INFORMATION SCIENCE & COMPUTING

International Book Series Number 5

Intelligent Technologies and Applications

Supplement to
International Journal "Information Technologies and Knowledge" Volume 2 / 2008

I T H E A
SOFIA, 2008

Krassimir Markov, Krassimira Ivanova, Ilia Mitov (ed.)

Intelligent Technologies and Applications

International Book Series "INFORMATION SCIENCE & COMPUTING", Number 5

Supplement to the International Journal "INFORMATION TECHNOLOGIES & KNOWLEDGE" Volume 2 / 2008

Institute of Information Theories and Applications FOI ITHEA

Sofia, Bulgaria, 2008

This issue contains collection of papers in which are presented applications of Artificial Intelligence in different areas of our life. Papers are selected from the International Conferences of the Joint International Events of Informatics "ITA 2008", Varna, Bulgaria.

International Book Series "INFORMATION SCIENCE & COMPUTING", Number 5

Supplement to the International Journal "INFORMATION TECHNOLOGIES & KNOWLEDGE" Volume 2, 2008

Edited by **Institute of Information Theories and Applications FOI ITHEA**, Bulgaria,
in collaboration with
- **V.M.Glushkov Institute of Cybernetics of NAS**, Ukraine,
- **Institute of Mathematics and Informatics, BAS**, Bulgaria,
- **Institute of Information Technologies, BAS**, Bulgaria.

Publisher: Institute of Information Theories and Applications FOI ITHEA, Sofia, 1000, P.O.B. 775, Bulgaria.
Издател: Институт по информационни теории и приложения ФОИ ИТЕА, София, 1000, п.к. 775, България
www.ithea.org, www.foibg.com, e-mail: info@foibg.com

General Sponsor: **Consortium FOI Bulgaria** (www.foibg.com).

Printed in Bulgaria

ISSN 1313-0455 (printed) **ISSN 1313-048X (online)** **ISSN 1313-0501 (CD/DVD)**

PREFACE

The scope of the International Book Series “Information Science and Computing” (**IBS ISC**) covers the area of Informatics and Computer Science. It is aimed to support growing collaboration between scientists from all over the world. IBS ISC is official publisher of the works of the members of the ITHEA International Scientific Society.

The official languages of the IBS ISC are English and Russian.

IBS ISC welcomes scientific papers and books connected with any information theory or its application.

IBS ISC rules for preparing the manuscripts are compulsory. The rules for the papers and books for IBS ISC are given on www.foibg.com/ibsisc. The camera-ready copy of the papers and books should be received by e-mail: info@foibg.com.

Responsibility for papers and books published in IBS ISC belongs to authors.

The Number 5 of the IBS ISC contains collection of papers in which are presented applications of Artificial Intelligence in different areas of our life. Papers are peer reviewed and are selected from the several International Conferences, which were part of the Joint International Events of Informatics “ITA 2008”, Varna, Bulgaria.

ITA 2008 has been organized by

Institute of Information Theories and Applications FOI ITHEA

in collaboration with:

- ITHEA International Scientific Society
- International Journal “Information Theories and Applications”
- International Journal "Information Technologies and Knowledge"
- Association of Developers and Users of Intelligent Systems (Ukraine)
- Association for Development of the Information Society (Bulgaria)
- V.M.Glushkov Institute of Cybernetics of National Academy of Sciences of Ukraine
- Institute of Mathematics and Informatics, BAS (Bulgaria)
- Institute of Information Technologies, BAS (Bulgaria)
- Institute of Mathematics of SD RAN (Russia)
- Taras Shevchenko National University of Kiev (Ukraine)
- Universidad Politecnica de Madrid (Spain)
- BenGurion University (Israel)
- Rzeszow University of Technology (Poland)
- University of Calgary (Canada)
- University of Hasselt (Belgium)
- Kharkiv National University of Radio Electronics (Ukraine)
- Astrakhan State Technical University (Russia)
- Varna Free University “Chernorizets Hrabar” (Bulgaria)
- National Laboratory of Computer Virology, BAS (Bulgaria)
- Uzhgorod National University (Ukraine)
- Sofia University “Saint Kliment Ohridski” (Bulgaria)
- Technical University – Sofia (Bulgaria)
- New Bulgarian University (Bulgaria)

The main ITA 2008 events were:

KDS	XIVth International Conference "Knowledge - Dialogue – Solution"
i.Tech	Sixth International Conference "Information Research and Applications"
MeL	Third International Conference "Modern (e-) Learning"
ISK	Second International Scientific Conference "Informatics in the Scientific Knowledge"
INFOS	International Conference "Intelligent Information and Engineering Systems"
GIT	Sixth International Workshop on General Information Theory
CS	Third International Workshop "Cyber Security"
eM&BI	Second International Workshop "e-Management & Business Intelligence"
IMU ICT	International Seminar "Information Models' Utility in Information and Communication Technologies"
ISSI	Second International Summer School on Informatics

More information about ITA 2008 International Conferences is given at the www.foibg.com.

The great success of ITHEA International Journals, International Book Series and International Conferences belongs to the whole of the ITHEA International Scientific Society.

We express our thanks to all authors, editors and collaborators who had developed and supported the International Book Series "Information Science and Computing".

General Sponsor of IBS ISC is the **Consortium FOI Bulgaria** (www.foibg.com).

Sofia, June 2008 *Kr. Markov, Kr. Ivanova, I. Mitov*

TABLE OF CONTENTS

Preface *3*
Table of Contents *5*
Index of Authors *7*

Papers in English

Hardware Implementations of Video Watermarking
Xin Li, Yonatan Shoshan, Alexander Fish, Graham Jullien, Orly Yadid-Pecht *9*

On the Feasibility of Steering Swallowable Microsystem Capsules Using Computer-Aided Magnetic Levitation
Billy Wu, Martin Mintchev *17*

Application of Fuzzy Logic on Image Edge Detection
Shashank Mathur, Anil Ahlawat *24*

Distance Matrix Approach to Content Image Retrieval
Dmitry Kinoshenko, Vladimir Mashtalir, Elena Yegorova *29*

Projective Methods of Image Recognition
Yevgeniy Putyatin, Vladimir Gorohovatsky, Alexey Gorohovatsky, Elena Peredriy *37*

On a Direct Approach to the Solution of Inverse Optical Problems
Peter Sharlandjiev, Georgi Stoilov *43*

Normal ECG Recognition for Express-diagnostics Based on Scale-space Representation and Dynamic Matching
Nataliya Bilous, Michael Bondarenko, Gleb Kobzar, Alexey Krasov, Artyom Rogozyanov *47*

Application of the Artificial Intelligence Algorithms with a Cognitive Graphic as a Tool for a System Analysis of Electrophysiological Processes
Nina Dmitrieva *53*

Minimum Mean Square Error Approach in Images Processing for Full-field Displacements and Deformation Measurements
Georgi Stoilov, Vasil Kavardzhikov, Dessislava Pashkouleva *57*

WLAN / WMAN Integration Architecture and Traffic Control for Voice Transmission
Boris Tsankov, Kiril Kassev, Radostin Pachamanov *62*

Improved Hybrid Model of HMM/GMM for Speech Recognition
Poonam Bansal, Anuj Kant, Sumit Kumar, Akash Sharda, Shitij Gupta *69*

Research of Influence of Medication Preparations on the Process of Renewal of the Broken Equilibrium of Man Organism on the Doctor of Phytotherapy Workstation
Anatoliy Bykh, Elena Visotska, Boris Samura, Olga Kozina, Andrei Porvan, Elena Chernikh, Sergey Parhomenko *75*

Smart Portable Tester for Bird Flue Express-Diagnostics: Principles of Design
Oleksandr Palagin, Volodymyr Romanov, Mykola Starodub, Igor Galelyuka, Oleksandra Skrypnyk, Kyryl Skyba *80*

KFTKGA: Knowledge Flow Tracer and Growth Analyzer for Community Learning
S.M.F.D. Syed Mustapha, Charles A.H. Chong *85*

Easy Creation of Semantics-enhanced Digital Artwork Collections
Darina Dicheva, Christo Dichev, Cristan Brown, Dhanya Veetil, William Thomas *92*

One Approach to Knowledge Mapping for International Student Portal
Tatiana Gavrilova, HaiNan Jin *99*

Emotions and a Prior Knowledge Representation in Artificial General Intelligence
Andrey Gavrilov *106*

Ontology-based Model of Representation of Knowledge about Language Mappings
Margarita Knyazeva, Vadim Timchenko *111*

Upgradable Tree Levels Editor of Metaontologies, Ontologies and Knowledge for a Chemistry
Irene Artemieva, Natalia Reshtanenko, Vadim Tsvetnikov *119*

Multiple-valued Structures of Intellectual Systems
Mikhail Bondarenko, Irina Vechirskaya, Grigoriy Chetverikov *125*

Development of Generalized Nature Environment Model for Emergency Monitoring
Oleksandr Kuzomin, Illya Klymov, Asanbek Toroev *131*

Leveraging Existing Plasma Simulation Codes
Anna Malinova, Vasil Yordanov, Jan van Dijk *136*

Cognitive Approach in Castings' Quality Control
Jürgen Bast, Valeriy Kamaev, Irina Polyakova, Andrey Tikhonin, Maxim Zabolotsky, Natalia Kudashov *143*

Papers in Russian

Распознавание сложных стерео и мульти–изображений в реальном времени
Адиль Тимофеев, Олег Дерин *149*

Структурная модель полутонового изображения и ее использование в задаче сегментации изображений
Владимир Калмыков, Виталий Вишневский, Татьяна Власова *153*

INDEX OF AUTHORS

Anil	Ahlawat	24
Irene	Artemieva	119
Poonam	Bansal	69
Jürgen	Bast	143
Nataliya	Bilous	47
Mikhail	Bondarenko	47, 125
Cristan	Brown	92
Anatoliy	Bykh	75
Elena	Chernikh	75
Grigoriy	Chetverikov	125
Charles	Chong	85
Christo	Dichev	92
Darina	Dicheva	92
Nina	Dmitrieva	53
Alexander	Fish	9
Igor	Galelyuka	80
Andrey	Gavrilov	106
Tatiana	Gavrilova	99
Alexey	Gorohovatsky	37
Vladimir	Gorohovatsky	37
Shitij	Gupta	69
HaiNan	Jin	99
Graham	Jullien	9
Valeriy	Kamaev	143
Anuj	Kant	69
Kiril	Kassev	62
Vasil	Kavardzhikov	57
Dmitry	Kinoshenko	29
Illya	Klymov	131
Margarita	Knyazeva	111
Gleb	Kobzar	47
Olga	Kozina	75
Alexey	Krasov	47
Natalia	Kudashov	143
Sumit	Kumar	69
Oleksandr	Kuzomin	131
Xin	Li	9
Anna	Malinova	136
Vladimir	Mashtalir	29
Shashank	Mathur	24
Martin	Mintchev	17
Radostin	Pachamanov	62
Oleksandr	Palagin	80
Sergey	Parhomenko	75
Dessislava	Pashkouleva	57
Elena	Peredriy	37
Irina	Polyakova	143
Andrei	Porvan	75
Yevgeniy	Putyatin	37
Natalia	Reshtanenko	119
Artyom	Rogozyanov	47
Volodymyr	Romanov	80
Boris	Samura	75
Akash	Sharda	69
Peter	Sharlandjiev	43
Yonatan	Shoshan	9
Oleksandra	Skrypnyk	80
Kyryl	Skyba	80
Mykola	Starodub	80
Georgi	Stoilov	43, 57
S.M.F.D.	Syed Mustapha	85
William	Thomas	92
Andrey	Tikhonin	143
Vadim	Timchenko	111
Asanbek	Toroev	131
Boris	Tsankov	62
Vadim	Tsvetnikov	119
Jan	van Dijk	136
Irina	Vechirskaya	125
Dhanya	Veetil	92
Elena	Visotska	75
Billy	Wu	17
Orly	Yadid-Pecht	9
Elena	Yegorova	29
Vasil	Yordanov	136
Maxim	Zabolotsky	143
Виталий	Вишневский	153
Татьяна	Власова	153
Олег	Дерин	149
Владимир	Калмыков	153
Адиль	Тимофеев	149

HARDWARE IMPLEMENTATIONS OF VIDEO WATERMARKING

Xin Li, Yonatan Shoshan, Alexander Fish, Graham Jullien, Orly Yadid-Pecht

***Abstract**: Various digital watermarking (WM) techniques for still imaging have been studied in the last several years. Recently, many new WM schemes have been proposed for other types of digital multimedia data, such as text, audio and video. This paper presents a brief overview of existing digital video WM. We classify WM techniques and discuss the properties of video WM. Since each WM application has its own specific requirements, WM design must take the intended application into consideration. Video WM applications are also discussed in the paper. The features of video WM implementations in software and hardware and their differences are presented through the description of four examples of existing work.*

***Keywords**: Digital video, watermarking, WM, hardware implementation, security.*

***ACM Classification Keywords**: B.0 Hardware - Conference proceedings*

***Conference**: The paper is selected from Sixth International Conference on Information Research and Applications – i.Tech 2008, Varna, Bulgaria, June-July 2008*

1. Introduction

Storing and transmitting digital multimedia data has become incredibly available throughout the world, especially with the advent of digital times. This has been a catalyst for the rapid growth of digital video technologies and applications [1]. Nowadays, the expansion of high speed digital computer networks all over the world and the advance of compression technologies have made the distribution of video data and applications much easier and faster. The amount of high quality digital video data is ready available on the internet so that users can conveniently be able to enjoy watching on-line video, transmit and exchange video files. Digital video is also useful in many other applications: surveillance video systems and broadcasting are good examples. However, at the same time a number of security problems have been introduced, since digital video sequences are very susceptible to manipulations and alterations using widely available editing software. This way video content is not reliable anymore. For example, a video shot from a surveillance camera cannot be used as a piece of evidence in a courtroom because it is not considered trustworthy enough. Therefore, authentication techniques are consequently needed in order to ensure the authenticity and integrity of video content. Till date, there have been various such techniques [2], of which digital watermarking (WM) is one of the most popular. Digital WM is a technique that embeds a secret, unnoticeable signal (called watermark) into the original multimedia object, like audio, image and video. The watermark can be detected or extracted later to claim the authenticity of the media content.

Several researchers have investigated digital WM with different contributions, implemented both on software and hardware platforms [3]-[11]. In 1990, the modern study of steganography and digital WM was started by Tanaka et al. [3]. They suggested hiding information in multi-level dithered images as a form of secured military communications. Following that work, digital image WM arose, and recently the development of video WM algorithms became a growing field of research. A relatively simple WM algorithm, working on raw video data, was presented in [4]. In [5], Wu proposed a method that adds a discrete cosine transform (DCT) transformed pseudo-random sequence (used as watermark) directly to the DC-DCT coefficients of the video frame to achieve better robustness against MPEG lossy compression. A spread spectrum method, described by Shan [6], was applied to watermark color video frames. According to this method, the mid-frequency DCT coefficients of a green component of the color frames were selected to embed the watermark because it was found to be the most robust after compression.

Although it might be easier to implement a WM algorithm on a software platform, there is a strong motivation for a move toward a hardware implementation [8]. The hardware implementation offers several distinct advantages over the software implementation in terms of low power consumption, less area usage and reliability. It features real time capabilities and compact implementations. In consumer electronic devices, a hardware WM solution is often more economical because adding the WM component only takes up a small dedicated area of silicon. Recently, a few hardware specific algorithms have been presented in the literature [7]-[11].

The objective of this paper is to provide an overall review of existing digital video WM solutions and applications. Background on video WM, such as different watermark classifications, applications and specifications are introduced. Following that, existing WM software and hardware implementations are also described.

The rest of the paper is organized as follows. Section II reviews the background on video WM. The WM software and hardware implementations are presented in section III. Conclusions are summarized in section IV.

2. Background on Video WM

2.1 Watermarks Classification

WM techniques can be divided into different categories according to various criterions [12]. The general classification of the currently available watermarks is shown in Figure 1. In [13] we have presented a decomposition of the variety of existing watermarks for still images and showed their features and possible applications, benefits and drawbacks. Since a video stream is regarded as a three-dimensional signal with two dimensions in space (called m x n frame) and one dimension in time, we can consider a video stream as a succession of still images. Therefore, most image WM techniques are equally applicable to video if the individual frames are treated as images [14]. However, contradictory to still image WM techniques, the video WM methods usually require that the WM encoding and decoding are processed in real time.

According to the domain in which video WM is performed, WM processing methods can be classified into two categories: spatial domain and frequency domain. In the spatial domain, directly applying minor changes to the values of the pixels in a minor way is mainly used. This technique makes the embedded information hardly noticeable to the human eye. For example, pseudo-random WM works by a simple addition of a small amplitude pseudo-noise signal to the original media data. In the frequency domain, the object first goes through a certain transformation, DCT or discrete wavelet transforms (DWT), the WM is embedded in the transform coefficients and then it is inversely transformed to receive the watermarked data. The frequency domain methods are more robust than the spatial domain techniques [15].

