:

1) Calculate the spectrogram of the given segment.
2) Find the largest component of the spectrogram, and determine the corresponding temporal location (n_i), frequency (f_i), amplitude (A_i) and phase (θ_i).
3) Calculate he time domain representation of the component based on the above mentioned parameters. The parameter σ_i can be optimized by exhaustive search on a predefined interval ([20 ms,75 ms]).
4) Subtract the calculated component from the signal.
5) If the energy of the residual signal per the energy of the original signal is greater than a given threshold (typically 5-10 %) then repeat steps 1-4. Otherwise the decomposition is complete.

An example for the MP based time-frequency decpomposition and reconstruction is shown in Fig. 1.

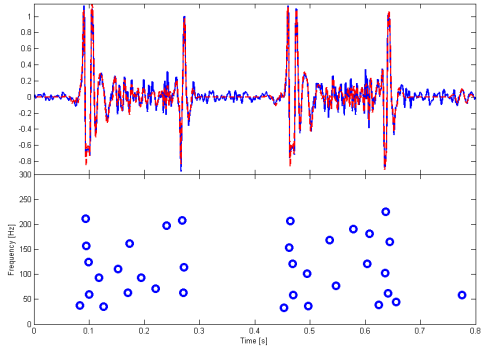


Fig. 1. Two heart cycles of a preterm with PDA (above) and the corresponding time-frequency representation calculated based on the described MP method (below). The original signal is shown with a solid line, the reconstructed signal with a dashed line. The stopping criteria for the decomposition was 5 %. It should be noted the the decomposition preserves the main heart sounds and the murmur with high accuracy, whereas only noise is rejected.

E. Feature enhancement based on the MP decomposition

Further feature enhancement and noise reduction can be achieved by superimposing different heart cycles in the time-frequency domain. In this way the important time-frequency atoms can be determined, as shown in Fig. 2.

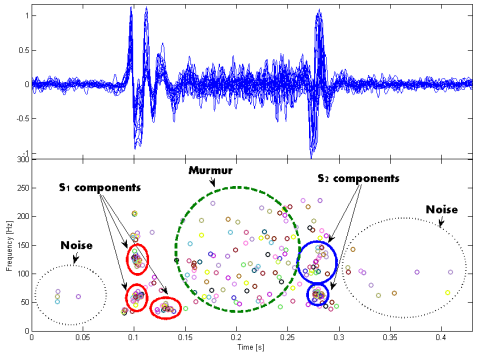


Fig. 2. Superimposed heart cycles of a preterm with PDA (above), alignment was performed based on the S1 heart sounds. The corresponding time-frequency representations, also superimposed, are shown on the bottom part of the figure. The time-frequency atoms can be classified based on their density and temporal location.

The noise reduction can be automated by investigating the density of the time-frequency atoms. That means that for each

atom the number of other atoms in a given radius has to be determined:

$$N_i(r) = \# \left\{ \text{atom}_j : \sqrt{(n_j - n_i)^2 + (f_j - f_i)^2} < r \right\} \quad (3)$$

The above shown calculation should be perform in a normalized time-frequency domain, that is $f_i \in [0,1]$ and $n_i \in [0,1], \forall i$. A given atom will be included in the denoised decomposition if N_i is greater than a given threshold, in this study this threshold was 60 % of the number of investigated heart cycles. The radius r was considered as 0.075.

III. Results

The above described methods have been applied to phonocardiographic records of those preterm infants with PDA who had murmur.

A. Results of parameter extraction from the time and frequency analysis

All extracted parameters were correlated with the medical parameters of the PDA, i.e. the diameter of the PDA (D_{PDA}), the maximal blood velocity through the PDA (v_{max}) and the left atrial to aortic root ratio (LA/Ao). A regression curve was fitted to each of the investigated data point sets.

Based on the NRMSE of the fitted regression curves the average mean, maximal and minimal instantaneous frequency parameters correlate the most with the medical parameters. These relationships are shown in Fig. 3.

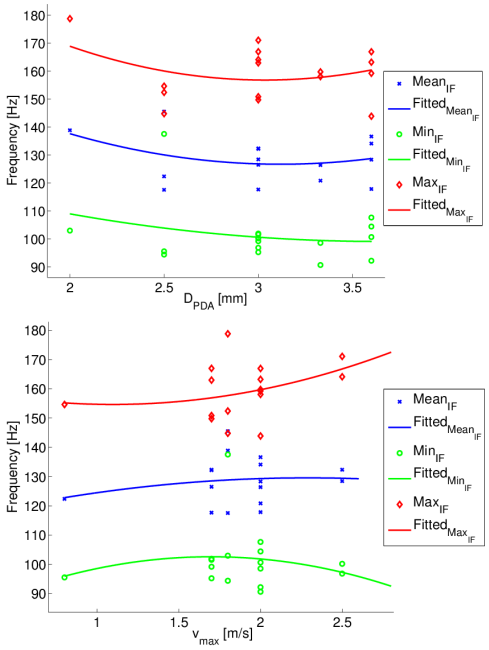


Fig. 3. Extracted murmur parameters vs. medical parameters of the PDA, i.e. diameter of the PDA (top) and maximal blood flow through the PDA (bottom). As observable the frequency parameters of the murmur rather decrease with the D_{PDA} and increase with the v_{max}.

Although these results are promising it should be noted that for application purposes further improvements are needed for more reliable parameter estimation, since the fitted regression curves had NRMSE $>$ 20 %. This is why also an other approach was investigated based on Matching Pursuit decomposition of the heart cycles since reliable estimation of PDA related parameters would be of great importance.

B. Results of the MP decomposition of heart cycles

As a preprocessing step the characteristic heart sound calculation method from a previous study [5] was applied for selecting the most typical heart cycles for the given record. These cardiac cycles were used as the input for the feature enhancement algorithm. The calculated time-frequency representations have been compared and evaluated. It has been found that different grades of PDA exhibit different time-frequency signatures, where the grade corresponds to the LA/Ao ratio. Unfortunately, due to the moderate amount of data, this relation could not be understood completely. Some case reports are presented hereunder.

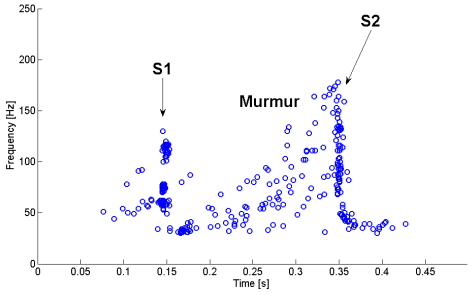


Fig. 4. The denoised time-frequency representation of the cardiac cycles of a preterm infant with PDA. The LA/Ao ratio was 1.34 ($D_{PDA} = 3.6$ mm, $v_{max} = 2$ m/s).

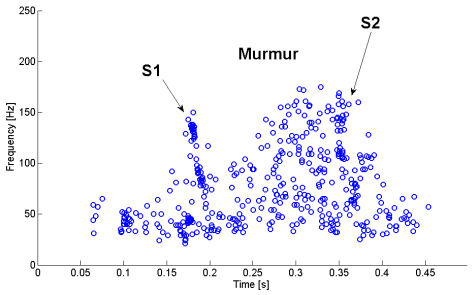


Fig. 5. The denoised time-frequency representation of the cardiac cycles of a preterm infant with PDA. The LA/Ao ratio was 2.08 ($D_{PDA} = 3$ mm, $v_{max} = 1.7$ m/s).

As observable in Figures 4-6 in case of greater LA/Ao ratios the time-frequency decomposition of the murmur is more dense in the time-frequency plane and also higher frequency components appear. The parameter extraction of this representation is the topic of further work.

IV. Conclusion

The monitoring of clinically relevant parameters of the PDA would be of great significance. Phonocardiography, as a simple and non-invasive tool could provide a way for solving

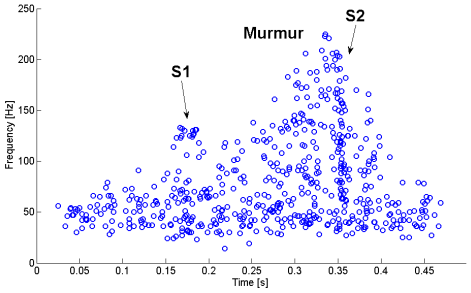


Fig. 6. The denoised time-frequency representation of the cardiac cycles of a preterm infant with PDA. The LA/Ao ratio was 2.21 ($D_{PDA} = 3.33$ mm, $v_{max} = 2$ m/s). Note the higher frequency components of the murmur compared to Fig. 5.

this problem. Based on the work so far it can be concluded that that the murmur associated to PDA can be sensitively identified with phonocardiography. In this work two methods have been shown for parameter extraction of this murmur for monitoring purposes. Promising results have been achieved but further measurements are needed for adequate verification of the conclusions. Nevertheless it seems to be possible to assess clinically significant parameters of the PDA in certain manner.

ACKNOWLEDGMENT

The author would like to thank Dr. Zoltán Molnár and Dr. Miklós Szabó from the Ist Department of Paediatrics, Semmelweis University of Medicine, Budapest, for the measurements and for their assistance and collaboration. This work was supported by the Swiss Contribution.

REFERENCES

[1] H. Allen, D. Discoll, R. Shaddy, and T. Feltes, *Moss and Adams' Heart Disease in Infants, Children, and Adolescents: Including the Fetus and Young Adults*, 7th ed. Philadelphia: Lippincott Williams & Wilkins, 2008, vol. 1.

[2] D. B. Knight, "The treatment of patent ductus arteriosus in preterm infants, a review and overview of randomized trials," *Seminars in Neonatology*, vol. 6, no. 1, pp. 63–73, 2001.

[3] D. J. Schneider and J. W. Moore, "Patent ductus arteriosus," *Circulation*, vol. 114, pp. 1873–1882, 2006.

[4] K. A. Hallidie-Smith, "Murmur of persistent ductus arteriosus in premature infants," *Archives of Disease in Childhood*, vol. 47, pp. 725–730, 1972.

[5] A. Balogh and F. Kovács, "Application of heart sound analysis in preterm neonates with patent ductus arteriosus," in *Applied Sciences in Biomedical and Communication Technologies, 2009. ISABEL 2009. 2nd International Symposium on*, 24-27 2009, pp. 1 –2.

[6] S. Hahn, *Hilbert transforms in signal processing*. Artech House, 1996.

[7] T. Leung, P. White, J. Cook, W. Collis, E. Brown, and A. Salmon, "Analysis of the second heart sound for diagnosis of paediatric heart disease," *Science, Measurement and Technology, IEE Proceedings*, vol. 145, no. 6, pp. 285–290, 1998.

[8] H. Tang, T. Li, Y. Park, and T. Qiu, "Separation of heart sound signal from noise in joint cycle Frequency–Time–Frequency domains based on fuzzy detection," *IEEE Trans. Biomed. Eng.*, vol. 57, no. 10, pp. 2438–2447, 2010.

Simulation of absorption based surface plasmon resonance sensor in the Kretschmann configuration

Ádám Fekete
(Supervisor: Dr. Áprád Csurgay)
fekad@digitus.itk.ppke.hu

***Abstract*—Surface plasmon resonance (SPR) has been efficiently employed for chemical- and bio-sensing. Fully understanding the relation among the parameters and the detection is instructive for optimal design. In this paper we present a numerical simulation of Kretschmann-type sensor.**

***Keywords*-surface plasmon resonance; Kretschmann configuration; Drude model; Lorentz model; simulation**

I. Introduction

Over the last 20 years Surface plasmon resonance (SPR) technology has been widely utilized in the field of biochemical or biophysical analyses. This rapidly growing field of nanoscience because this method able to real time monitoring of surface binding events and it has very high sensitivity [1], [2].

The monochromatic p-polarized light interacts with the free electrons inside the metal which lead to the oscillation of free electrons to excite evanescent waves that propagate along the surface of dispersive media and photoenergy is converted into surface wave energy, which is called SPR. The concentrate energy and resonance amplified electromagnetic field are directly responsible for ultra highly refractive index resolution of SPR sensors and several nonlinear phenomena, including surface-enhanced Raman scattering, enhanced fluorescence emission, high-harmonic generation.

The commonly used method for exciting surface plasmons is the Kretschmann prism configuration (Figure 1). The sensing layer made up of the immobilized receptors can couple with the ligands. For surface plasmon resonance phenomenon is extremely sensitive to small changes of the dielectric constant above the metal film based on measuring changes in reflection index. The excitation of a SPR will show up as a minimum in the reflected light.

There are available a several numerical electromagnetic simulator (CST Microwave Studio, Lumerical FDTD Solutions, etc.) which are open the door to making interactive design for optimize the parameters and predict the best experimental configuration.

II. Theoretical models

For simulating the Kretschmann configuration require different models for different layers. The optical properties of metals can be described by complex dielectric function that depends on the frequency of light. For wavelengths above 500 nm the Drude model is a good approximation of the metals. For wavelength below 500 nm interband transitions become significant. For the sensing layer the Lorentz model is suitable to simulate atomic system.

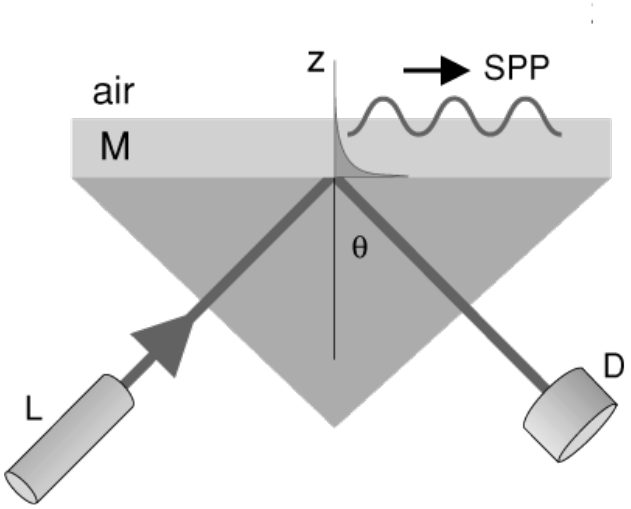


Fig. 1. Kretschmann configuration

A. Surface Plasmon Generation

The metal's free electron gas can sustain surface and volume charge density oscillations, called plasmon polaritons or plasmons with distinct resonance frequencies. We consider a plane interface between a metal and a dielectric. Consider p-polarized waves can generate an exponentially extenuating evanescent wave (Figure 2). The transformation of resonance energy takes place between evanescent wave an surface plasmons and influenced by the material absorbed into the thin metal film.

Kretschmann configuration is a thin metal film which decomposed on top of prism. If the metal too thin, the SPP will be strongly damped because of radiation damping into the glass. If the metal film is too thick the SPP can no longer be efficiently excite due to absorption in the metal [3].

For given energy $\hbar\omega$ the wavevector k_x is always larger than the wavevector of light in free space. Excitation of a

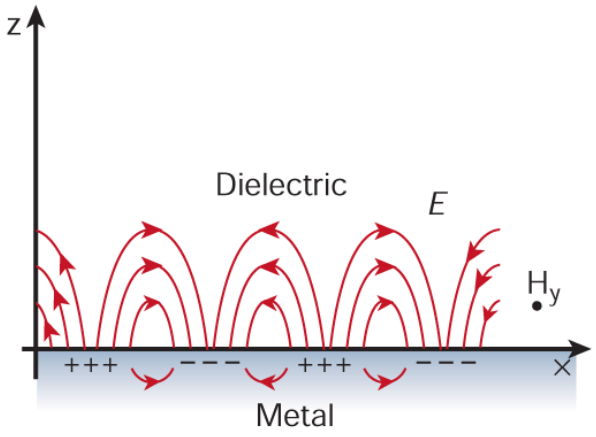


Fig. 2. Surface plasmon resonance

SPP by light is only possible if a wavevector component of the exciting light can be increased over its free-space value (Figure 3). Most simple solution to excite surface plasmon with reflective index $n > 1$.

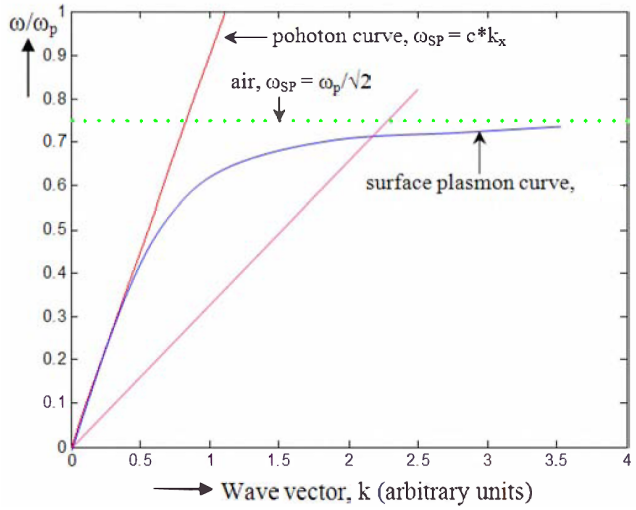


Fig. 3. Dispersion curve

B. Drude model of metal layer

The amplitude of the induced displacement $\mathbf{D}$ can be written in terms of the macroscopic polarization P according to

$$\mathbf{D}(\omega) = \varepsilon_0 \varepsilon(\omega)\mathbf{E}(\omega) = \varepsilon_0 \mathbf{E}(\omega) + \mathbf{P}(\omega)$$

where macroscopic polarization $\mathbf{P}$ can be expressed as

$$\mathbf{P} = \varepsilon_0 \chi_0(\omega)\mathbf{E}(\omega)$$

From this we get:

$$\varepsilon(\omega) = 1 + \chi(\omega),$$

the frequency-dependent dielectric function of the metal.

The Drude model, the free electron oscillate 180° out of phase relative to driving electric field.

$$\begin{aligned}\varepsilon_{Drude}(\omega) &= 1 - \frac{\omega_p^2}{\omega^2 + \mathrm{i}\Gamma^2\omega} \\ \varepsilon_{Drude}(\omega) &= 1 - \frac{\omega_p^2}{\omega^2 + \Gamma^2} + \mathrm{i}\frac{\Gamma\omega_p^2}{\omega(\omega^2 + \Gamma^2)}\end{aligned}$$

Most metal posses a large negative real part of the dielectric constant at optical frequencies with a small imaginary part (Figure 4). The light can penetrate a metal only to a very small extent since the negative dielectric constant leads to a strong imaginary part of the reflective index $n = \sqrt{\varepsilon}$. The imaginary part of ε describes the dissipation of energy (ohmic losses) associated with the motion of electrons in metal.

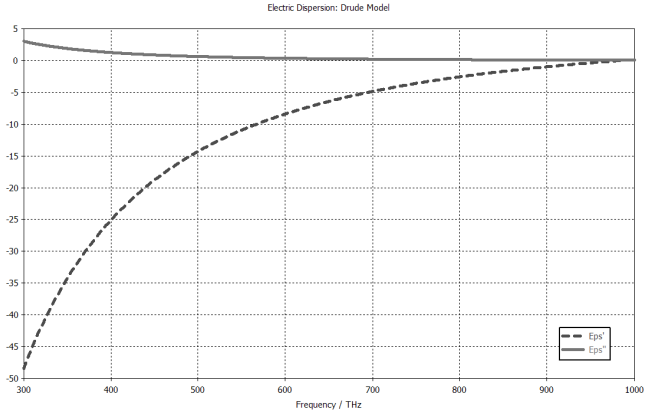


Fig. 4. Electric dispersion of gold Drude model

Intruduce a constant offset ε_∞, which accounts for the integrated effect of all higher-energy interband transition not considered int he present model.

$$\varepsilon_{Drude}(\omega) = \varepsilon_\infty - \frac{\omega_p^2}{\omega^2 + \mathrm{i}\Gamma^2\omega}$$

The gold, at a wavelength shorter than 500 nm, the measured imaginary part of dielectric function increases much more strongly as predicted by Drude model, because higher-energy photons can promote of lower-lying bands into the conduction band.

C. Interaction of radiation and atomic system

In the sensing layer the interband transitions become more significant. The optical absorption effect in sensing layer is included in the absorption-based SPR theory by the Lorentz model that expresses a damped harmonic oscillator. The optical absorption effect changes the RI imaginary part as well as the RI real part in sensing layer [4].

$$
\begin{aligned}
\varepsilon_{Lorentz}(\omega) &= \varepsilon_\infty + \frac{\tilde{\omega}_p^2}{(\omega_0^2 - \omega^2) - \mathrm{i}\gamma\omega} \\
\varepsilon_{Lorentz}(\omega) &= \varepsilon_\infty + \frac{\tilde{\omega}_p^2 (\omega_0^2 - \omega^2)}{(\omega_0^2 - \omega^2)^2 - \gamma^2\omega^2} \\
&+ \mathrm{i}\frac{\gamma\tilde{\omega}_p^2\omega}{(\omega_0^2 - \omega^2)^2 - \gamma^2\omega^2}
\end{aligned}
$$

where γ is the damping constant.

$$
\begin{aligned}
\tilde{\omega}_p &= \sqrt{\frac{Ne^2 f}{m_e \varepsilon_0}} \\
\omega_0 &= 2\pi c/\lambda_{max} \\
\gamma &= 2\pi c \frac{\lambda_{1/2}}{\lambda_{max}^2} \\
\varepsilon_0 &= 10^7/4\pi c^2 \\
N &= 10^3 N_A\,[Ab]
\end{aligned}
$$

where e is the elementary electric charge; m_e is the mass of an electron; ε_0 is the permittivity of vacuum; ω_0 and γ are the absorption frequency and the damping frequency, respectively; λ_{max} and $\lambda_{1/2}$ are the absorption maximum wavelength and the full width wavelength at half-maximum of absorption spectrum, respectively; f is the oscillator strength, which is related to the molar extinction coefficient; ε_∞ is the background dielectric constant of the sensing layer; N_A is Avogadro's constant; and $[Ab]$ is the molar concentration of absorption oscillators, such as dye molecules.

III. Numerical results

Two methods have been commonly used in these research works. One is Fresnel's equation of multilayered structures [5], [6]. The other one is the transfer matrix theory [7]. But for complex geometries we have to use numerical electromagnetic simulators [8], [9].

The CST Microwave Studio has been used for example for scattering for coated silica sphere and its capable to handle materials which has frequency-dependent complex dielectric functions (ie. Drude, Lorentz).

The goal is choose the optimal parameters for the following conditions:

- Excitation wavelengths: 532 nm (green SHG laser), 635 nm (red LD laser), 685 nm (red LD laser)
- Optimal optical properties of prism (ie. BK7 $n = 1.51$)
- The material (ie. Au, Ag) and the thickness of the metal layer
- The dielectric function of sensing layer

We use frequency domain solver with unit cell boundary condition and Floquet port excitation. The Floquet ports are used in place of the Perfectly Matched Layer (PML) absorbing boundary condition in order to obtain accurate results at very oblique angles of incidence. The change in the resonance condition was measured in the reflectivity in different angles.

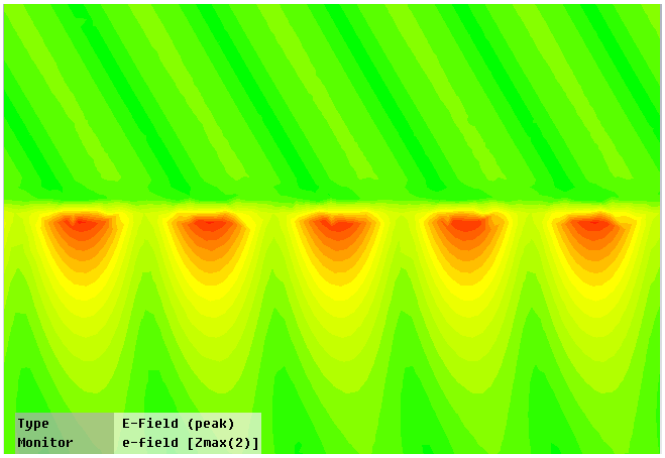


Fig. 5. Electric field of simulated Surface Plasmon Resonance

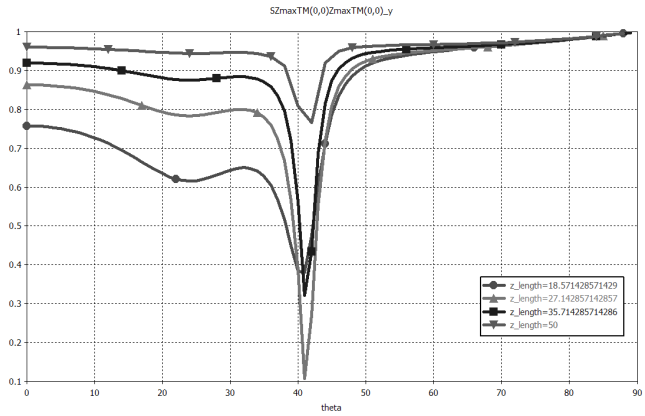


Fig. 6. Simulation of Surface Plasmon Resonance with different thickness of metal layer

We use the following parameters to simulate the gold metal layer:

- Dielectric of prism: $n = 1.51$
- Drude model:

$$
\begin{aligned}
\varepsilon_\infty &= 6 \\
\omega_p &= 13.8 \times 10^{15}\ \mathrm{s}^{-1} \\
\Gamma &= 1.075 \times 10^{14}\ \mathrm{s}^{-1}
\end{aligned}
$$

- Background material is vacuum

In the first simulation we optimize the thickness of the metal layer. Figure 6, show intensity of reflected light versus the angle of incident light with different thicknesses. The optimal solution is when the the Au layer is about 30 nm thick.

Figure 7, show the behavior of the surface plasmon resonance when changing dielectric function of the sensing layer. A small changes of the sensing layer dielectric constant change the minimum in the reflected light.

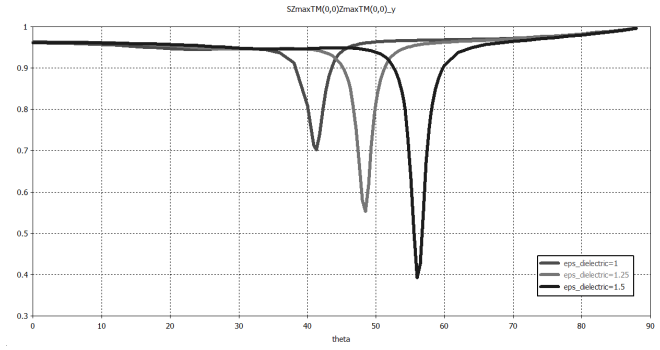


Fig. 7. Simulation of Surface Plasmon Resonance with different sensing layers

IV. Conclusion

In this paper we presented that the CST Microwave Studio is capable to optimize the parameters (reflection index of prism, thickness of metal layer) and illustrate the behavior of the sensor in an interactive design framework. For better approximation we could use more sophisticated multi-layer models for sensing layer.

References

[1] J. Homola, S. S. Yee, and G. Gauglitz, "Surface plasmon resonance sensors: review," *Sensors and Actuators B: Chemical*, vol. 54, no. 1-2, pp. 3–15, Jan. 1999.

[2] J. Homola, "Present and future of surface plasmon resonance biosensors." *Analytical and bioanalytical chemistry*, vol. 377, no. 3, pp. 528–39, Oct. 2003.

[3] L. Novotny and B. Hecht, *Principles of nano-optics*. Cambridge University Press, 2006.

[4] K. Kurihara and K. Suzuki, "Theoretical understanding of an absorption-based surface plasmon resonance sensor based on Kretchmann's theory." *Analytical chemistry*, vol. 74, no. 3, pp. 696–701, Feb. 2002.

[5] Y. Xinglong, W. Dingxin, and Y. Zibo, "Simulation and analysis of surface plasmon resonance biosensor based on phase detection," *Sensors and Actuators B: Chemical*, vol. 91, no. 1-3, pp. 285–290, Jun. 2003.

[6] M. S. Islam, A. Z. Kouzani, X. J. Dai, and W. P. Michalski, *Parameter sensitivity analysis of surface plasmon resonance biosensor through numerical simulation*. IEEE, Jul. 2010, vol. 00, no. c.

[7] X.-f. Luo and L. Han, *A universal model of surface plasmon resonance characteristics for isotropic multilayer films*. IEEE, Nov. 2010.

[8] X. Gao, Z. Xiao, and L. Ning, *Surface plasmons enhanced super-resolution focusing of radially polarized beam*. IEEE, Dec. 2010, no. 3.

[9] W. H. P. Pernice, F. P. Payne, and D. F. G. Gallagher, "An FDTD method for the simulation of dispersive metallic structures," *Optical and Quantum Electronics*, vol. 38, no. 9-11, pp. 843–856, Feb. 2007.

Designing a simple digital microfluidic device

Dániel Kovács
(Supervisor: Dr. Kristóf Iván)
kovacs.daniel@itk.ppke.hu

***Abstract*—Microfluidics, as a new branch of MEMS technology, has appeared in the last few decades and has already common applications in inkjet printers and fuel dispensers in spacecrafts. Lately, it has been more and more extensively used in biotechnology for diagnosis and screening tests. In digital microfluidics individual droplets of biological liquids are manipulated by electric fields in order to test samples for diseases or the presence of special analytes. We attempt to develop a digital microfluidic device based on the electrowetting-on-dielectric effect for demonstrational and educational purpose which is capable of moving and merging (blood) droplets and has an integrated detection apparatus to measure the concentration of a prescribed substance dissolved in it, probably sugar. Now I present two three dimensional numerical studies performed in COMSOL to simulate droplet movement and merging, then I summarize the results of real experiments carried out so far to create such a device.**

Keywords-microfluidics; digital microfluidics; electrowetting-on-dielectric; EWOD; MEMS; lab-on-a-chip (LOC)

I. INTRODUCTION

Microfluidics and digital microfluidics are relatively novel and emergent interdisciplinary technologies on the border of many branches of engineering and science such as electrical engineering, biology, medicine, physics and mathematics. Microfluidics deals with continuous biological liquid flows in miniature channels etched mainly in glass or PDMS (polydimethylsiloxane), a frequently used silicone elastomer, while digital microfluidics, as a subfield of microfluidics, studies the behavior of individual droplets placed on electrodes and exposed to electric field and makes use of it in handling biological samples mostly in medical examinations performed by autonomous and portable devices. Both disciplines, as a part of the lab-on-a-chip (LOC) and MEMS technologies, constitute a bridge over the gap between commercial inorganic silicon microelectronic technology and organic biological systems and seek for the possibility of miniaturizing known medical and biological measurements by downscaling theses processes onto a single chip and thereby lowering their cost, time and substance requirement, and making them easier to use — let the user be either a professional or a non-professional.