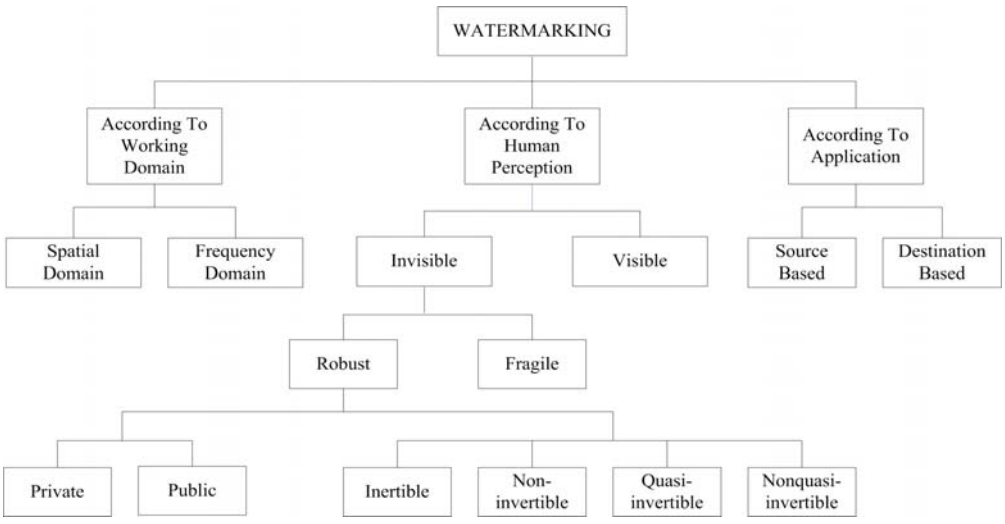


Figure 1 General classification of existing watermarking.

WM techniques can also be divided into three different types: visible, invisible robust and invisible fragile, according to human perception. Different applications have different requirements. Sometimes a certain application requires a WM to be visible, so that the embedded watermark appears visible to a casual viewer. Invisible robust WMs are primarily used in applications such as copyright protection, which require the algorithm to be as robust as possible so that severe modifications and degradations cannot remove the watermark. Conversely, invisible fragile is designed to reflect even slightest manipulation or modification of the media data, since the embedded watermark can easily become altered or destroyed after common attacks, such as lossy compression, cropping and spatial filtering. This WM method is a practical technique for content authentication.

From the application point of view, digital WM could be source based or destination based [12]. Source based WM can be used to authenticate whether a received media data has been manipulated and the destination based WM can trace the source of illegal copies.

2.2 Applications of Video WM

This section is consequently dedicated to the presentation of various applications in which digital WM can bring a valuable support in the context of video data. The following main watermarking applications are considered in the open literature and as commercial applications [16]. The reader is referred to [16]-[18] for a more thorough investigation. The applications presented have been gathered in table 1.

Table 1 Video WM: Applications and Purposes

Applications	Purpose
Copyright protection	Proof of ownership
Video authentication	Insure that the original content has not been altered
Fingerprinting	Trace back a malicious user
Copy control	Prevent unauthorized copying
Broadcast monitoring	Identify the video item being broadcasted

Copyright protection: For the protection of intellectual property, the video data owner can embed a watermark representing copyright information in his data. This watermark can prove his ownership in court when someone has infringed on his copyrights. For instance, embedding the original video clip by noninvertible WM algorithms during the verification procedure happens to prevent the multiple ownership problems in some cases.

Video authentication: Popular video editing software permit today to easily tamper with video content and therefore it is not reliable anymore. Authentication techniques are consequently needed in order to ensure the authenticity of the content. One solution is the use of digital WM.

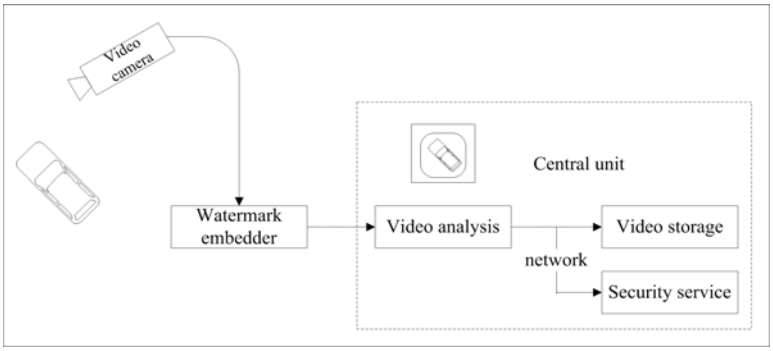


Figure 2 WM-based authentication for automatic VS.

In Figure 2, a sketch of a simple video surveillance (VS) system, in which WM is used to authenticate VS data, is given [17], [18]. Timestamp, camera ID and frame serial number are used as a watermark, embedded into every

single frame of the video stream. The central unit is in charge of analyzing the watermarked sequences and generating an alarm whenever a suspicious situation is detected, and then may either be sent to the security service or compressed for storage. When needed, the stored video sequence can be used as a proof in front of a court of law. It is possible to reflect any manipulation by detecting the watermarks.

Video fingerprinting: To trace the source of illegal copies, a fingerprinting technique can be used. In this application, the video data owner can embed different watermarks in the copies of the data that are supplied to different customers. Fingerprinting can be compared to embedding a serial number that is related to the customer's identity in the data. It enables the intellectual property owner to identify customers who have broken their license agreement by supplying the data to third parties.

A consumer can receive digital services, like pay TV, by cable using a set-top box and a smart card, which he has to buy and can therefore be related to his identity. To prevent other non-paying consumers from making use of the same service, the provider encrypts the video data and this protects the service during transmission. The set-top box of the consumer, who paid for the service, decrypts the data only if a valid smart card is used. Then, a watermark, representing the identity of the user, is added to the compressed video. The watermarked (fingerprinted) data can now be fed to the internal video decoder to view the video. A set-top box with WM capabilities is depicted in Figure 3.

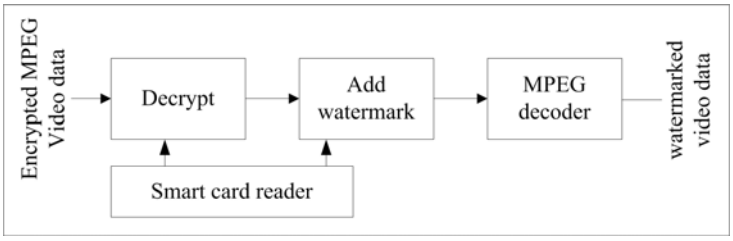


Figure 3 Set-top box with WM capabilities.

Copy control: The information stored in a watermark can directly control digital recording devices for copy protection purposes. In this case, the watermark represents a copy-prohibit bit and watermark detectors in the recorder determine whether the data offered to the recorder may be stored or not. For example, in the copy protection scheme using WM techniques shown in Figure 4, consumers can make copies of any original source, but they cannot make copies of copies.

This copy protection system checks all incoming video streams for a predefined copy-prohibit watermark. If such a watermark is found, the incoming video has already been copied before and is therefore refused by the recorder. If the copy-prohibit watermark is not found, the watermark is embedded and the watermarked video is stored. This means that video data stored on this recorder always contains a watermark and cannot be duplicated if the recorder is equipped with such a copy protection system.

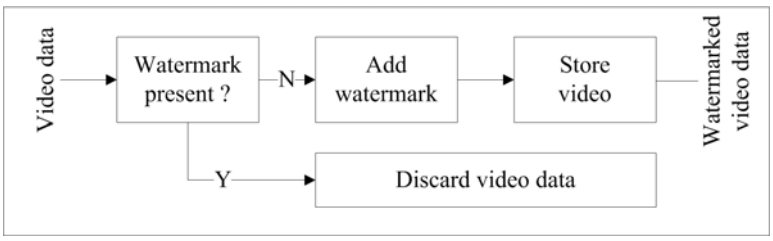


Figure 4 Video recorder with copy protection.

Broadcast monitoring: By embedding watermarks in commercial advertisements an automated monitoring system can verify whether advertisements are broadcasted as contracted. Not only commercials but also valuable TV products can be protected by broadcast monitoring. News items can have a value of over 100.000

USD per hour, which makes them very vulnerable to intellectual property rights violation. A broadcast surveillance system can check all broadcast channels and charge the TV stations according to their findings.

2.3 Requirements for Video WM

Different WM applications have specific requirements. Therefore, there is no universal requirement to be satisfied by all WM techniques. Nevertheless, some general directions can be given for most of the applications:

- Invisibility: WM should be imperceptible and invisible to a human observer.
- Transparency: WM embedding does not affect the quality of the underlying host data.
- Robustness: It should be impossible to manipulate the watermark by processing techniques or intentional operations such as filtering, addition of noises and cropping.
- Security: A WM technique is truly secure if knowing the exact algorithms for embedding and extracting the watermark does not help an unauthorized party to detect the presence of the watermark. It is very important, especially in authentication applications, that the watermark cannot be added or removed by an unauthorized user.
- Oblivious: It should be possible to extract watermark information without using the original multimedia data, since most receivers do not have the original data at their disposals.

Even though the requirements for the image and video WMs are very similar, they are not identical. New problems and new challenges have emerged in video WM applications. Apart from the basic requirements mentioned above, a WM technique should meet the following extra specific requirements to qualify as a real time technique for compressed video data:

- Low complexity: WM embedding and extracting should have low complexity, because they are to be processed in real time and if used in consumer products, they should also be inexpensive.
- Compressed domain processing: It should be possible to incorporate the watermark into compressed video (bit-stream).
- Constant bit-rate: WM should not increase the size of the compressed host video data and the bit-rate, at least for constant bit-rate applications where the transmission channel bandwidth has to be obeyed.

3. Video WM Implementations

Similar to image WM implementations, the video WM system can be implemented in either software or hardware, each having advantages and drawbacks. In software, the WM scheme can simply be implemented in a PC environment. The WM algorithm's operations can be performed as scripts written for a symbolic interpreter running on a workstation or machine code software running on an embedded processor. By programming the code and making use of available software tools, it can be easy for the designer to implement any WM algorithm at any level of complexity. However, such an implementation is relatively slow and therefore not suitable for real time applications.

In practical video storage and distribution systems, video sequences are stored and transmitted in a compressed format. Thus, a watermark that is embedded and detected directly in the compressed video stream can minimize computational demanding operations. Furthermore, frequency domain WM methods are more robust than the spatial domain techniques [15]. Therefore, working on compressed rather than uncompressed video is important for practical WM applications.

Before we describe the video WM techniques, we first briefly review the standards for video compression. All current popular standards for video compression, namely MPEG-x (ISO standard) and H.26x formats (ITU-T standard), are hybrid coding schemes and are DCT based compression methods [19]. Such schemes are based on the principles of motion compensated prediction and block-based transform coding. Table 2 resumes the features of commonly used video compression standards.

Table 2 Popular Video Compression Standards

Compression standards	Features
H.261	• Aimed at bit rates from 40 kbps to 2 Mbps. • Typically used in ISDN video conferencing.
MPEG-1	• Aimed for 1.5 Mbps data-rates and 352 x 240 resolutions. • Typically used for VCDs.
MPEG-2	• Outperforms MPEG1 at 3 Mbps • Below 1 Mbps, MPEG2 is similar to MPEG1. • Typically used for DVDs.
H.263	• Aimed at video coding for low bit rates (20 to 30 kbps). • Typically used for web video conferencing.
MPEG-4(H.264)	• 33% improvement over MPEG2. • 4 times frame size of MPEG4 part 2 at a given data rate. • Targeted for all media applications: mobile, internet, standard video, high definition, and full high definition.

In [4], Hartung presents a good example of software MPEG compressed video WM solution. The spread spectrum concept of communications is employed to watermark a compressed video stream, where the basic idea is embedding the watermark in the transform domain as represented in the entropy coded DCT coefficients. This is done in an MPEG-2 video signal, which currently is a mature and widely used video compression standard. Although an existing MPEG-2 bit-stream is partly modified, the scheme avoids visible artifacts by adding a drift compensation signal. This signal is needed because the P and B frames on the MPEG-2 compression format rely on information found on the I frame for encoding and decoding. For the retrieval of the WM, no original signal is needed. The system succeeds in achieving high data rate and a robust watermark scheme against malicious manipulations. Moreover, the computations involved in the embedding process are kept relatively basic, suggesting suitability for future hardware implementation as well.

3.1 Hardware Implementations

Over the last decade, numerous software WM algorithms have been invented [12]. However, WM implementation in hardware, especially for video stream, is a recent interest in the area. Prior to 1999, no work on video WM implementation in hardware had been shown [8]. However, the watermarking of video streams in real-time applications is mostly suitable for hardware implementations, thus motivating research efforts to that direction.

The implementation of hardware WM is usually done on custom-designed circuitry, i.e. application specific integrated circuits (ASICs) or field programmable gate arrays (FPGAs). The overall advantage of this scheme over the software implementation is in terms of lower power consumption, reduced area and reliability. It may be possible to add a small, fast and potentially cheap WM embedder as a part of portable consumer electronic devices, such as a digital camera, camcorder or other multimedia devices, so that the media data are watermarked at the origin. Therefore, it is most suitable for real time applications. On the other hand, hardware implementations of WM techniques demand the flexibility of implementation both in the computation and design complexity. The algorithm must be carefully designed, minimizing any unexpected deficiencies.

For example, in 2000, Strycker et al. proposed a real time video WM scheme, called Just Another Watermarking System (JAWS), for television broadcast monitoring [7]. JAWS is a well-known video WM algorithm and because it works on uncompressed real time video data, the author was allowed to concentrate on the watermark process and not on the compression issues. In the embedding procedure, a PR sequence is embedded in an uncompressed, real time video stream and the depth of the watermark insertion depends on the luminance value

of each frame. The implementation of JAWS is performed on a Trimedia TM-1000 VLIW processor with 4 BOPS (billion operations per second) developed by Philips Semiconductors. The results prove the feasibility of a professional television broadcast monitoring system. Mathai et al., present an ASIC implementation of the JAWS WM algorithm using 1.8V, 0.18μm CMOS technology for real time video stream embedding [8], [9]. The authors claim that their work is the first step toward analyzing the relationship between WM algorithmic features and implementation cost for practical systems. A WM embedder and detector have been demonstrated to process raw digital video streams at a rate of 30 frames/sec and 320×320 pixels/frame. The results show a chip with a core area of 3.53 mm^2, capable of operating at 75 MHz frequency, processing a peak pixel rate of over 3 Mpixels/sec and only consuming 60 mW of power for the embedder. The hardware employed in this implementation is comprised of video and WM RAM memories, adders/subtractors, registers and multipliers.

A new VLSI architecture of real time WM system for both spatial and transform domains is presented by Tsai and Wu [10]. In this scheme, the concepts of spread spectrum from the field of communications and the human visual system (HVS) are applied to create a robust WM system. The proposed design embeds a logo (used as a watermark) in uncompressed and compressed video streams efficiently. Performance is tested under real time conditions, using a video stream with a rate of 6 Mbits/sec and 65 bits/frame watermark sequence. They also claim that it could be combined with an MPEG encoder in a System-On-Chip (SOC) design to achieve real time intellectual property protection on digital video capturing devices.

To conclude, there is still much to be accomplished in the field of video WM hardware implementations. There are many potential applications and still not enough solutions at hand. The existing work is mainly focused on the adaptation of watermarking algorithms that were originally designed for still images software watermarking to the requirements of video and hardware. It is a great opportunity for new innovative watermarking solutions, specifically designed to accommodate the requirements of video applications including compression standards and real time operation.

4. Conclusions

In this paper, background on video WM techniques was provided. Common WM classification criterions and requirements, including general properties and specific constrains for video WM scheme, were presented. Various applications of video WM were discussed. Comparisons between software and hardware implementations have been presented from several points of view: major advantages, drawbacks and differences. Four examples of previous software and hardware WM implementations were also shown.

Bibliography

[1] V. M. Potdar, S. Han, E. Chang, "A survey of digital image watermarking techniques", 3rd IEEE International Conference on Industrial Informatics (INDIN '05), Aug. 2005, pp. 709- 716

[2] Piva A. & Barni M., “Managing Copyright in Open Networks,” IEEE Internet Computing, MAY-June 2002.

[3] S. P. Mohanty, "Digital Watermarking: A Tutorial Review", URL: http://www.csee.usf.edu/~smohanty/research/Reports/WMSurvey1999Mohanty.pdf

[4] Frank Hartung, and Bernd Girod. “Watermarking of Uncompressed and Compressed Video,” IEEE Transactions on Signal Processing. Vol. 66, No. 3, May 1998, pp. 283 - 302.

[5] T.L. Wu, S.F. Wu, “Selective encryption and watermarking of MPEG video” , International Conference on Image Science, Systems, and Technology, CISST’97, June 1997.

[6] Ambalanath Shan, and Ezzatollah Salari, “Real-Time Digital Video Watermarking,” 2002 Digest of Technical Papers: International Conference on Consumer Electronics, June 2002, pp.12 – 13.

[7] L. D. Strycker, P. Termont, J. Vandewege, J. Haitsma, A. Kalker, M. Maes, and G. Depovere, “Implementation of a real-time digital watermarking process for broadcast monitoring on Trimedia VLIW processor,” IEE Proc. Vision, Image Signal Processing, vol. 147, no. 4, pp. 371–376, Aug. 2000.

[8] Nebu John Mathai, Ali Sheikholesami, and Deepa Kundur, “Hardware Implementation Perspectives of Digital Video Watermarking Algorithms”, IEEE Transactions on Signal Processing. Vol. 51, Issue 4, April 2003. pp. 925 - 938.

[9] Nebu John Mathai, Ali Sheikholesami, and Deepa Kundur, "VLSI Implementation of a Real-Time Video Watermark Embedder and Detector". Proceedings of the 2003 International Symposium on Circuits and Systems, 2003. ISCAS '03. Vol. 2, May 2003 pp. II772 - II775.

[10] Tsai, T.H., Wu, C.Y, "An Implementation of Configurable Digital Watermarking Systems in MPEG Video Encoder," In: Proc. of Intl. Conf. on Consumer Electronics. (2003) 216–21

[11] Maes, M., Kalker, T., Linnartz, J.P.M.G., Talstra, J., Depovere, G.F.G., Haitsma, J, "Digital Watamarking for DVD Video Copyright Protection," IEEE Signal Processing Magazine 17 (2000) 47–57.

[12] Sin-Joo Lee, and Sung-Hwan Jung, "A survey of watermarking techniques applied to multimedia". IEEE International Symposium on Industrial Electronics, Korea, June 2001. Vol. 1, pp. 272 – 277.

[13] Y. Shoshan, A. Fish, X. Li, G. A. Jullien, O. Yadid-Pecht, "VLSI Watermark Implementations and Applications," IJ Information and Knowledge Technologies, Vol.2, 2008.

[14] Watermarking World, http://www.watermarkingworld.org.

[15] I. J. Cox, J. Kilian, F. T. Leighton, and T. Shamoon, "Secure spread spectrum watermarking for multimedia," IEEE Trans. Image Process., vol. 6, no. 12, pp. 1673–1687, Dec. 1997.

[16] G. Doërr and J.-L. Dugelay, "A guide tour of video watermarking," Signal Processing: Image Commun., vol. 18, no. 4, pp. 263–282, Apr. 2003.

[17] M. Barni, F. Bartolini, J. Fridrich, M. Goljan, and A. Piva, "Digital watermarking for the authentication of AVS data," in EUSIPCO00, 10th Eur. Signal Processing Conf., Tampere, Finland, Sept. 2000.

[18] F. Bartolini, A. Tefas, M. Barni, and I. Pitas, "Image authentication techniques for surveillance applications," Proc. IEEE, vol. 89, no. 10, pp. 1403–1418, Oct. 2001.

[19] K, Jack, "Video Demystified: a handbook for the digital engineer," 2rd ed., LLH Technology Publishing, Eagle Rock, VA 24085, 2001.

Authors' Information

Xin Lin *– ISL lab, ATIPS lab, ECE Department, University of Calgary, Calgary AB, Canada;*
e-mail: xinli@atips.ca

Yonatan Shoshan *– ISL lab, ATIPS lab, ECE Department, University of Calgary, Calgary AB, Canada;*
e-mail: shoshayi@atips.ca

Alexander Fish *– ISL lab, ATIPS lab, ECE Department, University of Calgary, Calgary AB, Canada;*
e-mail: fish@atips.ca

Graham Jullien *– ISL lab, ATIPS lab, ECE Department, University of Calgary, Calgary AB, Canada;*
e-mail: jullein@atips.ca

Orly Yadid-Pecht *– ISL lab, ATIPS lab, ECE Department, University of Calgary, Calgary AB, Canada;*
The VLSI Systems Center, Ben-Gurion University, Beer-Sheva, Israel; e-mail: orly@atips.ca

ON THE FEASIBILITY OF STEERING SWALLOWABLE MICROSYSTEM CAPSULES USING COMPUTER-AIDED MAGNETIC LEVITATION

Billy T. Wu, Martin P. Mintchev

***Abstract**: Swallowable capsule endoscopy is used for non-invasive diagnosis of some gastrointestinal (GI) organs. However, control over the position of the capsule is a major unresolved issue. This study presents a design for steering the capsule based on magnetic levitation. The levitation is stabilized with the aid of a computer-aided feedback control system and diamagnetism. Peristaltic and gravitational forces to be overcome were calculated. A levitation setup was built to analyze the feasibility of using Hall Effect sensors to locate the in-vivo capsule. CAD software Maxwell 3D (Ansoft, Pittsburgh, PA) was used to determine the dimensions of the resistive electromagnets required for levitation and the feasibility of building them was examined. Comparison based on design complexity was made between positioning the patient supinely and upright.*

***Keywords**: Computer-Aided Magnetic Levitation, Capsule Endoscopy, Real-Time Image Transmission*

***ACM Classification Keywords:** J.6 Computer-Aided Engineering*

***Conference**: The paper is selected from Sixth International Conference on Information Research and Applications – i.Tech 2008, Varna, Bulgaria, June-July 2008*

Introduction

Gastrointestinal (GI) disorders, including cancers, are common medical problems which affect up to 35% of the world population [1]. Fiberoptic endoscopy has become a preferred diagnostic method for the early detection of polyps and cancers in the GI tract, and particularly in the colon [2]. However, the invasive nature of this test makes its wide applicability for early screening and prevention purposes difficult, if not impossible [3].

Microsystem Diagnostic Capsules

Swallowable capsule endoscopy (CE) has been suggested as a non-invasive alternative to fiberoptic endoscopy for some GI organs to minimize the discomfort for patients and to enhance screening applicability of the test [4]. A capsule endoscope of the size of a medical pill may contain an imaging device and various sensors for monitoring the characteristics of the examined GI lumen. In addition, it is typically equipped with embedded battery and radio-frequency transmitter for power supply and real-time data transmission, respectively. CE has been utilized for diagnosing small intestinal bleeding and has been recently suggested for detecting abnormalities of the esophageal wall [4]. The relatively small lumens in both organs preclude the capsule from tumbling, and natural peristalsis provides a reasonable means of steering the capsule, albeit in uncontrolled fashion. Steering based on natural peristalsis is not feasible for larger-lumen GI organs, e.g. the stomach and the colon, because the capsule tumbles and correct recognition of the tested area becomes impossible, not to mention the fact that substantial segments of the tested organ could be missed simply because the capsule would fall through cylindrical spaces of larger diameter.

Steering Options

Presently, the movement of the capsule through the GI tract relies on the propulsive contractile activity of the smooth muscles of the given organ. The inability of CE to achieve positional stabilization and to have independent external steering precludes broader applicability of this innovative diagnostic technique, which could become a pivotal screening test for GI polyps and cancers. In addition, controlled steering of the capsule and future developments in the area of micro-electromechanical systems (MEMS) create the possibility for collecting biopsies concurrently with the luminal examination, and even for the removal and the collection of smaller polyps or growths. Therefore, independent, externally-controlled intraluminal steering of the capsule is of pivotal importance for exploiting the full potential of CE not only as a diagnostic, but also as a therapeutic technique.

Forces Acting on the Capsule

In order for the capsule to levitate in the lumen (the esophagus is chosen for this study because of its simple vertical structure), gravitational and peristaltic forces exerted on the capsule have to be overcome. Suppose the capsule is 10 mm in diameter and 25 mm in length. The two ends of the capsule are hemispheres, each with a surface area of 0.000157 m^2. With a miniature magnet, sensors, and circuitry inside, the capsule is estimated to weigh less than 11 g in total. The gravitational force on the capsule would be less than 0.108 N.

Peristaltic pressure is applied on the top of the hemispherical surface of the capsule as the contraction of the smooth muscles pushes the capsule towards the stomach. Table I lists the average and absolute maximum peristaltic forces for both supine and upright positions, calculated based on the hemispherical surface area of the capsule and published pressure data on esophageal contractility [5].

Table I. Average and absolute maximum pressures and peristaltic forces anticipated during a normal esophageal contraction.

	Supine	Upright
Average Pressure (kPa)	6.67	5.33
Absolute Max. Pressure (kPa)	10.5	8.66
Average Force (N)	1.05	0.837
Absolute Max. Force (N)	1.65	1.36

Methods for Magnetic Levitation

Earnshaw's theorem implies that a permanent magnet cannot maintain levitation in a magnetostatic field [6]. However, although magnetic levitation is intrinsically unstable, this does not necessarily mean that it cannot be artificially stabilized using external aiding sources such as particular sensors and/or diamagnetism.

Diamagnetic materials, including water and living tissues, develop persistent atomic or molecular currents which oppose externally applied magnetic fields. Even though the diamagnetic stabilizing force is much weaker than the magnetic lifting force, diamagnetically stabilized levitation of a miniature permanent magnet has been demonstrated [6]. The walls of the GI organ encompassing the capsule can fulfill the role of diamagnetic stabilizer, which can be supported further by shelling the capsule with an appropriate diamagnetic material (e.g. bismuth or pyrolytic graphite) [6].

If the position of the capsule (or the levitating magnet inside) can be determined in real time, the field of the electromagnets can be continuously adjusted via a feedback control system to keep the capsule in the desired position. A levitation kit [7] demonstrating this method of levitation is commercially available. It uses a Hall Effect sensor to detect the position of the levitating magnet, as the sensor produces a signal based on the field strength of the magnet. This signal controls the current supplied to the coil, which in turn determines the strength of the field.

A combination of these two levitation-aiding techniques can be utilized in steerable CE.

Aim of the Study

The aim of this study is to explore the feasibility of steering an endoscopic capsule and holding it at areas of interest in the GI tract using aided magnetic levitation. A conceptual design for controlling the movement of the capsule and for affixing it at particular locations is presented. Magnetic levitation is achieved by using powerful resistive electromagnets at least 30 cm apart, situated at the front and back sides of the patient. The levitation is stabilized with the aid of a feedback control system and diamagnetism. The feasibility of designing such system for a patient in supine and upright positions is examined.

Methods

The patient can be arranged in two possible positions, supine and upright, as depicted in Fig. 1.

Modeling Setup

The force between two permanent magnets or between a permanent magnet and an electromagnet is the result of non-uniform magnetic field [8]. Force calculations dealing with non-uniform magnetic fields and permanent magnets are complex, and involve finite element analysis. Software simulations were created using the computer-aided-design (CAD) system Maxwell 3D (Ansoft, Pittsburgh, PA) to determine the size and the arrangement of the electromagnets.

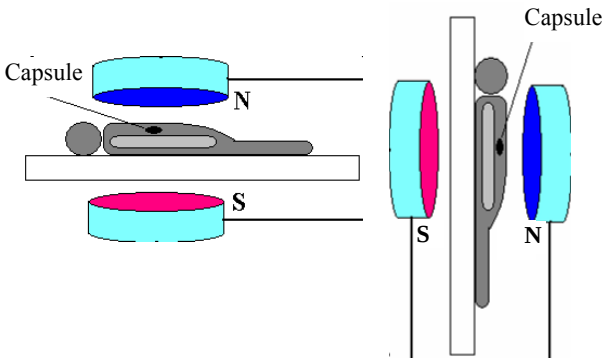


Fig. 1. Options for positioning the patient during diagnosis, supine (left) and upright (right).

Experimental Setup

A levitation kit [7], which can be thought of as a micro-model of a patient in supine position, was utilized to build a complete electronically-controlled system for establishing stable levitation with feedback control. The purpose of this setup was to analyze the detection of the magnet position and compare the experimental findings to the CAD system modeling. A hand-wound coil was used as the electromagnet, with roughly 5000 turns of American-wire-gauge (AWG) 20 copper wire. The core of the coil was a 1018-stainless-steel bolt. A spherical NdFeB permanent magnet was placed in a clay-filled capsule to give a combined mass of 10.9 g. The Hall Effect sensor was positioned directly below the coil on the center axis. The equilibrium position of the magnet at which it levitated was about 10 mm below the Hall Effect sensor for a coil current of 158 mA.

Steering in Supine Position

A simulation CAD model (Fig. 2) was created to verify the experimental setup. The upward force on the magnet given by Maxwell 3D was 0.106 N, which was the weight of the capsule. The second side of electromagnet and the peristaltic force were not considered at this time in order to simplify the experimental setup.

Ideally, the sensor should only detect the field from the magnet, but the magnetic field of the coil was inevitably picked up by the sensor as well. In the designed experimental setup, the magnet contributed much more to the magnetic field of the system compared to the contribution of the coil. This demonstrates the effectiveness of positioning the Hall Effect sensor directly below the coil.

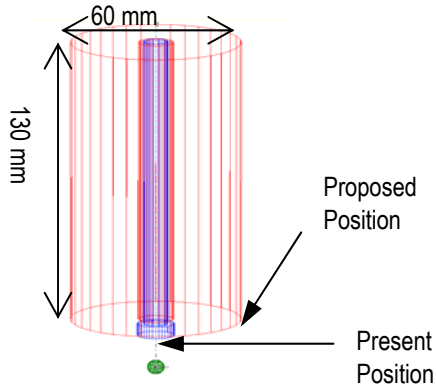


Fig. 2. Experimental Setup for the present and the proposed position of the Hall Effect sensor.

However, in an actual application, the distance between the coil and the levitating magnet should be far greater than 10 mm. The size of the coil and the current flowing through it would increase significantly as well. The present position of the Hall Effect sensor would not be applicable because the coil would produce a much greater magnetic field. Moving the sensor to the side at the bottom of the coil (proposed position in Fig. 2) was examined with simulation models. The respective contributions of the magnetic sources (the coil and the magnet) at the proposed position were studied.

Steering in Upright Position

Following the validation of the CAD model with the small-scale experimental setup, a different approach was pursued in the CAD modeling for the upright position. Rather than parameterizing a physical micro-model, the simulation CAD model was constructed based on the requirements of an actual application, i.e. actual maximal peristaltic force and a 30-centimeter separation between the electromagnets were utilized as parameters. This ensured that the electromagnets were of feasible size before more time and effort was spent on details of the feedback control system.

After some refining, the model in Fig. 3 was determined to be capable of producing an upward force greater than 1.47 N on the levitating magnet to counter the gravitational and peristaltic forces (0.108 N + 1.36 N).

In the model, the resistive electromagnets were designed using copper wire. The electromagnet on each side was separated into inner and outer coils. To keep things simple, the wire current (in amperes) and the wire gauge of the inner and outer coils were considered to be the same. The currents of each inner coil and each outer coil were 144,000 ampere-turns and 256,000 ampere-turns, respectively. Even with the additional costs of a power supply with precise regulation and a cooling system, resistive electromagnets were determined to be generally much cheaper and simpler to build than superconducting magnets [9]. Soft iron cores were used to concentrate the magnetic field generated.

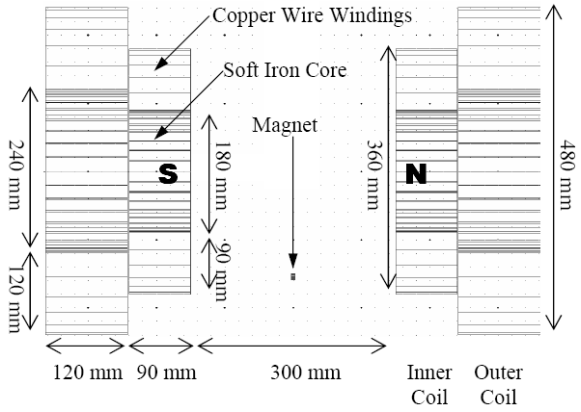


Fig. 3. Cross-sectional view of the simulation model for upright position in an actual application.

A cylindrical magnet with a diameter of 9.55 mm and a height of 5 mm was selected for levitation. It was larger and heavier than the spherical magnet used in the supine case, but the extra weight was much smaller compared to the peristaltic force. The trade-off of using a larger levitating magnet was the size reduction of the electromagnets required to achieve levitation.

It should be mentioned also, that additional stabilization of the capsule in both steering modalities can be facilitated by the organ walls (providing their proximity to the capsule is reasonable), or by implementing the entire or part of the outer shell of the capsule using diamagnetic material [6]. This stabilization aid can substantially ease the sensitivity requirements for the Hall Effect sensors.

Results

Steering in Supine Position

Fig. 4 shows that the field of the magnet was minimal at the proposed position and was overwhelmed by the coil field. The situation would worsen when the coil is enlarged with its field greatly increased in an actual application. Thus, keeping the Hall Effect sensor at a position immediately under the coil is not feasible.