Our work was motivated by the ever increasing number of biological applications of digital microfluidics. The boost took its start right after the millennium, when the usability of the principle of electrowetting-on-dielectric (EWOD) for biological samples such as blood, sweat, tears and urine was validated in the first decade of this century, see [1] and [2]. Since then a wide variety of biological, medical, chemical and environmental applications has appeared. A very promising one was developed in [3] and [4] for a fast and precise type of polymerase chain reaction (PCR) performed in as a small volume as that of only one droplet, while [5] gives some digital microfluidic solutions e.g. for glucose measurement from blood droplets, or detection of extremely low concentrations of trinitrotoluene (TNT) leaked out from land mines. A device capable of measuring the concentration of microscopic airborne particles is described in [6], which is example for an environmental application. The list could be continued endlessly and though such pieces of equipment are not yet available on the market, in the near future we expect to see real commercialization of these devices.

The support of OTKA grant PD73653 is greatly acknowledged.

II. THEORETICAL BACKGROUND

A. Basic Physics

The physical phenomenon underlying digital microfluidics is called electrowetting, i.e., the ability of a fluid to change its surface tension due to an electric field. The pioneer of this area was Gabriel Lippmann who observed that the presence of electric charges can alter the capillary rise of mercury in a tube. Almost every liquid contains freely moving electric charges, especially biological liquids such as blood and urine, on which external electric fields can act and so modify the wetting behavior of the fluid. Considering a sessile droplet on a flat surface of an electrode covered by dielectric (insulator), the contact angle of the droplet changes upon activation of the electrode. The governing equation relevant here is called the Lippmann–Young equation [7], describing this change by

$$\cos\theta = \cos\theta_0 + \frac{\varepsilon_0\varepsilon_D}{2\gamma_{LG}d}V^2 . \qquad (1)$$

Here θ_0 and θ denote the initial and final contact angles, respectively, d is dielectric thickness, γ_{LG} is the surface tension between the liquid and the gaseous phases, ε_0 and ε_D denote vacuum permittivity and relative permittivity of dielectric, respectively, and V is the potential of the activated electrode. From (1) it follows that for electrowetting to be effectively used in practice, i.e., to achieve a big contact angle change, a thin dielectric layer with a relatively high dielectric constant is required, otherwise the effect occurs only at high voltages (above 500 V). Contact angle change involves a change in the shape of the droplet which, if the former is significant, may lead to a switch from hydrophobic to hydrophilic type of contact with the substrate. This alteration of behavior facilitates various kinds of droplet manipulation techniques such as moving, merging, mixing and cleavage, which constitute

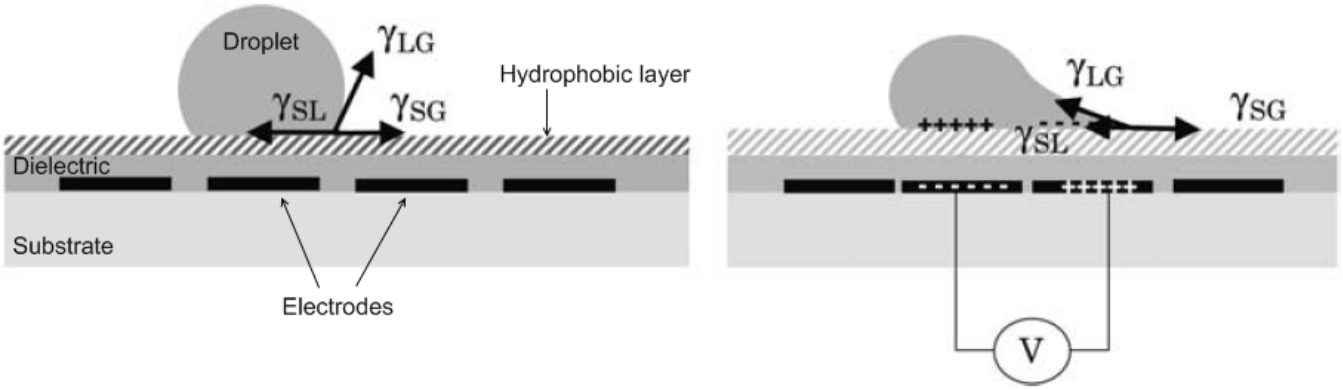


Figure 1. The electrowetting-on-dielectric (EWOD) principle and cross section of an open digital microfluidic system. The directions of surface tension forces arising along the triple contact line and also the changes thereof are depicted by thick black arrows. Source: [7].

common work phases in digital microfluidic based analytical systems.

B. Structure and Operation of an Open System

Our goal is to create a so-called *open* digital microfluidic system in which the droplet is placed on a flat, horizontal array of electrodes covered by dielectric (to eliminate electrolytic currents). The structure is open because there is no top electrode (closed system); the droplet is surrounded by ambient air. A cross section of such a system is shown by Fig. 1, also presenting the EWOD principle itself. The substrate can be silicon or plastic, e.g. in a printed circuit board (PCB). Electrodes are mostly made of alloys of gold, copper or aluminum. The dielectric layer is usually composed of silicon nitride (Si_3N_4) or Parylene C which are widely used materials because of their large permittivity allowing the application of thin layers of them, typically a few micrometers. Their drawback, however, is that their manufacturing needs a special technique called chemical vapor deposition (CVD), or one of its variants, with all relating pieces of equipment. Another solution is to use PDMS which is easy to handle and can be spin-coated in the desired thickness. Its relative permittivity, on the other hand, is lower which necessitates a thicker layer to be applied, at least 30-40 μm. The hydrophobic layer on top can be Teflon AF or CYTOP or other water repellent material, and is usually less than 1 μm in thickness.

Electrodes are fabricated in such a way that promotes droplet movement from one electrode to the other when potential difference is generated between them. Therefore, they are usually designed to form an interdigitating array where all electrodes have both protrusions and inlets ensuring that whatever the position of the droplet is, it is in touch with more than one electrode and can be moved by proper activation of them.

In a digital microfluidic platform droplet operations have a scientific goal, namely, assessing the presence and/or concentration of a specific chemical or component of a solution or some bodily fluid, see [5] again for examples. A conventional method to do this, after the dilution, merging and mixing steps, is to make use of a certain colorimetric reaction producing a light absorbing substance, the concentration of which can be measured by a pair of LED and photodiode by checking the absorbance spectrum at different wavelengths. The concentration of the original component is then deduced from the measured value by tracing back the reaction taken place.

III. NUMERICAL MODELS

For all simulation models we used the Phase Field Method from the Two-Phase Flow module of COMSOL Multiphyics 4.0 numerical software.

A. A 3D Model of Droplet Moving

We successfully modeled the moving of a three dimensional water droplet in air placed between two electrodes of 1.6 mm width and 0.8 mm length with a separating gap of 200 μm. A vertical symmetry plane along the array for the reduction of mesh cells was introduced which cut the electrodes in half so that only a 0.8 mm wide part of them is seen, as is shown in Fig. 2. Electric field was not directly included in the simulation because it would have caused an unreasonably long computation time even on the cluster of 8 processors that we used. Hence the interpretation of the result still contains two interrelated undetermined factor, namely, the electric field right under the droplet and the dielectric thickness above the electrodes. According to (1), an increase in thickness would necessitate an increase in applied voltage to keep the change in contact angle constant. In this manner, though electric field was not directly included in the model, the change in contact angle is realistic and can be attributed to different values of applied voltage from, say, 350 V to 600 V as dielectric thickness varies from 30 μm to 100 μm (if dielectric is assumed to have $\varepsilon_D = 3$).

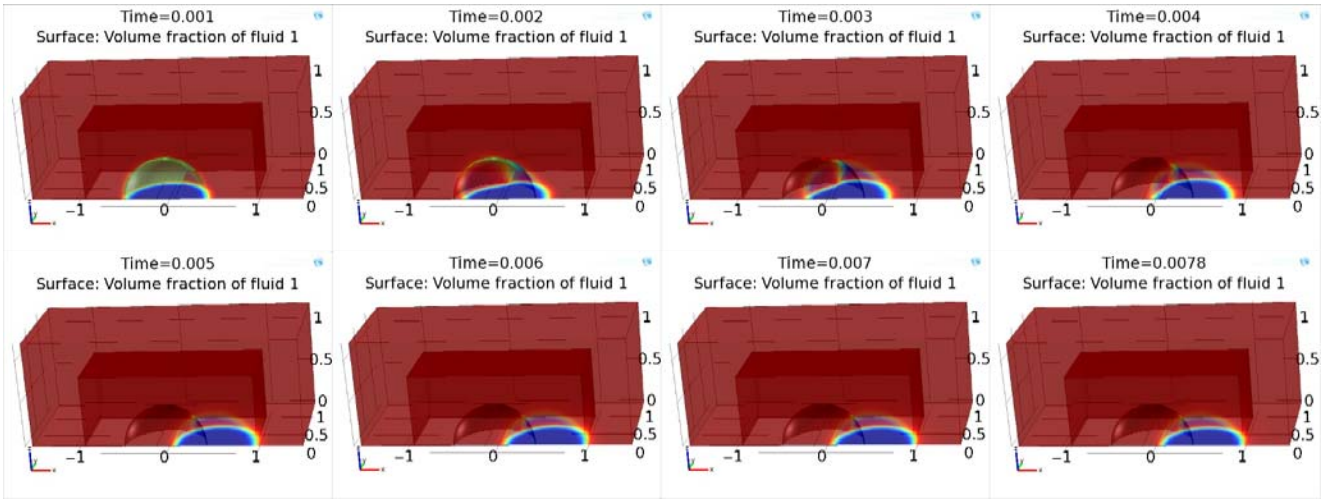


Figure 3. Snapshots from a time dependent numerical simulation of a moving water droplet in an open EWOD system. The droplet changes its contact angle from 110 degrees to 70 degrees due to electric field applied on the right electrode. This generates capillary forces which drag the droplet to the right. The movement is over in less than 4 milliseconds corresponding to a velocity of about 10 cm/s, in good agreement with literature.

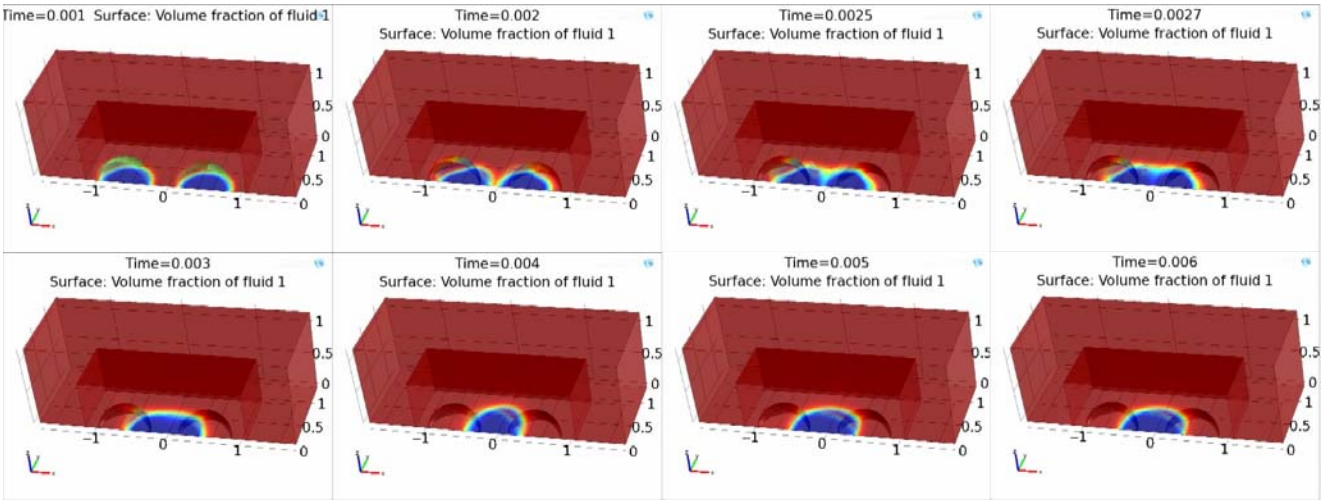


Figure 2. Snapshots from a time dependent numerical simulation of two water droplets merging together in an open EWOD system. The droplets change their contact angle from 110 degrees to 70 degrees due to electric field applied on the central electrode. This generates capillary forces which drag the droplets toward each other. The merging process took approximately 6 milliseconds.

The simulation was started from 1 millisecond when transient initialization had been finished. The time series in Fig. 2 shows that the effective transposition of the droplet took about 4 milliseconds and subsequent vibrations inside the droplet faded away in an additional 3 milliseconds. This corresponds to a velocity of 10 cm/s which is in good agreement with literature, [7].

B. A 3D Model of Droplet Merging

Next, we made a three dimensional model of the merging of two droplets placed above the separating gaps of three neighboring electrodes. All physical constants and geometrical settings were the same as in the previous model. The result is summarized by Fig. 3 in a series of snapshots. Merging started from 0.001 s and was complete by 0.004 s, with subsequent vibrations diminishing in the next 2 milliseconds. A speedup in the process is observable at the initial phase, between 0.002 s and 0.003 s.

IV. Experimental Work

After and along with numerical calculations we also carried out real measurements trying to reconstruct the predictions of the model. For this purpose we used several different PCBs, two of which is shown in Fig. 5. Small separation of electrodes is required for the electrowetting effect to operate. Unfortunately, the rigid PCB fabrication technology can only support electrode separations no smaller than 200 μm. This is much larger than typically used in digital microfluidic devices (20–50 μm) and which could only be achievable in a clean room with special etching techniques or, equivalently, by using elastic PCB technology which is not currently available in Hungary and is much more expensive.

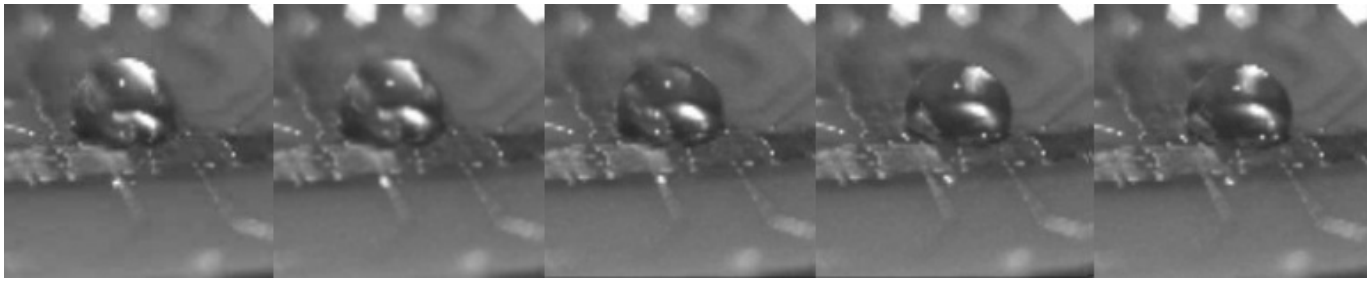


Figure 4. A sequence of video frames of the motion of a salt water droplet due to a voltage of about 300 V on a PCB. The motion took about 2 seconds indicating that at the actual dielectric thickness this voltage is just enough for the electrowetting effect to occur.

Despite this problem we managed to move salt water droplets on the rigid PCB shown on the left panel of Fig. 5. For covering the electrodes we spin-coated PDMS on them and applied a thin layer of Teflon AF on top of it. A time series from a movie in Fig. 4 presents the translocation process between two neighboring electrodes. We found that for a minimal electrowetting effect a voltage of 300–350 V was required and the motion lasted about 2 seconds. We expected that the big separation gap would hinder continuous droplet movement in the future therefore we have chosen another PCB type as well to carry out further experiments with, see the right panel of Fig. 5.

This array was big enough so that dielectric thickness could be measured approximately with a micrometer. We found that it was between 50 and 60 μm which corresponds to a 4–500 V minimal actuation voltage required. We made another movie showing droplet transport on this board and sometimes we could reach very fast and uncontrolled droplet movement which proves proper surface treatment — and a probable unwanted slope in the horizontal orientation, too. Even voltages above 1000 V did not lead to dielectric breakdown. For generating high voltages we used a high voltage supply of type 1.5M24-P1-WS from UltraVolt [8].

V. Further plans

In order to have a properly functioning device, we will have to assure that the surface of the chip is free of any kind of contamination and imperfection, i.e., we will have to guarantee sufficient smoothness for (almost) every board we spin PDMS onto. Then we will also need a controlling software and device by which the electrodes can be activated in a preprogrammed manner. We already have the C code framework and a programmable and USB attachable PIC18F87J50 microcontroller board from Microchip Technology and have done test runs with them. The main task will be the construction of a circuit which is to be controlled by this board and will have direct connection to the electrodes. A final step will be the attachment of an optical measuring device (essentially a laser diode–photodiode pair) that could determine the concentration of a chemical dissolved in water or a component of blood (sugar or some kind of protein) by absorbance of a small laser beam.

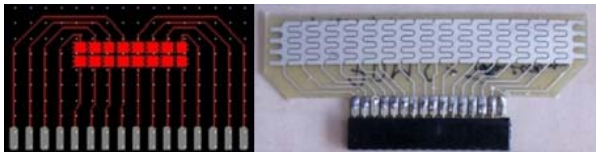


Figure 5. Two PCBs we used to experiment with the appropriate dielectric thickness. The one on the left has square-shaped electrodes with 2.5 mm side length and 200 mm separation gaps, on the right one the electrodes are 12 mm in width and 8 mm in length.

Acknowledgment

We are grateful to Dr. Attila Tihanyi for his technical support in designing the circuitry required for electrode activation and his assistance in programming the PICDEM Board. We also would like to thank András Laki for his work with the printed circuit boards and Péter Hajdu for making the EyeRIS camera available for us.

References

[1] V. Srinivasan, V. K. Pamula, M. G. Pollack and R. B. Fair, “Clinical Diagnostics on Human Whole Blood, Plasma, Serum, Urine, Saliva, Sweat, and Tears on a Digital Microfluidic Platform”, Proc. 7th Int’l Conf. Micro Total Analysis Systems (MicroTAS 03), Transducers Research Foundation, 2003, pp. 1287–1290.

[2] L. Li, H. Hu, H. Lin, and D. Ye, “Electrowetting of the blood droplet on the hydrophobic film of the EWOD chips”, Proceedings of the 2005 IEEE Engineering in Medecine and Biology 27th Annual Conference, Shanghai, China, 1–4 September, 2005.

[3] Y. H. Chang, G. B. Lee, F. C. Huang, Y. Y. Chen, and J. L. Lin, “Integrated polymerase chain reaction chips utilizing digital microfluidics”, Biomed. Microdevices, vol. 8, pp. 215–225, 2006.

[4] M. G. Pollack, P. Y. Paik, A. D. Shenderov, V. K. Pamula, F. S. Dietrich, R. B. and Fair, “Investigation of electrowetting-based microfluidics for real-time PCR applications”, Proceedings of the 2003 MicroTas Conference, Squaw-Valley, Ca, USA, 5–9 October, 2003.

[5] R. B. Fair, A. Khlystov, T. D. Tailor, V. Ivanov, R. D. Evans, P. B. Griffin, V. Srinivasan, V. K. Pamula, M. G. Pollack, J. Zhou, “Chemical and Biological Applications of Digital–Microfluidic Devices"”, Design & Test of Computers, IEEE, vol. 24, pp. 10–24, January-February 2007.

[6] Y. Zhao, and S. K. Cho, “Microparticle sampling by electrowetting-actuated droplet sweeping”, Lab Chip, vol. 6, pp. 137–144, 2006.

[7] J. Berthier, Microdrops and Digital Microfluidics, William Andrew, 2008.

[8] http://www.ultravolt.com/products/1-single-output-high-voltage-modules/27-microsize-micropower/58-m/

An integrated LOC hydrodynamic focuser with a CNN-based camera system for cell counting application

András Laki

(Supervisors: Dr. Kristóf Iván and Dr. Danilo Demarchi)
lakanjo@digitus.itk.ppke.hu

***Abstract*—A microfluidic analyzer system and a cell detection algorithm were developed to analyze biomedical fluids. The obtained microfluidic device is based on the integration of different fluidic systems: the SensoNor glass/silicon/glass multilayer technology and ThinXXS plastic slide technology. A hydrodynamic focuser was designed to sort and analyze cells/particles in a $100\mu m$ wide channel. The advantages of this Lab-On-a-Chip (LOC) structure are the easy interfaceability with electrodes and optical systems, biocompatibility and ability of optical analysis and morphologic recognition. The proposed CNN-based (Cellular Neural Network) algorithm is real-time and scalable. This constructed microfluidic and optical system is able to analyze and measure any biological liquid, which contain less than $10\mu m$ size particles or cells, and count the number of morphologically well-separated different elements in the focused liquid flow using image processing algorithms.**

I. Introduction

The present-day requirements of biomedical sciences are to construct reliable, efficient, and smart Total Analysis Systems (μTAS) in micro scale. These devices have to be small, effective - they must carry out complex procedures - and, in the meantime, they should be cheap and large-scale products. With miniaturization processes, not only the characteristic size of the product changed, but the price is also reduced. In biomedical engineering research we have to remember that our products need to be marketable and large-scale producible, they have to follow international regulations, and perform complex analyses to support diagnosis [1].

Several previous cell-counting realizations exist, which work with electrostatic field [2] [3], light diffraction [4] or optical analyzer [5] [6]. The specialities of our observation system are the real-time image processing of the CNN-based camera and the recognition or parallel counting of the morphologically different shaped cells/particles [7]. Following these requirements, a possible solution is to create an analysis system onto the surface of a microchip, defined as Lab-On-a-Chip (LOC) technology, to integrate the different analysis steps in a minimal space. To obtain a micro-Total-Analysis-Systems (μTAS), it is inevitable to work with fluids in the micro scale.

The aim of this present work is to construct an analysis microfluidic system for biomedical use and recognize or count the cells/paricles through optical detection. The designed chip with the core microfluidic system [1] is based on the SensoNor glass/silicon/glass multilayer microchip [8] and ThinXXS plastic slide technology (the complete system is shown in Fig.1).

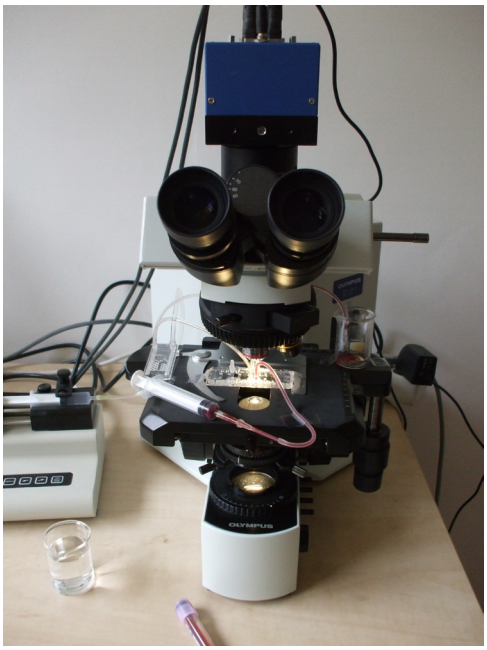


Fig. 1. The constructed biomedical fluid analyzer hydrodynamic focuser with CNN-based camera system. The flow velocity is controlled by precision syringe pump system.

II. The Hydrodynamic focuser for biological use

The hydrodynamic focusing effect is an efficient and simple microfluidic tool to guide the sample fluid and define the wide of the flow inside the output channel. The chosen materials are biocompatible and low cost. The microfluidic channels are created on silicon surface by Dry Reactive Ion Etching (DRIE) and closed with glass slides by anodic bonding [1] [8].

A. The hydrodynamic concept for cell counting

The designed microfluidic SensoNor multilayer microchips contain a rectangular cross-section hydrodynamic focusing geometry with 90° angle between the inlet channels. There are

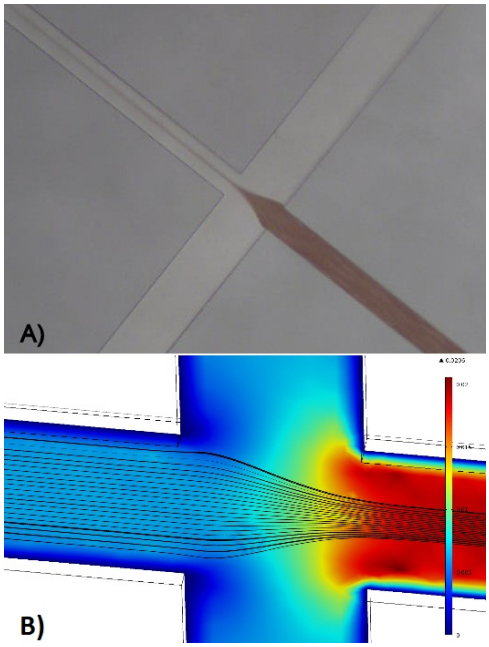


Fig. 2. Hydrodynamic focusing in micron scale.(a) Intravenous blood sample focused to $5-7\mu m$ wide stream inside the output channel. (b) Virtual simulation of the particle racing in velocity field inside the cross-section of the hydrodynamic focuser.

three inlets and one outlet with external-world connections. The four rectangular microchannels meet each other in the cross-section. The angles between the central channel and the two lateral inlets are $90°$.

The concept of the microchip was designed to create a one cell wide focused sample flow and the size of the output channel is $10\mu m$ deep and $50\mu m$ wide. The width of the two lateral channels are $100\mu m$, the depth is $10\mu m$ and the contact angle between them is $90°$. During the research work the microfluidic focuser was tested, first with paint and ink, after with biological liquids, like hemolyzed blood, and thirdly with native canine blood (Fig. 2(a)). With the hemolyzed blood the focuser follows the Lee's model [9]:

$$\frac{w_f}{w_o} = \frac{Q_i}{\gamma(Q_i + Q_{s1} + Q_{s2})} \tag{1}$$

where:
w_f is the width of the hydrodynamically focused stream (m),
w_o is the width of the output channel stream (m),
Q_i is the volumetric flow rate of the inlet channel (μl/min),
Q_{s1} is the volumetric flow rate of side channel 1 (μl/min),
Q_{s2} is the volumetric flow rate of side channel 2 (μl/min),
γ is the velocity ratio ($\overline{v_f}/\overline{v_o}$).

$$\gamma = \frac{\overline{v_f}}{\overline{v_o}} = \frac{\frac{2}{w_f}\int_0^{w_f/2}\overline{u}(y)dy}{\frac{2}{w_o}\int_0^{w_o/2}\overline{u}(y)dy} \tag{2}$$

The Lee's model was simulated also with particle tracing by COMSOL software for our application (Fig. 2(b)). The simulations were verified by experimental results (Fig. 2(a)).

B. CNN-based cell counting algorithm

The proposed CNN-based algorithm is able to count particles in the liquid flow using different image processing steps. The used EyeRIS camera is a special instrument, which contains a Qeye SIS (Smart Image Sensor). These sensors include signal processors with local connection. The main advantages of the FPP (Focal Plane Processor) system are the image recording and the image processing capabilities at the same place and real-time. The cell analyzer algorithm is executed by a 144x176 pixel resolution processor network and behind these pixels there are the CNN cells. The image processing functions are supported by 7 Local Analogical Memories (LAM) and 4 Local Digital Memories (LDM). The fist type of memory can store the image in grey scale, while the digital one in binary scale. The image recording time of the binary images can be as high as 10000 frames per second. The camera is able to record in HDR (High Dynamic Range) also in 80 dB scale.

The EyeRIS camera contains not only the sensor system, but also other hardware. The NIOS II soft core processor is the most important element after the sensor, which is based on FPGA technology. This component makes the connection with the PC, and communicates with the other devices. The camera also contains serial flash memory (64 Mb DDR2 type RAM). To help and accelerate the digital image processing functions a DICop (Digital Image CoProcessor) is also integrated, which is able to execute geometric transformations and PtP (Pixel-to-Pixel) transformations.

III. Results and Discussion

The image processing is a commonly used method to analyze and record movements. In the industry production these systems are also frequently used for quality control of the fast production steps. The main difficulty of these processes is the object-background segmentation, which are not simple methods to discriminate. In general the static object detection is insufficient because of the high variance of the background.

First of all the recorded image contains noises form the sensor and external cases. The elimination of these errors need correction, especially if the diameter of the objects are just few pixels. After the image filtration, which can help to eliminate this problem, the main goal is to determine the shape of objectives from the background. Previously many other articles were published in this field, but also important to deal with the speed of the different techniques. Our image processing steps are the following:

1) Image recognition
2) Gaussian filter
3) Global Threshold
4) Morphologic erosion
5) Morphologic dilatation
6) Morphologic centering
7) Cell/object counting

The proposed algorithm works with grey-scale and binary images and contains four main parts. First of all it starts

with the image recognition, continues the preprocessing part with filtration. Thirdly the algorithm perform the binary image processing steps, which start with the threshold measurements, and terminate with one pixel in the middle of recognized cells. Finally the cells/particles are counted from the result images.

A. Image recognition

This algorithm sequence performs the optical image acquisition that represents in the field of view of the Eye-RIS Vision System. During the measurement the integration time (also called exposure time) is highly correlated to the light intensity ($expTime = 0.7ms$). In addition, the acquired image can be amplified during the sensing process. The gain of this amplification process is also indicated as a parameter ($gain = 2$).

The output channel of the hydrodynamic focuser is shown on the sensed image (*SensedImg N*, showed on Fig. 3(a)). The camera has a limited 144x176 resolution, with 20x magnification objective, the work field is around $190\mu mx230\mu m$ and the blood cells are around 5-6 pixels. The technological velocity limit of this system, which comes from relation of the recognition time and the velocity of the liquid flow with particles, is around 0.560 mm/s. If the flow velocity is higher than this value the particle is faster than the image recognition ($T_{Rec} = T_{exposure} + T_{readout} + T_{movedata} = 0.7ms+16ms+322ms = 338.7ms$). In general our algorithm works with 0.020 mm/s flow velocity.

B. Preprocessing

The sensed image (*SensedImg N*, Fig. 3(a)) could be infected with different noises by the sensor or external causes. The proposed algorithm starts with a **Gaussian filter** to reduce the one pixel noise (*GaussianImg N*, Fig. 3(b)). In our case this preprocessing function implements a low-pass filter that emulates a Gaussian filter using the Resistive Grid module available in the SIS Q-Eye. The resistive grid module is extremely efficient in both speed and power consumption. The spatial bandwidth of the filter is specified as the sigma parameter of the equivalent Gaussian filter ($sigmaValue = 0.4$).

C. Image processing algorithm

The conversion from gray-scale (*GaussianImg N*) to binary image (*ThreshImg N*, Fig. 3(c)) is made by the **Global Threshold Value** (GTV). The histogram of the gray-scale image is not flattened, the values of the pixels are between 100 and 145, but anyway we can consider a stable a microfluidic system with fix light illumination ($TresholdValue = 115$). During this step the algorithm uses only one function, thus it is optimal in time, but not in quality. The fluctuation of the light can cause significant errors, if size of the noise is not just few pixels.

The **erosion** function on a binary image (*ErosionImg N*, Fig. 3(d)) eliminates or reduces the noise. Before this step the image is inverted because the following functions work with white objects on black background. Two main methods exist for image erosion. The first is to use a predefined constant that allows to select between 4-neighbor connection and 8-neighbor connection or use a 3x3 pattern that completely defines the structuring element. Our algorithm works on the first way with the 4-neighbor connection case and erases 1 pixels to open morphologically the objects and eliminate the one pixel mistakes.