The fundamental problem of detecting the magnet position is not in finding an appropriate location for the sensor. Instead, it is the field of the magnet decreasing significantly with distance. If locating the magnet is to be based on magnetic field detection, the sensor must be in close proximity to the magnet.

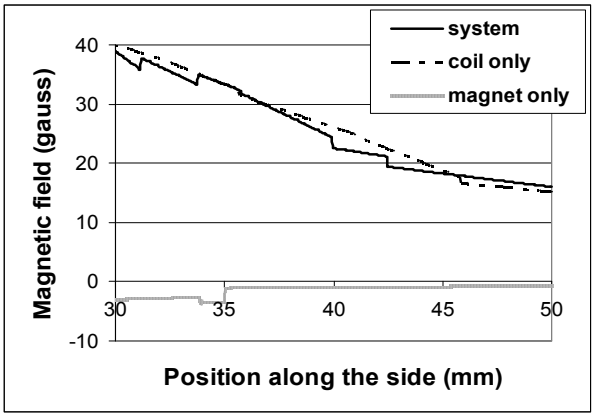


Fig. 4. Breakdown of the field contributions near the proposed position (supine case).

An alternative would be to measure the magnetic field of the coil from inside the capsule. With the coil field at a fixed location outside the body serving as a reference, the capsule can be located with respect to the coil. This two-sensor design is illustrated in Fig. 5.

The unprocessed signal from sensor #1 consists of contributions from the magnet and the coil. Since the magnet and sensor #1 are both inside the capsule, the field from the magnet remains the same with respect to sensor #1, regardless of the capsule position. Thus the field contribution of the coil inside the capsule can be obtained by adaptively subtracting a constant from the unprocessed signal.

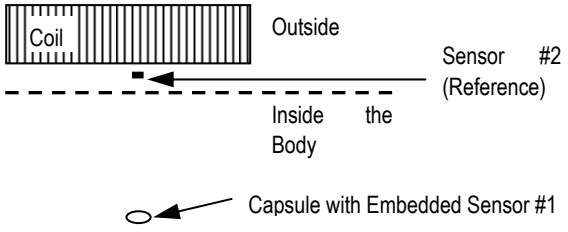


Fig. 5. Two-sensor concept for supine position.

Steering in Upright Position

In the developed simulation model, the magnetic field produced by the electromagnets was about 2.1 T at the pole faces of the inner coils, and 0.5 T at the point of levitation.

The size of the electromagnets was considered acceptable. In order to show the feasibility of building such electromagnets, it is necessary to determine their weights and heating rate. These two factors are dictated by the wire gauge selection, since the dimensions of the electromagnets have been predicted using simulation. Equations (1) and (2) for designing MRI resistive electromagnets [9] were utilized:

$$dT/dt \approx \frac{1}{c_p \rho \sigma}\left(\frac{I_w}{A_w}\right)^2 \quad \textbf{(1)}$$

where dT/dt is the rate of temperature increase in °C/s, I_w is the current of each winding, and A_w is the cross-sectional area of the wire. The parameters c_p, ρ, and σ are the specific heat, density, and the electrical conductivity of copper, respectively.

Given a coil with N turns, the mass is:

$$m = V\rho = 2\pi A_w \rho \sum_{n=1}^{N} a_n \quad \textbf{(2)}$$

where V is the volume of the coil (with copper windings), and a_n is the average radius of the n^{th} winding. Since the cross-sectional area is a square matrix of coil windings, N_S can be defined as the number of turns per side, where $N_S = \sqrt{N}$. a_n can be quantified as:

$$a_n = a_{n-1} + d_w = a_{n-2} + 2d_w = ... = a_1 + (n-1)d_w \quad \textbf{(3)}$$

where d_w is the diameter of the wire and a_1 is the radius of the innermost winding. Then the mass of the coil is:

$$m = 2\pi A_w \rho N_S \sum_{n=1}^{N_S} (a_1 + (n-1)d_w) \quad \textbf{(4)}$$

$$m = 2\pi A_w \rho N_S \left[N_S (a_1 - d_w) + d_w \sum_{n=1}^{N_S} n \right] \quad \textbf{(5)}$$

The masses of the soft iron cores can be estimated using the density of iron. AWG 14 wire was selected because of the reasonable balance between the winding current and the heating rate. The winding current was calculated to be 53 A, while the heating rate was 3.15°C/s. The heating rate is significant, but this value is for the case when the electromagnets are running at full capacity (i.e., when there is a contraction). The contraction only lasts few seconds [5], then the current will drop back to a much lower value. Air or water cooling will probably be required. The number of turns and the masses of each coil and core are listed in Table II.

Table.II Electromagnet characteristics for upright position in an actual application.

	Inner	Outer
# of Turns	52	70
Coil Mass (kg)	42.3	102.6
Core Mass (kg)	17.9	42.3

Discussion

In the search for optimal computer-aided magnetic levitation for capsule endoscopy, the proposed two-sensor, diamagnetically-enhanced concept for the supine position could be implemented for the upright position as well. A magnetic measurement device capable of measuring substantial field with considerable precision, such as a NMR gaussmeter, should replace sensor #2, because the field at the location of sensor #2 is much greater in an actual application. An additional magnetic measurement device outside the body would be required for the second electromagnet at the back side of the patient. The control circuit would adjust the currents of the inner coils with the magnetic sensing system. A contraction could be detected with pressure sensors mounted on each end of the capsule. Depending on the pressure exerted on the capsule, the currents of the outer coils could be increased accordingly to maintain levitation.

When the second side of the electromagnet and the peristaltic force are taken into account for the supine position, the magnetic forces are not symmetrical because the gravitational and peristaltic forces are in orthogonal directions. The design for controlling the asymmetrical field strength of the electromagnets would be complicated and not very feasible. When the patient is positioned upright, the peristaltic and gravitational forces are in the same direction, leading to symmetrical magnetic forces. This simplifies the design of the field control system.

Utilization of this diagnostic technique in larger-lumen organs of the GI tract (e.g. in the colon) is highly feasible. Although colonic pressures are complex and variable, the maximum peristaltic force in the colon is not significantly greater compared to that in the esophagus [5], [10]. In addition, pharmacological agents for reducing colonic contractility could temporarily be used during the CE test.

An interesting discussion point is the potential mechanical resistance of a collapsed organ wall to the controlled steering of the capsule in the absence of contractile activity. Similarly to present-day fiberoptic endoscopy, this problem can be easily bypassed by appropriately inflating the investigated organ with air prior to the test. However, this procedure itself could jeopardize the non-invasive nature of the test. Thus, additional work is needed to estimate the related forces and incorporate them into the design. The issue of the maximal steering velocity of the capsule and its relationship to the levitation characteristics needs quantification as well.

Conclusion

Simulation models demonstrated that electromagnets in a set of inner and outer coils with soft iron cores can be built to levitate an endoscopic capsule and to counteract the peristaltic force. Stabilization can be achieved with the aid of a computer-aided feedback control system and diamagnetism.

Bibliography

[1] O. Y. Lee, M. Schmulson, and E. A. Mayer, "Common functional gastrointestinal disorders: Nonulcer dyspepsia and irritable bowel syndrome," *Clinical Cornerstone*, vol. 1, no. 5, pp. 57-71, 1999.

[2] Medline Plus, U.S. National Library of Medicine. (March, 2005). Endoscopy. Bethesda, U.S. [Online]. Available: http://www.nlm.nih.gov/medlineplus/ency/article/003338.htm

[3] R. Frankenthaler et al., "Virtual Otoscopy," *Otolaryngologic Clinics of North America*, vol. 31, pp. 383-392, 1998.

[4] B.S. Lewis, "The utility of capsule endoscopy in obscure gastrointestinal bleeding," *Techniques in Gastrointestinal Endoscopy*, vol. 5, no. 3, pp. 115-120, July, 2003.

[5] H. J. Stein, S. Singh, and T. R. DeMeester, "Efficacy of esophageal peristalsis: A manometric parameter to quantify esophageal body dysfunction," *Diseases of the Esophagus*, vol. 17, no. 4, pp. 297-303, December, 2004.

[6] M. D. Simon, L. O. Heflinger, and A. K. Geim, "Diamagnetically stabilized magnet levitation," *American Journal of Physics*, vol. 69, no. 6, pp. 702-713, June, 2001.

[7] G. Marsden, "Levitation! Float objects in a servo controlled magnetic field," *Nuts & Volts Magazine*, pp. 58-61, September, 2003.

[8] H. D. Young and R. A. Freedman, "Force and torque on a current loop," in *University Physics*, 9th ed., J. Berrisford, Ed. Boston: Addison-Wesley, 1996, pp. 865-902.

[9] P. N. Morgan, S. M. Conolly, and A. Macovski, "Resistive homogeneous MRI magnet design by matrix subset selection," *Magnetic Resonance in Medicine*, vol. 41, pp. 1221-1229, 1999.

[10] S. S. C. Rao, P. Sadeghi, J. Beaty, and R. Kavlock, "Ambulatory 24-Hour Colonic Manometry in Slow-Transit Constipation," *American Journal of Gastroenterology*, vol. 99, no. 12, pp. 2405-2416, December, 2004.

Authors' Information

Billy T. Wu *– Graduate Student, Department of Electrical and Computer Engineering, University of Calgary, Calgary, Alberta, Canada T2N 1N4; e-mail: wu@enel.ucalgary.ca*

Martin P. Mintchev *– Professor, Department of Electrical and Computer Engineering, University of Calgary, Calgary, Alberta, Canada T2N 1N4; e-mail: mintchev@ucalgary.ca*

APPLICATION OF FUZZY LOGIC ON IMAGE EDGE DETECTION

Shashank Mathur, Anil Ahlawat

***Abstract**: In this paper a novel method for an application of digital image processing, Edge Detection is developed. The contemporary Fuzzy logic, a key concept of artificial intelligence helps to implement the fuzzy relative pixel value algorithms and helps to find and highlight all the edges associated with an image by checking the relative pixel values and thus provides an algorithm to abridge the concepts of digital image processing and artificial intelligence. Exhaustive scanning of an image using the windowing technique takes place which is subjected to a set of fuzzy conditions for the comparison of pixel values with adjacent pixels to check the pixel magnitude gradient in the window. After the testing of fuzzy conditions the appropriate values are allocated to the pixels in the window under testing to provide an image highlighted with all the associated edges.*

***Keywords**: Fuzzy logic, digital image processing, feature extraction.*

***ACM Classification Keywords**: I.2 Artificial Intelligence, I.4 Image processing and computer vision*

***Conference**: The paper is selected from International Conference "Intelligent Information and Engineering Systems" INFOS 2008, Varna, Bulgaria, June-July 2008*

Introduction

Over the years, several methods have been proposed for the image edge detection which is the method of marking points in a digital image where luminous intensity changes sharply for which different type of methodology have been implemented in various applications like traffic speed estimation [I], Image compression [II], and classification of images [III]. Most of the traditional edge-detection algorithms in image processing typically convolute a filter operator and the input image, and then map overlapping input image regions to output signals which lead to considerable loss in edge detection [IV]; however there is no such loss in the fuzzy based method described here. Research has clearly demonstrated that methods involving Gaussian filtering suffer from problems such at edge displacement, vanishing edges and false edges [V]. Another problem faced by few methods like the anisotropic diffusion lies in obtaining the locations of semantically meaningful edges at coarse scales generated by convoluting images with Gaussian kernels [VI]. Methods that involve simple scan line approach are not able to detect all the edges due to limitation of the methodology to trace only the horizontal and vertical neighbors [VII] of a point.

The method described does not implement any thresholding unlike few published methods [VIII] which helps to detect each and every edge associated with the image but introduces fuzzy logic which derives its origin from approximate reasoning for highlighting all the edges associated with an image. The fuzzy relative pixel value algorithm has been developed with the knowledge of vision analysis with low or no illumination [IX], thus making this method optimized for application requiring such methods. The method helps us to detect edges in an image in all cases due to subjection of pixel values to an algorithm involving host of fuzzy conditions for edges associated with an image.The purpose of this paper is to present a new methodology for image edge detection which is undoubtedly one of the most important operations related to low level computer vision, in particular within area of feature extraction with plethora of techniques, each based on a new methodology, having been published. The method described here uses a fuzzy based logic model with the help of which high performance is achieved along with simplicity in resulting model [X]. Fuzzy logic helps to deal with problems with imprecise and vague information and thus helps to create a model for image edge detection as presented here [XI] displaying the accuracy of fuzzy methods in digital image processing [XII].

Methodology

The algorithm described below is based on the subjection of a set of nine pixels, part of a 3x3 window of an image to a set of fuzzy conditions which help to highlight all the edges that are associated with an image. The fuzzy conditions help to test the relative values of pixels which can be present in case of presence on an edge. So the relative pixel values are instrumental in extracting all the edges associated to an image.

The Fuzzy Relative Pixel Value algorithm

The Algorithm begins with reading an MxN image. The first set of nine pixels of a 3x3 window are chosen with central pixel having values (2,2). After the initialization, the pixel values are subjected to the fuzzy conditions for edge existence shown in Fig.1.(a-h). After the subjection of the pixel values to the fuzzy conditions the algorithm generates an intermediate image. It is checked whether all pixels have been checked or now, if not then first the horizontal coordinate pixels are checked. If all horizontal pixels have been checked the vertical pixels are checked else the horizontal pixel is incremented to retrieve the next set of pixels of a window (refer to flowchart shown in Fig.4.). In this manner the window shifts and checks all the pixels in one horizontal line then increments to check the next vertical location.

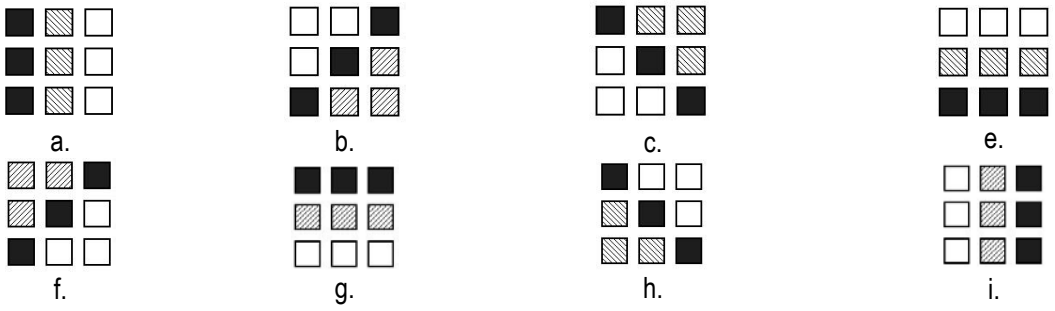


Fig 1(a-h). Fuzzy conditions have been displayed.

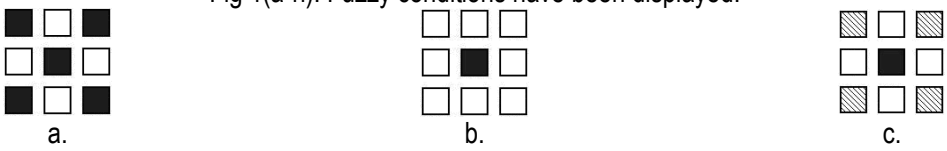


Fig 2. (a,b)Type of unwanted edge pixels (c) Fuzzy condition for removal of unwanted edge pixels.

Legend:

Pixel Checked Edge pixel Unchecked Pixel

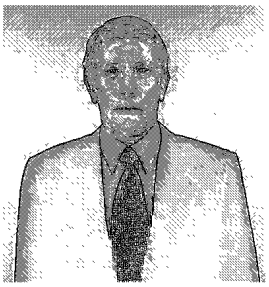
After edge highlighting image is subjected to another set of condition with the help of which the unwanted parts of the output image of type shown in Fig.2.(a-b) are removed to generate an image which contains only the edges associated with the input image. Let us now consider the case of the fuzzy condition displayed in Fig.1. (g). For an input image A and an output image B of size MxN pixels respectively we have the following set of conditions that are implemented to detect the edges pixel values.

```
Input: An image A of MxN pixels
Output: An image B of MxN pixels
Edge Detection (A, B)
For I←2 to M-1
    For J←2to N-1
      If A (I-1, J)>A (I-1, J+1)
        Then If A (I-1, J-1)>A (I, J)
          Then If A (I, J-1)>A (I+1, J-1)
            Then
              B (I-1, J+1) ←0
              B (I, J) ←0
```

```
                    B (I+1, J-1) ←0
    End For
End For
For I←2 to M-1
    For J←2to N-1
        If B(I-1,J)=255& B(I,J)=0& B(I+1,J)=255& B(I,J-1)=255
            Then B (I, J) ←255
    End For
End For
```

We can observe in the above algorithm written for a particular fuzzy condition that the nesting of statements is done in a manner that only the edge associated pixels are granted black pixel values. The application of fuzzy conditions on the image helps to highlight all the edges associated with it but do leave unnecessary pixel values which only distort the edge values. To eliminate these unwanted edge pixels another fuzzy condition is implemented to enhance the working of the fuzzy relative pixel value algorithm. With the help of these set of conditions the algorithm is able to eliminate all the noisy pixels and filters out the edges to provide us with a clean output image with all the distinct edges associated with that image.

a. Input Image



b. Intermediate Image

c. Output Image

Experimental Results

The fuzzy relative pixel value algorithm for image edge detection was tested for various images and the outputs were compared to the existing edge detection algorithms and it was observed that the outputs of this algorithm provide much more distinct marked edges and thus have better visual appearance than the ones that are being used. The sample output shown below in Fig.5.(a-c) compares the "Sobel" Edge detection algorithm and the fuzzy relative pixel value algorithm. It can be observed that the output that has been generated by the fuzzy method has found out the edges of the image more distinctly as compared to the ones that have been found out by the "sobel" edge detection algorithm. Thus the Fuzzy relative pixel value algorithm provides better edge detection and has an exhaustive set of fuzzy condition s which helps to extract the edges with a very high efficiency.

a. Input Image

b. Sobel algorithm

c. Fuzzy relative pixel value algorithm

Fig 5(a-c). Comparison of Fuzzy relative pixel value algorithm and Sobel edge detection algorithm.

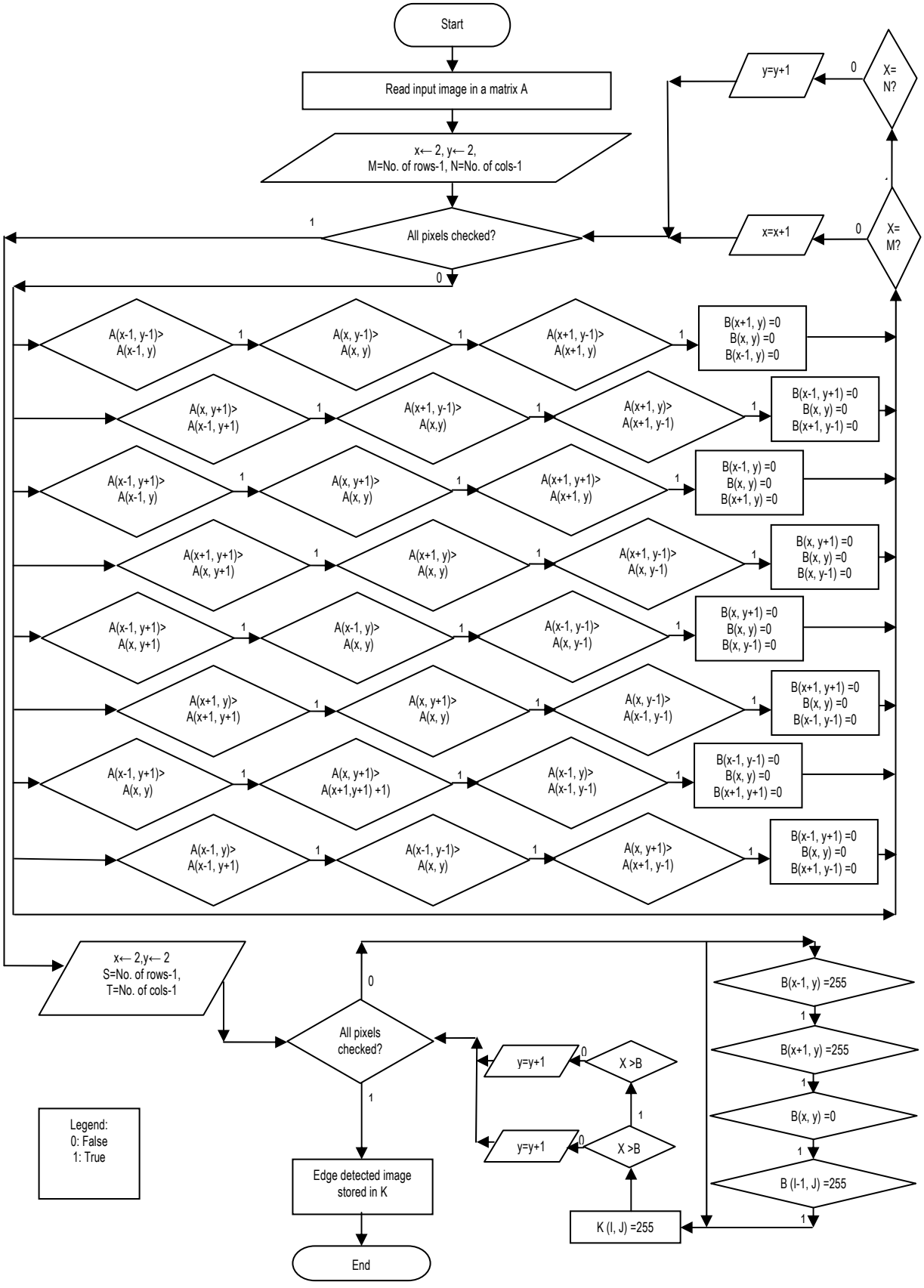


Fig.3.Flowchart

Conclusion

In this paper, the algorithm to find the edges associated with an image had been introduced which has been instrumental to abridge the concepts of artificial intelligence and digital image processing. Comparisons were made amongst the various other edge detection algorithms that have already been developed and displayed the accuracy of the edge detection using the fuzzy relative pixel value algorithm over the other algorithms which has tremendous scope of application in various areas of digital image processing. The image edge detection using fuzzy relative pixel value algorithm has been successful in obtaining the edges that are present in an image after the implementation and execution of the algorithms with various sets of images. Sample outputs have been shown to make the readers understand the accuracy of the algorithm and the display that the algorithm can find image edges even in case of minor pixel value gradients.

Bibliography

[I.] Dailey D. J., Cathey F. W. and Pumrin S. 2000. An Algorithm to Estimate Mean Traffic Speed Using Uncalibrated Cameras. In proceedings of IEEE Transactions on intelligent transport systems, Vol. 1.

[II.] Desai U. Y., Mizuki M. M., Masaki I., and Berthold K.P. 1996. Edge and Mean Based Image Compression. Massachusetts institute of technology artificial intelligence laboratory .A.I. Memo No. 1584.

[III.] Rafkind B., Lee M., Shih-Fu and Yu C. H. 2006. Exploring Text and Image Features to Classify Images in Bioscience Literature. In Proceedings of the BioNLP Workshop on Linking Natural Language Processing and Biology at HLT-NAACL 06, pages 73–80, New York City.

[IV.] Roka A., Csapó Á., Reskó B., Baranyi P. 2007.Edge Detection Model Based on Involuntary Eye Movements of the Eye-Retina System. Acta Polytechnica Hungarica Vol. 4.

[V.] Basu M. 2002.Gaussian-Based Edge-Detection Methods—A Survey. In proceedings of IEEE Transactions on systems, man and cybernetics-Part C: Applications and reviews, Vol. 32, No. 3.

[VI.] Perona P., Malik J. 1990. Scale-Space and Edge Detection Using Anisotropic Diffusion. In proceedings of IEEE Transactions on pattern analysis and machine intelligence, Vol. 12. No. 7.

[VII.] Jiang X., Bunke H. 1999. Edge Detection in Range Images Based on Scan Line Approximation. Computer Vision and Image Understanding Vol. 73, No. 2, February, pp. 183–199.

[VIII.] Meer P. 2001. Edge Detection with Embedded Confidence. IEEE Transactions on pattern analysis and machine intelligence, Vol. 23, No. 12.

[IX.] Wang C. W. 2006. Real Time Sobel Square Edge Detector for Night Vision Analysis. ICIAR 2006, LNCS 4141, pp. 404 – 413.

[X.] Desai Y. U, Mizuki M. M., Masaki I., Berthold Horn K. P .The application of fuzzy logic to the construction of the ranking function of information retrieval system n.o.rumbens. Computer Modeling and New Technologies, 2006, Vol.10, No.1, 20-27, Artificial intelligence laboratory, Massachusetts Institute of Technology.

[XI.] Stefano B. D., Fuks' H. and Lawniczak A. T. 2000. Application of fuzzy logic in CA/LGCA models as a way of dealing with imprecise and vague data, IEEE 2000

[XII.] Chen Y., 2002.A region-based fuzzy feature matching approach to content based image retrieval. In proceedings of IEEE Transactions on pattern analysis and machine intelligence, VOL 24, NO. 9.

Authors' Information

Shashank Narain Mathur *– Bachelors of Technology (Computer Science Engineering), Amity School of Engineering and Technology affiliated to Guru Gobind Singh Indraprastha University, New Delhi, India.*
Email: shashanknarainmathur@yahoo.com

Anil Kumar Ahlawat *– Department of Computer Science Engineering, Amity School of Engineering and Technology affiliated to Guru Gobind Singh Indraprastha University, New Delhi, India.*
Email: a_anil2000@yahoo.com

DISTANCE MATRIX APPROACH TO CONTENT IMAGE RETRIEVAL

Dmitry Kinoshenko, Vladimir Mashtalir, Elena Yegorova

***Abstract**: As the volume of image data and the need of using it in various applications is growing significantly in the last days it brings a necessity of retrieval efficiency and effectiveness. Unfortunately, existing indexing methods are not applicable to a wide range of problem-oriented fields due to their operating time limitations and strong dependency on the traditional descriptors extracted from the image. To meet higher requirements, a novel distance-based indexing method for region-based image retrieval has been proposed and investigated. The method creates premises for considering embedded partitions of images to carry out the search with different refinement or roughening level and so to seek the image meaningful content.*

***Keywords**: content image retrieval, distance matrix, indexing.*

***ACM Classification Keywords**: H.3.3 Information Search and Retrieval: Search process*

***Conference**: The paper is selected from XIVth International Conference "Knowledge-Dialogue-Solution" KDS 2008, Varna, Bulgaria, June-July 2008*

Introduction

For the image retrieval from large scale database traditionally queries 'ad exemplum' are used. There are many approaches developed considering similarities of the query and images in database based on the distance between feature vectors which contain image content descriptors, such as color, texture, shape, etc [Greenspan et al., 2004; Yokoyama, Watanabe, 2007]. The most effort at present is put on the problem of putting the image retrieval on a higher semantic level to perform the search based on the image meaningful content, and not just on image own properties.

There are many metrics developed and widely used in image processing applications: Minkowski-type metric (including Euclidean and Manhattan distances), Mahalanobis metric, EMD, histogram metric, metric for probability density functions, sets of entropy metrics, pseudo metrics for semantic image classification [Rubner et al., 2000; Cheng et al., 2005; Wang et al., 2005]. Yet, by virtue of their limitations these metrics cannot give the desirable results, so a new metric was introduced and extended for considering the embedded partitions and so it was effectively used for the content image retrieval [Kinoshenko et al., 2007]. Due to the embedded structure it will become possible to perform the search with different level of refinement or roughening.

As volume of multimedia databases grows exponentially a great need of means of fast search arises. Many of multidimensional indexing methods used in the field of text retrieval were modified and improved in order to index high-dimensional image content descriptors. Among them X-trees, VA-file and I-Distance approaches are the most promising [Bohm et al., 2001]. However, in case of comparing images as embedded partitions we do not have the features to describe complex objects and only information about distances between them is available, and so-called 'distance-based' indexing methods come to the aid [Chavez et al., 2001; Hjaltason, Samet, 2003]. In this work existing 'distance-based' indexing methods are analyzed and improved and their possible application for the region-based image retrieval is considered.

Theoretical background of distance matrix based content image retrieval

Let $\mathrm{X} = \{x_1, x_2, ..., x_n\}$ be a set characterizing images in the database. Each element of this set can represent as an image $\mathrm{B}(z), z \in \mathrm{D} \subset \mathrm{R}^2$ (D is a sensor field of view); as feature vector $p \in \mathrm{R}^k$ (Z^k); as some combination of image processing results and features (e.g. segmentation, detected edges, shape features).

Further, keeping the generality of consideration, $\mathrm{X} \subseteq \mathrm{U}$ where U is some universum ensuring introducing of

similarity functional (specifically metric), we shall understand as an image database putting aside the indexing. The task consists in the search for correspondence of elements $x_i \in \mathrm{X}$ in the best way to the query $y \in \mathrm{U}$. Under 'the best way' we shall understand the minimum of distance $\rho(y, x), y \in \mathrm{U}, \ x \in \mathrm{X}$. Using metric as a similarity criterion provides adequacy of the search result to the query and the triangular inequality makes premises for excluding from the consideration whole sets of images without calculating distances to them.

We shall note, that there are 2 ways to perform the search with limited matches number: either by using preliminary clustering in image or feature spaces, or based on methods analyzing values of pre-calculated distance matrix for all images collection elements. Ex altera parte, all search algorithms can be classified as follows: search of k most similar images ordered according to their similarity; search of the images which differ from the query on not more than given δ (range queries), and combination of these two approaches.

Definition 1. The result of (δ)- search on query $y \in \mathrm{U}$ is any element (all the elements) $x_i \in \mathrm{X}$, if $\rho(y, x_i) \le \delta$ for given $\delta \ge 0$, called as the search radius.

It is clear that choice of range δ is a non-trivial task. Moreover, choice of rational value δ much depends on the database objects configuration (mutual location regarding the chosen metric). However often the choice of this value is dictated by the practical application, i.e. required extent of image similarity.

Definition 2. The result of (k)- search on query $y \in \mathrm{U}$ are elements of set $\mathrm{X}^k = \{x_{i_1}, x_{i_2}, ..., x_{i_k}\} \subseteq \mathrm{X}$, for which $\forall x_{i_j} \in \mathrm{X}^k, \forall x \in \mathrm{X} \setminus \mathrm{X}^k, \forall y \in \mathrm{U} \ \ \rho(y, x_{i_j}) \le \rho(y, x), \ \rho(y, x_{i_j}) \le \rho(y, x_{i_{j+1}}), \ j = \overline{1, k-1}$.

It is necessary to indicate a rather important special case: $y \in \mathrm{X}, k = 1$. It means that it is needed to find exact coincidence of query with a database element, i.e. practically identify query and detect image characteristics connected to it, e.g. to identify a person according to his fingerprints.

Definition 3. The result of (δ, k)-search on query $y \in \mathrm{U}$ are elements of set $\mathrm{X}^m = \{x_{i_1}, x_{i_2}, \ldots, x_{i_m}\} \subseteq \mathrm{X}, \ m \le n$, for which $\forall x_{i_j} \in \mathrm{X}^m, \forall x \in \mathrm{X} \setminus \mathrm{X}^m, \forall y \in \mathrm{U}, \rho(y, x_{i_j}) \le \delta \ \Rightarrow \rho(y, x_{i_j}) \le \rho(y, x), \rho(y, x_{i_j}) \le \rho(y, x_{i_{j+1}}), \ j = \overline{1, m-1}$.

We shall call a search successful if there are elements satisfying definitions 1, 2 and 3. Otherwise, the feedback coupling is needed i.e. query object or search parameters (for instance radius δ) refinement what is closely connected to the image presentation by feature descriptions and their matching. We shall emphasize that formally (k)- search is always successful as query refinement decision should always being made on the base of obtained distances analysis and solving task requirements. Notice that each successive search type is more complicated to solve than the previous one. Thus procedures of handling the (δ)- search are to be used as pre-processing during (k)- search, (δ, k)- search should exploit (δ)- search and (k)- search results.