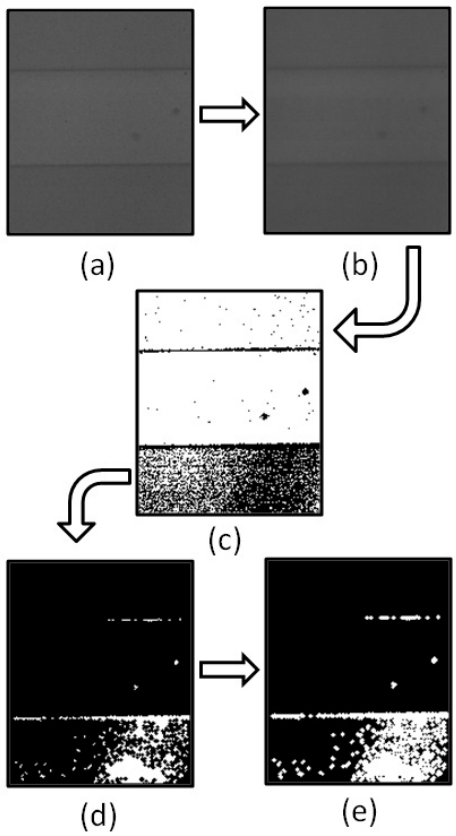


Fig. 3. Cell detection in the output channel of the hydrodynamic focuser. (a) Grey-scale sensed image from the cell flow (*SensedImg N*). (b) Gaussian filtration on the *SensedImg N* (*GaussianImg N*). (c) Binary image is the result of use of threshold (*ThreshImg N*). (d) The erosion function eliminate the noise from the image (*ErosionImg N*). (e) Dilatation fills the holes on the cells (*DilatationImg N*).

The erosion function erases not only the noise and mistakes, but also consumes pixels from the objects, which in the next step the algorithm compensates. The **dilatation** is complementary to the morphological closing, it dilates a binary image in which objects are white and the background is black. After the dilatation function the cells have the same diameter on the result image (*DilatationImg N*, Fig. 3(e)) like before the erosion. The second importance of dilatation is colligated to the next function.

The last step of the image precessing is the **centering**. This function gets the centroid positions of the objects (*CentroidImg N*, Fig.4(b)). The morphological centroid peels the image one pixel off as many times as indicated in an input parameter. In our case it iterates until no change occurs between iterations.

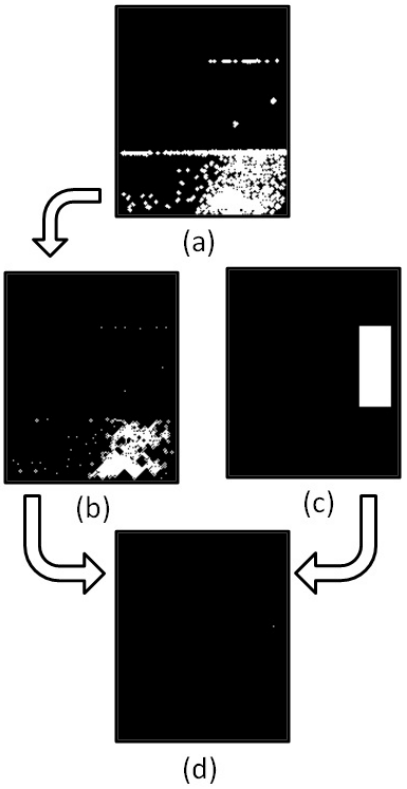


Fig. 4. Termination steps of the image processing. (a) Results of the dilatation function (*DilatationImg N*). (b) Centering function lets one pixel remain in the centroid positions of the objects (*CentroidImg N*). (c) Region-Of-Interest (ROI) window mask (*MaskImg*).(d) The final result image (*ResultImg N*) from the camera, just one pixel in the centroid positions of the focused cells inside the ROI area.

D. Counting Particles

The termination part of the algorithm counts the cells/particles inside the **Region-Of-Interest** (ROI), which is determined by a predefined binary mask (*MaskImg*, Fig.4(c)). The result image (*ResultImg N*) is generated from a **logical AND function** of the *CentroidImg N* and the *MaskImg*. The flow velocity is constant inside the output channel and in generally it is 0.020mm/s. If the flow velocity is fix, in that case also the waiting time ($T_{waiting} = 2370ms$) is well-known between two subsequent images (*ResultImg N*, *ResultImg N+1*). The number of the white pixels in the *ResultImg N* Images describes the number of the cells in the focused liquid flow. The efficiency of this algorithm was more than 90 percent.

IV. Conclusion

The main goals of this work were to realize a stable focusing device with the use of different microfluidic parts, combining SensoNor multilayer glass/silicon/glass microchip with ThinXXS polymeric microfluidic platform and to count the number of the cells in the focused flow with a CNN-based camera system. The microchannels are $50\mu m$ and $100\mu m$ wide, they are constructed in silicon by DRIE process. The hydrodynamic focuser was designed and tested by biomedical fluids: yeast solution, intravenous blood, hemolyzed blood, and parasite infected blood (Dirofilaria repens). The microfluidic channels are not transparent, the reflected light is used to analyze the liquid flow. The image flow is processed by CNN-based EyeRIS camera. The cell detector and counter algorithm works with more than 90 percent efficiency.

In the following steps the algorithm will be extended with special parts like the quality enhancement of the sensed image [10] and other important further measuements (viscosity of the fluidics, dynamic velocity observation) [11].

Acknowledgment

We like to thank the microBUILDER consortium who supported this work and also the support of the Hungarian National Research Fund OTKA (grant number: PD73653) is kindly acknowledged. The biological samples were kindly offered by the Szent István University. Finally also we would like to thank to Alessandro Sanginario, Niccolo Piacentini and Daniel Lapadatu for the kindly help of the microchip design and thank to Olga Jacsó for the blood samples and the Péter Hajdu's fundamental works.

References

[1] A. Laki, I. Rattalino, A. Sanginario, N. Piacentini, K. Ivan, D. Lapadatu, J. Taylor, D. Demarchi and P. Civera, *An integrated and mixed technology LOC hydrodynamic focuser for cell counting application*, Biomedical Circuits and Systems Conference (BioCAS), 2010 IEEE, pp.74-77, 2010.

[2] K. Roberts, M. Parmeswaran, M. Moore and R.S. Muller *A silicon microfabricated aperture for counting cells using the aperture impedance technique*, Engineering Solutions for the Next Millennium. 1999 IEEE Canadian Conference on Electrical and Computer Engineering, pp.1668-1673, 1999.

[3] M. L. Shuler, R. Aris and H. M. Tsuchiya, *Hydrodynamic Focusing and Electronic Cell-Sizing Techniques*, Appl. Environ. Microbiol., pp.384-388, 1972.

[4] L. Cui, T. Zhang and H. Morgan, *Optical particle detection integrated in a dielectrophoretic lab-on-a-chip" in Journal of Micromechanics and Microengineering*, 12, 7-12, 2002.

[5] Y. Fainman, L. Lee, D. Psaltis and C. Yang, *Optofluidics: Fundamentals, Devices, and Applications*, McGraw Hill Professional, 2009.

[6] D. Di Carlo, D. Irimia, R. G. Tompkins and M. Toner, *Continuous inertial focusing, ordering, and separation of particles in microchannels*, Proceedings of the National Academy of Sciences, 104, 48, pp. 18892-18897, 2007.

[7] F. Corinto, F. Biey and M. Gilli,*Non-linear coupled CNN models for multiscale image analysis*, International Journal of Circuit Theory and Applications, pp.77-88, 2006.

[8] D. Lapadatu, *MultiMEMS Design handbook version 4.2*, SensoNor Technologies, 2009.

[9] G. B. Lee, C. C. Chang, S. B. Huang and R. J. Yang, *The hydrodynamic focusing effect inside rectangular microchannels* in Journal of Micromechanics and Microengineering, 16, p. 1024-1032, 2006.

[10] M. Brendel and T. Roska, *Adaptive image sensing and enhancement using the cellular neural network universal machine*, International Journal of Circuit Theory and Applications, pp.287-312, 2002.

[11] F. Sapuppo, M. Bucolo, M. Intaglietta, L. Fortuna and P. Arena, *A cellular nonlinear network: real-time technology for the analysis of microfluidic phenomena in blood vessels*, Nanotechnology, pp.54-63, 2006.

Nonparametric High Resolution Image Segmentation

Balázs Varga
(Supervisor: Dr. Kristóf Karacs)
varga.balazs@itk.ppke.hu

***Abstract*—In this paper, we present a fast and effective method of image segmentation. Our design follows the bottom-up approach: first, the image is decomposed by nonparametric clustering; then, similar classes are joined by a merging algorithm that uses color, and adjacency information to obtain consistent image content. The core of the segmenter is a parallel version of the mean shift algorithm that works simultaneously on multiple feature space kernels. Our system was implemented on a many-core GPGPU platform in order to observe the performance gain of the data parallel construction. Segmentation accuracy has been evaluated on a public benchmark and has proven to perform well among other data-driven algorithms. Numerical analysis confirmed that the segmentation speed of the parallel algorithm improves as the number of utilized processors is increased, which indicates the scalability of the scheme. This improvement was also observed on real life, high resolution images.**

***Keywords*-High resolution imaging; parallel processing; image segmentation; multispectral imaging; computer vision**

I. Introduction

Our interest in this paper is the task of image segmentation in the range of quad-extended, and hyper-extended graphics arrays. We have designed, implemented and numerically evaluated a segmentation framework that works in a data parallel way, and which can therefore efficiently utilize many-core mass processing environments. The structure of the framework follows the bottom-up paradigm and can be divided into two main sections. During the first, clustering step, the image is decomposed into sub-clusters. The core of this step is based on the mean shift segmentation algorithm, which we embedded into a parallel environment, allowing it to run multiple kernels simultaneously. The second step is a cluster merging procedure, which joins sub-clusters that are adequately similar in terms of color and neighborhood consistency. The framework has been implemented on a GPGPU platform and numerical evaluation was run on miscellaneous devices with different numbers of stream processors to demonstrate algorithmic scaling of the clustering step and speedup in segmentation performance.

A. The Mean Shift Segmentation Scheme

The basic principles of the mean shift algorithm were published by Fukunaga and Hostetler [1] in 1975. These were then enhanced by Cheng [2] in 1995 and refined further by Comaniciu and Meer [3] in 1999. The mean shift technique considers its feature space as an empirical probability density function. A local maximum of this function (namely, a region over which it is highly populated) is called a mode. Mode calculation is formulated as an iterative scheme of mean calculation, which takes a certain number of feature points and calculates their weighted mean value by using a kernel function. The algorithm can handle various different types of feature spaces, such as edge maps or texture, but in most cases of still image segmentation, a composite feature space consisting of topographical (*spatial*) and color (*range*) information is used. Consequently, each feature point in this space is represented by a $\chi = (x_r; x_s)$ 5D vector, which consists of the 2D position $x_s = (x, y)$ of the corresponding pixel in the spatial lattice, and its 3D color value x_r in the applied color space (in the current paper, we use $x_r = (Y, Cb, Cr)$ coordinates).

The iterative scheme for the calculation of a mode is as follows: let χ_i and z_i be the 5D input and output points in the joint feature space for all $i \in [1, n]$, with *n* being the number of pixels in color image *I*. Then, for each *i*

1) Initialize $\chi_i^{k=0}$ with the original pixel value and position;
2) Compute a new weighted mean position using the iterative formula

$$\chi_i^{k+1} = \frac{\sum_{j=1}^{n} \chi_j g\left(\left\|\frac{x_{r,j} - x_{r,i}^k}{h_r}\right\|^2\right) g\left(\left\|\frac{x_{s,j} - x_{s,i}^k}{h_s}\right\|^2\right)}{\sum_{j=1}^{n} g\left(\left\|\frac{x_{r,j} - x_{r,i}^k}{h_r}\right\|^2\right) g\left(\left\|\frac{x_{s,j} - x_{s,i}^k}{h_s}\right\|^2\right)}, \quad (1)$$

where g denotes the Gaussian kernel function, with h_s and h_r being the spatial and range bandwidth parameters respectively, until

$$\left\|\chi_i^{k+1} - \chi_i^k\right\| < thresh \quad (2)$$

that is, the *shift* of the *mean* positions (effectively a vector length) falls under a given threshold (referred to as *saturation*).
3) Allocate $z_i = \chi_i^{k+1}$.

Clusters are formulated in such a way that those z_i modes that are adequately close to each other are concatenated, and all elements in the cluster inherit the color of the contracted mode, resulting in a non-overlapping clustering of the input image.

Despite the listed advantages, the algorithm has a notable downside. The naïve version, as described above comes with a computational complexity of $\mathcal{O}(n^2)$. The fact that runtime

is quadratically proportional to the number of pixels makes it slow, especially when working with high definition images.

Several techniques were proposed in the past to speed up the procedure, such as [4], [5], however due to the limitation of this report's extent, we are unable to discuss them in detail, but they are evaluated in [6] among several other acceleration alternatives.

II. Computational Method

Our framework is devoted to accelerate the segmentation speed of the mean shift algorithm with a major focus on its performance on high resolution images. We designed and utilized three main speed up strategies.

The first aimed to reduce the computational complexity via recursive sampling of the feature space. Practically, this means that after a mean shift kernel has been initialized from χ_i feature space point, and the iterative scheme described above led to saturation, an χ_j feature space element is assigned to $z_i = \chi_i^{k+1} = (x_{r,i}^{k+1}; x_{s,i}^{k+1})$ mode if, and only if

$$\|x_{s,j} - x_{s,i}^{l}\| < h_s, \tag{3}$$

and

$$\|x_{r,j} - x_{r,i}^{l}\| < h_r, \tag{4}$$

where $l \in [0, k+1]$ denotes the mean shift iterations. In case a pixel is covered by more than one kernel, it is associated to the one with the most similar color. If unclustered pixels remain after the pixel-cluster assignment, resampling is done in the joint feature space, and new mean shift kernels are initialized in those regions, in which most unclustered elements reside.

The second strategy was, to gain speedup through the parallel inner structure of the segmentation. Parallel inner structure in a nutshell means, that more than one kernel is initiated at a resampling iteration, and all of their subsequent mean shift steps are calculated simultaneously until saturation.

The third acceleration technique was, to reduce the number of mean shift iterations by decreasing the number of saturated kernels required for termination (referred to as abridging). Among other issues detailed in [6], the parallel implementation induces an important change in the mean shift scheme, as in this case it is not feasible to isolate saturated modes and replace them with new kernels in a "hot swap" way, due to the characteristics of block processing. On the other hand the ratio of kernel saturation follows an exponential pattern; such that a remarkable fraction of the kernels saturates in the first few shifting steps, so that in their case, each additional iteration is superfluous. This phenomenon is suppressed by the abridging technique, which is controlled by a single constant called the *abridging parameter* $A \in [0, 1]$. It specifies the minimum proportion of kernels that is required to saturate before a new resampling step is started. thus it gives us a simple tool to terminate the mode seeking procedure after a reasonable amount of steps. The practical effect of the technique was studied with subject to numerous aspects (please refer to [6] for details).

Fig. 1 reveals the flowchart of the segmentation framework.

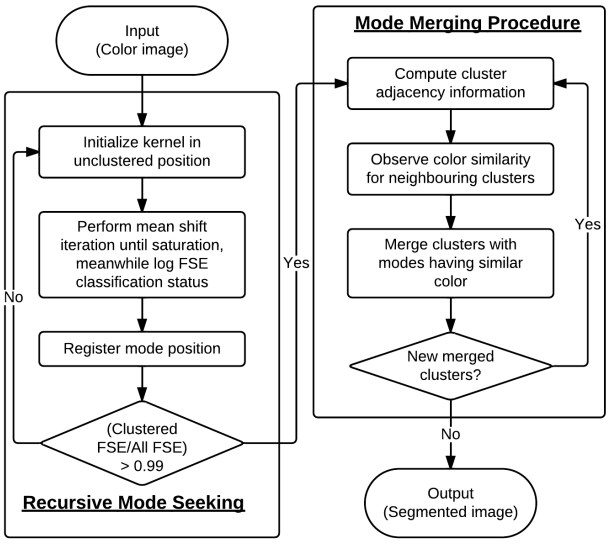


Fig. 1. Flowchart of the segmentation framework. The result of recursive mode seeking procedure is a clustered output that is an over-segmented version of the input image. The step of mode seeking is therefore succeeded by the merging step that concatenates similar clusters such that a merged output is obtained. The term FSE refers to feature space element.

III. Experimental Design

A. Hardware Specifications

The measurements were performed on five nVidia GPGPUs with various characteristics. As a reference, the framework was also tested on a single PC equipped with 4GB RAM and an Intel Core i7-920 processor clocked at 2.66GHz, running Debian Linux. The technical specifications of the hardware are summarized in Table I. Note that in the case of the nVidia S1070, only a single GPU was utilized (for this reason it is referred later on as S1070SG).

TABLE I
Parameters of the used GPGPU devices.

Device name	No. of stream processors	Clock frequency	Device memory	Compute capability
8800GT	112	1500 MHz	1024 MB	1.1
GTX280	240	1296 MHz	1024 MB	1.3
S1070SG	240	1440 MHz	4096 MB	1.3
C2050	448	1500 MHz	3072 MB	2.0
GTX580	512	1544 MHz	1536 MB	2.0

B. Measurement Specifications

In the case of the scaling and timing experiments, the measurements were made on five different image sizes. The naming conventions and corresponding resolutions are summarized in Table II.

C. Quality Measurement Design

For output quality analysis, we used the Berkeley Segmentation Dataset and Benchmark [7] in order to provide comparable quantitative results. The "test" set consisting of 100 pictures was segmented multiple times using the same

TABLE II
NAMING CONVENTION AND RESOLUTION DATA OF THE IMAGES USED FOR THE TIMING AND SCALING MEASUREMENTS.

Name of extended graphics array	Abbreviation	Resolution	Resolution in megapixels (MP)
Wide Quad	WQXGA	2560×1600	4.1 MP
Wide Quad Super	WQSXGA	3200×2048	6.6 MP
Wide Quad Ultra	WQUXGA	3840×2400	9.2 MP
Hexadecatuple	HXGA	4096×3072	12.6 MP
Wide Hexadecatuple	WHXGA	5120×3200	16.4 MP

parametrization for each image in a run. Three parameters were alternated among two consecutive runs: h_r taking values between 0.02 and 0.05, h_s with values in the interval of 0.02 and 0.05, both utilizing a 0.01 stepsize, and the abridging parameter ranging from 0.4 to 1.0 with a stepsize of 0.2. In each case, the segmenter was started with 100 initial kernels, and in every resampling iteration 100 additional kernels were utilized.

D. Timing Measurement Design

Timing measurements aimed at registering the runtime of the algorithm on high resolution real life images. We formulated an image corpus consisting of 15 high quality images that were segmented in five different resolutions, using the parameter settings "speed" and "quality", obtained during the quality measurements (see subsection IV-A). In each case, the segmenter was started with 10 initial kernels, and in every resampling iteration, 10 additional kernels were utilized.

IV. RESULTS

A. Quality Results

As a result of alternating h_r, h_s and the abridging parameter, the framework was run with 64 different parametric configurations for each image of the 100 image test corpus.

The obtained average F-measure values for the different bandwidths and abridging parameters are displayed in Fig. 2.

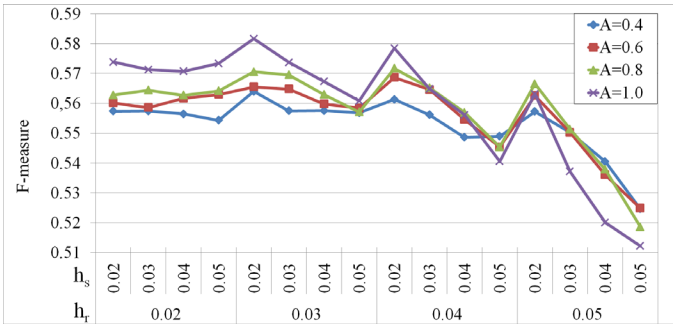


Fig. 2. F-measure values obtained for the different parametrizations of the segmentation framework. h_s and h_r denote the spatial and range kernel bandwidths respectively, A values stand for different abridging constants.

The highest F-measure value was 0.5816 for parameters $h_r = 0.03$ and $h_s = 0.02$ without any abridging, which fits in well among purely data-driven solutions [8]. It can be observed on Fig. 2 that the output quality remained fairly consistent when relatively small bandwidths were selected. The system is more robust to changes made to the spatial bandwidth, while the effect of a high range bandwidth parameter decreases output quality. As one may expect, abridging has a negative effect on quality, but it can be seen that for certain parameter selection (namely, for $h_s \in [0.02, 0.03]$ and $h_r \in [0.03, 0.04]$) even an abridge level of 0.6 results in acceptable quality. Based on these results, we selected two parametric settings for the timing measurements:

1) the **Quality Setting** was selected to be $h_r, h_s, A = (0.03, 0.02, 1)$, $(F = 0.581)$ while
2) the **Speed Setting** was selected to be $h_r, h_s, A = (0.04, 0.03, 0.6)$, $(F = 0.565)$.

B. Runtime Results

The average running times measured on the 15 image corpus are summarized in Fig. III using the speed setting and the quality setting.

TABLE III
RUNTIME VALUES OF THE ALGORITHM RUN ON IMAGES WITH DIFFERENT SIZES USING FIVE DIFFERENT GPGPUS AND THE CPU AS THE REFERENCE. RESULTS USING THE SPEED SETTING (TOP HALF) AND THE QUALITY SETTING (BOTTOM HALF) ARE DISPLAYED IN SECONDS. EACH MEASUREMENT REPRESENTS AN AVERAGE VALUE OBTAINED FROM RUNNING THE ALGORITHM ON 15 IMAGES. "N/A" MEASUREMENT VALUES INDICATE AN OVERRUN IN DEVICE MEMORY.

Speed Setting					
	WQXGA	**WQSXGA**	**WQUXGA**	**HXGA**	**WHXGA**
I7_920	82.95	127.92	163.05	220.17	261.52
8800GT	23.76	38.02	53.65	88.37	N/A
GTX280	7.99	11.48	16.55	28.64	N/A
S1070SG	8.14	10.97	14.16	19.34	23.39
C2050	7.79	9.24	11.25	15.78	18.46
GTX580	6.04	7.30	9.23	14.45	18.23
Quality Setting					
	WQXGA	**WQSXGA**	**WQUXGA**	**HXGA**	**WHXGA**
I7_920	160.46	224.63	359.60	471.86	565.09
8800GT	53.56	78.67	110.92	122.69	N/A
GTX280	23.95	29.96	37.26	39.80	N/A
S1070SG	27.27	27.75	30.84	38.16	42.80
C2050	19.32	23.62	28.85	33.73	41.38
GTX580	17.89	19.57	21.76	20.51	33.37

In the case of the measurements made on the GPGPUs, the displayed values include all operations and memory accesses that have been performed in order to obtain the merged output image.

When using either the GTX580 or the C2050, the average time demand for segmenting a 16 megapixel image was just above 18 seconds in the case of the speed setting, and a bit more than 33 seconds on the GTX580 using the quality setting. Compared to the runtimes of the CPU using the same, 16 megapixel setup as the GTX580, this means an acceleration of 16.93 times in the case of the quality setting, and an acceleration of 14.34 times in the case of the speed setting.

Fig. 3 shows an example of the high quality input images from the 15 image segmentation corpus and the segmented output before and after the merging procedure.

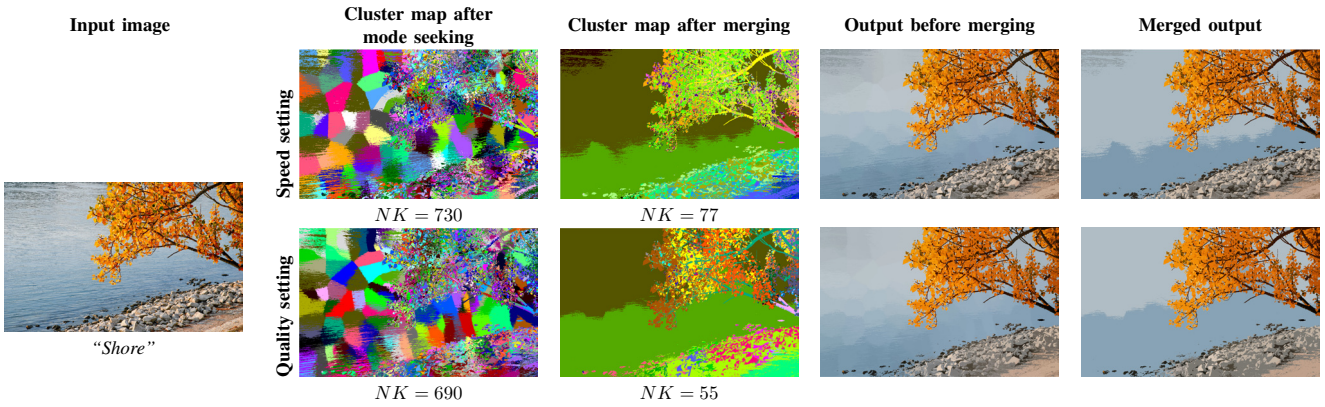
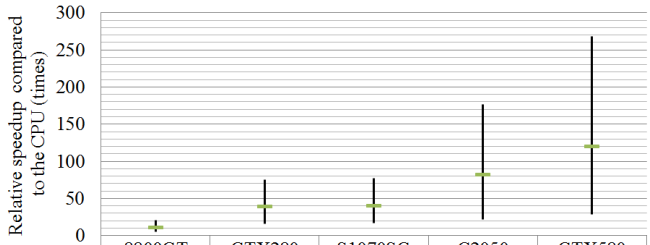


Fig. 3. Three segmentation examples from the 15 image corpus. For the sake of better visibility, the extent of the clusters is also displayed in the form of cluster maps before and after the merging procedure. NK refers to the number of clusters.

C. Scaling Results

Finally, we investigated the runtime of calculating the mean shift iteration on the different devices in proportion to the runtime of the same task measured on the CPU. Fig. 4 displays an overview of the speedup that is obtained by taking into account all of the different parametrizations of $h_s \in [0.02, 0.05]$ and the number of kernels being 1, 10 and 20.



	8800GT	GTX280	S1070SG	C2050	GTX580
Maximum speedup	20.90	75.71	77.14	176.61	268.59
Minimim speedup	5.09	16.24	16.46	21.54	28.17
Average speedup	11.35	38.88	40.00	82.34	120.06

Fig. 4. Speedup results obtained for different devices by pairwise comparison to the CPU. The basis of comparison were the runtime values representing the time demand of calculating new position(s) of mean(s) with all combinations of $h_s \in [0.02, 0.05]$ with number of kernels being 1, 10 and 20.

As one may expect, the fastest performance was observed on the GTX580: compared to the CPU, the speed increase was greater than 28 for all parameter settings, with an average speedup of around 120. One may ask why the speedup of the mean shift iteration differs from the overall speedup of the framework. The answer to this question is that in the case of the former, only arithmetic operations are involved, so that these results represent more closely the speed of the GPGPU processing units. In contrast, the overall speedup – with all the data transfers, memory read and write operations that are involved – represent the integrated performance of the device.

V. Conclusion

The details and design of an image segmentation framework have been presented in this paper. The core of the system is given by the parallel extension of the mean shift algorithm, which we accelerated by utilizing a recursive sampling scheme and an abridging technique. The framework was implemented on a many-core computation platform, and a common segmentation benchmark was used to evaluate the output quality, and to demonstrate its robustness concerning parameter selection. Segmentation performance was analyzed on different high resolution real life images, using five GPGPUs with miscellaneous specifications. The runtime of a parallel mean shift iteration was measured on the different devices in order to observe the scaling of the data parallel scheme. The algorithm has proven to work fast and to provide good quality outputs.

References

[1] K. Fukunaga and L. Hostetler, "The estimation of the gradient of a density function, with applications in pattern recognition," *IEEE Trans. Inf. Theory*, vol. 21, no. 1, pp. 32–40, 1975.

[2] Y. Cheng, "Mean shift, mode seeking, and clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 17, no. 8, pp. 790–799, 1995.

[3] D. Comaniciu and P. Meer, "Mean shift analysis and applications," in *Proc. 7th IEEE Int. Computer Vision Conf. The*, vol. 2, 1999, pp. 1197–1203.

[4] M. A. Carreira-Perpiñán, "Acceleration strategies for gaussian mean-shift image segmentation," in *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition*, vol. 1, 2006, pp. 1160–1167.

[5] D. Freedman and P. Kisilev, "Fast mean shift by compact density representation," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition CVPR 2009*, 2009, pp. 1818–1825.

[6] B. Varga and K. Karacs, "High resolution image segmentation using fully parallel mean shift," *EURASIP Journal on Applied Signal Processing*, 2011, to Appear.

[7] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proc. 8th Int'l Conf. Computer Vision*, vol. 2, July 2001, pp. 416–423.

[8] D. R. Martin, C. C. Fowlkes, and J. Malik, "Learning to detect natural image boundaries using local brightness, color, and texture cues," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 5, pp. 530–549, 2004.

Simulation and Measurement of Memristor Crossbar Devices

Balázs Jákli
(Supervisors: Dr. György Cserey and Dr. László Négyessy)
jakli.balazs@itk.ppke.hu

***Abstract*—This paper introduce a simulation and a measurement method of a memristor crossbar device. I designed a PCB memristor package and the appropriate measurement board. Technical details of these circuits are presented. Cellular like topology of this crossbar device can provide high density and local connectivity, which later can be used to store the state of an artificial neural network in a prosthesis. A memristor simulation module for PSPICE circuit simulator is presented, based on a previous model.**

***Keywords*-memristor, artificial neural network, prosthesis.**

I. Introduction

In 1971, Professor Leon O. Chua suggested [1] that due to functional symmetries between the three fundamental circuit elements – the resistor, the inductor and the capacitor – a fourth one should exist. Due to its unique properties, he named the missing element as memristor (memory resistor). In 2008, members of the HP Labs published [2] the successful realization of a nano-sized electrical device, of which overall behavior was perfectly explained by the memristor theory. The existing publications using memristor at the area of programmable logic [3], signal processing [4], control systems [5] and neural networks [6] confirmed the significance of this discovery. My PhD work is based around an artificial neural network, which can be trained to adaptively control a human hand prosthesis. This application requires a large number of neurons, so a compact representation of the connections – based on a memristor – would be beneficial.

A previous publication [7] presented a SPICE memristor model, which successfully simulated the newly introduced memristor realization, and could be utilized at electrical circuit designs. The computation of this memristor macro-model is relatively fast, an average personal computer can simulate about 100 devices at a time. It is also stable, if the parameters are in the permitted range. Comparing these simulation results to real hardware devices [11], [12], generally simulations show a more clear behavior which hardly can fit to the hardware results.