We shall analyze the possibility of reducing the number N of calculative values $\rho(y, x_i)$ which can be rather computationally expensive especially in the image space. With that purpose a symmetrical matrix of all pairwise distances of all database elements

$$d(\mathrm{X}) = \begin{pmatrix} 0 & \rho(x_1, x_2) & \rho(x_1, x_3) & \ldots & \ldots & \rho(x_1, x_n) \\ & 0 & \rho(x_2, x_3) & \ldots & \ldots & \rho(x_2, x_n) \\ & & 0 & \ldots & \ldots & \ldots \\ & & & \ldots & \ldots & \ldots \\ & & & & 0 & \rho(x_{n-1}, x_n) \\ & & & & & 0 \end{pmatrix}. \qquad (1)$$

is created. Let $y \in \mathrm{U}$ be a query image. We shall fix some image $x^* \in \mathrm{X}$ called pivot object or vantage point or simply pivot and consider the triangular inequality involving one more image $x_i \in \mathrm{X},\ i \in \{1, 2, ..., n\}$ (remind that the distance $\rho(x^*, x_i)$ is known)

$$\rho(y, x_i) \le \rho(y, x^*) + \rho(x^*, x_i), \tag{2}$$

$$\rho(x_i, x^*) \le \rho(y, x^*) + \rho(y, x_i), \tag{3}$$

$$\rho(y, x^*) \le \rho(y, x_i) + \rho(x_i, x^*). \tag{4}$$

From inequalities (1) – (3) it follows that knowing two distances, notably $\rho(y, x^*)$ and $\rho(x^*, x_i)$, it is not hard to obtain low and upper distance bounds

$$\rho(x_i, x^*) - \rho(y, x^*) | \le \rho(y, x_i) \le \rho(y, x^*) + \rho(x_i, x^*). \tag{5}$$

Thus the implication $\forall y \in \mathrm{U},\ \forall x_i, x^* \in \mathrm{X} : \rho(x_i, x^*) \ge 2\rho(y, x^*) \Rightarrow \rho(y, x_i) \ge \rho(y, x^*)$ is true what can be used in the (k)- search. Let exact value of distance $\rho(x^*, x_i)$ be unknown and it can be evaluated as

$$\varepsilon_{min} \le \rho(x^*, x_i) \le \varepsilon_{max}. \tag{6}$$

Then if for objects x^*, x_i the inequality (6) is fulfilled the evaluation of low and upper distance bounds

$$max\{\rho(y, x^*) - \varepsilon_{max}, \varepsilon_{min} - \rho(y, x^*), 0\} \le \rho(y, x_i) \le \rho(y, x^*) + \varepsilon_{max} \tag{7}$$

is true. Indeed, according to the triangular inequality rule and taking into consideration (6) we have

$$\rho(y, x^*) \le \rho(y, x_i) + \rho(x_i, x^*) \le \rho(y, x_i) + \varepsilon_{max}$$

then

$$\rho(y, x^*) - \varepsilon_{max} \le \rho(y, x_i). \tag{8}$$

On the other hand, for object s x_i and x^* it is true that $\varepsilon_{min} \le \rho(x_i, x^*) \le \rho(y, x^*) + \rho(y, x_i)$ from where

$$\varepsilon_{min} - \rho(y, x^*) \le \rho(y, x_i). \tag{9}$$

Inequalities (8) and (9) are the low bounds evaluations $\rho(x, x_i)$, and for narrowing the inequality conditions we chose the maximal value. It also should be considered that both evaluations can simultaneously become negative what is reflected in formula (7). Finally, evaluation of the upper bound $\rho(y, x_i)$ directly follows from the triangular inequality (2) and condition $\rho(x^*, x_i) \le \varepsilon_{max}$.

Let an object x_i be situated 'closer' to the pivot x_1^* than to x_2^*, i.e.

$$\rho(x_i, x_1^*) \le \rho(x_i, x_2^*). \tag{10}$$

Then the following inequality takes place

$$max\left\{\left(\rho(y, x_1^*) - \rho(y, x_2^*)\right)/2,\ 0\right\} \le \rho(y, x_i). \tag{11}$$

Indeed, from the triangular inequality we have $\rho(y, x_1^*) \le \rho(y, x_i) + \rho(x_i, x_1^*)$, hence $\rho(y, x_1^*) - \rho(y, x_i) \le$ $\le \rho(x_i, x_1^*)$. Then $\rho(x_i, x_2^*) \le \rho(y, x_2^*) + \rho(y, x_i)$ is hold. Using condition (10), from the inequalities above we get $\rho(y, x_1^*) - \rho(y, x_i) \le \rho(y, x_2^*) + \rho(y, x_i)$ what with account of the expression $\rho(y, x_1^*) - \rho(y, x_2^*)/2$ possible negativity gives relationship (11).

Metric search models

Let us consider some approaches for (δ)- search support, which make premises for creating effective indexing system for reducing calculation operations on the search stage.

Suppose the distance matrix $d(\mathrm{X})$ is calculated on the pre-processing stage. We shall denote set X as X_0 and its cardinality $card(\mathrm{X}_0)=n_0$. On the initial search stage we calculate the distance $\rho(y,x^{(0)})$ between the searched object y and some object $x^{(0)}\in \mathrm{X}_0$ which is chosen arbitrary or using some criterion. If $\rho(y,x^{(0)})\le\delta$ we add object $x^{(0)}$ to a resulting set Y. From inequality (5) it directly follows that the distance $\rho(y,x_j^{(0)})$, $j=\overline{1,n_0}$, $x_j^{(0)}\ne x^{(0)}$ does not satisfy the search criterion δ if $|\rho(x_j^{(0)},x^{(0)})-\rho(y,x^{(0)})|>\delta$. Applying this criterion to all elements X_0 we get a set $\mathrm{X}_1=\{x_j^{(0)}\in \mathrm{X}_0:|\rho(x_j^{(0)},x^{(0)})-\rho(y,x_j^{(0)})|\le\delta\}\subseteq \mathrm{X}_0$. If $n_0>card(\mathrm{X}_1)=n_1$ we shall chose by analogy element $x^{(1)}\in \mathrm{X}_1$ and repeat the procedure of evaluating $\rho(y,x_j^{(1)})\le\delta$ and distance filtration for all $x_j^{(1)}\in \mathrm{X}_1$, $x_j^{(1)}\ne x^{(1)}$ getting in result $\mathrm{X}_2\subseteq \mathrm{X}_1$, $card(\mathrm{X}_2)=n_2$. The procedure is carried out recursively till step l when $n_{l-1}-1=card(X_l)$, $X_l\subseteq X_{l-1}$. In this case distances $\rho(x_j^{(l)},y)$, $j=\overline{1,n_l}$ are calculated and evaluated directly. Thus the matches number will be equal to $N(\delta)=l+n_l$ where $n_l=card(\mathrm{X}_l)$.

In practice storing distance matrix $d(\mathrm{X})$ 'in whole' is insufficient due to considerable preprocessing time and especially quadratic memory space requirements. One of the ways to solve this problem is to use its 'sparse' form where for some limited set of paired indices $\{k,l\}$, $0\le k,l\le n, k\ne l$ value $\rho(x_k,x_l)$ is calculated on the pre-processing stage. The natural demand of the methods choosing a set of given combinations is a compromise between storage expenses and number of distance calculating operations on the search stage, which should tend to $N(\delta)=l+n_l$. At the same time one should note that value $N(\delta)$ is a random one in a sense of being dependent on objects space configuration, pivots $x^{(1)},x^{(2)},\ldots,x^{(l)}$ choice order and location of the query object y. For instance $N(\delta)$ can be decreased if at first the point $x^{(2)}$ and then the point $x^{(1)}$ are chosen. Thus, indexing methods using the sparse form $d(\mathrm{X})$ (and, therefore operating is limited by information volume), theoretically can perform less matching operations than methods on 'complete' distance matrix $d(\mathrm{X})$.

We shall introduce a set of pivots $\mathrm{X}^*=\{x_1^*,x_2^*,...,x_k^*\}$. From (5) it follows that low distance value is $\rho(y,x_i)\ge\rho_{\mathrm{X}^*}(y,x_i)$ where $\rho_{\mathrm{X}^*}(y,x_i)=\max\limits_{x^*\in \mathrm{X}^*}|\rho(y,x^*)-\rho(x^*,x_i)|$. This is the simplest filtering method for the sparse distance matrix, where $d(\mathrm{X})$ after corresponding index rearrangement of indexes takes form

$$d(\mathrm{X})_{\mathrm{k}}^*=\begin{pmatrix}0&0&0&\ldots&\rho(x_1,x_{k+1})&\ldots&\rho(x_1,x_n)\\ &0&0&\ldots&\rho(x_2,x_{k+1})&\ldots&\rho(x_2,x_n)\\ &&0&\ldots&\ldots&\ldots&\ldots\\ &&&0&\rho(x_k,x_{k+1})&\ldots&\rho(x_k,x_n)\end{pmatrix}=\begin{pmatrix}\rho(x_1,x_{k+1})&\ldots&\rho(x_1,x_n)\\ \rho(x_2,x_{k+1})&\ldots&\rho(x_2,x_n)\\ \ldots&\ldots&\ldots\\ \rho(x_k,x_{k+1})&\ldots&\rho(x_k,x_n)\end{pmatrix}. \quad (12)$$

We will emphasize that we do not store distances between pivots in $d(\mathrm{X})_{\mathrm{k}}^*$ since $\rho(y,x_j^*),j=\overline{1,k}$ are calculated directly during the search. It also should be noted that the introduced approach can be interpreted as a mapping $(\mathrm{X},\rho)\to(\mathrm{R}^k,L_\infty)$ and search in k-dimensional space.

Another way of creating index structure without calculating and storing $d(\mathrm{X})$ 'in whole' is to analyze the structure of the data set on the base of distances between objects and then create a partition (possibly nested) of

$\mathrm{X} = \{\mathrm{X}^{(j)}\}$, $j = \overline{1,m}$, where

$$\left.\begin{array}{r} \forall j, j' \in \{1,\ldots,m\} : j \neq j' \Rightarrow \mathrm{X}^{(j)} \bigcap \mathrm{X}^{(j')} = \varnothing, \\ \mathrm{X}^{(1)} \bigcup \ldots \bigcup \mathrm{X}^{(m)} = \mathrm{X}, \end{array}\right\} \quad (13)$$

is fulfilled. Here the equivalence relation built on the base of function ρ can be exploited on the search stage: we do not consider those sets $\mathrm{X}^{(j)}$ which element $x_i^{(j)} \in \mathrm{X}^{(j)}$ is not equivalent to the query object y. Here important role plays determination of low and upper bounds (7), (12) $\rho(y, x_i)$ which allow to estimate the distance from y to the elements of $\mathrm{X}^{(j)}$, $j = \overline{1,m}$, separately or using distances to other sets $\mathrm{X}^{(j')}$, $j' = \overline{1,k}$, $j' \neq j$.

Let us consider another way of partitioning X. We shall choose pivot x_j^* which has index j in matrix $d(\mathrm{X})$, and determine the distance to all the rest of the objects $d(\mathrm{X})_{j,1}, d(\mathrm{X})_{j,2}, \ldots, d(\mathrm{X})_{j,n}$. We shall sort values of raw j in ascending order reassign indices $d^*(\mathrm{X})_{j,1}, d^*(\mathrm{X})_{j,2}, \ldots, d^*(\mathrm{X})_{j,n}$ and define the distance to the median object $\rho(x_j^*, d^*(\mathrm{X})_{j,k}) = \mathrm{M}$, $k = \lceil n/2 \rceil$. We shall introduce partition of X into two equivalence classes $\mathrm{X}_{\leq}$ and $\mathrm{X}_{>}$ where $\mathrm{X}_{\leq} = \{x_i \in \mathrm{X} : \rho(x_j^*, x_i) \leq \mathrm{M}\}$, $\mathrm{X}_{>} = \{x_i \in \mathrm{X} : \rho(x_j^*, x_i) > \mathrm{M}\}$. In this case on the search stage under $\rho(x_j^*, y) \leq \mathrm{M} - \delta$ it is necessary to search only class $X_{\leq}$ and under $\rho(x_j^*, y) > \mathrm{M} + \delta$ only class $\mathrm{X}_{>}$. Thus producing such a partition can allow us excluding from the consideration half of the set elements. But in the worse case when $\mathrm{M} - \delta < \rho(x_j^*, y) \leq \mathrm{M} + \delta$ is true, search algorithm has to consider both branches $\mathrm{X}_{\leq}$ and $\mathrm{X}_{>}$. It should be emphasized that this method is to be used recursively.

We shall introduce into consideration equivalence relation based on the closeness to the pivot. As before let us chose set $\mathrm{X}^* = \{x_1^*, x_2^*, \ldots, x_k^*\}$. Then elements x_j^* produce partition $\mathrm{X} = \{\mathrm{X}_s^*\}$, $s = \overline{1,k}$ such that

$$\mathrm{X}_s^* = \{x_i \in \mathrm{X} : \forall t = \overline{1,k}, t \neq s \; \rho(x_i, x_s^*) < \rho(x_i, x_t^*)\}. \quad (14)$$

Such criteria coincide with definition of Voronoi cell in Euclidean space. Partition introduced this way allows to use evaluation (11) of the low distance bound. Indeed, the criterion $\rho(x_i, x_s^*) \leq \rho(x_i, x_t^*)$ for $x_i \in \mathrm{X}_s^*$, $t \neq s$ is fulfilled by (14). Then from (11) for k pivots it follows that the low bound of $\rho(y, x_i)$ for $x_i \in \mathrm{X}_s^*$ will be

$$\rho_{min}^{(1)}(s) = max\Big\{ \max_{t=\overline{1,k}(t \neq s)} \{(\rho(y, x_s^*) - \rho(y, x_t^*)/2)\}, 0\Big\}. \quad (15)$$

Let $\varepsilon_{max}(s) = \max\limits_{x_i \in \mathrm{X}_t^*} \rho(x_s^*, x_i)$ be a cover radius for the partition X_s^*. Then according to (7) the evaluation low distance bound $\rho_{min}^{(2)}(s) = \rho(y, x_s^*) - \varepsilon_{max}(s)$ is true. Let minimal and maximal distance evaluations between partitions be calculated on the preprocessing stage for $s, t = \overline{1,k}$.

$$\varepsilon_{min}(s,t) = \min_{x_i \in \mathrm{X}_t^*} \rho(x_s^*, x_i), \; \varepsilon_{max}(s,t) = \max_{x_i \in \mathrm{X}_t^*} \rho(x_s^*, x_i)$$

Then $\varepsilon_{min}(s,t)$ and $\varepsilon_{max}(s,t)$ under $s \neq t$ are evaluations of ε_{min} and ε_{max} distance $\rho(x_s^*, x_t^*)$ in (7). For $s = t$ $\varepsilon_{max}(s,s) = \max\limits_{x_i \in \mathrm{X}_s^*} \rho(x_s^*, x_i) = \varepsilon_{max}(s)$. Hence it is legitimate to claim that evaluation $\rho_{min}^{(2)}(s)$ is a special case of evaluation $\rho_{min}^{(3)}(s) = max\{\max\limits_{t=\overline{1,k}} \{\rho_{min}^{(3)}(s,t)\}, 0\}$ where

$$\rho^{(3)}_{min}(s,t) = max\{\rho(y,x^*_s) - \varepsilon_{max}(s,t), \varepsilon_{min}(s,t) - \rho(y,x^*_s)\}. \quad (16)$$

The final maximal low bound of distance $\rho(y,x_i)$, $x_i \in X^*_s$ is defined as $\rho_{min}(s) = max\{\rho^{(1)}_{min}(s), \rho^{(3)}_{min}(s)\}$.

To make search algorithm on partition index structure more optimal, we propose to carry out the following steps during the search. Let E be a set of not processed pivots x^*_s which form corresponding regions X^*_s, $s = \overline{1,n}$, and T be a set of regions X^*_s which cannot be dropped by partition index structure. Then iteratively get the first pivot $x^*_s \in E$, calculate $\rho(x^*_s, y)$, estimate the low bound of all $\rho(x^*_t, y)$, $x^*_t \in E$ using (16) and remove from E those $x^*_t : \rho^3_{min}(s,t) > \delta$ or put corresponding to x^*_t set X^*_t to T. Also after each iteration remove considered pivot x^*_s from E, thus iterative procedure stops when some pivots $x^*_s, s = \overline{1,t}$ are eliminated by (16) and distance to the rest of pivots is calculated. After it we argue to apply (15) criteria for all pairs $x^*_s, x^*_t \in E$, $s \neq t$ since a range $(\varepsilon_{min}(s,t), \varepsilon_{max}(s,t))$ could be large and therefore could produce large low bound of $\rho(y,x_i)$, $x_i \in X^*_t$ or $x_i \in X^*_s$.

Results of experiments and conclusion

A number of tests have been performed on set of points in R^2 space with L_2 metric. We used two configurations of data distribution: uniform and with formed clusters. We implemented the following index structures to preprocess the initial data set: i) indexation on full matrix (1); ii) indexation on sparse matrix (12); iii) indexation via binary tree with branches $X_{\leq}$ and $X_{>}$; iv) indexation via compact partition using criteria (14) and iterative procedure which exploits (15) and (10) low bounds.

The purpose was to calculate the matches number during the search on uniform and clusters distribution of objects and its dependency on the query object position and data set configuration.

Below the results of the tests with the following experiment parameters are presented. The data for uniform distribution consisted of 1024 points, for the clusters one there were 16 clusters, with cluster cardinality mean equal to 64 and variance equal to 10 of (1024 elements in total). Both sets of points coordinated were within the range [0;256] (data square here and after). Variance of single cluster points location was set to 6 for both coordinates. It was allowed that clusters could overlap. We used $k = \sqrt{n}$ parameters for indexation ii), and created one-level partition with $k = \sqrt{n}$ number of pivots in indexation iv).

First experiment examined dependency of the index structure on the position of the query object. We generated uniform and cluster data structures, created all index structures and randomly chose query object from the data square 500 times, tracking the number of matches of each data structures for all queries. We then calculated the mean and variance of matches count for each index structure (Table 1).