The memristor can give interesting new dimensions to my work with artificial neural networks. But as long as the memristor is simulated in SPICE on a PC, and the rest of the neural network is running on FPGAs in the prosthesis, utilizing memristors leads to large overhead. Utilizing a memristor chip next to the FPGA could solve this problem, integrating the whole system on one printed circuit board. This is the reason I try to implement an operational memristor read-write circuit board.

II. Memristor crossbar network

I worked with an experimental memristor chip, which contained a memristor crossbar network. Its size is 5x5 mm. The chip has three main layers, a golden layer for wiring and two special layers for memristive behavior. The rows and columns of its crossbar topology are made of gold and all of the crossings of the rows and columns a memristor can be found. This method can provide the maximal number of memristors in a given area. Moreover because there is only one wiring layer, therefore theoretically more memristor layers can be placed on each other. The experimental chip is for testing purposes, therefore the memristors are relatively far from each other. Using this technology for higher density memristor arrays, more than 10000 memristors could be implemented on a chip of this size.

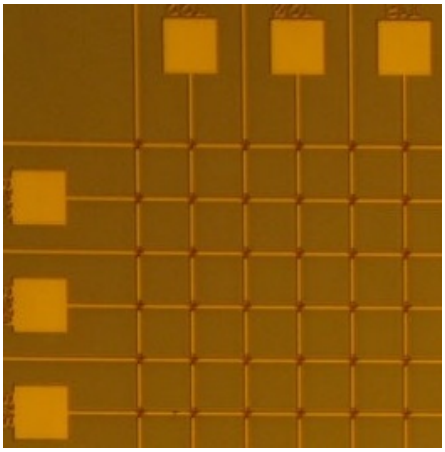


Fig. 1. Photo of a section of the memristor chip device.

The memristor chip was manufactured in two different architectures *[Fig. 2.]* There is a low density diagonal layout with 23 devices, and a higher density crossbar architecture with 23x23 devices. The former one is easier to control because of the lack of crosstalk between the neighboring devices, and the latter one has much more capacity. The designed measurement board was prepared for both architectures.

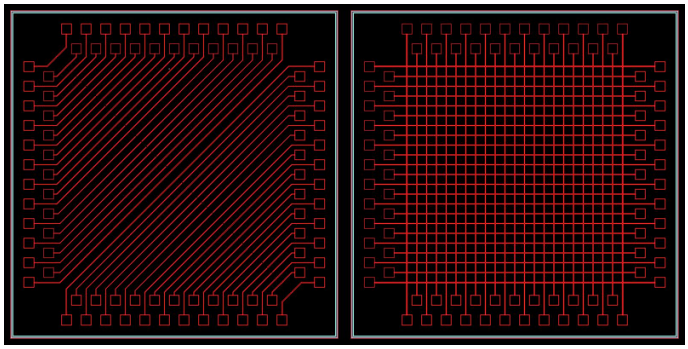


Fig. 2. The diagonal (left) and the crossbar (right) layout of the memristor chips. *[T. Prodromakis (Imperial College, London)]*

III. PCB MEMRISTOR PACKAGE AND MEASUREMENT BOARD

In order to place the memristor chips into complex electronic circuits, we developed and built a pin grid array type of package for the 92 pad chips. The printed circuit board is manufactured with a 4 mil accurate technology, with chemical gold cover layer. The substrate is fixed on an 8x8mm pad at the center of the package. In order to accurately measure some characteristics of the memristors, the substrate has a dedicated terminal to connect it to the reference voltage. The 92 golden pads of the chip were bonded on the small surface terminals along the center pad. The connections are conducted from here to the precision pin terminal arrays along the PCB edge.

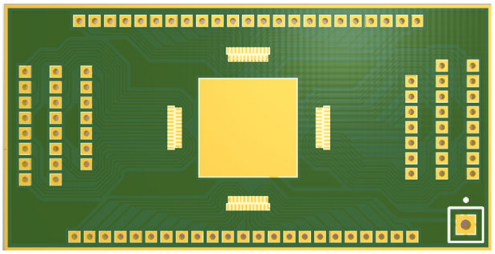


Fig. 3. The 92 pin PCB package of the memristor chips

The bonding was accomplished with an ultrasonic bonding machine using aluminium wires. In order to safely and easily bond the thin wires, the pads of the chip and the pads of the PCB were perfectly synchronized, so the aluminium wires are running parallel from the memristor chip to the PCB.

The bonded chips were covered with a layer of silicone in order to physically protect the connections.

Due to the rigid requirements of determining the characteristics of proprietary devices, we developed an automatic measuring instrument to correctly find out the values of the different memristor devices. To read out the values set in the memristors, we only need to measure the resistivity of the device. Due to the physical structure of the memristor, we need to measure it quickly, because flowing current through the device alters its state, messing up the proper measurement. The scheme of the measurement is on the following image:

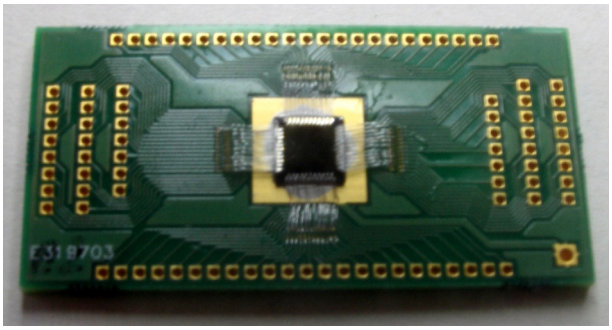


Fig. 4. Bonded chip on the PCB package

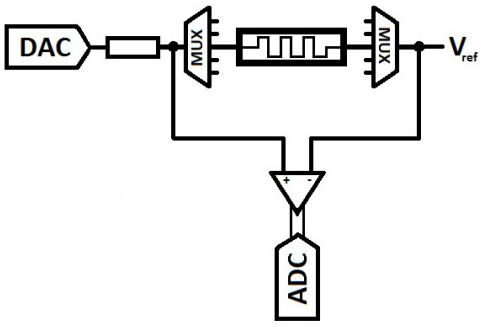


Fig. 5. The circuit schematic of the measurement process

The Digital-To-Analog converter device (DAC from now), creates a short pulse of current that flows through the memristor. We use a differential Analog-To-Digital converter (ADC) to accurately measure the voltage drop on the memristor. By knowing the voltage and current values, the actual resistivity of the device can be calculated simply by Ohm's law. Due to the proprietary nature of the devices under measurement, we cannot know for sure how much energy it can dissipate without damage. So we use a series resistor to protect the memristor from over current. This is a fixed resistor on the measurement board, which can be changed according to the changing parameters of the fabricated chips.

The Digital-To-Analog Converter is Linear Technology's 16 bit LTC2642 device. This DAC is connected to a precision operational amplifier which boosts its power to the required range. This section can be powered from internal 5V power, or through external V- and V+ pins. The latter option can be used to quickly adopt the measurement unit to the required power needs with any laboratory power supply.

The other end of the memristor is connected to a reference voltage, which has the value of ((V+) + (V-)) / 2. The reference is in the middle of V+ and V-, so by changing the DAC output voltage between V+ and V-, we can flow current both ways through the memristor devices. With the direction of the current, we can increase or decrease the resistivity of the memristor chip. The Analog-To-Digital Converter is Linear Technology's LTC2393 device. Its differential input is connected to the two ends of the memristor devices with a precision operational amplifier stage for better isolation. The

ADC is capable of 16 bit sampling with 1 Msps speed.

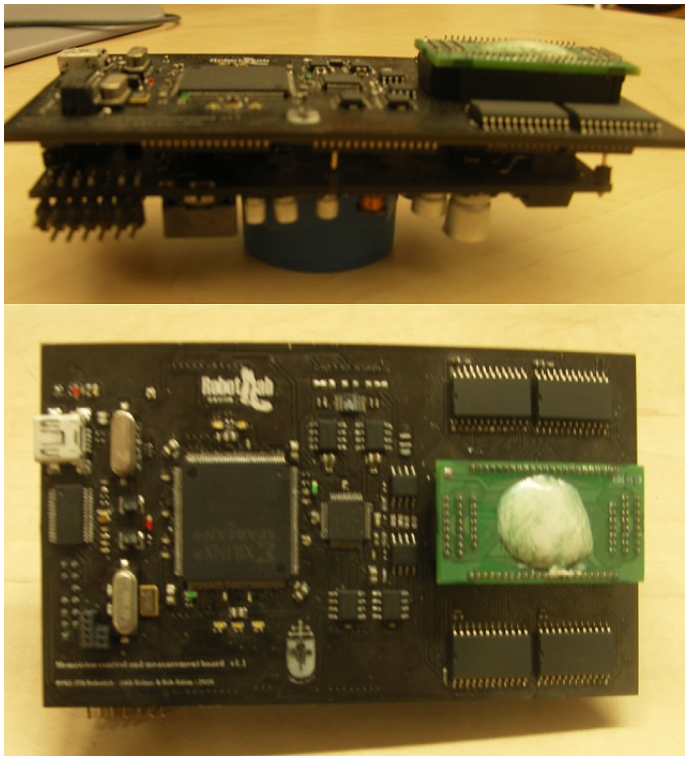


Fig. 6. The measurement board, with the memristor package at the right side

Because we have to measure multiple memristors on the chip, we connect the memristors to the ADC endpoints with multiplexer and demultiplexer stages. So the board is capable of controlling and measuring a 23x23 memristor array.

The ADC and DAC units and the multiplexers are controlled by a Xilinx Spartan3A FPGA device. The FPGA is programmed by an Atmel ATmega microcontroller at startup with the bit code supplied by a computer through USB connection. The FPGA also has a dedicated USB connection to quickly transfer data to the controlling computer. The Spartan has an external 32 bit wide SRAM memory block to store the data temporarily. This setup is also capable of preprocessing the measured data before submitting it to the computer.

Due to technical problems on the memristor chip, unfortunately we have not been able to complete clear real device measurements with the FPGA board yet. The yield of the devices on the chip was only around 50%. The working devices had large differences between their parameters. The life-span of the memristors were around 3-4 read-write cycles, with the hysteresis loop decaying in every cycle *[Fig. 9]*. The continous linearization of the hysteresis loop grounded every attempt to use the memristor as a memory. The measurements in the laboratory although revealed some design errors at the masking process and some mem-capacitive behavior as well. The next generation of the chip is under design, and I hope I can measure a working memristor chip in the near future.

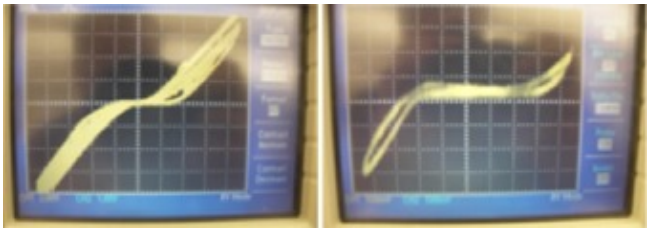


Fig. 7. Hysteresis loops of the memristor devices

IV. Simulation of the memristor device

Due to the lack of working hardware memristors, a simulation method was introduced in PSPICE circuit simulator. Previous implementations of the memristor SPICE simulator [7] used the ngSPICE engine, which runs under UNIX/Linux systems. Because most of the electronic design tools are running under Windows system, this model had to be transformed into a PSPICE model.

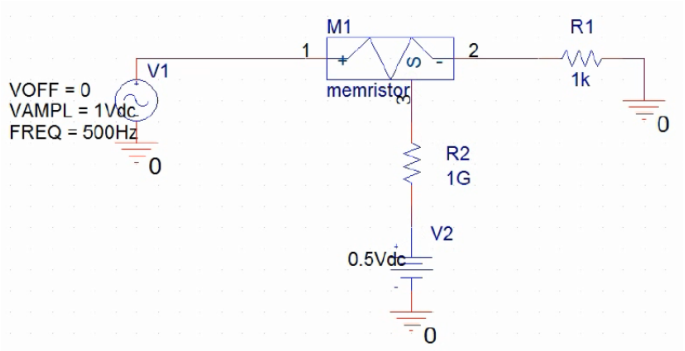


Fig. 8. Schematic circuit of the memristor simulation

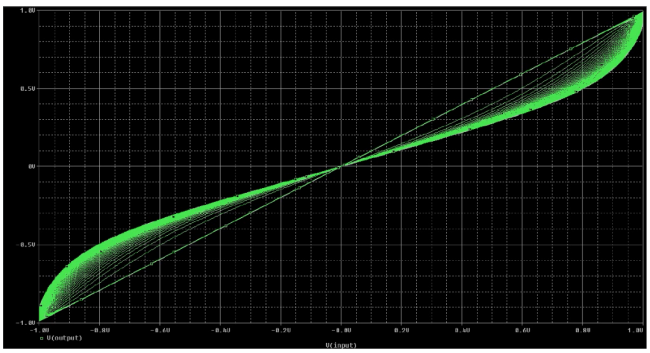


Fig. 9. Hysteresis loop with initial transient

This model was also used for teaching purposes. An artificial neural network now can be implemented with this model, using the PSPICE simulator to implement the memristors.

V. Conclusion

In order to measure the memristor crossbar device, we designed a PCB memristor package and appropriate measurement board. Unfortunately, we could not complete our

hardware measurements. We designed and implemented a simulation software based on earlier SPICE memristor model. Future plans contains a memristor simulation with an FPGA, to dispose of the overhead of PC simulation.

With the functional neural network, a control mechanism will be implemented, to control a simulated hand prosthesis.

Most of this work was published at a technical conference in Athens [13].

REFERENCES

[1] L. Chua, "Memristor-The missing circuit element," *Circuits and Systems, IEEE Transactions on [legacy, pre-1988]*, vol. 18, no. 5, pp. 507–519, 1971.

[2] D.B. Strukov, G.S. Snider, D.R. Stewart, and R.S. Williams, "The missing memristor found," *NATURE*, vol. 453, no. 7191, pp. 80, May 2008.

[3] G.S. Snider, "Architecture and methods for computing with reconfigurable resistor crossbar," Apr. 10 2007, US Patent No. US 7,203,789.

[4] B.L. Mouttet, "Programmable corssbar signal processor," Nov. 27 2007, US Patent No. US 7,302,513.

[5] B.L. Mouttet, "Operational amplifier with resistance switch crossbar feedback," Dec. 11 2008, US Patent App. US 2008/0307151.

[6] G.S. Snider, "Molecular-junction-nanowire-crossbar-based neural network," Apr. 15 2008, US Patent No. US 7,359,888.

[7] A. Rak and G. Cserey, "Macromodeling of the Memristor in SPICE," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 29, no. 4, pp. 632–636, 2010.

[8] Leon O. Chua and Tamás Roska, "The CNN paradigm," *IEEE Transactions on Circuits and Systems*, vol. 40, pp. 147–156, 1993.

[9] Tamás Roska and Leon O. Chua, "The CNN Universal Machine," *IEEE Transactions on Circuits and Systems*, vol. 40, pp. 163–173, 1993.

[10] Tamás Roska, "Computational and computer complexity of analogic cellular wave computers," in *Proceedings of the 7th IEEE International Workshop on Cellular Neural Networks and their Applications, CNNA 2002*, Frankfurt, Germany, July 2002, pp. 323–335.

[11] T. Prodromakis, K. Michelakis, and C. Toumazou, "Fabrication and electrical characteristics of memristors with TiO2/TiO2+x active layers," in *Circuits and Systems (ISCAS), Proceedings of 2010 IEEE International Symposium on*. IEEE, 2010, pp. 1520–1522.

[12] T. Prodromakis, K. Michelakis, and C. Toumazou, "Practical micro/nano fabrication implementations of memristive devices," in *Cellular Nanoscale Networks and Their Applications (CNNA), 2010 12th International Workshop on*. IEEE, 2010, pp. 1–4.

[13] Gy. Cserey, Á. Rák, B. Jákli, T. Prodromakis, "Cellular nonlinear networks with memristive cell devices," in *Proceedings of 17th IEEE International Conference on Electronics, Circuits and Systems*, ICECS 2010, (Athens, Greece), Dec. 2010.

Human like semantic models for object detection and classification

Attila Stubendek
(Supervisors: Dr. Kristóf Karacs and Dr. Tamás Roska)
stubendek.attila@digitus.itk.ppke.hu

***Abstract*—In this paper I introduce a human like object detection and classification system implemented in a banknote recognition system as a part of the Bionic Eyeglass Project. A morphological preprocessor selects candidates for patterns in the banknote. Then based on a statistical shape descriptor extracted from the face portrait a hierarchical classification is made. Finally the results are verified by the color of the pattern.**

***Index Terms*—object recognition, banknote recognition, classification, verification, morphology, shape descriptors**

I. Introduction

A. The Bionic Eyeglass

The Bionic Eyeglass Project is one of the three research branches of the Hungarian Bionic Visual Center. The Bionic Eyeglass [1],[2] is a portable device, to help blind and visually impaired people in everyday navigation, orientation and recognition tasks that require visual input. The banknote recognition system presented in the paper is the next experimental application for the Bionic Eyeglass.

B. The human like recognition system

Human brain processes the incoming visual information in multiple steps. The photo-receptor cells on the retina are not uniformly positioned. The rodes and cones are concentrated in the central are called fovea, the density of the cells is decreasing towards the edge of the retina. The eye is moving to the areas representing the highest attention. The peripherical visual field provides a low resolution image and also the attention information where to focus with the fovea. When searching for specific objects the peripherical output is enough to select possible locations of the object. Then the object is classified and verified by focusing with the fovea.

II. Banknote recognition

The banknote-recognition as the part of the Bionic Eyeglass Project helps visually impaired people in everyday situations using a portable mobile device.

The task of the banknote recognition system is to classify the banknote placed in front of the camera in real time. The end user tolerance towards misclassification is extremely low due to the immediate financial consequences. Therefore achieving the highest possible recognition and precision rate is essential even if this leads to lower recall rates. We made no restriction on the environment. The background can be inhomogeneous, lighting conditions may vary between frames and may not be constant within frame, and shadows may also appear on the image.

Figure 1. The 2000 Ft banknote. The patterns are in color rectangles: green is the face, light blue the coat of arms, red is the number, yellow the metal strip, blue the blank area with the watermark, pink is the sing of the National Bank

In many countries the banknotes can be distinguished by their size. The size of the Hungarian Forints banknotes does not depend on the represented value. There are tactile markers on the back side of the banknote designed to help visually impaired people, but due to usage and softness of the banknotes the marker provides inadequate means for visually impaired people to distinguish them.

Forint banknotes consist of darker characteristic patterns on lighter background. The front side is divided into two parts by a metal strip (except the 500 Ft note on which the strip is missing). On the left side there is a blank area containing the watermark and the sign of the Bank. The bigger, right side consists of the denomination of the banknote printed in numbers and in letters, the coat of arms, smaller texts and a cutaway face portrait of famous historical figures. The backside also consists of two parts. On the right side can be found the blank area with the watermark and the tactile mark, on the left side the denomination in numbers and a longish cutaway image of a building or a famous painting. An sample banknote can be seen in Figure 1.

The color of the Forint banknotes is characteristic for every banknote type. The color of the portrait image, the coat of arms on the front side and the backside image are the same, but it may differ from the number and text color.

The banknote recognition system is composed of modules pattern extracting and candidate selection, classification based on the face portrait mask and confirmation with the color information.

A. Morphological pattern extraction

Foreground objects are extracted using morphological operators. Due to the properties of the environment local adaptive threshold is needed. On the grayscale image a region of interest approximation is made, and a corrected threshold value is computed based on the ROI area.

Different pattern types, such as the portrait and the number are categorized based on morphological characteristics, like number of holes, size, absolute area, orientation.

The morphological pattern extraction is the result of Mihály Radványi's work. [3]

B. Classification

The classification is performed based on the portrait shape on the front side and by the tactile markers on the back side. The backside image is longish and detailed (except for the 1000 Ft banknote) resulting in a highly varying binary mask after thresholding. This makes inviable any classification based on the shape of the backside mark.

The current work focuses on the recognition of shapes. These include front side portraits and backside image of the 1000 Ft banknote. Wherever Times is specified, Times Roman or Times New Roman may be used. If neither is available on your system, please use the font closest in appearance to Times. Avoid using bit-mapped fonts if possible. True-Type 1 or Open Type fonts are preferred. Please embed symbol fonts, as well, for math, etc.

1) Pre-classification: The output of the morphological module are binary masks with the face or the backside image. The masks, as seen in Figure 2. are only candidates, that may also in addition to the desired shapes, other patterns from the banknote or from the inhomogeneous background.

To eliminate potential outliers before classification, a morphological filter is applied based on the mask size and the ratio of the minor and major axis length. The standard distance between the camera and the banknote determines a lower bound for the portrait size in the image. Based on experimental measurements we have determined that object masks smaller than 0.66% of the total area of the original image can be flagged as other objects. We also assumed that shapes of banknote portraits are typically not oblong, thus candidates having a major axis-minor axis ratio larger than 1.8 are also neglected. Objects with an axis ratio higher than 1.8 are typically the backside images. In both cases threshold values were set to reduce the false positive error rate. Applying the filters reduces the complexity of the machine learning task and results in shorter computation time. For the remaining candidates a feature vector is computed and a machine-learning model is used to make a decision detailed in the following section.

2) Face contour description: Since the image is taken by a mobile camera hold in hands, the relative position of the banknote and the camera can vary in distance and angle. In addition, due to different lighting and background conditions, the thresholding may cause that some parts of the portrait may be missing, holes might open up, and neighboring patches may be joined to the face mask.

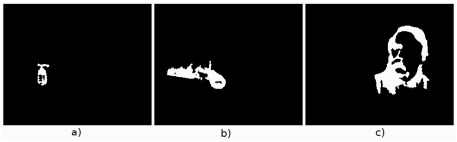


Figure 2. Examples of morphological outputs: a) area 0.59%, axes ratio 2.04; b) area 2.88%, axes ratio 2.82; c) area 9.02%, axes ratio 1.42

To provide scale and rotation invariant features for the classifier which are also robust to minor shape deformations, the Zernike moments of the shapes are used.

The Zernike Moments Descriptor considers the image canvas and the inner pixels of the object as a statistical space and a statistical set of 2D points, respectively. The shape is characterized by the statistical moments of the inner points. The higher the number of the moments used, the better the description, however by increasing the number of moments the shape descriptor is going to be more specialized, thereby loosing generality.

The $(nm)^{th}$ element of the Zernike Moments Descriptor is given by the following formulas [4],[5]:

$$A_{nm} = \frac{m+1}{\pi} \sum_{n} \sum_{m} P_{xy} \lfloor V_{nm}(x,y) \rfloor *$$

$$V_{nm}(r,\Theta) = \left[\sum_{s=0}^{\frac{m-|n|}{2}} (-1)^s F(m,n,s,r) e^{jn\Theta} \right]$$

$$F(m,n,s,r) = \frac{(m-s)!}{s! \left(\frac{m+|n|}{2} - s\right)! \left(\frac{m-|n|}{2} - s\right)!}$$

Zernike moments provide rotation-invariance by their absolute value: if an object is rotated, the magnitude of the complex values still remain the same. We have used Zernike Moment Descriptor with a maximum order of 24 and a square window size of 64 pixels, generating a feature vector of 173 moment magnitudes. The parameters of the descriptor were chosen based on the results described in the section III. A.

3) Hierarchical training and classification: The training and the setup of the training set are essential parts of the algorithm. We used a labeled training set with 600 input masks extracted during the development and the live tests. The face candidates may vary in the same class and also high similarity could be observed between certain classes and incorrect morphological outputs. Figure 3. shows some examples for morphologic outputs. To avoid high complexity and potential overfitting, we have defined sub-classes representing different binarizations of the same input class, two for each class (super-class). For training positive examples were the elements of a given sub-class, the negative examples were the samples in all other super-classes, whereas examples from the same super-class but from different sub-class have been left out. We have used k-means clustering on the Zernike moment feature vector

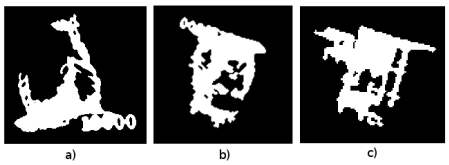


Figure 3. The a) and b) masks are extracted from HUF 10 000, the mask c) from the backside of the HUF 5000.

to split the classes, that resulted in 8% gain in the precision rate. We also performed a segmentation manually, but that led to inferior performance in precision (-4% respected to the automatic method). The advantage of the automatic method might be due to the better ability to take into account distances in the feature space.

4) Face vote: The characteristics of the task and the mobile platform implies complexity restrictions. The algorithm has to reply in real-time using slower mobile architectures, while the power consumption has to be minimalized also.

Due to the restrictions, only one artificial neuron is used with a sigmoid function for every decision class. As the results in the Section III. show. the perceptrons represent high separation ability for the given feature vectors.

Due to the characteristics of the platform, computation time and power are critical factors. Hence only one perceptron is used for every decision class.

The portrait recognition vote is given to the class giving the first positive decision in the order.

5) Decision: To achieve high robustness and reduce the false-positive error rate, two consistent votes have to arrive from the classifier modules to verify the vote. Due to the recognition time of one frame it can be assumed that correct backside and front side votes do not follow each other immediately, between votes from different sides 4 frames has to be elapsed.

C. Verification

The color of the banknotes are characteristic for each denomination, but as seen in Figure 4. due to the low quality of the mobile cameras and to the varying light conditions, the color values may differ radically even in consecutive frames. Moreover incorrectly extracted patterns from the background may have similar colors to portraits. Hence the color information by itself is not a reliable feature for classification.

However the color represents independent information from shape and color values of the patterns despite the inconstant lighting conditions stay between certain limits providing the ability to verify the classification.

If objects are extracted based on one type of information, such as luminance, other characteristics, like color, texture may not be constant in the whole area.

The color of a pixel can be described in different color spaces and using statistical functions on the color channel values of pixels in the object. RGB, HSV and YCbCr color spaces were used and the mean, the standard deviation and the

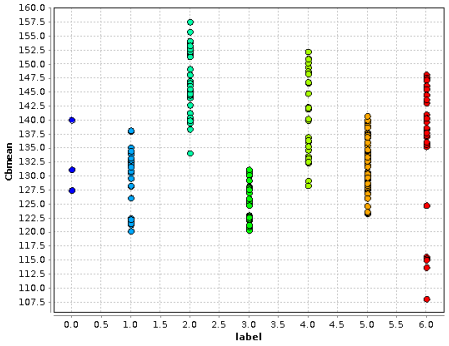


Figure 4. The mean of the cB chanell in the YcBcR space

median values and also RGB standard deviation ratios were computed. The HSV space did not fulfill our expectations, the other two spaces possess higher description capability.

To determine the limit values of the color functions a training set 240 elements has been used. The training images were taken in the various possible environments and lighting conditions. To remove outliers, we took the middle 95% of the data points into consideration for every function.

Verification showed high rate in refusing falsely classified objects, but naturally reduced the recall rate. Hence for the 18 channel one outlier is accepted.

III. Experimental results

A. Face shape recognition

Different kind of offline tests were performed in order to determine the optimal window size, highest moment of the moment descriptor and to prove the adequacy of the hierarchical classification.

I used a test set with 1541 binary masks as the outputs of the morphological module including large number of false examples.

As the metrics of correctness I used three functions, the accuracy (a), precision (p) and recall (r), computed as follows:

$$a = cc/ac$$

$$p = cc/pc$$

$$r = cc/pe$$

where cc is the number of correctly classified objects, ac the every classified objects, pc the classified object in the class, pe the number of the elements in the actual class.

The Figure 5. shows the results for the different window size using 24. highest moment. The results show that the 64 pixel window size is the most adequate in representing the face shape, bigger size do not increase significantly the efficiency, while the computation complexity would cause longer recognition time.

The Figure 6. shows the efficiency depending on different maximal moments. The results shows that using the 24.

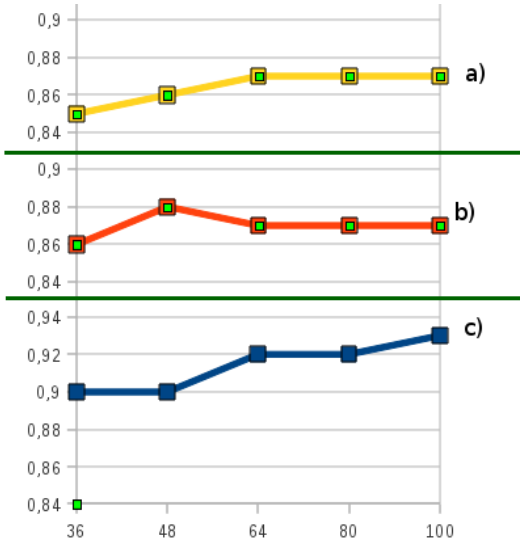


Figure 5. Accuracy (a), recall (b) and precision (c) dependency of the window size

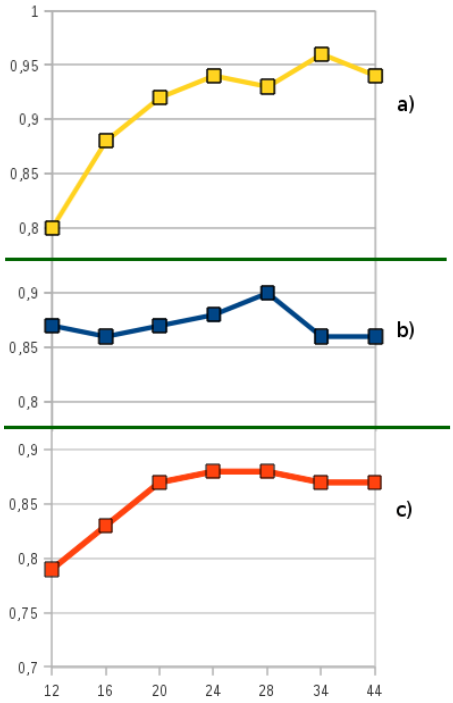


Figure 6. The accuracy (a), recall (b) and the precision (c) dependency of the maximal moment

maximal order provides good results while the evaluation time is acceptable. In the tests I used 64 pixel window size.

The Figure 7. shows the results of the hierarchical classification. For some classes defining sub-classes caused large increase in precision.

B. Human test results

We performed experiments with three visually impaired subjects who were given a few basic instructions about the optimal

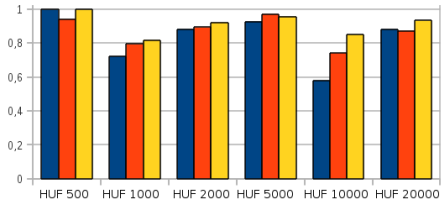


Figure 7. Precision rate of portrait recognition for the individual classes without sub-classes (blue), with manually selected sub-classes (red) and with sub-classes generated by clustering (yellow).

measurement technique and the 6 Hungarian banknotes. Their task was to tell which note they are holding with the help of our method installed to a mobile device. Each person made three series of tests. On the test at 21th of May the users could correctly classify the banknotes in the 94% of the cases. The application made 1649 decision, 1552 were correct, the results can be seen in Table I.

Table I
RESULTS OF THE TEST PERFORMED ON 10^{th} OF MAY

	500	1000	2000	5000	10000	20000
500	312	0	2	0	1	0
1000	0	375	0	0	0	0
2000	0	0	97	0	0	0
5000	0	2	1	263	26	0
10000	0	5	2	3	161	0
20000	0	7	0	4	5	158

IV. CONCLUSION

The banknote recognition system developed in the Bionic Eyeglass Project achieved high recognition and precision rate in real time. The developed method of selecting objects, classifying and verifying was successful in the special task and predicts usage in other task in the future.

REFERENCES

[1] K. Karacs, A. Lazar, R. Wagner, D. Balya, T. Roska, and M. Szuhaj, "Bionic eyeglass: An audio guide for visually impaired," in *Biomedical Circuits and Systems Conference, 2006. BioCAS 2006. IEEE*, 29 2006-dec. 1 2006, pp. 190 –193.

[2] K. Karacs, A. Lazar, R. Wagner, B. Balint, T. Roska, and M. Szuhaj, "Bionic eyeglass: The first prototype a personal navigation device for visually impaired - a review," in *Applied Sciences on Biomedical and Communication Technologies, 2008. ISABEL '08. First International Symposium on*, oct. 2008, pp. 1 –5.

[3] M. Radvanyi, "Visual feature detection and classification for banknote detection on low resolution images," in *PPKE ITK PhD Proceedings 2011, Budapest*, 2011.

[4] A. Khotanzad and Y. H. Hong, "Invariant image recognition by zernike moments," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 5, pp. 489–497, 1990.

[5] R. B. Yadav, N. K. Nishchal, A. K. Gupta, and V. K. Rastogi, "Retrieval and classification of objects using generic fourier, legendre moment, and wavelet zernike moment descriptors and recognition using joint transform correlator," *Optics and Laser Technology*, vol. 40, no. 3, pp. 517 – 527, 2008.

Genomic arrangement of bacterial genes involved in intercellular communication

Zsolt Gelencsér
(Supervisor: Dr. Sándor Pongor)
gelzs@digitus.itk.ppke.hu

Abstract— The study of bacterial communications is very important for better understanding bacterial communities found in nature and in certain disease situations. I carried out a preliminary survey of the key genes of acyl-homoserine based communication of bacteria often referred to as quorum sensing (QS). I used Hidden Markov Models to locate the known QS genes in the available genomes and prepared a survey of the topology of the genomic layouts.

Keywords—Intercellular communication in bacteria; regulators, autoinducers; quorum sensing

I. INTRODUCTION

Quorum Sensing (QS) is a method of bacterial communication and a basic mechanism of cooperation. Many species of bacteria use QS to coordinate their gene expression according to the local density of their population [1] This mechanism uses chemical signal molecules (autoinducers) to exchange information between cells to carry out this task, bacteria need a synthase i.e.: a protein that produces the signal and a censor protein that binds the signal and modulates the behavior of other genes.

From the possible regulatory network topologies that one can imagine for two genes, only a few are taxonomically mentioned in the literature of bacterial communication. These are known to be different for Gram-negative and Gram positive bacteria. Figure 1 show the typical genomic layouts [2-3].

Given the amount of fast growing genomic information we started a survey of potential and observed topological arrangements of the canonical density sensing genes in bacterial genomes. In order make this survey manageable in size, we first concentrated on the best known class of density sensing genes, the Acyl Homoserine Lacton (AHL) regulatory circle of Gram negative bacteria. Our long-term goal is to develop an automated, self-updating annotation system for genes of bacterial communication.

II. METHODS

A. Potential topologies and methods of description

Topological arrangements, same as other genomic patterns can be defined in terms of entities (genes) and relationships (distance within the chromosome and orientation). In order to find a time efficient search methodology, we first started with a survey of the potential arrangements. The simplest way to describe genes is to consider them as segments on the chromosome and in this sense we can speak of disjunct,

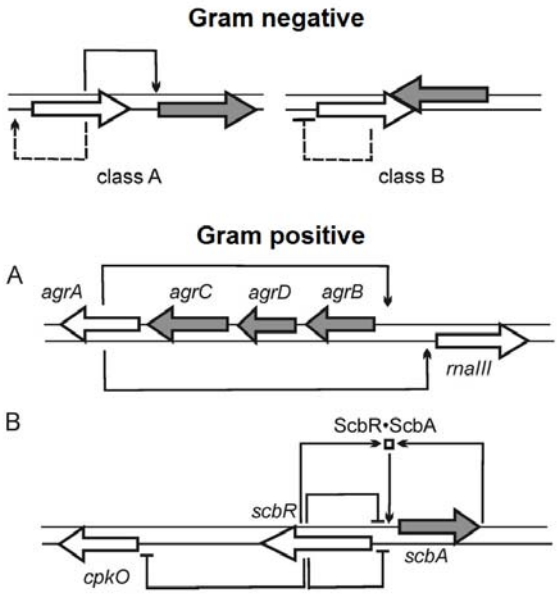


Figure 1. The genomic layout of the Gram negative (luxR/I) and the Gram positive QS bacteria. White arrows indicate QS transcription regulator genes; the gray arrows depict AI synthase genes (Gram negative) or other regulatory genes (Gram positive) (after ref.s. [2-3])

partially overlapping, genes. In terms of orientation, the direction of translation to protein in bacteria is parallel for genes located on the same DNA strand, and is anti-parallel for genes located on opposite strand. Figure 2 shows the potential arrangement of two genes on the bacterial chromosome with the common terminology of biologists. It is important to note that not all of these arrangements exist in nature. In bacteria, the overwhelming majority of genes are disjunct, and there are very few overlaps. Some arrangements (e.g. identical genes) can only occur due to errors in gene annotation.

We also developed a simple method for describing the described the neighborhood as set of genes, and used the Jacquard coefficient for comparing such neighborhood.

B. Searching QS genes

In order to find the canonical quorum sensing genes we first developed recognizers for the QS genes using validated QS genes obtained from Dr. Venturi's group at ICGEB Trieste (The International Centre for Genetic Engineering and Biotechnology). In agreement with the literature, we term the

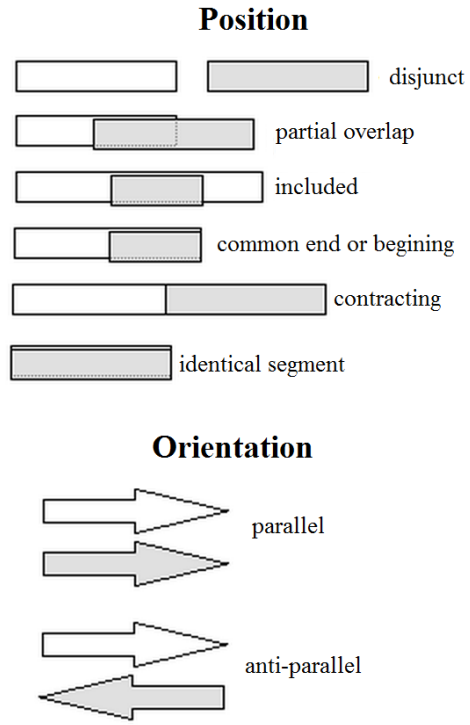


Figure 2 Positional and orientational arrangements used in the terminology of biologists. The total number amounts to 12 arrangements if we make no distinctions between the two genes (R or I) and 24 if we do. Out of these we observed only 3 arrangements in the RI gene pairs. Se Figure five for additional topologies.

regulator and synthase genes by R and I, respectively. We prepared a multiple alignments from the genes, and constructed Hidden Markov Model (HMM) recognizers using the HMMER package. The HMM recognizers were then used to scan the translated proteomes (faa files) of the 1346 full bacterial genomes found at the NCBI FTP site (the National Center for Biotechnology Information) [4]. The positional and orientation information was retrieved from the respective ptt data files. In the R genes we also identified those that contained a sequence pattern believed to be necessary for AHL binding.

The R and I genes were then validated by visual inspection, based on their HMM scores, their sequence pattern as well as by their distance to other R or I genes. In this way we identified 624 R genes 269 I genes. The majority was located in canonical clusters, however we also noted unusual clusters not described before.

For the qualitative description of the gene neighborhoods we first retrieved the 10 flanking genes (proteins) on both sides of the suspected R and I genes. We found that a large number of the flanking genes were not annotated, so we tentatively identified them by sequence similarity searching against the COG database [5] using the SBASE search engine [6]. This way we could annotate only one third of the unidentified genes, the rest was clustered into groups using BLASTCLUST [7] and the cluster numbers were added as tentative identifiers. Table 1. shows the distribution of the used annotating methods.

TABLE 1. Table Type Sty This table shows the distribution of the annotation methods in the genes of the neighborhoods. COG and SBASE: Have COG value and the SBASE confirm it. COG but SBASE: Have COG value, but the SBASE result is different. (don't change COG value).There is genes where there isn't COG value, but the SBASE prediction was successfully. In this case we run a BLASTclust clusterisation too, that result is: the genes are alone (Alone in cluster) or not alone in their cluster. If there are more genes are in a cluster, and one of them have COG value, all genes in cluster annotate this COG values.

COG		**BLASTclust**		
COG and SBASE	COG but SBASE	COG value	New cluster	Alone in cluster
8589	485	60	719	2778
68,0 %	*3,8%*	*0,5%*	*5,7%*	*22,0%*

For data processing I developed Perl scripts and the results were stored in a specific MySQL database that simplifies the data-mining, data-processing and data-visualization. The database is suited to receive more genes and neighborhoods if later the used NCBI database contains more bacteria fully annotated. On the stored data can execute complex queries, and if in the future is required, it is possible to supplement the database with the needed fields.

III. RESULTS

We found ~200 genes, all in Gram negative bacteria, as expected. The typical chromosomal arrangements found are shown in Figure 3. The tandem arrangement found in the canonical QS circuit RI of V. fischeri is the most frequent, however we found a good number of convergent or class B arrangements, and to our surprise a number of divergent RI genes as well that were previously known only in Gram positive bacteria. Another novel and relatively frequent arrangement we found was the RXI topology. Where the orientation of the intervening X gene is mostly parallel either with the R or with the I gene. But there was a few genomes

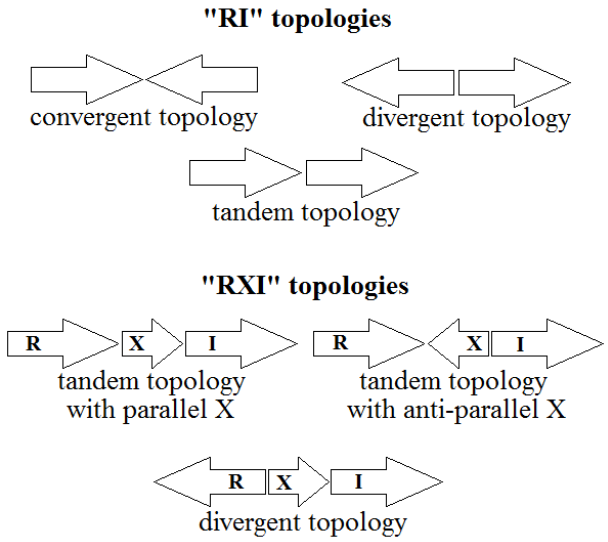


Figure 3. The possible topologies of the QS genes. "RI" topology means there is not any gene between R and I gene. In this case we don't make a distinction between the genes. "RXI topology means there is only one gene (called X) between QS genes. In this case we make a distinction between the genes: we see the topology by the side of R gene.

TABLE **2**. This table shows the distribution of the possible topologies. In the case of RI the tandem topology is significant, in the case of RXI the divergent.

RI topologies		
convergent	divergent	tandem
34,4%	*7,3%*	*58,3%*

RXI topologies		
tandem with parallel X	tandem with anti-parallel X	divergent
31,3%	*4,2%*	*64,5%*

where the X gene was anti-parallel to QS genes (e.g.: Pseudoalteromonas atlantica) The X position was filled with at least 10 different kinds of genes (identified by BLASTclust), the most numerous of these were the so called rsaL [8] and rsaM [9] genes that we will call here L and M, for brevity. Since both of these genes are known to be of biologically interesting but still not sufficiently characterized, we built HMM recognizers from them using validated L and M homologs and searched all the complete proteomes. It was found that M occurs only in genomes with validated QS machinery, and is invariably located next to an I gene (parallel), in divergent topology. The L gene, on the other hand, also occurs in genomes that have no identified QS genes. In QS genomes, it occurs exclusively in tandem RXI topology and always anti-parallel with QS genes. Table 2. shows the distribution of the founded topologies.

IV. Conclusion

The preliminary survey of the existing 143 genomes confirmed the genomic layouts known for QS genes but it also showed some novel arrangements. After optimizing several steps we could come up with a semi-automated annotation of protein similarity groups (sequences related to each other by sequence similarity). In this workflow only the multiple alignment step requires operator attention, the rest relies on mapping sequences to the UNIPROT database and to genomic databases. For many of the current datasets this mapping already exists, so the annotation requires relatively litter operator-time. The novel step is that of topological filtering which includes a comparison of the candidate genes I) to known (or recurrent) topological arrangements and II) to taxonomic groups in which the genes are expected to occur. Currently we use brute force heuristics at this final point. Reliable automation of this step requires future work.

ACKNOWLEDGEMENTS

This project was developed within the PhD program of Multidisciplinary Doctoral School, Faculty of Information Technology, Pázmány Péter Catholic University (Budapest).

I thank my supervisor Dr. S. Pongor (PPKE, ICGEB) for his help and guidance throughout the project, Dori Bihary and Borisz Galbats (PPKE) for their help in using the HMM programs and Dr. V. Venturi and his group for their advice.

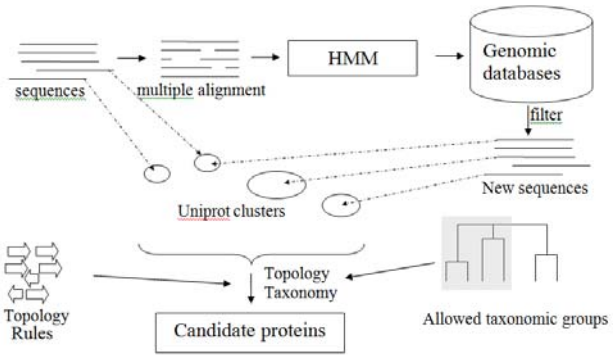


Figure 4. The block diagram of our workflow: I build a HMM from well-known QS genes, then search the whole database for new sequences, then I examine the Uniprot clusters of the founded genes (candidate proteins), and check every genes in the clusters using the topology and taxonomic rules.

References

[1] C. Fuqua and E. P. Greenberg, "Cell-to-cell communication in Escherichia coli and Salmonella typhimurium: they may be talking, but who's listening?," *Proc Natl Acad Sci U S A,* vol. 95, pp. 6571-2, Jun 9 1998

[2] A. B. Goryachev, "Understanding bacterial cell-cell communication with computational modeling," *Chem Rev,* vol. 111, pp. 238-50, Jan 12 2011.

[3] A. B. Goryachev, "Design principles of the bacterial quorum sensing gene networks," *Wiley Interdiscip Rev Syst Biol Med,* vol. 1, pp. 45-60, Jul-Aug 2009.

[4] NCBI FTP side: *ftp.ncbi.nlm.nih.gov*

[5] R. L. Tatusov, *et al.*, "The COG database: an updated version includes eukaryotes," *BMC Bioinformatics,* vol. 4, p. 41, Sep 11 2003.

[6] SBASE on ICGEB homepage: http*://hydra.icgeb.trieste.it/sbase/*

[7] BLASTclust bioinformatic tool on Max Plank Institute homepage: *http://toolkit.tuebingen.mpg.de/blastclust*

[8] T. de Kievit, *et al.*, "RsaL, a novel repressor of virulence gene expression in Pseudomonas aeruginosa," *J Bacteriol,* vol. 181, pp. 2175-84, Apr 1999.

[9] M. Mattiuzzo, *et al.*, "The plant pathogen Pseudomonas fuscovaginae contains two conserved quorum sensing systems involved in virulence and negatively regulated by RsaL and the novel regulator RsaM," *Environ Microbiol,* vol. 13, pp. 145-62, Jan 2011.

Analysis of bacterial communities using agent based methods

Dóra Bihary
(Supervisor: Dr. Pongor Sándor)
bihdo@digitus.itk.ppke.hu

Abstract— **Agent-based models can be used to simulate the growth of communities containing multiple species. In this work I present numerical indices developed to characterize the spatial distribution and the fitness of competing bacterial species in an agent-based model. Statistical results are also presented in order to provide deeper information about competing bacterial populations.**

***Keywords*—quorum sensing, *Pseudomonas aeruginosa*, hybrid model, statistics, segregation, fitness**

I. INTRODUCTION

Multispecies communities are believed to be the a major form of bacterial life. These communities contain more than one species; the interaction between individual bacteria is based on diffusible signals, the best known example of which is a mechanism called quorum sensing (QS). In this mechanism signaling materials are secreted by the bacteria and diffuse in the environment and the concentration of signals is correlated with density of bacteria, or loosely speaking, with the size of the consortium. This is obviously true only in certain cases, however a small community can also be wedged in a small place where high signal concentration can be reached by a small community.

A typical behavior of bacterial communities is swarming, an intense and coordinated motion of the entire community. Bacterial swarming can be explained by the mechanism of quorum sensing. When the concentration of the secreted signal is greater than a certain threshold bacteria switch from low metabolic activity to high activity level which means that they secret a higher amount of signaling molecules and they also start to secret other molecules, so-called public goods (e.g. surfactants, enzymes, siderophores, all together factors) which facilitates their movement and their nutrient uptake[1].

Bacterial behavior is often discussed in terms of communication, cooperation and competition between species. This approach is interesting not only because bacterial consortia are now known as a major form of life on the earth, but also because their organization shows analogies with human societies.

Recently we developed a computational model for describing quorum sensing in the bacterium *Pseudomonas aeruginosa* [2]. The model was first developed to show QS signaling is able to explain swarming movement and other fundamental properties of bacterial colonies. Later results showed that the model can be extended to the study of multiple bacterial species. Mutant species were also studied with the simulation model and it was shown that some types of mutants can collapse a cooperating community[3, 4].

II. MODEL TYPES FOR BACTERIAL COMMUNICATION

In recent studies there are various approaches for modeling bacterial communities. Individual-based models use individual entities (agents, cells of cellular automata) to represent a bacterium, a group of bacteria, or a species. Continuous models represent bacteria as a continuously diffusing material. Not all of these models have been applied to intercellular signaling. These are basic types of models, each is good for some aspect of the bacterial communication but this complex system can not be described by a single model type. The main idea was to combine the existing models into a hybrid model. In the next paragraphs an overview of these model types can be found.

A. Agent-based model

Agents are individual objects in an environment. They can percieve some aspects of this environment and they can respond to its changes. This is an individual-based, and also a discrete model which means that time and space are unually both represented in discrete steps. There are well defined states between which the agents can switch in every time step based on its actual inner state and the observed change in the surrounding environment. This is usually defined by rules (simple first order logic or more complex algorithms). A simple rule set can be for example that (1) steer to avoid crowding local agents, (2) steer towards the average heading of local agents and (3) steer to move toward the average position of local agents[5].

B. Cellular automata

Cellular automaton is also a discrete, individual-based model which consists of regular grid of cells. The model has a constant size, in every time step bacteria play pair wise games, the “better” survives, and replaces the “worst” one as well. In a recent study of QS the mechanism is represented by a set of genes.

Agent-based and cellular automaton models aim to describe the behavior of bacteria but they do not explicitly represent bacterial motion or the materials present in the environment.

C. Continuous model

In continuous models for bacterial communication bacteria and the diffusible environment can be represented by diffusion equations which describe the propagation of chemicals.

A more realistic model can be given with reaction-diffusion equations[6]. These are capable of describing the behavior of diffusible material and the interactions between them. At a QS system this is necessary for the explanation of decay in time. A reaction-diffusion equation can be written as follows:

$$\frac{\partial u}{\partial t} = D\left(\frac{\partial^2 u}{dx^2} + \frac{\partial^2 u}{dy^2}\right) - Ru \qquad (1)$$

where u is the concentration of the chemical, D is the diffusion coefficient and R is the decay coefficient. This means that the time derivative of the concentration is equal with an expression of two terms, the first is proportional to the second order space derivative of the concentration, and the second to the concentration itself. When this second term is negative it means decay (of signals and factors), when it is positive we can talk about production (signal and factor secretion).

If we represent the bacterial communication with a continuous-only model bacteria loose their ability of decision making (e.g. "where should I go?").

D. Hybrid model

As mentioned above neither agent-based/cellular automaton nor continuous models can properly describe some aspects of QS communities. Hybrid models were developed with the aim of solving some of these problems

In a hybrid model bacteria are represented as agents that interact with each other (secret signals and factors) and the environment (eat nutrient), they respond to the signal concentration[7, 8]. The environment of the bacteria consists of nutrient, signals and public goods. The distribution of these materials in space and time is described by reaction-diffusion equations that show how materials are produced and consumed by the bacteria and how they decay. The aim of my work in this year was to develop numerical indices that characterize the spatial distribution of multispecies bacterial communities.

III. NUMERICAL TESTS

In the simulation model the space is represented as a longitudinal surface where the bacterial community starts from the bottom and proceeds upwards while consuming the nutrients found on the surface. We can follow the collapse or survival of a species by counting the number of cells at each step. On the other hand the distribution of cells also changes in space. This visible information is however quite meaningful. Fig. 1 shows that some populates separate in space while they move, others remain together.

Our main goal was to represent these distributions by numerical data. In the next sections I will provide an overview of the methods and our main results of statistical analysis in bacterial communities.

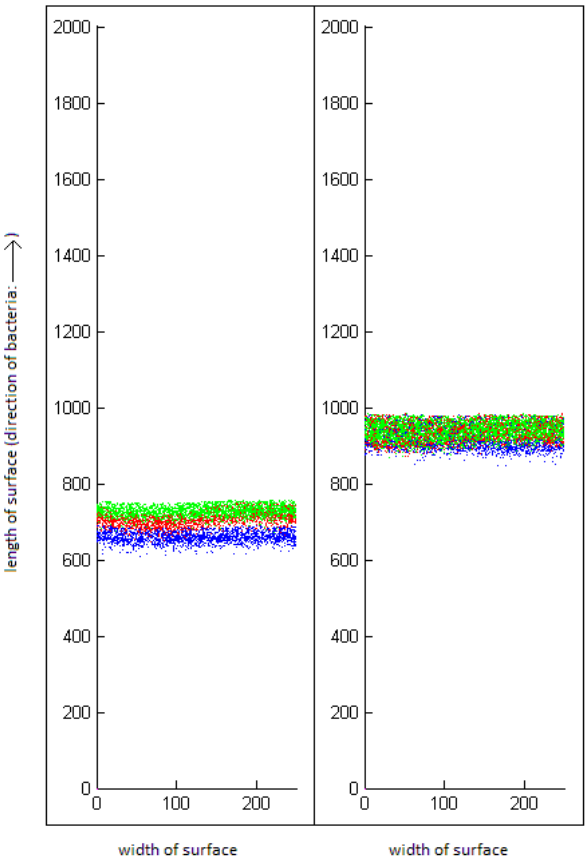


Figure 1. Simulation of segregation(left) and mixed (right) swarming communities

A. Segregation index

We call a population segregated when the different species are separated from each other (see Fig. 1, left side). For numerically measuring this quantity we developed a so called segregation index. This quantity was originally defined in such a way that a certain number of nearest neighbors were identified for all bacterial cells, and the largest percentage of a given species was determined for each cell. The segregation index was then calculated as the average of these materials [Xavier]. Calculation of this index is time consuming, especially since our bacterial communities are quite large.

In order to develop a segregation index that can be calculated in a more time efficient manner, we take advantage of the fact that space in our simulation is divided to cells that form a matrix-like lattice. In each row we can count the number of bacteria from each species. E.g. for three species we can have $n_1(i)$, $n_2(i)$, $n_3(i)$ in the i^{th} row. The sum of these numbers in the i^{th} row is N(i). So the fraction of these values gives the ratio of each species in the population. If a population is segregated, this fraction is almost 1 for one of the species, and almost 0 for the two other species. For computing segregation index we average the maximal elements (the size of the largest species) in each row. We get a more representative measurement if we weight these values with the total number

of bacteria in the actual row. By this step we avoid counting fractions for each row, we simply have to add the maximal numbers in each row, and divide this value with the total number of bacteria in the given step.

The number we get with this computation is in the range [1/number of species; 1], so for later comparisons it is better to use the normalized version of the value where total segregation corresponds to 1.0 and no segregation is a value close to zero.

Equation (2) shows this normalized computation:

$$S = \left(\frac{S_{sumPerRow}}{N_{allBacteria}} - \frac{1}{N_{species}} \right) \Bigg/ \left(1 - \frac{1}{N_{species}} \right) \qquad (2)$$

where S is the segregation index in a certain step, $S_{sumPerRow}$ is the sum of the segregation numbers in each row, $N_{allBacteria}$ is the size of colony and $N_{species}$ is the number of species.

B. *Computation of fitness*

Bacterial agents can be mutant species, the simplest mutants are for example when a bacterium does not secret signals, but can respond for external signals, or when a bacterium is not only unable to secret signals but can not respond to any kind of signal either. In normal case bacteria can show both of these phenomena, these are called wild type entities.

An interesting question is that in a multispecies community how species behave according to the case when only wild type entities are there in the population. This can also be represented by numerical data.

For measuring this property we need to calculate the fitness of the bacterial species. In principle the fitness of a species can be computed from the size of the population taken at the beginning and at the end of simulation (3).

$$F = \frac{1}{\Delta t} \log_2 \frac{N_{end}}{N_{start}} \qquad (3)$$

where F is the fitness value, Δt denotes the elapsed time, N_{end} and N_{start} are the size of the population at the beginning and end of the simulation respectively. By taking the logarithm of the fraction we get an expression that's sign depends on whether an increasing or decreasing population is there in the simulation. For increasing population the logarithm is positive, however for decreasing it becomes negative.

Fitness is a dimensionless number which is often represented on a relative scale, in comparison with the fitness of a reference species. If we want to compare a mutant species with wild type population we have to divide the fitness of the two species. Equation (4) describes this computation.

$$F_{rel} = \log_2 \left(\frac{N_{end}}{N_{start}} \right) \Bigg/ \log_2 \left(\frac{N_{end,wt}}{N_{start,wt}} \right) \qquad (4)$$

where F_{rel} is the relative fitness, $N_{end,\ wt}$ and $N_{start,\ wt}$ are the reference values for wild type population. The Δt terms are cancelled by the division.

IV. RESULTS

Figure 2. shows the segregation index index as a function of simulation time. The blue line shows two segregating species, the green line shows two non-segregating species. In both cases segregation index starts from a value close to zero, which means that the starting population is a well-mixed community. In the case of segregating species, the value of the index reaches a plateau near 1.00. In the non-segregating case the index fluctuates around a small value. This significant difference is visible on Fig. 2.

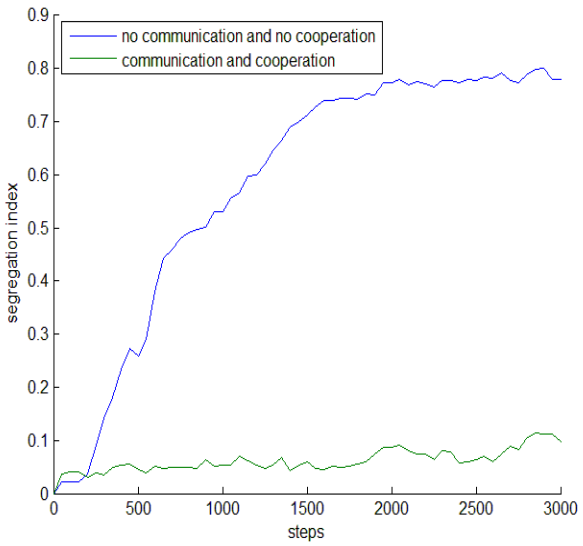


Figure 2. Visualization of the growth of segregation index in communicating and cooperating population (green) and non-communicating and non-cooperating community (blue)

Results of relative fitness computations can be found in Table 1. Rows represent simulations, columns represent the existing species. In a population where only wild type agents are present the relative fitness is 1 for the first race, because this is the reference population as well, species number two and three does not exist.

In these simulations there were three species one-by-one with same initial bacterium number as in the reference simulation ensuring the same initial arrangement. In well mixed (cooperating) population there are more agents than the reference species. For segregating, non-cooperating species the results are close to one. In each simulation, the values for each of the 3 species are near to each other, and the differences reflect the reproducibility of the calculation. We estimate that the reproducibility of the relative fitness is more than 98% in both cases. This value was calculated from 25-25 results for each type of simulation.

TABLE 1. Comparison of relative fitness in communicating, cooperating populations and non-communicating, non-cooperating communities

	Relative fitness of first race	Relative fitness of second race	Relative fitness of third race
Single wild type population alone	1	-	-
Well-mixed (cooperation) species	1.2402	1.2304	1.3063
Segregated (no cooperation) species	1.0062	1.0188	1.0094

V. CONCLUSION

In this paper I presented numerical indices for characterizing the results of a model simulating communicating bacterial communities. This model is a so-called hybrid model with an agent-based part for representing bacteria, and a reaction-diffusion part for describing the environment consisting of diffusible materials. I described numerical induices for segregation and relative fitness for describing differences in well-studied population types. In the near future I plan to use these statistical methods to analyze complex bacterial communities containing multiple species.

ACKNOWLEDGEMENT

This project was developed within the PhD program of Multidisciplinary Doctoral School, Faculty of Information Technology, Pázmány Péter Catholic University, Budapest. Thanks are due to my supervisor, Prof. Sándor Pongor and to Ádám Kerényi, who is the major developer of the agent-based simulation program.

REFERENCES

[1] V. Venturi and S. Subramoni, "Future research trends in the major chemical language of bacteria," *HFSP J,* vol. 3, pp. 105-16, 2009.

[2] S. Netotea*, et al.*, "A simple model for the early events of quorum sensing in Pseudomonas aeruginosa: modeling bacterial swarming as the movement of an "activation zone"," *Biol Direct,* vol. 4, p. 6, 2009.

[3] V. Venturi*, et al.*, "Locality versus globality in bacterial signalling: can local communication stabilize bacterial communities?," *Biol Direct,* vol. 5, p. 30, 2010.

[4] V. Venturi*, et al.*, "Co-swarming and local collapse: quorum sensing conveys resilience to bacterial communities by localizing cheater mutants in Pseudomonas aeruginosa," *PLoS One,* vol. 5, p. e9998, 2010.