Table 1. Dependency of the matches number on different data configurations

Data configuration	Full matrix		Sparse matrix		Partition search		Binary search	
	μ	σ	μ	σ	μ	σ	μ	σ
uniform	23.56	4.82	50.98	4.58	107.78	36.18	90.08	23.7
clusters	30.63	29.73	55.19	26.27	71.2	71.39	74.362	48.49

As it was expected indexation which exploits a full distance matrix i) outperforms the other ones while indexation on sparse matrix ii) keeps the second place. Indexations iii) and iv) have approximately equal efficiency. We can claim that when objects of database tend to form clusters the number of matches does not vary dramatically for all indexing methods, they only differ notably on variance and therefore in some cases performance of methods

iii) and iv) can take much more time than it was expected. On the other hand, when distribution of objects is uniform then indexation i) perform 4 times better results than iii) and iv) while indexation ii) performs 2 times better results.

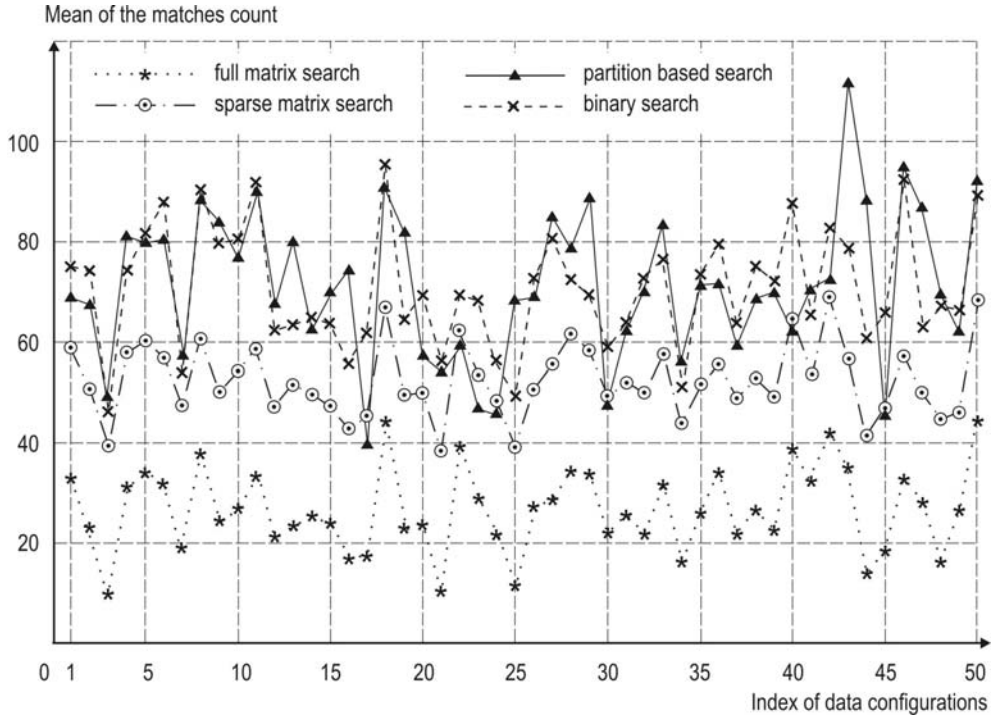


Figure 1. Results of matches count dependency on change of cluster data configuration

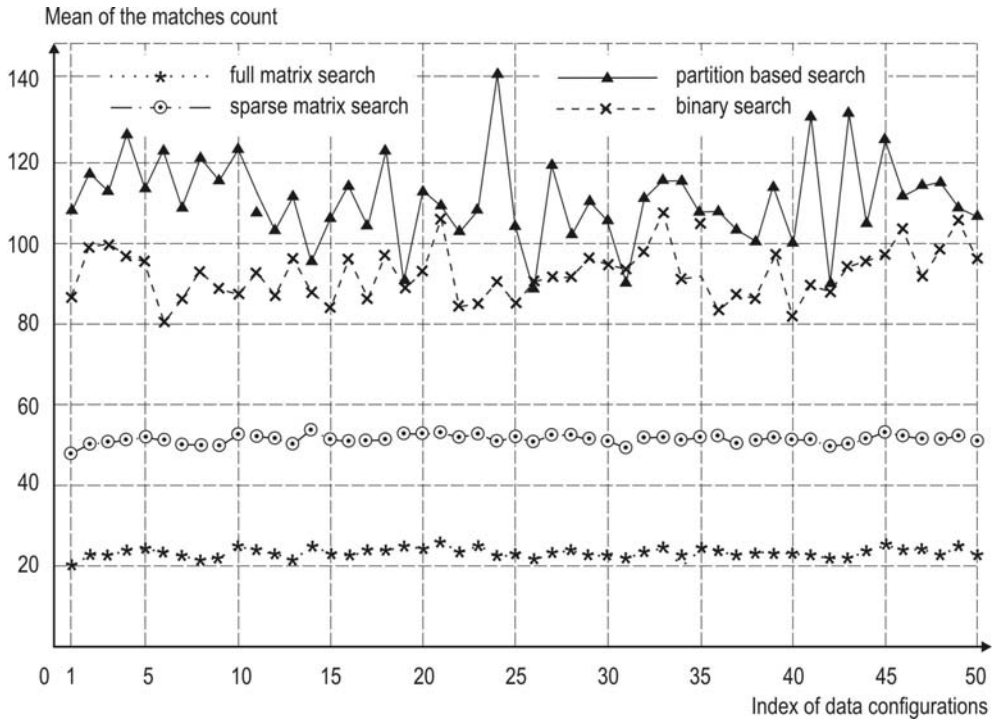


Figure 2. Results of matches count dependency on change of uniform data configuration

In the second experiment we tried to track dependency of index structures efficiency on different data configurations. We created 50 random configurations of uniform and cluster data distributions and ran routines of the first experiment changing query object position 20 times. We found out that statistics of the indexations efficiency was almost the same compared to first experiment results, only variance grew a bit. On Figures 1, 2 an average number of distance computations on 50 iterations for indexation algorithms i) – iv) is given for uniform and cluster data configuration. From here we can conclude that only index structures i) and ii) can guarantee almost/rather constant low bound of matches count on uniformly distributed data set solely.

We have proposed a novel indexing structure using sparse distance matrices for the image search with queries 'ad exemplum' which considering embedded partitions of the images. The experimental exploration of the method has proved it to be fast and efficient. The future work will be directed for investigation of the pivots choice method and possibility to use clustering methods for distance matrices analysis which would provide effective using of the partition metric for content image retrieval.

Bibliography

[Greenspan et al., 2004] H. Greenspan, G. Dvir, Y. Rubner. Context-Dependent Segmentation and Matching in Image Database. Computer Vision and Image Understanding. Vol. 93, No. 1, 2004, pp. 86-109.

[Yokoyama, Watanabe, 2007] T. Yokoyama, T. Watanabe. DR Image and Fractal Correlogram: A New Image Feature Representation Based on Fractal Codes and Its Application to Image Retrieval. In: Lecture Notes in Computer Science. Berlin Heidelberg: Springer-Verlag. Vol. 4351, 2007, pp. 428-439.

[Rubner et al., 2000] Y. Rubner, C. Tomasi, L.J. Guibas. The Earth Mover's Distance as a Metric for Image Retrieval. International Journal of Computer Vision. Vol. 40, No 2, 2000, pp. 99-121.

[Cheng et al., 2005] W. Cheng, D. Xu, Y. Jiang, C. Lang. Information Theoretic Metrics in Shot Boundary Detection. In: Knowledge-Based Intelligent Information and Engineering Systems. Khosla R., et al. (eds.). Lecture Notes in Computer Science, Springer-Verlag, Berlin Heidelberg. Vol. 3683, 2005, pp. 388-394.

[Wang et al., 2005] D. Wang, X. Ma, Y. Kim. Learning Pseudo Metric for Intelligent Multimedia Data Classification and Retrieval. Journal of Intelligent Manufacturing. Vol. 16, No 6, 2005, pp. 575–586.

[Kinoshenko et al., 2007] D. Kinoshenko, V. Mashtalir, V. Shlyakhov A Partition Metric for Clustering Features Analysis // International Journal 'Information Theories and Applications'. Vol. 14, No 3, 2007, pp. 230-236.

[Bohm et al., 2001] C., Berchtold S., Keim D. Searching in High-Dimensional Spaces: Index Structures for Improving the Performance of Multimedia Databases. ACM Computing Surveys. Vol. 33, No. 3, 2001, pp. 322–373.

[Chavez et al., 2001] Chavez E., Navarro G., Baeza-Yates R., Marroquin J.L. Searching in Metric Spaces. ACM Computing Surveys. Vol. 3, No. 3, 2001, pp. 273–321.

[Hjaltason, Samet, 2003] Hjaltason G., Samet H. Index-driven Similarity Search in Metric Spaces. ACM Transactions on Database Systems. Vol. 8, No. 4, 2003, pp. 517-580.

Authors' Information

Dmitry Kinoshenko *– Researcher; Kharkov National University of Radio Electronics, Lenin ave. 14, Kharkov 61166, Ukraine; e-mail: Kinoshenko@kture.kharkov.ua*

Vladimir Mashtalir *– Professor, Dean of Computer Science faculty; Kharkov National University of Radio Electronics, Lenin ave. 14, Kharkov 61166, Ukraine; e-mail: Mashtalir@kture.kharkov.ua*

Elena Yegorova *– Senior researcher; Kharkov National University of Radio Electronics, Lenin ave. 14, Kharkov 61166, Ukraine; e-mail: Yegorova@kture.kharkov.ua*

PROJECTIVE METHODS OF IMAGE RECOGNITION

Yevgeniy Putyatin, Vladimir Gorohovatsky, Alexey Gorohovatsky, Elena Peredriy

***Abstract**: We propose a method for image recognition on the base of projections. Radon transform gives an opportunity to map image into space of its projections. Projection properties allow constructing informative features on the base of moments that can be successfully used for invariant recognition. Offered approach gives about 91-97% of correct recognition.*

***Keywords**: Radon transforms, invariant features, pattern recognition, image projections.*

***ACM Classification Keywords**: I.4.7 Feature Measurement - Invariants*

***Conference**: The paper is selected from XIVth International Conference "Knowledge-Dialogue-Solution" KDS 2008, Varna, Bulgaria, June-July 2008*

Introduction

The modern computer vision problems, such as pattern recognition and normalization [1-5] are actual and still unsolved problems. Additional difficulties arise when the recognized objects are subject to geometrical transformations and noises.

The primary tasks of pattern recognition at presence of geometrical distortions are related to formation of highly-informative invariant systems of attributes, choice of metrics for their comparison, and construction of decision-making criterion. The type of the metric frequently depends on properties, ranges of formed attributes values, and also on possible noise types. While solving the specific problems, the chosen decision criterion together with the system of attributes determines significantly parameters of speed and reliability of recognition. One of the effective ways to solve the recognition problem is the construction and analysis of attributes on the basis of set of image projections. The projections are formed by applying Radon transform (RT), as well as similar Hough transform, and trace transform, etc. [6-10].

The main advantages of projective transformations are the following, i.e., high self-descriptiveness, noise protection ratio, and possibility of realization in the real-time systems. Displaying the two-dimensional image as a combination of one-dimensional functions enables to simplify and accelerate the procedures of normalization and recognition.

Projective transforms

RT represents integral of brightness function $B(x, y)$ [4,8]:

$$R(p,\theta) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} B(x,y)\delta(p - x\cos\theta - y\sin\theta)dxdy, \quad (1)$$

where $\delta(.)$ is Dirac delta function, which determines belonging of image points to the straight line with parameters (p,θ) along which integration is carried out. Here $p \in P$ is the distance from the beginning of the coordinates up to the line of integration, $\theta \in \Theta$ is the angle between the line of integration and the X-axis, P, Θ are ranges of p, θ values respectively.

The spatial mapping which is carried out with RT is possible to write as:

$$R : L^2(R^2) \to L^2(L^2(R),[0,2\pi]),$$

where R, R^2, L^2 are classical spaces of functions [10].

As a result of the discrete transformation the image $B(x, y)$ is put in correspondence with the image $R(p, \theta)$ from space of projections. $R(p, \theta)$ is the matrix of size $P \times \Theta$ determined by parameters (p, θ).

Hough transform is close enough to RT and in discrete variant is its special case [8]. Originally it was developed for curve figures identification on the image and is especially effective if the image contains a few points. The main difference between Radon and Hough transforms for direct line model is that RT puts in correspondence to each straight line on the image the fixed quantity of the points laying on it while Hough transform examines every possible straight line which passes through each point [8].

Trace transform is related to the generalizations of RT [9]. In it, similar to RT, functionals are calculated along the lines. Thus RT is reduced to calculation of the fixed functional – integral, while in trace transform any functional from the image is applied. Various of functionals enable to construct variety of the trace transforms, underlining some or other properties of the object on the image. In view of identical general view of trace functionals transition from one transform to another can be carried out rather easily [9].

Invariant features on the base of the Radon transform

Properties of projective transforms allow to construct system of features which are invariant to the number of geometrical transforms. Moment invariants are convenient and reliable system of features for visual objects recognition [1,5]. By processing one-dimensional projections on the basis of the central moments it is possible to construct system of invariant attributes for each projection.

Let's write down expressions for the classical m_k and central moments μ_k for the one-dimensional functions received on basis of RT:

$$m_k = \int_P R(p, \theta) p^k dp \ , \ \mu_k = \int_P R(p, \theta)(p - \frac{m_1}{m_0})^k dp \ , \ k = 0, 1, 2, \ldots . \qquad (2)$$

In view of an estimation of influence of geometrical transforms of the initial image on its Radon mapping image, the following systems of invariant attributes is offered to use.

Invariants to moving and scale transforms of the image will be functions of the variable θ:

$$\gamma_k(\theta) = \frac{\int_P R(p, \theta)^{k+1} (p - \frac{m_1}{m_0})^k dp}{(\mu_0)^{k+1}} . \qquad (3)$$

Features, which are invariant to the rotating and moving transforms, are the following:

$$\xi_k = \int_\theta \int_P R(p, \theta)(p - \frac{m_1}{m_0})^k dp d\theta . \qquad (4)$$

For the most general transforms of metric group (includes scaling, rotation and moving) invariant features turn out by integration of functions γ_k on space Θ and look like:

$$\eta_k = \int_\theta \gamma_k(\theta) d\theta . \qquad (5)$$

Invariant comparison methods

It is possible to present the general approach for pattern recognition on the basis of the invariant attributes constructed with the use of projective transforms as follows. The input is the image which can be deformed by manipulation of metric group transforms (include rotation, scaling, moving). RT $R(p, \theta)$ calculates for it, in discrete realization it represents a matrix of projection values. For each projection the set of n invariant attributes

is calculated. The quantity of attributes and projections used in recognition as a whole is defined by the etalon set and required level of correct recognition probability.

Recognition is realized by comparison of invariant attributes values for each projection with the values of reference attributes from database.

Let $\omega_1 = \{\gamma_i(\theta)\}$, $\omega_0 = \{\gamma_i^0(\theta)\}$ be sets or vectors of invariant features, constructed on the basis of projections for the image and the etalon respectively. Number i of an attribute can be directly connected to the order of the moment k in (3) - (5). Comparison of sets ω_1, ω_0 can be constructed by comparison of values at the fixed size k as well as by comparison based on parameter θ at fixed k.

For invariant (3) as a function $\gamma_k(\theta)$ the distance in the space of attributes on the basis of the popular metrics of the module of differences can be calculated as

$$\rho(\omega_1, \omega_0) = \sum_k \sum_\theta |\gamma_k(\theta) - \gamma_k^0(\theta)|. \quad (6)$$

The criterion of recognition is implemented by minimization of value (6) on the set of etalons $\Omega_0 = \{\omega_0\}$.

One of the variants of recognition by invariant attributes is also the estimation of deviations of calculated invariant features from etalon values, for example, on the basis of average and mean square deviation. Even such simple rules as «three sigma» yield encouraging results [3].

Recognition procedure with the use of invariant attributes is submitted on Fig. 1.

Figure 1 – Recognition procedure on the base of RT invariant features

Voting methods

Other important opportunity provided by the application of RT in consequence of its independence of calculation of its values for separate values of parameters p, θ is an independent decision-making for each of used projections. The recognition criterion thus is reduced to optimization on set Ω_0 of more simple values, than (6):

$$\rho(\omega_1, \omega_0, \theta) = \sum_k |\gamma_k(\theta) - \gamma_k^0(\theta)|, \quad (7)$$

that enables to make steady decisions, for example, on the majority of local decisions of used projections that corresponds to the voting procedure [5]. We shall notice that there are practical classes of images, whose recognition needs only one projection.

On application of the rule (7), for example, for 4 projections, natural way of acceptance of the reliable decision is agreement of the decisions at least for 3 of 4 projections. At the same time, decisions of separate projections can be not equal for concrete an application that allows carrying out recognition on separate, most important components of set of projections.

During the recognition using only one projection among all values of deviations the minimal size $\rho(\omega_1, \omega_0^i)$ gets out. The number i which corresponds to the minimum is a considered number of the etalon from the database.