[5] C. W. Reynolds, "Flocks, herds, and schools: A distributed behavioral model.," *Computer Graphics,* vol. 21:25– 34, 1987.

[6] K. Kawasaki*, et al.*, "Modeling spatio-temporal patterns generated by Bacillus subtilis," *J Theor Biol,* vol. 188, pp. 177-85, Sep 21 1997.

[7] E. Ben-Jacob, "Spatio-selection in expanding bacterial colonies.," *Physica A,* 1999.

[8] E. Ben-Jacob*, et al.*, "Complex bacterial patterns," *Nature,* vol. 373, pp. 566-7, Feb 16 1995.

Improved force field for Vector Field Convolution method

Andrea Kovács
(Supervisors: Dr. Tamás Szirányi and Dr. Zoltán Vidnyánszky)
kovacs.andrea@itk.ppke.hu

***Abstract*—Parametric active contours are efficient tools for boundary detection. However, existing external-energy-inspired methods have difficulties when detecting high curvature, noisy or low contrasted contours and they often suffer from initialization sensitivity. To address these issues, this paper introduces Harris-based Vector Field Convolution (HVFC), operating with the modified characteristic function of Harris corner detector used in the feature map of the external force component. Initial contour is calculated as the convex hull of the most salient points of the map. Experimental results show that HVFC outperforms other state-of-the-art methods, when tested on high curvature, noisy or low-contrasted contours.**

***Keywords*-Boundary analysis, vector field convolution, Harris characteristic function**

I. Introduction

Since the introduction of active contour (snake) approach in [1], deformable models were proved to be efficient for robust object detection. Generally, in these methods the evolution of the snake is an energy minimizing method controlled by internal and external forces. Internal force is responsible for the conformation (elasticity and rigidity) of the snake; while external force represents the constraints of the image. To compensate the drawbacks of the original algorithm, like initialization sensitivity and convergence to concave boundaries, several modifications have been published, parametric [2], [3], [4] and non-parametric [5], [6] methods as well.

Parametric approaches suffer from noise, parameter and initialization sensitivity, topological changes and have weaknesses when detecting high curvature boundaries. However non-parametric models are insensible to initial location and can handle complex boundaries with sharp corners and changes of topology; they have problems when detecting broken or open edges. Additionally their convergence rate is slower, they are more sensitive to noise than the parametric approaches and they are not robust against intensity changes inside the object.

To address difficulties concerning high curvature, noisy boundary points along with smooth edges and initialization sensitivity, we called for corner detector approaches. Harris corner detector [7] is reliable and invariant to rotation [8]. This paper introduces a parametric active contour representation, called Harris based Vector Field Convolution (HVFC), which applies a modified characteristic function of the Harris corner detector. Defining the cornerness and edginess in one function, this modification can be used to emphasize edge and corner points of the object boundary equally and attain a balanced saliency map. The most attractive points of the map are used to initialize a starting curve around the object, while the whole saliency map is applied to determine a new feature map in the external energy expression.

The modified characteristic function was applied efficiently for improving the accuracy of parametric Gradient Vector Flow (GVF) active contour method [9] and [10]; for shape based recognition of flying targets [11] and for building localization in aerial images [12].

The performance of our proposed algorithm has been evaluated on the Weizmann segmentation evaluation dataset [13] and the results were compared to three state-of-the-art techniques [2], [3], [5]. Regarding the accuracy of detecting high curvature, noisy object boundaries, our method performed better.

II. Parametric Active Contour Theory

Parametric active contour model was first published in [1]. The aim is to find the curve, denoted by $\mathbf{x}(s) = [x(s), y(s)]$, $s \in [0, 1]$, that minimizes the following energy function:

$$E = \int_0^1 \frac{1}{2} \left[(\alpha \left|\mathbf{x'}(s)\right|^2 + \beta \left|\mathbf{x''}(s)\right|^2) + E_{\text{ext}}(\mathbf{x}(s)) \right] \, \mathrm{d}s, \quad (1)$$

where α and β are the weighting parameters of the elasticity and rigidity components of the internal energy; $\mathbf{x'}(s)$ and $\mathbf{x''}(s)$ are the first and second order derivatives with respect to arclength s. While elasticity component results smooth contours, the rigidity component is responsible for detecting curvature, setting $\beta = 0$ allows the snake to develop a corner.

E_{ext} is the external energy derived from the image, it gives smaller values at features of interest (like edges and ridges) and pushes the snake toward an optimum in the feature space. This energy represents the constraints of the image itself and is usually calculated as a function of edge information over the intensity distribution.

A. Vector Field Convolution

To address some disadvantages of the former parametric methods (like Gradient Vector Flow (GVF) [2]), as high computational cost, noise and parameter sensitivity, Vector Field Convolution (VFC) was introduced as a new external force in [3]. The field is calculated as the convolution of a vector field kernel and the edge map derived from the image:

$$\mathbf{f}_{\mathrm{VFC}}(x,y) = f(x,y) * \mathbf{k}(x,y), \tag{2}$$

where the edge map generated from the image. One of the generally used forms is:

$$f(x,y) = \left|\nabla(G_\sigma(x,y) * I(x,y))\right|. \tag{3}$$

The $\mathbf{k}(x,y)$ kernel is defined as a multiplication of a unit vector pointing to the kernel origin and a magnitude function, which should be chosen as a decreasing positive function of the distance from the origin [3].

The f edge map used in the external component of VFC does not emphasize the high curvature corners of the boundary (see Figure 1(a)). As intensity changes largely from only a few direction in these areas, the gradient function gives lower values to these regions on the edge map compared to sharp edges. As a result of this effect, these methods are not able to detect such corners accurately (see Figure 2(e)). Therefore, our aim was to construct a feature map that emphasizes sharp corners and edges equally and the iterative active contour method can detect high curvature boundaries precisely.

III. Harris based Vector Field Convolution

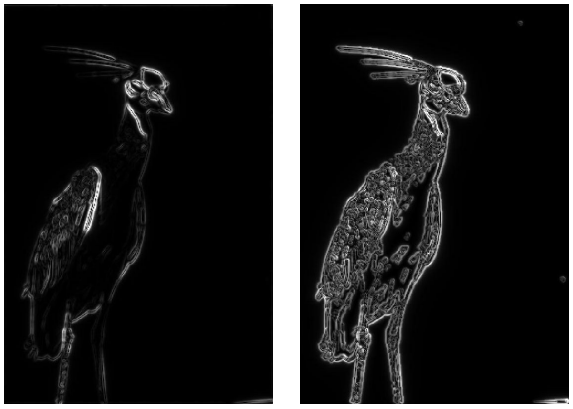
Our process generates feature points based on the proposed modification of the Harris detector's characteristic (saliency) function [9], then the convex hull of these points is used to initialize the snake. As a novelty, this modification of the characteristic function is used instead of intensity function in the feature map (see Equation 3) for generating VFC snake [14].

A. Harris based feature map

Harris detector is based on the second moment (or auto-correlation) matrix (M), which is often used to describe the local image structure or to characterize the curvature behavior around a keypoint.

$$M = \begin{bmatrix} A & C \\ C & B \end{bmatrix}, \tag{4}$$

where $A = \dot{x}^2 \otimes w$, $B = \dot{y}^2 \otimes w$, $C = \dot{x}\dot{y} \otimes w$. $\dot{x}$ and $\dot{y}$ denote the approximation of the first order derivatives, w is a Gaussian window. The eigenvalues of M represent two principal signal changes in the neighborhood of a point and will be proportional to the principal curvatures, which feature could be used efficiently when external forces are calculated to measure salient points of any boundaries as the principal curvatures describe well the fine details of shapes. By denoting the two eigenvalues of M by λ_1 and λ_2, they can be applied to create a novel characteristic (saliency) function emphasizing the curvature around image pixels. The eigenvalues separate the following cases: both of them are large in corner regions, only one of them is large in edge regions and both of them are small in flat (homogeneous) regions [7]. When constructing a feature map, we need to emphasize edges and corners equally. Considering the aforementioned terms for the eigenvalues, one of the eigen values is large both in corner and edge regions, therefore the $\max(\lambda_1, \lambda_2)$ function separates flat and non-flat regions. The range of this function is quite wide, so to get a balanced distribution, the dynamics should be compressed while preserving the original proportion of the values. The natural logarithmic ($\log$) function satisfies these conditions and results a balanced output with emphasizing corners and edges, but may result negative results when $\max(\lambda_1, \lambda_2) < 1$ (flat regions). As we would like to construct a characteristic function for a feature map, the target set have to be the positive domain. Therefore, the $\log(\max(\lambda_1, \lambda_2))$ function should be further modified. For edges and corners $\max(\lambda_1, \lambda_2) >> 1$, so negative values of small lambdas in the $\log$ function can be replaced with zeros without losing any corner and edge information and the proposed saliency function looks as follows:



(a) (b)

Fig. 1. The two different feature maps: (a) is the original, f_{VFC} intensity based map (Eq. 2); (b) is the proposed $\mathbf{f}_{\mathrm{HVFC}}$ map based on R_{logmax} (Eq. 6).

$$R_{\mathrm{logmax}} = \max(0, \log[\max(\lambda_1, \lambda_2)]). \tag{5}$$

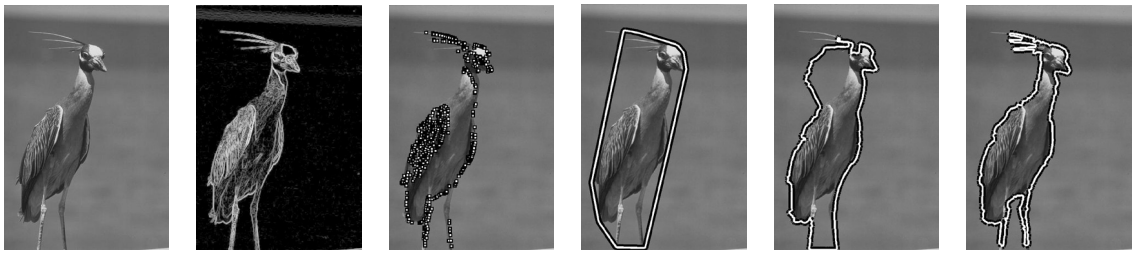
With the constructed R_{logmax} characteristic function (see Figure 2(b)) our proposed feature map for the Harris based Vector Field Convolution (**HVFC**) method:

$$\mathbf{f}_{\mathrm{HVFC}} = \left|\nabla(G_\sigma(x,y) * R_{\mathrm{logmax}}(x,y))\right| * \mathbf{k}(x,y). \tag{6}$$

Figure 1(b) shows the modified feature map with the equally emphasized edge and corner points of the boundary, the result of the contour detection based on the modified feature map can be seen in Figure 2(f).

B. Harris based initial contour

Similarly to the operation of the original Harris detector, feature points (edge and corner points) can be calculated as the local maxima of the characteristic function (R_{logmax}). [9] introduced an initialization approach based on the convex hull of the feature points inside the region of interest. The initial interest region was modified by the local surroundings of saliency points to avoid the poor definition of smooth transition or multi-directional saddle effects of edges around corners. The optimal size of the local neighborhood was determined by applying automatic scale selection theory. In our case, a local



(a) (b) (c) (d) (e) (f)

Fig. 2. Effect of R_{logmax} characteristic function: (a) is the original image, (b) is the generated characteristic function, (c) shows the generated corner points, (d) is the initial contour based on the convex hull of the corner points, (e) and (f) show the segmentation results of original VFC [3] and the proposed HVFC methods.

area with a fixed, 3 pixel radius around the feature point should be considered as part of the region of interest (ROI), where relevant local structures can be detected. This fixed radius was proved to work as efficiently as the automatic in [9], while the computational time decreased. Points representing the outline of the supported area should be added to the Harris feature point set (Figure 2(c)). After this, the initial contour is defined as the convex hull of the extended set of points (see Figure 2(d)).

IV. Experiments

In the first part, our proposed method was tested on 23 selected images from the Weizmann dataset [13]. The selected images had high curvature boundaries. In the second part, 3 specific images were selected with high curvature, noisy or low contrasted boundary parts. The performance of our HVFC algorithm was compared to GVF snake [2], VFC [3] and active contour without edges (ACWE) [5], based on the Chan-Vese model. In both parts of the experiments, the initial ROI was marked as an ellipse and the starting contour was calculated as described in Section III-B (see the first column of Figure 3). The parameter setting for the compared methods (GVF, VFC and ACWE) was chosen from [2], [3] and [5].

In the first part of the experiments the single object average F-measure score was calculated, it can be seen in Table I. The traditional F-measure score is the weighted, harmonic mean of precision and recall values and calculated as:

$$\mathbf{F} = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}.$$

Like it was also mentioned in [4], evaluation results showed that – due to the intensity variation inside the object –, this database is not suitable for ACWE method. Therefore, the table contains results for GVF, VFC and HVFC algorithms. In the second part, specific images were selected to test the accuracy of the methods (see Figure 3). The execution time of different methods for images in Figure 3 can be seen in Table II. These experimental results were based on an Intel(R) Core(TM) i7 CPU with 4 GB RAM and MATLAB R2010b. ACWE [5] method follows a level-set representation, therefore it uses intensity homogeneity constraints instead of gradient based edge map. ACWE can successfully identify object boundaries even with high curvature parts, if intensity is homogeneous inside the object and the contour is closed properly (see image 'leaf', first row of Figure 3). Else, ACWE converges to object parts representing homogeneous regions which differ largely from the estimated background ('egret','plane', 'mirage' images in Figure 3). Beside suffering from high curvature and low contrasted boundaries, GVF [2] also fails when the initial contour is further from the real boundary. In case of larger concavities of the object boundaries the convexity feature of the contour initialization step results in a distant initial contour, therefore the method is trapped in local minima. VFC [3] has the advantage to be less sensitive to initialization than GVF, due to the calculated vector field kernel. Therefore large concave outlines do not cause challenge (like image 'egret'), but the method fails to detect high curvature and low contrasted boundary parts due to the existing problems of the feature map (see Section II-A for further details). The proposed HVFC method benefits from the advantages of traditional VFC algorithm and the introduced feature map of modified Harris function and detects the aforementioned complex boundaries accurately (see last column of Figure 3).

TABLE I
Average F-measure Score (mean ± standard deviation) for GVF [2], VFC [3] and the proposed HVFC algorithms for 23 images [13].

Algorithm	Average F-measure Score
GVF	0.79 ± 0.09
VFC	0.86 ± 0.07
HVFC	$\mathbf{0.91 \pm 0.06}$

TABLE II
Execution Time of the different methods for images in Figure 3. IC indicates the initial contour; ACWE [5], GVF [2], VFC [3] and HVFC (proposed) are the compared methods.

	Execution Time [seconds]				
Images	IC	ACWE	GVF	VFC	HVFC
'leaf'	0.36	13	5.8	3.7	4.8
'egret'	0.44	66	6.2	4.2	5.1
'plane'	0.38	12	3.9	3.2	3.7
'mirage'	0.85	68	9.4	6.3	6.9

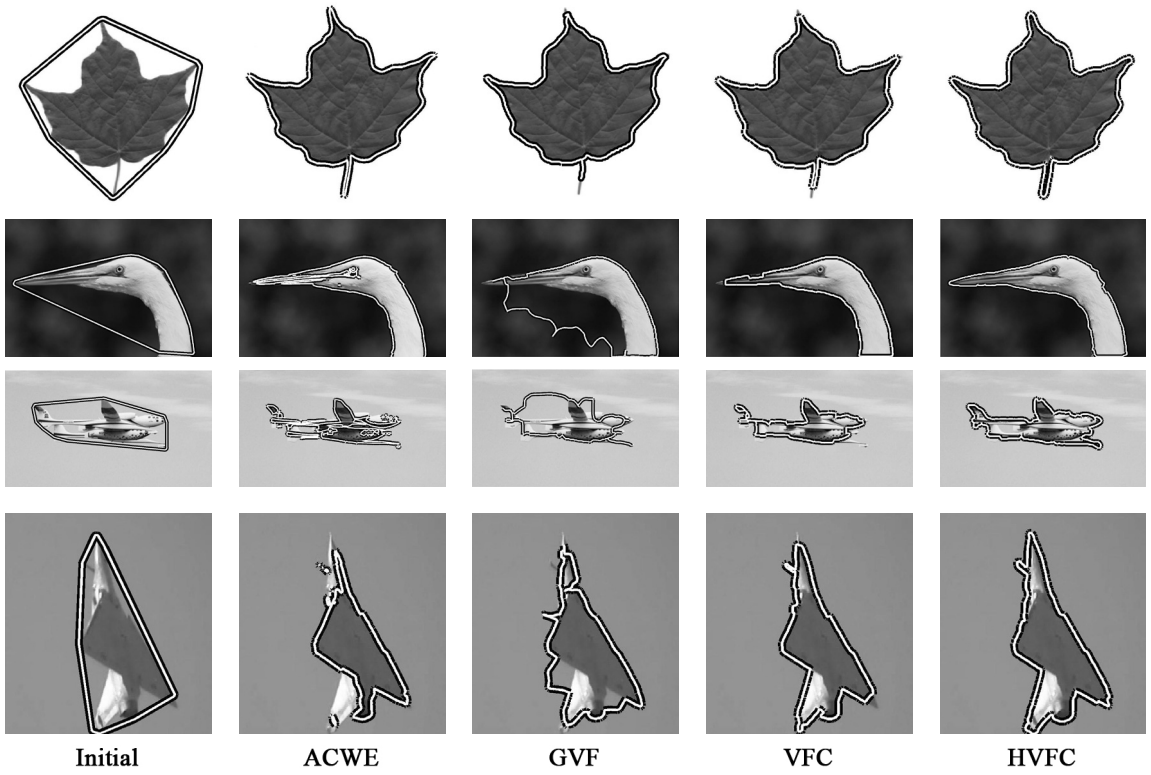


Fig. 3. Examples of contour detection (row-by-row: 'leaf', 'egret', 'plane', 'mirage': The first column shows the calculated initial contour (see Section III-B). Second, third, fourth, fifth and sixth columns present the results for ACWE [5], GVF [2], HGVF (proposed), VFC [3] and HVFC (proposed) methods.

V. Conclusion

In this paper, a modified feature map has been introduced for Vector Field Convolution method, which is able to emphasize sharp corners and edges equally with the Harris detector's modified characteristic function. Initialization is based on the most attractive points of the characteristic map. According to the evaluation results, due to the modification in the external force, the Harris-based Vector Field Convolution (HVFC) method can converge to high curvature and low contrasted boundary parts and performs better than the compared state-of-the-art methods.

References

[1] M. Kass, A. P. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *International Journal of Computer Vision*, vol. 1, no. 4, pp. 321–331, 1988.

[2] C. Xu and J. L. Prince, "Gradient vector flow: A new external force for snakes," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 1997, pp. 66–71.

[3] B. Li and T. Acton, "Active contour external force using vector field convolution for image segmentation," *IEEE Transactions on Image Processing*, vol. 16.

[4] A. K. Mishra, P. W. Fieguth, and D. A. Clausi, "Decoupled active contour (DAC) for boundary detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 2, pp. 310–324, 2011.

[5] T. F. Chan and L. A. Vese, "Active contours without edges," *IEEE Transactions on Image Processing*, vol. 10, no. 2, pp. 266–277, 2001.

[6] X. Bresson, S. Esedoglu, P. Vandergheynst, J.-P. Thiran, and S. Osher, "Fast global minimization of the active contour/snake model," *Journal of Mathematical Imaging and Vision*, vol. 28, no. 2, pp. 151–167, 2007.

[7] C. Harris and M. Stephens, "A combined corner and edge detector," in *Proceedings of the 4th Alvey Vision Conference*, 1988, pp. 147–151.

[8] C. Schmid, R. Mohr, and C. Bauckhage, "Evaluation of interest point detectors," *International Journal of Computer Vision*, vol. 37, no. 2, pp. 151–172, 2000.

[9] A. Kovacs and T. Sziranyi, "High definition feature map for GVF snake by using Harris function," in *Proceeding of Conference on Advanced Concepts for Intelligent Vision Systems (ACIVS), LNCS 6475, Part 1*, 2010, pp. 163–172.

[10] ——, "Harris function based external force field for snakes," *Pattern Recognition Letters*, under revision.

[11] A. Kovacs, A. Utasi, L. Kovacs, and T. Sziranyi, "Shape and texture fused recognition of flying targets," in *Proceedings of Signal Processing, Sensor Fusion and Target Recognition XX*, vol. 8050.

[12] A. Kovacs, C. Benedek, and T. Sziranyi, "A joint approach for building localization and outline extraction," in *Proceedings of IASTED International Conference on Signal Processing, Pattern Recognition and Applications (SPPRA)*, 2011.

[13] S. Alpert, M. Galun, R. Basri, and A. Brandt, "Image segmentation by probabilistic bottom-up aggregation and cue integration," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1–8.

[14] A. Kovacs and T. Sziranyi, "Improved force field for vector field convolution method," in *Proceeding of International Conference on Image Processing*, 2011.

Electrophysiological Correlates of Object-specific Processing Deficits in Amblyopia

Petra Hermann
(Supervisor: Dr. Zoltán Vidnyánszky)
hermann.petra@gmail.com

Abstract — **Electrophysiological research on animal models of amblyopia suggests that timing and synchronization of visual cortical responses is disturbed when stimuli are presented in the amblyopic eye. However, very little is known about the changes in temporal properties of human visual cortical responses in amblyopia. Using a single trial EEG analysis approach here we show that early visual cortical responses to foveal stimulation of the amblyopic eye are delayed and less time-locked to stimulus presentation as compared to responses from the fellow eye. Importantly, the magnitude of the latency shift in the two hemispheres show stimulus specificity: it is more pronounced and correlates with interocular visual acuity in the right and left hemispheres in the cases of face and letter stimuli, respectively, thus implying that the latency shift reflects impaired higher, object-level visual cortical processing. We also find that it is the larger inter-trial latency variance in the amblyopic eye that accounts for the decreased amplitudes of the N170 component of the averaged ERP responses in the amblyopic eye relative to the fellow eye. These findings provide evidence for increased latency and timing uncertainty of visual cortical responses evoked by the stimuli presented in the amblyopic eye, which might be the primary cause of the impaired binocular vision in amblyopia.**

Keywords-amblyopia; fovea; single-trial EEG analysis; N170; timing uncertainity; object-level visual processing

I. INTRODUCTION

Ablyopia is a known developmental visual disorder characterized by loss of visual acuity and decreased contrast sensitivity. Currently it is most accepted that the earliest functional physiological abnormalities occur in cortical area V1 [1][2] - but see [3], where the author suggests that amblyopic eyes relative to eyes of control subjects have reduced innervations of comparable retinal areas and [4], which shows functionally relevant grey matter reduction in the LGN of amblyopes relative to controls. Several neuroimaging studies have found deficits in areas downstream from V1: loss in V2 comparable to that of V1 [5], deficits up to V3a and V4 along the dorsal and ventral path, respectively [6] or selective abnormality in the face-related as opposed to building-related cortical areas of higher-order occipitotemporal cortex [7][8]. Since amblyopia is believed to affect mostly foveal processing, these latter results are in line with findings that show a foveal preference for face-selective and peripheral preference for object-selective higher-order areas [9][10].

So far, however, single cell recordings in squinting cats and monkeys have failed to disclose a clear relationship between the amblyopic deficits and modifications of neuronal response properties in the visual cortex. Neurons dominated by the amblyopic eye are as numerous as neurons dominated by the normal eye and exhibit similar receptive field properties [11][12][13][14][15]. Surprisingly, even the spatial frequency tuning of neurons activated through the amblyopic eye is comparable to that of neurons activated through the normal eye, in spite of the reduced visual acuity of the amblyopic eye [13][15]. The only consistent abnormalities that have been described are prolonged latencies of responses evoked through the deviating eye [13][16]. In contrast to the results obtained with single cell recordings, studies based on pattern evoked potentials have not only demonstrated increased latencies but also reduced amplitudes of potentials evoked through the affected eye in cats [11][17][14] and humans [18][19].

This study was aimed at accounting for the well-known amblyopic effects while trying to resolve the apparent discrepancy between the absence and presence of amplitude reduction in single-cell and VEP studies, respectively. Therefore, we analyzed single trial P1 and N170 peaks and evaluated the amblyopia induced change in their distribution.

II. EXPERIMENTAL PROCEDURES

A. Subjects

Twelve amblyopic subjects (5 females, 9 right-handed, mean age: 31 years) participated in this experiment. In six cases the amblyopic eye was their right eye. None of them had any history of neurological or psychiatric diseases and all had normal or corrected-to-normal visual acuity of the dominant fellow eye (see TABLE 1. for more details).

B. Stimuli and Procedure

1) Visual Stimuli: Participants viewed images of human faces, Gabor patches and letters and performed a gender, orientation and vowel/consonant categorization task. Face-stimuli consisted of front view grayscale photographs of four female and four male neutral faces that were cropped and covered with a circular mask to eliminate the outer features. All images were equated for luminance and contrast. Male and female images were paired and warped into each other along the gender axis by a morphing algorithm (Winmorph 3.01) [20][21][22] to create intermediate images. In the Face condition (F) we presented these morphed face images with a constant morph level (5/95% gender content) across subjects and eyes. The other conditions consisted of Gabor patches with a spatial frequency of 1 cycle/degree oriented slightly to the right and left of the vertical meridian (with 5, 6, 7, 8 degrees in a random order) (Orientation condition, O) and single 55 pt Courier New font black letters ('A' 'E' 'O' 'U' 'K' 'P' 'C' 'V') on a uniform grey background also subtending 2 visual degrees (Letter condition, L).

TABLE 1. OBSERVER CHARASTERISTICS

Subject	Age	Sex	Etiolology	ref RE	ref LE	VA RE	VA RE logMAR	VA LE	VA LE logMAR	Dominant Hand	Dominant Eye
AA	38	F	strab	plus1.5 +1.75 91°	plus2.5 +1.0 84°	1	0	0.5	0.3	R	R
CJ	34	F	strab	plus1.25 -1.5 53°	plus0.25 +0.25 62°	0.2	0.7	1	0	R	L
CN	44	F	strab	plus0.75	plus2.0	1	0	0.63	0.2	L	R
HK	32	F	aniso	-0.5	plus0.5 +1.75 129°	1.6	-0.2	0.25	0.6	R	R
JL	38	M	strab	plus 1.5 -0.5 141	plus 1.75 -0.75 175°	0.5	0.3	1	0	R	L
KJ	22	M	strab	plus0.25	-0.25 -0.5 58°	0.25	0.6	2	-0.3	L	L
MR	20	F	aniso	plus1.75 +1.25 101°	-1.0 +0.75 82°	0.25	0.6	1	0	R	L
SA	39	M	strab	plus1.25 -1.25 11°	plus0.5 +1.5 95°	1.6	-0.2	0.8	0.1	R	R
SI	23	M	strab	plus1.5 +1.25 100°	plus2.75 +0.5 63°	0.5	0.3	1.6	-0.2	R	L
TK	24	M	strab	plus2.25 +1.0 177°	plus3.75 +1.75 117°	1.25	-0.1	0.63	0.2	L	R
VeA	36	M	aniso	plan	plus2.5	1.6	-0.2	0.33	0.5	R	R
VoA	24	M	aniso	-0.25 -1.75 97°	-3.0 -0.75 73°	0.25	0.6	1.25	-0.1	R	L

C. Stimuli and Procedure

1) Visual Stimuli: Participants viewed images of human faces, Gabor patches and letters and performed a gender, orientation and vowel/consonant categorization task. Face-stimuli consisted of front view grayscale photographs of four female and four male neutral faces that were cropped and covered with a circular mask to eliminate the outer features. All images were equated for luminance and contrast. Male and female images were paired and warped into each other along the gender axis by a morphing algorithm (Winmorph 3.01) [20][21][22] to create intermediate images. In the Face condition (F) we presented these morphed face images with a constant morph level (5/95% gender content) across subjects and eyes. The other conditions consisted of Gabor patches with a spatial frequency of 1 cycle/degree oriented slightly to the right and left of the vertical meridian (with 5, 6, 7, 8 degrees in a random order) (Orientation condition, O) and single 55 pt Courier New font black letters ('A' 'E' 'O' 'U' 'K' 'P' 'C' 'V') on a uniform grey background also subtending 2 visual degrees (Letter condition, L).

All stimuli were presented centrally on a uniform gray background and subtended 2 visual degrees, matching approximately the size of the fovea.

2) Procedure: Gender, orientation and letter categorization was measured by a two-alternative forced choice procedure. Subjects were required to judge the gender of the face images (female/male), the orientation of the Gabor patches (left/right) and the category of the letters (vowel/consonant) as accurately and fast as possible, indicating their choice with one of the mouse buttons. Button assignment was left for one category and right for the other category for half of the subjects and was reversed for the other half. Each stimulus was presented for 250 ms followed by a response window which lasted until the subjects responded but was maximized in 2 s. The inter-trial interval (ITI) was randomized in the range of 1800–2200 ms. The fixation point was present throughout the entire trial. The conditions were presented in separate blocks in random order. Viewing was monocular with the amblyopic eye (AE) in one block and with the dominant fellow eye (FE) in another while the unused eye was patched. Each participant completed four runs for each eye yielding 192 trials altogether for each condition per eye.

Stimulus presentation was controlled by MATLAB 7.1. (The MathWorks Inc., Natick, MA) using the Cogent 2000 toolbox (www.vislab.ucl.ac.uk/Cogent/) and were presented on a 26" LG LCD monitor at a refresh rate of 60 Hz and were viewed from 56 cm.

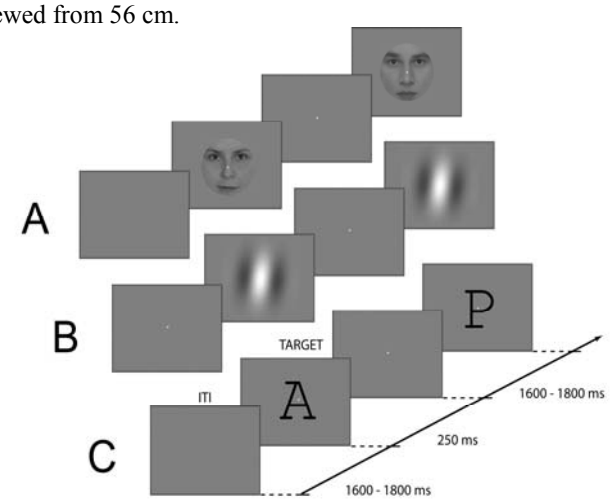


Fig 1. Experimental design. (A) Morphed face images with constant morph level (5/95% gender content) in the gender categorization task (Face condition). (B) Gabor patches with spatial frequency of 1 cycle/degree oriented slightly to the right and left of the vertical meridian (5,6,7,8 degrees) in the orientation categorization task (Orientation condition). (C) Single 55 pt Courier New font black letter sequence showing the vowel/consonant categorization task (Letter condition).

D. Data Analysis

1) Behavioral Data Analysis: Responses and reaction times were collected during the experiment and entered into 2-way repeated-measures ANOVAs with condition (F vs. O vs. L) and eye (FE vs. AE) as within subject factors. Post-hoc t-tests were computed using Tukey HSD tests.

2) Electrophysiological Recording and Analysis: EEG acquisition and processing. EEG data were acquired using a

BrainAmp MR (Brainproducts GmbH., Munich, Germany) amplifier from 60 Ag/AgCl scalp electrodes placed according to the extended 10-20 international electrode system and mounted on an EasyCap (Easycap GmbH, Herrsching-Breitbrunn, Germany) with four additional periocular electrodes placed at the outer canthi of the eyes and above and below the right eye for the purpose of recording the electrooculogram. All channels were referenced to joint earlobes online; the ground was placed on the nasion. All input impedance was kept below 5 kΩ. Data were sampled at 1000 Hz with an analog bandpass of 0.016–250 Hz and re-referenced offline using a Laplacian transform on spherical spline interpolated data (4th order splines, maximum degree of Legendre polynomials:10, lambda: 10-5) to generate scalp current density (SCD) waveforms. The SCD data is reference independent and displays reduced volume conduction eliminating raw EEG contamination from saccadic potentials [23]. Moreover its peaks and troughs are sharper and larger than those of the original scalp potential [24], which makes it better suited for single-trial peak detection compared to raw surface potentials [25][24]. Subsequently, a digital 0.1 Hz 12 dB/octave Butterworth Zero Phase high-pass filter was used to remove DC drifts, and a 50 Hz notch filter was applied to minimize line-noise artifacts. Finally, a 24 dB/octave low-pass filter with a cutoff frequency of 30 Hz was applied. Data was segmented (see below) and trials that contained voltage fluctuations exceeding ±100 μV, or electro-oculogram activity exceeding ±50 μV were rejected. Data processing was done using BrainVision Analyzer (Brainproducts GmbH., Munich, Germany).

3) ERP Data Analysis: The trial-averaged EEG waveform – i.e. the event-related potential (ERP) – was computed as follows. Data was segmented into 800 ms epochs starting from 200 ms preceding the stimuli. Segments were baseline corrected over a 200 ms pre-stimulus window, artifact rejected and averaged to obtain the ERP waveforms for each subject for each condition. Subject ERPs were averaged to compute the grand average ERP for visualization purposes. Statistical analysis was performed on early component peaks (P1, N170) of the averaged ERP waveform. Early peak amplitudes were computed as follows: peak latency was determined individually on pooled electrodes from left and right clusters separately, while mean peak amplitudes were measured over the individual electrodes in the above clusters in a 10 ms window centered on the peak latencies. The clusters included electrodes where P1 and N170 showed their maxima (PO7, PO9, P7 and P9, and PO8, PO10, P8 and P10 for left and right clusters, respectively), which happened to coincide due to the SCD transform. Amplitude values were analyzed by four-way repeated-measure ANOVAs with condition (F vs. O vs. L), eye (FE vs. AE), side (2) and electrode (4) as within-subject factors separately for each component. Greenhouse-Geisser correction was applied to correct for possible violations of sphericity. Post-hoc t-tests were computed using Tukey HSD tests.

Single-trial peak detection was also performed in a similar manner. For P1 and N170 minima and maxima, respectively were detected on each trial for each electrode in a 100 ms time window centered on the individual peak latency of the respective component measured on the averaged ERPs. The amplitude and corresponding time of the extrema were taken as the amplitude and latency of the component on the given trial. The trial was rejected if the detected extrema was located at the beginning or end of the time window. The single trial amplitude and latency values were pooled from the four electrodes on each side and the distribution of the values was characterized by calculating the median and the interquartile range (IQR). The IQR is used as a measure of spread and is computed as the difference of the upper and lower quartile of the data, and thus describes the middle 50% of the data values. The median and IQR values were entered into two-way repeated-measures ANOVAs with eye (FE vs. AE) and side (2) as within-subject factors separately for each component. Post-hoc t-tests were computed using Tukey HSD tests.

The purpose of the single-trial peak detection analysis was to account for the observed changes in the averaged ERPs and behavioral results between the two eyes. For this we calculated an amblyopia index (AI) for each measure as follows:

$$AI = FE - AE$$

and used this index in correlational and regressional analyses. We assessed the relationship between the relative changes (AI) in median and IQR of the distributions between eyes and interocular visual acuity (VA) expressed in logMAR values obtained at a distance of 5 m with the best refractive correction. For this Pearson correlation was used and to correct for multiple comparisons (c=X), significance threshold was set to pBonf =0.05 (punc=0.00X).

III. Results

A. Behavioral Results

The tasks were designed so that they could be carried out viewing with either eyes (FE and AE) and thus, performance was at ceiling. Nevertheless, accuracy significantly differed between eyes in all cases the fellow eye performing better compared with the amblyopic eye (main effect of eye: $F(1,11)=6.37$, $p=0.028$). Moreover, accuracy between conditions also significantly differed from one another, overall accuracy in the face condition being slightly worse (main effect of condition: $F(2,22)=7.16$, $p=0.014$). This was the result of patents performing this task significantly worse only in the case of amblyopic viewing (condition × eye: $F(2,22)=5.67$, $p=0.032$), see Figure 1A. Reaction times also differed between eyes and conditions, the responses with amblyopic viewing being significantly slower (main effect of eye: $F(1,11)=45.37$, $p<0.0001$) and the orientation condition being significantly faster than the other two (main effect of condition: $F(2,22)=22.67$, $p<0.0001$), see Figure 1B.

B. Amblyopic Effects on Amplitude and Latency

Electrophysiological results revealed that amblyopia has a profound effect on the amplitude and latency of the early event-related potential (ERP) components similarly to those of the visual-evoked potentials (VEPs) of clinical studies [18][19]. In all cases, viewing with the amblyopic eye resulted in reduced amplitudes (main effect of eye: $F(1,11)=12.78$, $p=0.0044$ and $F(1,11)=15.03$, $p=0.0026$ for the components P1 and N170, respectively) and increased latencies (main effect of eye: $F(1,11)=44.37$, $p<0.0001$ and $F(1,11)=47.63$, $p<0.0001$ for the components P1 and N170, respectively) compared with

the fellow eye for both early ERP components (see Figure 3A).

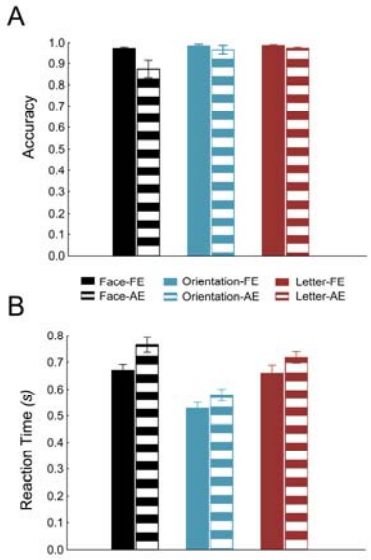


Fig 2. Behavioral results. (A) Accuracy in the three categorization task. Although in all conditions the fellow eye performed better compared with the amblyopic eye, only in the face categorization task did the accuracy differ significantly (post-hoc t-test: $p<0.005$) (see *Behavioral Results* for more details). (B) The reaction times differed between eyes and conditions: the responses with amblyopic viewing being significantly slower and the orientation condition being significantly faster than the other two.

C. *Amblyopic Effects on Amplitude and Latency*

Electrophysiological results revealed that amblyopia has a profound effect on the amplitude and latency of the early event-related potential (ERP) components similarly to those of the visual-evoked potentials (VEPs) of clinical studies [18][19]. In all cases, viewing with the amblyopic eye resulted in reduced amplitudes (main effect of eye: $F(1,11)=12.78$, $p=0.0044$ and $F(1,11)=15.03$, $p=0.0026$ for the components P1 and N170, respectively) and increased latencies (main effect of eye: $F(1,11)=44.37$, $p<0.0001$ and $F(1,11)=47.63$, $p<0.0001$ for the components P1 and N170, respectively) compared with the fellow eye for both early ERP components (see Figure 3A). Moreover, the ERP evoked by facial stimuli significantly differed from the rest of the conditions. More specifically, both components were larger in amplitude ($F(2,22)=18.47$, $p<0.0001$ and $F(2,22)=4.71$, $p=0.045$ for the components P1 and N170, respectively) in the face condition compared with the other two conditions in the case of the P1 component (post-hoc t-test: $p<0.005$ both for F vs. O and F vs. L), while for the N170 it only differed from N1 evoked by the Gabor patches but not from that evoked by letters ($p=0.021$, $p=0.078$ for F vs. O and F vs. L). The evoked components in the case of the letter condition on the other hand, differed from the rest in respect their faster latency. In both components, the component evoked in the case of viewing with the fellow eye was significantly faster compared to other stimulation to the same eye. However, the N170 latencies of the different conditions evoked with amblyopic viewing didn't differ from one another (condition × eye interaction: $F(2,22)=3.76$, $p=0.042$ and $F(2,22)=7.97$, $p=0.0071$ for the components P1 and N170, respectively).

D. *Single-trial EEG Analysis*

Importantly, however, the amplitude of the trial-averaged ERP is dependent on two factors: the amplitude of the individual peaks and their jitter over time. The higher the individual amplitudes and/or the smaller the latency jitter, the higher and sharper the peak of resulting averaged ERP will be. The ERP peak latency on the other hand can be modulated by either the lag of the individual peaks and their jitter over time. The more delayed the peaks and/or the larger the latency jitter – especially if the distribution is skewed to longer latencies –, the later the resultant average ERP will peak. Therefore, we analyzed single trial P1 and N170 peaks and evaluated the amblyopia induced change in their distribution (see *Experimental Procedures* for more details). In general in the case of the P1 component, there was no or only slight – in the case of face stimuli – amblyopia related change in the amplitude distribution, while there was a significant shift in the latency distribution of the amblyopic relative to the fellow

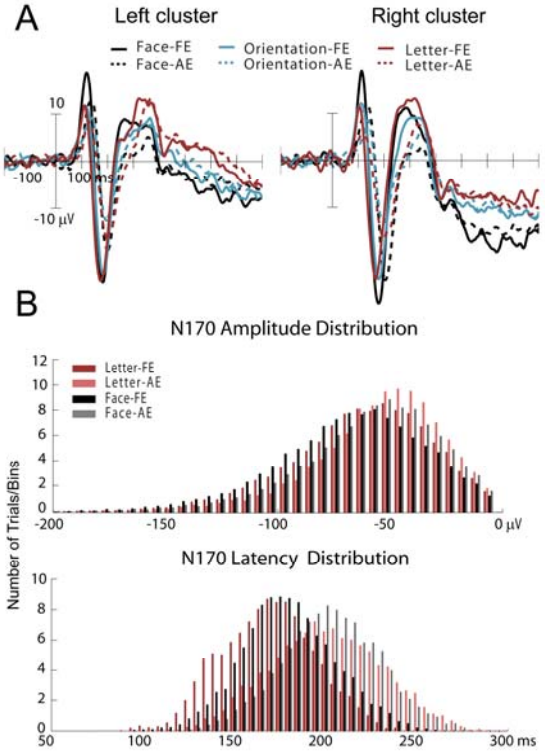


Fig 3. Electrophysiological results. (A) Amblyopic effects on the grand average ERPs in the left and right cluster (see *Experimental Procedures* for more details). The amblyopic eye resulted in reduced amplitudes and increased latencies of the early visual ERP components (including P100 and N170) compared with the fellow eye in all conditions (see *Amblyopic Effects on Amplytude and Latency* for more details). (B) Amblyopia related change in the N170 amplitude and latency distribution. The amblyopic eye displayed overall smaller amplitudes and a significant shift towards longer latencies and increased spread compared with the fellow eye in all conditions (see *Single-trial EEG Analysis* for more details).

eye along with a consistent increase in spread, indicating bigger inter-trial latency jitter, thus worse time-locking to the stimulus, which decreases the amplitude of the averaged ERPs (see TABLE 2A). On contrary to the P1, the N170 amplitude distribution significantly differed between the two eyes: the

amblyopic eye displayed overall smaller amplitudes compared with the fellow eye in all conditions (see Figure 3B), while only in the face conditions they were overall larger over right-side electrodes irrespective of eyes (see TABLE 2B). The difference in the N170 latency distributions was similar to that of observed in the case of P1: a significant shift towards longer latencies and increased spread in the amblyopic compared to the fellow eye (see Figure 3B). Importantly, in the case of face stimuli, the latencies in the amblyopic eye were more delayed relative to the fellow eye over right-side compared with left-side electrodes.

TABLE 2A

Condition	Effect	P1 Latency IQR		P1 Latency Median	
		F(1,11)	p	F(1,11)	p
Face	Eye (AF>FE)	18.2853	0.001307	22.155	0.000643
Orientation	Eye (AF>FE)	18.7634	0.001191	65.3356	0.000006
Letter	Eye (AF>FE)	13.3846	0.003764	67.969	0.000005

a. Significant main effect of 'eye' on the P1 latency distribution in all conditions.

b. See *Experimental Procedures* for more information on IQR and Median.

TABLE 2B

Condition	Effect	N1 Latency IQR		N1 Latency Median	
		F(1,11)	p	F(1,11)	p
Face	Eye (AF>FE)	11.1593	0.006587	34.393	0.000109
Orientation	Eye (AF>FE)	10.823	0.007206	29.904	0.000195
Letter	Eye (AF>FE)	9.7917	0.009592	69.035	0.000005

a. Significant main effect of 'eye' on the N1 latency distribution in all conditions.

b. See *Experimental Procedures* for more information on IQR and Median.

E. Correlations

To evaluate the right-laterized amblyopic effects on the N170 in the case of facial stimuli that are know to be processed dominantly in the right hemisphere, we correlated the single trial measures with interocular visual acuity (VA logMAR). In agreement with the ANOVA findings, the amblyopic deficits in face processing were confined to the right hemisphere: Spearman correlations were only significant with ERP measures derived form the right hemisphere (interocular VA vs. R latency median difference: r=6.35, p=0.026), while there were absolutely no correlations with those derived from the left hemisphere. Interestingly, in the case of letter stimuli that are known to be processed dominantly in the left hemisphere the correlation pattern was different from that observed in the case of face stimuli: interocular latency median differences in both hemispheres correlated with interocular VA (r=6.16, p=0.033 and r=7.13, p=0.009 for left and right latency median, respectively), while in the case of latency spread only the left hemisphere showed correlation (r=5.77, p=0.050).

IV. CONCLUSIONS

Taken together, the above findings underline timing uncertainty of visual processing starting from the amblyopic eye, which manifests itself in delayed responses that are also less time-locked to stimulus presentation. This larger inter-trial latency jitter completely accounts for the decreased amplitudes of P1 in the amblyopic eye and also contributes to the N170 amplitude decrease. Importantly, the above deficits are object-specific, i.e. tied to the specific higher-level processing that the brain is engaged with. Finally, significant timing differences between the eyes might also be the primary cause of the impaired binocular vision in amblyopia.

REFERENCES

[1] L. Kiorpes and S. P. McKee, "Neural mechanisms underlying amblyopia," Current Opinion in Neurobiology, vol. 9, no. 4, pp. 480-486, Aug. 1999.

[2] L. Kiorpes and J. A. Movshon, "Neural Limitations on Visual Development in Primates," in The visual neurosciences, MIT Press, 2003.

[3] P. Lempert, "Retinal area and optic disc rim area in amblyopic, fellow, and normal hyperopic eyes: a hypothesis for decreased acuity in amblyopia," Ophthalmology, vol. 115, no. 12, pp. 2259-2261, Dec. 2008.

[4] G. R. Barnes, X. Li, B. Thompson, K. D. Singh, S. O. Dumoulin, and R. F. Hess, "Decreased gray matter concentration in the lateral geniculate nuclei in human amblyopes," Investigative Ophthalmology & Visual Science, vol. 51, no. 3, pp. 1432-1438, Mar. 2010.

[5] G. R. Barnes, R. F. Hess, S. O. Dumoulin, R. L. Achtman, and G. B. Pike, "The cortical deficit in humans with strabismic amblyopia," The Journal of Physiology, vol. 533, no. 1, pp. 281-297, May. 2001.

[6] X. Li, S. O. Dumoulin, B. Mansouri, and R. F. Hess, "Cortical deficits in human amblyopia: their regional distribution and their relationship to the contrast detection deficit," Investigative Ophthalmology & Visual Science, vol. 48, no. 4, pp. 1575-1591, Apr. 2007.

[7] Y. Lerner et al., "Area-specific amblyopic effects in human occipitotemporal object representations," Neuron, vol. 40, no. 5, pp. 1023-1029, Dec. 2003.

[8] Y. Lerner et al., "Selective fovea-related deprived activation in retinotopic and high-order visual cortex of human amblyopes," NeuroImage, vol. 33, no. 1, pp. 169-179, Oct. 2006.

[9] U. Hasson, I. Levy, M. Behrmann, T. Hendler, and R. Malach, "Eccentricity bias as an organizing principle for human high-order object areas," Neuron, vol. 34, no. 3, pp. 479-490, Apr. 2002.

[10] U. Hasson, M. Harel, I. Levy, and R. Malach, "Large-scale mirror-symmetry organization of human occipito-temporal object areas," Neuron, vol. 37, no. 6, pp. 1027-1041, Mar. 2003.

[11] W. Singer, M. von Grünau, and J. Rauschecker, "Functional amblyopia in kittens with unilateral exotropia. I. Electrophysiological assessment," Experimental Brain Research. Experimentelle Hirnforschung. Expérimentation Cérébrale, vol. 40, no. 3, pp. 294-304, 1980.

[12] G. D. Mower, J. L. Burchfiel, and F. H. Duffy, "Animal models of strabismic amblyopia: physiological studies of visual cortex and the lateral geniculate nucleus," Brain Research, vol. 281, no. 3, pp. 311-327, Nov. 1982.

[13] Y. M. Chino, W. H. Ridder 3rd, and E. P. Czora, "Effects of convergent strabismus on spatio-temporal response properties of neurons in cat area 18," Experimental Brain Research. Experimentelle Hirnforschung. Expérimentation Cérébrale, vol. 72, no. 2, pp. 264-278, 1988.

[14] R. D. Freeman, G. Sclar, and I. Ohzawa, "An electrophysiological comparison of convergent and divergent strabismus in the cat: visual evoked potentials," Journal of Neurophysiology, vol. 49, no. 1, pp. 227-237, Jan. 1983.

[15] D. P. Crewther and S. G. Crewther, "Neural site of strabismic amblyopia in cats: spatial frequency deficit in primary cortical neurons," Experimental Brain Research. Experimentelle Hirnforschung. Expérimentation Cérébrale, vol. 79, no. 3, pp. 615-622, 1990.

[16] G. W. Eschweiler and J. P. Rauschecker, “Temporal integration in visual cortex of cats with surgically induced strabismus,” The European Journal of Neuroscience, vol. 5, no. 11, pp. 1501-1509, Nov. 1993.

[17] M. W. von Grünau and W. Singer, “Functional amblyopia in kittens with unilateral exotropia. II. Correspondence between behavioural and electrophysiological assessment,” Experimental Brain Research. Experimentelle Hirnforschung. Expérimentation Cérébrale, vol. 40, no. 3, pp. 305-310, 1980.

[18] R. E. Manny and D. M. Levi, “The visually evoked potential in humans with amblyopia: pseudorandom modulation of uniform field and sine-wave gratings,” Experimental Brain Research. Experimentelle Hirnforschung. Expérimentation Cérébrale, vol. 47, no. 1, pp. 15-27, 1982.

[19] V. Parisi, M. E. Scarale, N. Balducci, M. Fresina, and E. C. Campos, “Electrophysiological detection of delayed postretinal neural conduction in human amblyopia,” Investigative Ophthalmology & Visual Science, vol. 51, no. 10, pp. 5041-5048, Oct. 2010.

[20] G. Kovács, M. Zimmer, I. Harza, A. Antal, and Z. Vidnyánszky, “Position-specificity of facial adaptation,” Neuroreport, vol. 16, no. 17, pp. 1945-9, Nov. 2005.

[21] G. Kovács, M. Zimmer, E. Bankó, I. Harza, A. Antal, and Z. Vidnyánszky, “Electrophysiological correlates of visual adaptation to faces and body parts in humans,” Cerebral Cortex (New York, N.Y.: 1991), vol. 16, no. 5, pp. 742-53, May. 2006.

[22] G. Kovács, M. Zimmer, I. Harza, and Z. Vidnyánszky, “Adaptation duration affects the spatial selectivity of facial aftereffects,” Vision Research, vol. 47, no. 25, pp. 3141-9, Nov. 2007.

[23] L. Melloni, C. M. Schwiedrzik, E. Rodriguez, and W. Singer, “(Micro)Saccades, corollary activity and cortical oscillations,” Trends in Cognitive Sciences, vol. 13, no. 6, pp. 239-245, Jun. 2009.

[24] F. Perrin, O. Bertrand, and J. Pernier, “Scalp Current Density Mapping: Value and Estimation from Potential Data,” Biomedical Engineering, IEEE Transactions on, vol. 34, no. 4, pp. 283-288, Apr. 1987.

[25] L. Trujillo, M. Peterson, A. Kaszniak, and J. Allen, “EEG phase synchrony differences across visual perception conditions may depend on recording and analysis methods,” Clinical Neurophysiology, vol. 116, no. 1, pp. 172-189, 2005.

Parameter Estimation of LTI Systems Using Dynamic Bayesian Networks

Zoltan Tuza
(Supervisors: Dr. Gábor Szederkényi and Dr. Katalin Hangos and Dr. Kristóf Karacs)
tuza.zoltan@itk.ppke.hu

***Abstract*—In this paper the parameter estimation of Linear Time Invariant systems will be evaluated in the Dynamic Bayesian Network (DBN) framework. For parameter estimation the Expectation Maximization algorithm will be used which can cope with missing or hidden data - commonly occur in dynamical systems where not all the states are directly measured.**

***Index Terms*—LTI systems, parameter estimation, Dynamic Bayesian Network, EM**

I. Introduction

System identification, particularly Linear Time Invariant (LTI) system identification has a long standing history [1], several traditional approaches like least square estimation or instrumental variables are widely accepted and used by engineers. However, these algorithms have difficulties to incorporate restrictions and *a priori* information. In this paper the Dynamic Bayesian Network (DBN) [2] model will be evaluated for LTI system modelling and parameter estimation. This model is a graph where each node represents a random variable (r.v.) and the graph structure expresses the conditional independence between them. In this way the posterior probabilistic function (PDF) is represented over the modelled domain. With DBN several well-known algorithm can be described, some example of them Principal Component Analysis (PCA), Hidden Markov Model (HMM), Independent Component Analysis (ICA), Kalman filtering, factor analysis, etc. A detailed review about these algorithms and their representation in the BN framework can be found in [3].

Our short term goal is to estimate the parameters of an uncertain LTI system with fixed structure using the BN framework. In this way many different information can be incorporated in the network about the LTI system. The long term goal is to identify the structure of complex dynamical systems including non-linear systems. The first step towards these goals is a case study presented in this paper where two simple LTI systems are evaluated as DBNs.

II. LTI State Space Models and Dynamic Bayesian Networks

A. State Space Model

The following discrete linear time invariant dynamical system model with additive Gaussian noise will be used in this paper.

$$\begin{aligned} x_{t+1} &= \Phi x_t + w_t \\ y_t &= C x_t + v_t \end{aligned} \tag{1}$$

where $x_t \in \mathcal{R}^n$ n dimensional vector represents the state of the dynamical system at time t. The time invariant $\Phi \in \mathcal{R}^{nxn}$ matrix governs the dynamic of the system. $y_t \in \mathcal{R}$ is the scalar output of the system, which is the linear function of the state through the $C \in \mathcal{R}^{1xn}$ matrix. Both state and measurement noise, w_t and v_t, are zero-mean normally distributed r.v. with covariance matrices Q and R, respectively [4]. All the matrices are assumed time invariant. Rather then regarding state as a deterministic value corrupted by random noise, we combine the state variable and the state noise variable into a single Gaussian r.v. We form a similar combination for the output as well. In this way the measurement and the state transition form the following conditional probability [5].

$$\begin{aligned} &P(y_t|x_t) = \\ &\quad exp\{-\tfrac{1}{2}[y_t - Cx_t]R^{-1}[y_t - Cx_t]\}(2\pi)^{-1/2}|R|^{-1/2} \\ &P(x_{t+1}|x_t) = \\ &\quad exp\{-\tfrac{1}{2}[x_{t+1} - \Phi x_t]Q^{-1}[x_{t+1} - \Phi x_t]\}(2\pi)^{-1/2}|Q|^{-1/2} \end{aligned} \tag{2}$$

B. Parameter estimation

Let $\mathcal{M}$ be the set of possible models of a dynamical system, then the model set is decomposed into structural and parametric levels. Let $\mathcal{S}$ be a set of model structures. A model structure $S \in \mathcal{S}$ is a set of possible dynamical system model structures. The unknown parameter vector θ belongs to a parameter set Θ_S, which is a subset of of a finite-dimensional Euclidean vector space. This defines the model set

$$\mathcal{M} = \{M = (S, \theta) | S \in \mathcal{S}, \theta \in \Theta_S\} \tag{3}$$

A comprehensive introduction for parameter estimation can be found in [1]. In the rest of the paper, we assume that the model structure S is known and fixed in advance.

C. Graph models and Bayesian networks

A graphical model or graph model is a family of probability distributions defined in terms of a directed or undirected graph. The nodes in the graph are referred to r.v. In the directed case, which is called Bayesian Network, let $\mathcal{G}(\mathcal{V}, \mathcal{E})$ be a directed acyclic graph where $\mathcal{V}$ are the nodes and $\mathcal{E}$ are the edges of the graph. The edges denote direct causal influences, defining the model structure, too. Let $X_v : v \in \mathcal{V}$ be a collection of r.v. indexed by the nodes of the graph. To each node $v \in \mathcal{V}$, let $pa(X_v)$ denote the subset of indices of its parents. Given a collection of conditional PDF $P(x_v|pa(x_v)) : v \in \mathcal{V}$ that sum (in the discrete case) or integrate (in the continuous case) to one (with respect to x_v) we define a joint probability distribution as follows:

$$P(x_v) = \prod P(x_v|pa(x_v)) \tag{4}$$

Regardless of the functional form of $P(p_v|pa(x_v))$ the factorization of this equation implies a set of conditional independence statements can be obtained from a polynomial time reachability algorithm based on the graph. This algorithm is called D-separation [6]. The structure of the graph can be exploited by algorithms for probabilistic inference [7].

1) Bayesian Networks: Bayesian Network (BN) is a graphical way to represent a particular factorization of a joint PDF. The directed graph allows us to represent the causal structure of the modelled domain or system. In this model the nodes represent the uncertain quantities such as state of the system or noisy measurement. The multifaceted nature of BNs originates from that this representation addresses jointly three autonomous levels of the domain: the causal model, the probabilistic dependency-independency structure, and the distribution over the uncertain quantities. Additionally, the BN, as a complete probabilistic domain model, can be applied as an input-output system model [8].

The semantics of a BN is simple: each node is conditionally independent from its non-descendants given its parents. More generally, two disjoint set of nodes A and B are conditionally independent given C, if C D-separates A and B.

The nodes in a BN can be divided into two categories. The first one is called observable node, this represent our direct measurement or our known input of the system. The second type of node is the hidden node, which we cannot measure directly rather than we just measure the effect of this node through one or more observable node(s). If the causality of events have temporal ordering, then a DBN is needed where the joint PDF can change over time.

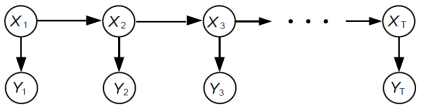


Fig. 1. Dynamic Bayesian network representation of the state space model. X_i nodes are hidden, and Y_i nodes are observed

2) Dynamic Bayesian Networks: Dynamic Bayesian Networks (DBNs) are directed graph models of stochastic processes. The direction of edges points toward the flow of time. In this way a wide range of spacio-temporal phenomena can be modelled. A DBN is Bayesian network where the structure is repeated over time as it can be seen in Figure 1 and in eq (5).

$$P(X_{1:T}, Y_{1:T}) = P(X_1)P(Y_1|X_1)\prod_{t=2}^{T} P(X_t|X_{t-1})P(Y_t|X_t) \quad (5)$$

where $X_{1:T}, Y_{1:T}$ denote the sequences form $t = 1...T$. If we consider additive Gaussian noise, the state transition probability $P(x_t|x_{t-1})$ and the measurement probability $P(y_t|x_t)$ is the same as in eq (2). The graph representation of the state-space model is shown in Figure 1. The first time-slice of the state-space DBN is important because the *a priori* knowledge and the initial condition are stored there. It is possible that the connection between the first time-slice and the rest of the model will be different. Although its called DBN, the model structure is not allowed to change over time other than in the first transition. Two possible advantages of DBN over traditional LTI model and identientification techniques are the *a priori* knowledge we can incorporate into the network and the possibility of maintaining control over the computational cost associated with DBN via the structure of the graph.

D. Learning and Inference

Learning and inference are the two major operations associated with BNs. Learning is the system identification in BN terminology. Structure learning is the model selection problem, which is a super-exponential problem in the number of nodes in the graph. For that reason proper *a priori* information is crucial to keep the complexity of the problem low. With fixed model structure the learning process is the parameter estimation, where the properties of nodes (random variables) are estimated. Inference is a process in which the known data propagated through the network and the output is measured. A detailed introduction can be found in [2].

The Bayesian system identification in nutshell is described below. First we assume a prior distribution over the possible model structures $P(\mathcal{M})$, a prior distribution over parameters for each model structure $P(\theta|\mathcal{M})$ and a data set $\mathcal{D}$. For a given model structure, we can compute the posterior distribution over the parameters:

$$P(\theta|\mathcal{M}, \mathcal{D}) = \frac{P(\mathcal{D}|\theta, \mathcal{M})P(\theta|\mathcal{M})}{P(\mathcal{M}|\mathcal{D})} \quad (6)$$

For eq (6) it is necessary to compute the likelihood of data set, which is

$$P(\mathcal{D}|\theta, \mathcal{M}) = \prod_{i=1}^{N} P(Y_i|\theta, \mathcal{M}) \quad (7)$$

we will use the Maximum Likelihood (ML) estimation to estimate the parameters of the PDF. In our case the parameters are Gaussian r.v., which can be fully characterized with their first two moments. If the observation vector includes all the variables in the BN, then the log likelihood factors as:

$$\mathcal{L}(\theta) = \sum_{i=1}^{N}\sum_{j} logP(Y_j^{(i)}|Y_{pa(j)}^{(j)}, \theta_j) \quad (8)$$

where $j \in \mathcal{V}$, pa(j) is the parents of j. If we using this ML estimation with Gaussian r.v., it can be shown that it is equivalent with the least squares estimate.

If not all nodes are directly measured, then the factorization in eq (8) cannot be made, rather we find:

$$\mathcal{L}(\theta) = log\, P(Y|\theta) = log \sum_{X} P(Y, X|\theta) \quad (9)$$

where X is the set of hidden variables, and $\sum_X$ is the sum (or integral) over X required to obtain the marginal probability of the data. This likelihood function can be computed using the Expectation Maximization algorithm [9]. It contains two steps, the first one is expectation, which calculates the expectation of likelihood condition on the data set.

$$\mathcal{Q} = E[logP(x_{1:T}, y_{1:T})|y_{1:T}] \quad (10)$$

In the second, maximization step the partial derivatives of the expected log likelihood are calculated in order to find the minimum of the negative log likelihood. In case of the state space model, the Φ matrix is calculated as follows:

$$\begin{aligned} \frac{\partial \mathcal{Q}}{\partial \Phi} &= -\sum_{t=2}^{T} Q^{-1} P_{t,t-1} + \sum_{t=2}^{T} Q^{-1} \Phi P_{t,t-1} = 0 \\ \Phi^{new} &= \left(\sum_{t=2}^{T} P_{t,t-1} \right) \left(\sum_{t=2}^{T} P_{t,t-1} \right)^{-1} \end{aligned} \quad (11)$$

where $P_{t,t-1} = E[x_t x_t | y_{1:T}]$, which is the conditional expectation of the covariance matrix of the state vector conditioned on the data set. The detailed description of the EM algorithm in case of state space model can be found in [9], [5].

III. Related Works

The Bayesian approach for state estimation and control was introduced by Y. Ho [10], this paper shows the Bayesian viewpoint of Kalman Filtering. Later on the DBN representation of Kalman Filter was given. Chenś review paper gives a comprehensive tutorial on Bayesian Filtering where the details of graph state-space model can be found [11]. With this connection an LTI system can be represented as a BN either with vector valued nodes or with scalar value nodes. K. Murphyś PhD thesis considers the DBN in general, where state-space specific DBN have been studied [2]. This thesis also summarizes the fundamental inference and learning algorithms for DBNs, with exact and approximate algorithms. Shumway et al. investigated the EM algorithm for parameter estimation of LTI systems [9]. With the EM algorithm the state transition matrix Φ, the measurement R and process noise Q covariance matrices were estimated. Zoubin Ghahramai also used the EM algorithm, where in the Expectation step the Kalman-Rauch recursion was used [5]. ML estimate of conditional linear gaussian distribution is described by K. Murphy, an EM extension of this method is applied in BNet Matlab Toolbox for parameter estimation [12], [13]. The literature of System Modelling using DBN is very limited, R. Deventerś PhD thesis introduces the concept of modeling and control of dynamical systems using DBN. However his thesis just exploited the Kalman Filter and DBN equivalence. H. Xiong used an extended version of the EM algorithm where biologically motivated structural constraints were included are used to limit the parameter search space [14].

IV. Case Studies

Two different autonomous LTI systems were examined from parameter estimation perspective. The structure of these systems were known in advance which simplifies the learning process.

A. BNet Toolbox

BNet is a Matlab toolbox is written by Kevin Murphy and many others [13]. It supports discrete and continuous r.v., many inference and learning techniques with exact and sampling based version. To perform inference the variables elimination and the junction tree algorithms are available [2]. Structure and parameter learning are supported. Our experiments were carried out using BNet.

B. Experiment Setup

First, we wrote a Matlab representation for each LTI system and we used them as learning examples. Since the DBN implementation in BNet deals with only discrete systems, the continuous model was sampled with first order hold sampling to keep the original structure of the A matrix. In the experiments first vector valued, then scalar valued nodes were used, and all nodes were vector-valued Gaussian r.v. In case of the scalar valued nodes the structure of the DBN resembles the structure of the LTI. In the simulation part random samples were generated at the first time slice and they were propagated through the DBN. In each time-slice hidden nodes represent the state trajectory and the observed nodes are the measured output. For the parameter estimation part, a blank DBN was created with fixed structure. Several sets of the output of the LTI system with different initial state was collected. Then the BNetś built-in EM algorithm was applied to estimate the Φ, Q and R matrices.

C. Mass-Spring-Damper system

The following mass-spring-damper system [4] were used for simulation and parameter estimation. The equation (12) shows the differential equation of the mass-spring-damper system.

$$M\ddot{x}_1 = -kx_1 - k_d\dot{x}_1 \quad (12)$$

M is the weight of the mass, k is the spring constant and k_d is the damping constant. The corresponding state-space model is the following:

$$\begin{aligned} \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} &= \begin{bmatrix} 0 & \frac{1}{M} \\ -\frac{k}{M} & -\frac{k_d}{M} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + w \\ y &= Cx + v \end{aligned} \quad (13)$$

Since no external force was applied, the Γ matrix is zero. The matrix C will be defined at the actual experiment. The w and v were defined in eq. 1. In both representation the fixed parameters were the following: $Q = 1e-4 \times I$, $R = 1e-4 \times I$, $x(0) = \begin{bmatrix} 1 & 1 \end{bmatrix}^T$ and the initial covariance matrix was $V = 1e-4 \times I$. The parameters of the LTI system are: $k = 4$, $k_d = 1$ and $M = 1$. The sampling time was $T_s = 0.1s$.

Simulations First, the LTI-DBN equivalence was tested. Figure 2. shows the states and the outputs of the DBN and the LTI. The state trajectory is almost the same in the two cases, the difference of trajectories is shown in Figure 3. this error is due to numerical error and the additive Gaussian noise. The C matrix was an identity matrix to compare the state trajectory in both representation.

Learning An empty DBN was created with the same structure as in the simulation part. For learning examples 10-20 outputs were generated with different initial states. We used different initial states for every learning example and we also fixed the structure of the C matrix as the identity matrix.

During the learning process the convergence rate was rapid in the first few iteration. However later on the convergence became very slow though. We calculated the matrix square norm of the difference between the Φ matrix and the counterpart in the DBN representation. Using the above mentioned initial values, we got $1.2672e-4$ for the difference norm, which is a very low

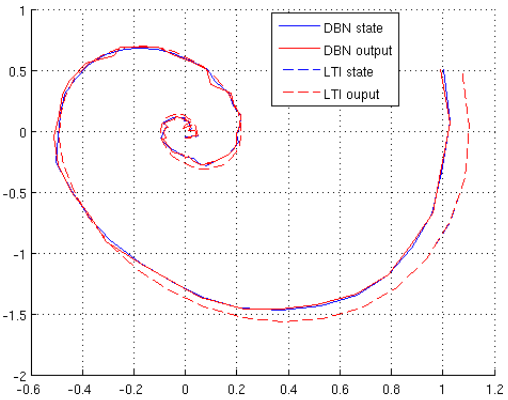


Fig. 2. Mass-spring-damper simulation with state space model and with DBN. The initial conditions and the applied disturbances were the same in both representation

diffence. We also tested the system with scalar valued random variables where we got $6.5097e-04$. The resulting covariances

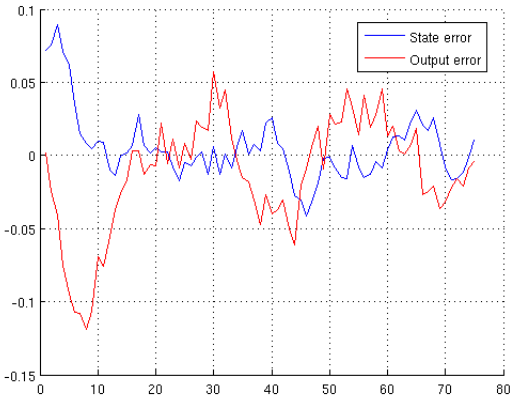


Fig. 3. The differences between the two representations under the same conditions

matrices we obtained had non-zero off-diagonal elements after convergence, despite that in the original system there was no off-diagonal elements. One explanation for this is that the learning examples did not provide enough information for the EM algorithm about the noise process.

D. Reaction Kinetic Network

Our second system was a simple reaction kinetic network with the following A matrix

$$A = \begin{bmatrix} -k_1 & k_2 & k_4 \\ k_1 & (-k_2 - k_3) & 0 \\ 0 & k_3 & -k_4 \end{bmatrix} \quad (14)$$

We simulated the model as a DBN similarly to the mass-spring-damper system. We used the following parameter setup: $k_1 = 2$, $k_2 = 0.5$, $k_3 = 0.9$ and $k_4 = 4.3$.

After convergence we calculated the matrix square norm of the difference between, which was $9.9895e-04$. Our goal was to exploit the fact that an LTI structure can be mapped onto a DBN. We used scalar valued nodes and with the connection structure we incorporated the zero elements in the DBN and run the EM algorithm. The resulting continuous model was the following:

$$A_{est} = \begin{bmatrix} -2.017 & 0.51 & 4.29 \\ 1.978 & 0.901 & -0.043 \\ -0.009 & -0.6282 & -4.314 \end{bmatrix} \quad (15)$$

The difference norm was $3.1176e-04$, which is in the same magnitude as in the vector-valued case. The increased number of nodes in the same layer did not cause significant increase in the run-time of the EM algorithm, however there were 3 times more nodes.

V. Conclusion and Future Work

The parameter estimation of LTI systems was evaluated in the DBN framework. As it can be seen from the experiments the advantage of DBN approach over traditional parameter estimation is the ability to incorporate *a priori* information to maintaine the associated computational cost with the structure of the graph. Two simple LTI systems were evaluated in the DBN framework. DBNs with vector and scalar valued node types were tested and it was shown that similar result in the two cases with no significant computational over head. This result is important because structural constraints can be exploited in the scalar case. Considering the results above several directions of future work emerged. One direction is to apply external input to enhance the result of parameter estimation. Another direction is associated with the concept of identifiably where on result of the parameter estimation can be guarantied on rigorous mathematical foundation.

References

[1] L. Ljung, *System identification: theory for the user*. Prentice Hall PTR, 1987.
[2] K. Murphy, "Dynamic bayesian networks: Representation, inference and learning," Ph.D. dissertation, UC Berkeley, Computer Science Division, 2002.
[3] S. Roweis and Z. Ghahramani, "A unifying review of linear gaussian models," *Neural Comput.*, vol. 11, pp. 305–345, February 1999.
[4] E. H. et al., *Linear systems control: deterministic and stochastic methods*. Springer, 2008.
[5] Z. Ghahramani and G. Hinton, "Parameter estimation for linear dynamical systems," University of Toronto, Tech. Rep., 1996.
[6] J. Pearl, *Probabilistic Reasoning in Intelligent Systems:Networks of Plausible Inference*. Morgan Kaufmann, 1988.
[7] M. I. Jordan, "Graphical models," *Statistical Science*, vol. 19, pp. 140–155, 2004.
[8] P. Antal, "Integrative analysis of data, literature, and expert knowledge," Ph.D. dissertation, K.U.Leuven ESAT, 2008.
[9] R. H. Shumway and D. S. Stoffer, "An approach to time series smoothing and forecasting using the em algorithm," *Journal of Time Series Analysis*, vol. 3, pp. 253–264, 1982.
[10] Y. Ho, "A Bayesian approach to problems in stochastic estimation and control," *IEEE Transactions on Automatic Control*, vol. 9, no. 4, pp. 333–339, Oct. 1964.
[11] Z. Chen, "Bayesian filtering: From kalman filters to particle filters, and beyond," *Statistics*, pp. 1–69, 2003.
[12] K. Murphy, "Fitting a constrained conditional linear gaussian distribution," MIT, Tech. Rep., 1998.
[13] ——, "The Bayes net toolbox for matlab," *Computing Science and Statistics*, vol. 33, 2001.
[14] H. Xiong and Y. Choe, "Structural Systems Identification of Genetic Regulatory Networks," *Bioinformatics*, vol. 24, no. 4, Jan. 2008.

Analysis of controlled dynamic systems using coloured Petri nets

János Rudan
(Supervisors: Dr. Katalin Hangos and Dr. Gábor Szederkényi)
rudan.janos@itk.ppke.hu

***Abstract*—In this paper a coloured Petri net representation of discrete event systems is used for dynamic reachability analysis in order to construct a supervisory controller for the system. The examined system is a pressurizer of a nuclear power plant, where the temperature dynamics and the effect of the heating controller is investigated.**

***Index Terms*—Petri net, reachability analysis, supervisory control**

I. Introduction

Considering the dynamic system as a discrete event system (DES) theories emerging from the computer science gives us the opportunity to discover important qualitative properties of the given system. DES is a well-known approach in theoretical computer science where the system is described as a state machine (i.e. with states and transitions) and current state of the considered system is obtained as a state of the automaton.

The main advantage of the DES-based description is the ability to extract functional and qualitative information from model and to get know the structure of the state space [1]. Based on these a special way of controller design can be completed often used for supervisory or hierarchical controller design [2].

The aim of this work is to use coloured Petri net (CPN) representation of DESs for dynamic reachability analysis, that enables to construct a supervisory controller for the system. The examined system is a pressurizer of a nuclear power plant, where the temperature dynamics and the effect of the heating controller is investigated.

II. Related work

At the beginning the results from graph theory based deterministic system analysis were applied to DES description. [3] uses a Boolean matrix equation model to represent the state changes occurring in the system on a deterministic way. This enables the examination of basic properties of a discrete system, for example, if it has any deadlocks or cyclic behavior.

Petri nets (PNs) are introduced by Carl Adam Petri in 1966 as a special types of DES, originally to describe chemical processes [4]. In his work PN are introduced as graphical tools for the description and analysis of concurrent processes which arise in systems with many components. The structure of Petri nets is a biparite graph, its nodes can only be places (denoting substances, states or conditions) and transitions (denoting reactions, state transformations) connected into a directed graph. Tokens are introduced to represent the occurrence of a logical value at a given place. Later these type of directed graphs were extended with extra parameters: adding time parameters to the transitions leads to timed PN, while extending the pool of available tokens to an unordered set leads us to coloured PN.

All types of Petri nets can be used effectively in several fields of research and applications, a comprehensive survey of the properties and usage of Petri nets is written by T. Murata [5] and K. Jensen [6].

Main areas of using Petri nets are the following: task scheduling and parallel processing [7], investigating liveness problem of event based systems [8], supervisory control [2], software analysis and validation and urban traffic control [9], [10]. Several theoretical and practical results are published according to the usage of Petri nets in hybrid systems, namely where the described system contains continuous-time and discrete components, too [11].

As it is shown in several papers, PN is a proper tool to examine a given system's deadlock-free operation (liveness), discover the places in concurrency or in conflict, analyse bottleneck nodes and resource allocation [12].

Completing reachability analysis of the system - determining the set of the reachable states - can be done with the help of PN [13], which capability has crucial importance in the presented work. In most cases, reachability analysis is done by search algorithms exploring the state space [14].

III. Coloured Petri nets for discrete event system description

In the presented work coloured Petri nets (CPNs) are used for describing a dynamic system. First the framework of the PN is described, after that the algorithms used for PN analysis - especially for reachability analysis - are introduced.

A. Coloured Petri nets

A coloured Petri net is biparite multigraph described by a 6-tuple $< P, T, C, I^-, I^+, M_0 >$ where P is a set of places, T is a set of transitions and S and T are disjoint. C is a color function defined from P into a finite and non-empty sets. I^- and I^+ are the backward and forward incidence functions defined on $(P \times T)$ that $I^-(p,t), I^+(p,t) \in [C(t) \rightarrow C(p)], \forall (p,t) \in (P \times T)$. M_0 is a marking function defined on P describing the initial marking such that $M_0(p) \in C(p), \forall p \in P$. Arcs representing I^- and I^+ are often called input arcs are output arcs.

The operation of a PN is as follows: firing a transition t in a marking M consumes $I^-(s,t)$ tokens from each of its input places s, and produces $I^+(s,t)$ tokens in each of its

output places s. A transition is enabled (it may fire) in M if there are enough tokens in its input places (noted by s) for the consumptions to be possible, i.e. iff $\forall s : M(s) > I^-(s,t)$. Note that transitions may fire in arbitrary order.

B. Reachability and the reachability analysis algorithms

Reachability is a fundamental property of a dynamic system. Considering a known system dynamics and given constraints, a system is reachable, if it is possible to reach the final state S_1 starting the system from a given S_0 initial state. This definition is similar to the concept of controllability in control theory.

Considering a PN the concept of reachability deals with the following problem: given a PN called N, an initial marking M_0 and a marking M_1, is it possible that $M_N(M_0) = M_1$. This means that starting the PN from the initial state, using the transition rules of a PN the M_1 final state can be reached.

Because of a standard coloured PN has a finite number of different states, the reachability set of a given initial state can be calculated using a search algorithm exploring all descendant states. The search algorithm could be either a depth-first or a breath-first one with a loop-rejecting capability. Starting the algorithm for each and every possible states the advance of the states in the state space can be reconstructed as a set of trees.

IV. Pressurizer tank - a case study

A simplified dynamic model of a pressurizer tank in the primary circuit of a nuclear power plant was used in this work. The model introduced below was used as a case study to show the capabilities of the coloured PN-based system analysis.

A. Dynamic model of the pressurizer tank

The simplified scheme of the primary circuit of a pressurised water nuclear power reactor includes the following parts. The reactor vessel contains the fuel- and control rods and it is filled with water which is circulated in the circuit by the coolant pump. The hot water goes through the steam generator while heating up the secondary circuit. The water in the primary circuit is kept on high pressure with the help of the pressurizer in order to increase the boiling temperature of the water (up to 200-220 bar and 350-380 °C). Evidently in this closed system the water temperature and the pressure is strongly connected to each other.

The detailed structure of the pressurizer is shown in Fig. 1. The pressurizer controls the primary circuit pressure with active heating, passive cooling and active emergency cooling according to the current pressure and the operation limits. The tank contains a constant volume of water. Due to the natural heat loss, the tank is cooling constantly, while extra cooling is provided by a valve which allows to let cold water into the tank (at the same time the same amount of hot water is released from it to keep the water level constant). The heating is provided by four electric heater divided into two groups.

The basic operation flow of the pressurizer is as follows. When the water temperature is low, the first heater group is switched on, while in case of an extra-low temperature an emergency actuation mechanism activates both group of heaters. If such a high temperature is reached that the natural heat loss is not enough to keep the temperature in the normal range,

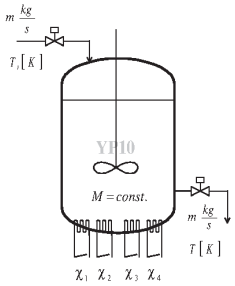


Fig. 1. Detailed structure of the pressurizer.

the active cooling is switched on which is considered as an emergency actuation.

Based on these the following discrete-time model of the pressurizer tank is proposed:

$$x(k+1) \quad = \quad x(k) + \kappa_h * \lambda_h - \lambda_l + \delta \qquad (1)$$

where x denotes the temperature in a given timestep, κ_h is a boolean value representing the state of the active heating, λ_h and λ_l are proportional tags weighting the heating and heat loss effect, respectively. The variable δ represents the disturbance in the model.

B. Coloured Petri net model of the pressurizer

The coloured Petri net (CPN) model of the pressurizer tank can be seen in Fig. 2. Places are represented with yellow circles (named p_i) while the blue rectangles represent the transitions (named t_i). Tokens are represented with coloured dots in the places.

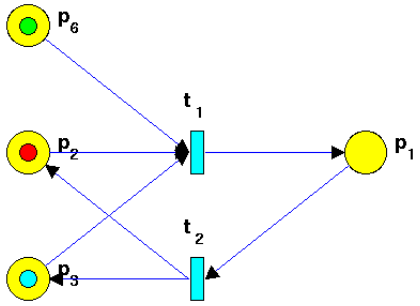


Fig. 2. CPN model of the pressurizer.

The model is representing the pressurizer grabbed out from the primary circuit. The setup focuses on the temperature control of the tank using the active heating and considers the temperature of the primary circuit as a disturbance.

The main idea behind the model setup was the following. Discretizing the value range of the dynamic system (i.e. the temperature) and using a PN-based system description we obtain a model of the dynamic system discretized both in time- and value-range. Step size of discretization has direct effect on the size of the state space so it has to be selected carefully.

In this model some token types represent different temperature intervals and other token types represent the operational state of the active heating. We defined an arithmetic over the set of tokens representing temperature intervals. In order to keep

the model simple the token set was mapped to the set of positive integers and the normal integer arithmetic was used.

The CPN has the following structure: p_2 and p_1 are the places representing the temperature of the pressurizer in the current and in the next time step, respectively. Tokens at place p_6 symbolize the primary circuit temperature which is handled as the disturbance. p_3 stands for the state of the switch of the active heating.

Transition t_1 realizes the dynamics of the pressurizer and computes the solution of the model equation Eq. 1. and the corresponding token type. The generated token appears at p_1. Transition t_2 is representing the heating controller algorithm. It is a simple threshold-based logic: if the temperature (token at p_1) is over a given threshold, the heating is switched on, otherwise it is off. If the temperature is extremely low or high, tokens representing emergency cooling or heating is used. According to this, temperature control is represented by four type of tokens symbolizing emergency cooling, heating off, heating on, emergency heating operations. The output of the transition is the new tank temperature (transferring the token from p_1 to p_2) and the state of the active heating represented by the token at p_3.

C. Simulation results

In order to validate the behaviour of the proposed model, we completed several simulations. The temperature range was divided into five intervals as follows: E^-, L, N, H, E^+ representing *emergency low*, *low*, *normal*, *high* and *emergency high* temperature intervals. The value of the disturbance was also represented by a set containing 5 different types of tokens. The weight of heating was set to have larger effect on the system as the natural heat loss, namely $\lambda_h = 2, \lambda_l = 1$.

For the simulation we used the PetriSimM MATLAB toolbox [15]. The toolbox has been largely improved and extended to fit our purposes. Alltogether 9 colour was in the CPN model: 5 colours for temperature interval description and 4 for temperature control actuation. In order to test the correct operation of the heating control, we used random disturbance values in each step.

As it can be seen in Fig. 3., the controller can react and actuate correctly according to the given state of the system. The simulation was run through 50 steps. On the first subplot the pressurizer temperature, on the second the disturbance values while on the third subplot the heating control values can be seen.

D. Reachability analysis of the pressurizer model

Using the proposed CPN model, a reachability analysis was completed in order to determine the set of states which can be reached by the system. The algorithm of the analysis was the following: each possible state was considered as an initial state. A simulation was started from all initial states using all possible disturbance values, completing a depth-first exploration of the successive states. If a state is visited by the algorithm, it was marked and never revisited in order to avoid infinite loops. The resulting set of trees was merged into one graph.

In Fig. 4. the reachability graph of the system can be seen. The states are represented by circles and directed arcs represents

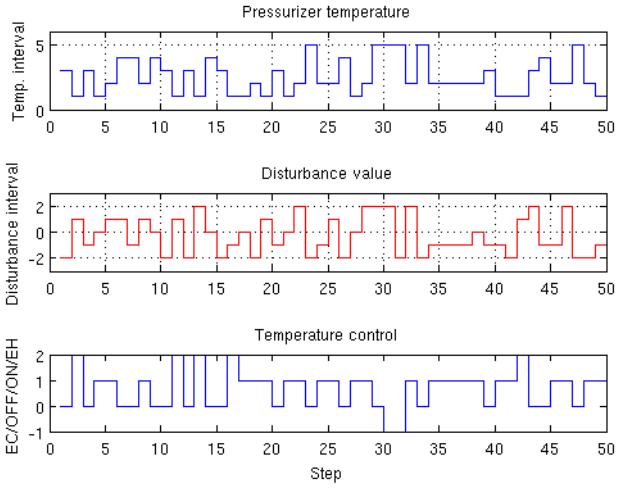


Fig. 3. Simulation results of the CPN model of the pressurizer under random disturbance.

that the system can move from one state to an other. The arc colors symbolize the disturbance value which is needed to move between two states. Colour coding at the nodes is designated to show arcs have the same starting and ending node. Every disturbance value has an own color stated by the colorbar.

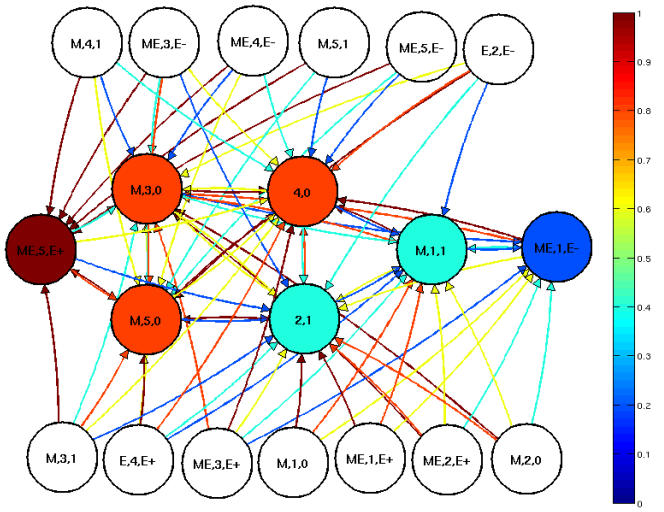


Fig. 4. Reachability graph of the CPN model of the pressurizer tank.

E. Partitioning the state space based on reachability analysis

Investigating the graph generated as a result of the reachability analysis the following was concluded: the state space can be divided into three parts. This partitioning is represented in Fig. 4, where node labels depict the class of a node.

The first set of nodes belong to emergency states. This class marked with an extra *E* character in the node label. If the system reach one of these states, an emergency actuation is initialized to force the system back into a nominal temperature range.

The second set called *margin* contains nodes from which the states in the emergency set are reachable in one step. Nodes belong to this class marked with *M* in the node label. This set plays a crucial role from the control point of view: if the

system steps into one of the states in the margin set, we could be able to determine the proper control input to avoid to reach an emergency state.

The third set is the so called *core* part containing states that are not in neighbourhood with states from the emergency set. While the system is moving between states in this set the normal operation of the system elements can be performed.

F. Designing a supervisory controller

As it is shown previously, the state space of the given dynamic model can be partitioned into three parts: the set of emergency states, the margin and the core. The information obtained from the discovering of the structure of the state space can be used to design a supervisory controller. This controller is created to drive the original controller, and by using the predictions based on the reachability graph avoid the emergency states resulting a better system performance.

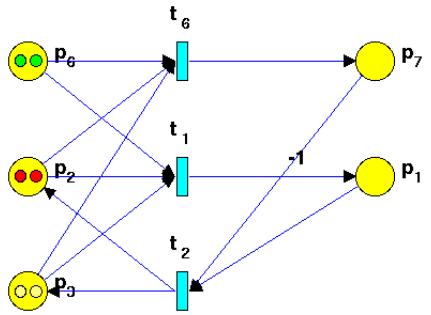


Fig. 5. CPN model of the pressurizer with the supervisory controller.

Technically the supervisory controller designed as follows. Based on the reachability analysis a lookup table is created containing the states, and the disturbance values which takes the system into the *margin* set or into an emergency state. An extra transition is introduced into the CPN model depicting the supervisory controller. The transition gets the state of the pressurizer, the disturbance and the heating control values, and produces a token describing the recommended heating control value. This token is processed by an extended version of the original heating controller. The resulting CPN model can be seen in Fig. 5 containing the supervisory controller at transition t_6.

The supervisory controller recommends heating control inputs in order to avoid that the system gets into the *margin* set. This control strategy gives more chance to avoid the emergency states because it has capability to predict the future states using the information extracted from the reachability analysis.

V. Conclusion and Future Work

In this paper a coloured Petri net based model of a dynamic system is presented. After introducing the Petri net as a special type of discrete event systems, it is shown that a simple dynamic model of a pressurizer can be represented with this tool.

The formulated CPN model is used to map the structure of the state space completing a reachability analysis of the system. The obtained reachability graph is partitioned into several sets to separate the states into functionally meaningful groups.

The created sets and the corresponding data were utilized as a knowledge base of a supervisory controller, which tries to help to avoid the emergency states using the prediction capability gained from the reachability analysis.

The future work can have three main directions. First, the model of the pressurizer can be developed using more sophisticated dynamic model and considering more effect appearing in the real system. As second direction, reducing the size of the temperature intervals by increasing the number of tokens representing them leads to more realistic simulation results. The expanded token set is also necessary at the examination the proper operation of the supervisory controller. Finally the model have to be tested using real data sets in order to prove that an increased performance can be achieved with the help of the proposed architecture.

References

[1] A. Fanni and A. Giua, "Discrete event representation of qualitative models using Petri nets," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 28, pp. 770–780, 1998.

[2] A. Giua, "Petri nets as discrete event models for supervisory control," Ph.D. dissertation, Rensselaer Polytechnic Institute, Troy, New York, Jul. 1992.

[3] R. L. Aveyard, "A boolean model for a class of discrete event systems," *IEEE Trans. on Systems, Man, and Cybernetics*, vol. SMC-4, no. 3, pp. 249–258, May 1974.

[4] C. A. Petri, "Kommunikation mit automaten." *New York: Griffiss Air Force Base, Technical Report*, vol. 1, pp. 1–Suppl. 1, 1966, english translation.

[5] T. Murata, "Petri nets: Properties, analysis and applications." *Proceedings of the IEEE*, vol. 77, no. 4, pp. 541–580, April 1989.

[6] K. Jensen, *Coloured Petri nets*. Springer Berlin / Heidelberg, 1997, vol. 1.

[7] Z. Hanzalek, "Parallel algorithms for distributed control a Petri net based approach," Ph.D. dissertation, Czech Technical University, Prague, Jan. 2003.

[8] H.-Y. Su, W.-M. Wu, and J. Chu, "Liveness problem of Petri nets supervisory control theory for discrete event systems," *Acta Automatica Sinica*, vol. 31, no. 1, Jan. 2005.

[9] G. F. List, "Modeling traffic signal control using Petri nets," *IEEE Transactions on Intelligent Transportation Systems*, vol. 5, no. 3, Sep. 2004.

[10] M. Dotoli and M. P. Fanti, "An urban traffic network model via coloured timed Petri nets," in *Discrete event systems*, 2004.

[11] C. Vázquez, H. Sutarto, R. Boel, and M. Silva, "Hybrid Petri net model of a traffic intersection in an urban network," in *2010 IEEE Multiconference on Systems and Control*, Yokohama, Japan, 09/2010 2010.

[12] M. V. Iordache and P. J. Antsaklis, *Supervisory Control of Concurrent Systems, a Petri net Structural Approach*. Birkhauser, Apr. 2005.

[13] J. Wang, Y. Deng, and G. Xu, "Reachability analysis of real-time systems using time Petri nets," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 30, pp. 725–736, 2000.

[14] K. Hangos, R. Lakner, and M. Gerzson, *Intelligent Control Systems: An Introduction with Examples*. Kluwer Academic Publisher, 2001.

[15] *PetriSimM toolbox*, http://seth.asc.tuwien.ac.at/petrisimm/. [Online]. Available: http://seth.asc.tuwien.ac.at/petrisimm/