While using several projections the result of comparison represents some set of numbers of etalons in which each of the used projections denotes. At ideal recognition all numbers of etalons should coincide among themselves. Recognition can be carried out also on a maximum of quantity of identical numbers of etalons. If there are some identical values of maxima, we can pick of them that one which has the least deviation among all

the other maxima from the etalon. The strictest rule is the choice of that etalon which simultaneously specifies all projections.

Informative properties of the spectrum

Recognition procedure can be constructed also directly on the basis of the analysis of spectrum $R(p,\theta)$ values. The most informative part of RT for some classes of images is the set of values of local maxima [7,9]. Such characteristics of maxima as value, quantity, site, relative positioning in the general structure and others allow constructing the system of attributes having sense of planimetric representation, displaying property of the form of initial visual object [6]. The circuit of recognition on the basis of comparison of RT spectrums is submitted in Fig. 2.

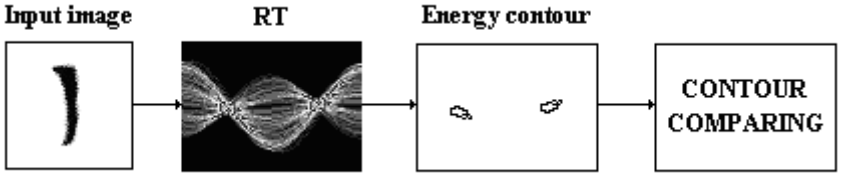


Figure 2 – Recognition procedure using local maximums of RT

Experiments and conclusion

Research of methods efficiency of recognition was carried out on several test sets. Sets of images of animals, hieroglyphs and examples of images of fingerprints are shown in Fig. 3. Entrance test images were formed from etalons by transforms of scale and movement. Recognition results are submitted in Table 1.

The type and complexity of recognition procedure depends on the considered problem and the prospective type of geometrical transforms. Our research have shown, that it is possible to reach high enough probability (is higher 0,95) at the specified transformations for complex images such as fingerprints (see Fig. 3) with the use of only 3-4 projections. The influence of rotation transformation yields the mixing of values for separate projections, and thus the required quality of recognition is achieved either by increasing the number of projections used for recognition or constructing the rotational invariants.

For the case of one-parametrical transformations (displacement within the limits of a vision field, or rotating about $\pm 15°$, or scale changes within the limits of 0,85-1,15) recognition of fingerprints is reached with probability 1 on the base of only one projection. With the presence of combinations of two listed transformations on three projections for images of fingerprints the probability of correct recognition has made about 0,95.

In some cases RT allows to narrow essentially the base of etalon images with a view of using any other alternative criterion of recognition. This purpose is achieved, for example, by cutting off the maximal deviation using some threshold. In such a way recognition procedure can be reduced to search not only one but also several etalons closest to the input image.

Table 1 – Comparative results of test sets recognition

№	*Test set*	*Amount of etalons /amount of experiments*	*Invariant features/ Amount of projections*	*Recognition level*
1	*Animals*	*10 / 200*	*6 features / 1*	*1,0*
2	*Animals*	*26 / 390*	*6 features / 1*	*0,9871*
3	*Hieroglyphs*	*30 / 450*	*6 features / 1*	*0,8688*
4	*Hieroglyphs*	*30 / 450*	*12 features / 2*	*0,9377*
5	*Fingerprints*	*17 / 255*	*6 features / 1*	*0,95*

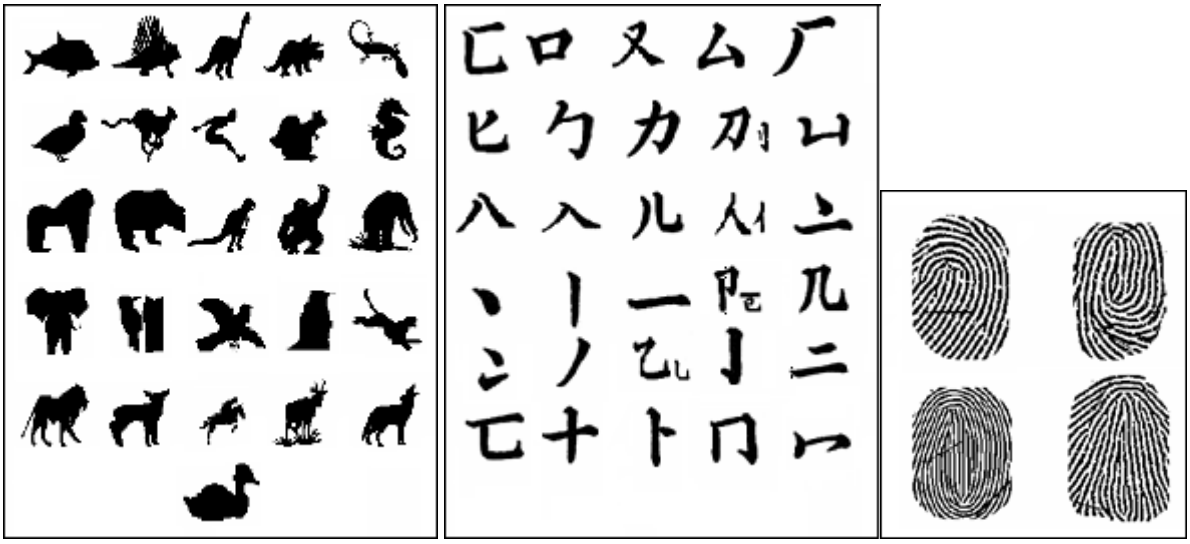


Figure 3 – Test images: animals, hieroglyphs, fingerprints

As one can see from the table, the probability is slightly decreasing with the increasing of etalons quantity in a database (Table 1, experiments №1-2). Besides the probability grows with the increasing quantity of the used projections (Table 1, experiments №3-4). For fingerprint images each of two projections (at $\theta = 0°$ and $\theta = 90°$) yields approximately identical results.

The suggested approach was tested also on the problem of hand-written signatures verification, examples of which are shown in Fig. 4. For this problem the algorithm with an estimation of mean square deviations and dispersions for set of invariant values of type (3) was used. The probability of correct verification was about 0,97.

Figure 4 – Test images of hand-written signatures: a – etalon genuine signatures; b – genuine signatures; c – forgery signatures

Experiments have shown, that time of calculation of the suggested invariant features (3) and, respectively, the time of recognition is essentially less, than for classical moment invariants. For example, computer modeling time of calculation of two-dimensional invariant attributes at the size of the image 256x256 pixels was about 1,3 sec. while the time of invariants (3) calculations was near 0,3 sec., which was four times less.

Projective transforms are one of the effective ways of image analysis and recognition. Properties of projections allow receiving informative features that can be successfully used for invariant recognition. Result of representation the two-dimensional image as the projections is reducing to one-dimensional space of features that allows reducing time of calculations essentially. Projective transforms can be successfully applied to the practical problems of recognition and identification of the complex visual data. The practical level of recognition is within the limits of 91-97 %.

The problem moments of application of projection methods are necessity of image segmentation concerning a difficult background, and also the maintenance of sufficient accuracy at performance of displaying working with discrete angles.

Bibliography

1. Путятин Е.П., Аверин С.И. Обработка изображений в робототехнике. Москва: Машиностроение, 1990. – 320 с.

2. Гороховатский А.В., Путятин Е.П. Применение преобразования Радона для нормализации изображений. // Бионика интеллекта, 2005. №1(62).- С.3-9.

3. Путятин Е.П., Гороховатский А.В. Построение инвариантных моментных признаков изображений с использованием одномерных проекций // Вестник НТУ «ХПИ». Системный анализ, управление и информационные технологии, 2006,.№39.- С.125-132.

4. Herman G.T. Image reconstruction from projections. Berlin, Springer,1980.

5. Shapiro L., Stockman G. Computer vision. Prentice Hall, 2001. -625 p.

6. Sytnik O. V., Gorohovatsky A. V. Signal processing algorithms in identification of subsurface objects. Radioelectronics and Communications Systems, Vol. 50, 2007, № 10, pp. 557-563.

7. Srisuk S., Petrou M., Kurutach W., Kadyrov A. Face authentication using the trace transform. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, 2003, pp. 305 – 312.

8. M. van Ginkel, C.L. Luengo Hendriks and L.J. van Vliet. A short introduction to the Radon and Hough transforms and how they relate to each other. Technical Report QI-2004-01, Quantitative Imaging Group, Delft University of Technology.- Delft, The Netherlands, 2004.

9. Kadyrov A., Petrou M. The Trace Transform and Its Applications.-IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.23, 2001, No.8, pp. 811-828.

10. A. Majumdar. Bangla basic character recognition using digital curvelet transform.-Journal of Pattern Recognition Research, Vol.2, 2007, No.1. pp. 17-26.

Authors' Information

Yevgeniy Putyatin *– Doctor of Technical Sciences, Professor of Informatics Department, Kharkiv National University of Radio Electronics, Lenin Ave., 14, Kharkiv, Ukraine, 61166.*

Vladimir Gorohovatsky *– Candidate of Technical Sciences, Informatics Department, Kharkiv National University of Radio Electronics, Lenin Ave., 14, Kharkiv, Ukraine, 61166.*

Alexey Gorohovatsky *– Ph.D. student of Informatics Department , Kharkiv National University of Radio Electronics, Lenin Ave., 14, Kharkiv, Ukraine, 61166, e-mail: gorohovatsky@rambler.ru*

Elena Peredriy *– Ph.D. student of Informatics Department, Kharkiv National University of Radio Electronics, Lenin Ave., 14, Kharkiv, Ukraine, 61166, e-mail: elena_peredriy@ukr.net*

ON A DIRECT APPROACH TO THE SOLUTION OF INVERSE OPTICAL PROBLEMS

Peter Sharlandjiev, Georgi Stoilov

***Abstract**: The evaluation from experimental data, of physical quantities, which enter into the electromagnetic Maxwell equations, is described as inverse optical problem. The functional relations between the dependent and independent variables are of transcendental character and numeric procedures for evaluation of the unknowns are largely used. Herein, we discuss a direct approach to the solution, illustrated by a specific example of determination of thin films optical constants from spectrophotometric data. New algorithm is proposed for the parameters evaluation, which does not need an initial guess of the unknowns and does not use iterative procedures. Thus we overcome the intrinsic deficiency of minimization techniques, such as gradient search methods, Simplex methods, etc. The price of it is a need of more computing power, but our algorithm is easily implemented in structures such as grid clusters. We show the advantages of this approach and its potential for generalization to other inverse optical problems.*

***Keywords**: Thin films; Materials and process characterization*

***ACM Classification Keywords**: J.2 Physical Sciences and Engineering (Physics)*

***Conference**: The paper is selected from Sixth International Conference on Information Research and Applications – i.Tech 2008, Varna, Bulgaria, June-July 2008*

Introduction

A 'direct problem' in mathematics and physics is referred as an evaluation of a dependent variable from a set of independent variables, which enter in some functional relations, usually partial differential equations. In optics, this direct problem can be defined as finding the response of an optical object to electromagnetic radiation, forming a solution of the Maxwell equations and their boundary conditions [1]. The so-called 'inverse optical problem' is the evaluation of the object characteristics from experimentally observed data. Herein, we focus on a specific inverse problem of great importance to fundamental and applied science, namely the determination of the complex refractive index (N = n – i*k) and physical thickness (D) of very thin films from spectrophotometric quantities, such as transmittance (T), front side reflectance (R) and backside reflectance (R').

In optics, for many cases of practical interest N is an unknown function of the radiation wavelength (λ). The functional relations between {n(λ), k(λ), D} and {T(λ), R(λ), R'(λ)} are of transcendental character, so inverse optical problems are nonlinear and they are solved by numeric optimization techniques, which have some disadvantages [2]. The derivative methods (Gauss – Seidel, steepest descent, etc.) are not reliable, because of their dependence on the initial guess and they fail when the numeric evaluation of the gradient is not effective. The Simplex or direct search methods are derivative free, but their dependence on the initial guess of the unknowns is very substantial. The principle difficulty for finding a robust and efficient solution to the inverse problem that we shall discuss below comes from the fact that for small D (when D/λ $\rightarrow$ 0) the optical response becomes less sensitive to {n, k} values. We propose a new algorithm for the direct numeric solution of the above optical problem and explore its potential for generalization and application to other inverse problems.

Algorithm and Computational Procedures

We consider a very thin homogeneous film with wavelength dependence of the complex refractive index in the visible spectral range. Spectrophotometric experiments are simulated numerically at normal incidence of light. That is, measurable quantities {T(λ), R(λ), R'(λ)} are calculated for specific parametric values {n(λ), k(λ), D} by the help of the Abeles characteristic matrix, which gives an exact solution of the Maxwell differential equations [1].

To those 'true' measurable values, experimental noise is added. We assume random experimental uncertainty with zero mean and variance, equivalent to that of high precision equipment, i. e. Cary 5E (Varian corp.).

Once the 'experimental data' is obtained we proceed to the evaluation of the parameters {n(λ), k(λ), D} at each wavelength, meaning that the evaluation is wavelength orientated and independent one from each other. The criteria for successful evaluation are that the three parameters under estimation must result in measurable quantities that satisfy simultaneously the uncertainty conditions on the observables. First, we define the range in the parametric space, where solutions will be searched for. This can be done on the basis of some preliminary information of general character, say whether the film material is a metal, dielectric or semiconductor. The range is increased by 30 – 50 % of the minimum to maximum expected values for each parameter. Second, a mesh grid is defined, where the step for a change of each variable must still have a physical meaning, simulating a quasi continuous change in the measurable quantities. These two inputs to the calculation procedure are not relevant for the numeric approach; they just limit the number of necessary computer evaluations. Third, for each point of the parametric mesh, direct calculations of the measurable quantities are done. Those calculations are rigorous, no approximations are maid, and accuracy is limited by the machine precision. Next, all values of the unknown parameters that give observables, which simultaneously fulfill the uncertainty conditions on the measured quantities, are accepted and stored. In other word, we keep all {n(λ), k(λ), D} from the parametric space that result in residuals that fall in the experimental joint confidence region in the observation space {T(λ), R(λ), R'(λ)}. This procedure is computer time consuming but that is the price for making no linearization, neither on the goal function, nor on the model function of the dependent variable.

Last step in the algorithm is to remove the 'outliers' from the sample parametric data and make a robust estimate of the unknowns. Here the main difficulty is that these outliers cannot be rejected as due to random errors. Besides they do not have some pre-determined statistical distribution, i.e. they are not just extreme values of random variables that occur naturally but infrequently [2]. To deal with that problem we have incorporated two independent and parallel decision making subroutines, referred bellow as Fit1 and Fit2. The first one is related to the distribution-free tests. The median of the parametric sample vector is evaluated as an estimate of the ensemble mean. Then, by the help of interquartile range of values of the sample realization, the upper and the lower 25% of the data are eliminated. Keeping the difference of 75% to 25% percentiles, we obtain an estimate of the spread of the body of the parametric values. The second subroutine makes use of the mode function for the estimation of each of the parametric realizations. The mode function for a continuous probability distribution evaluates its density function. The two subroutines are different mathematical implementations of the intuitive picture that when we get closer to the 'true value' of an unknown parameter, the relative frequency of the parameter realization is asymptotically increased.

Model, Results and Discussion

We define a model of a thin homogeneous film with physical thickness of 15 nm. The wavelength dependence of the film complex refractive index is that of bulk gold in the range 400 – 800 nm. This is expensively studied case, because very thin gold films are largely used in nanotechnologies, and, on the other hand, the evaluation of the complex refractive index is difficult in the red part of the spectrum due to strong correlation of front and backside reflectance. 'True values' for {T(λ), R(λ), R'(λ)} are generated and 'experimental noise' is added. We defined a relative uncertainty in T(λ) as derived from random values with zero mean and standard deviation of 0.0025. Reflection measurements are less accurate, so we used a standard deviation for R(λ) and R'(λ) of 0.0050. Then, a mesh grid is defined in the parametric space. The range of n(λ) is between 0.1 and 1.7, the range of values for k(λ) is between1.5 and 5.5 and that of D is from 8 to 25 nm. The step in the D parameter is chosen to be 1 nm, the step in n(λ) is $5*10^{-4}$ and that for k(λ) is $1*10^{-4}$. The film is supported by a substrate of refractive index 1.5.

In Figure 1 we show results of the fit for D = 12 nm. The length of realization by Fit1 (see text above) is 14 estimations, and for Fit2 – 15 estimations of the parameters n and k out of needed realizations of 21 (that is the number of wavelength estimations). It is obvious that for certain λ the fit results in no acceptable solution. The

values of the refractive index are overestimated, in order to catch up with the underestimated physical thickness. The normalized norm of the fit for n(λ) is 0.0144 (Fit1) and 0.0145 (Fit2). The results for k(λ) is 0.115 (Fit1) and 0.117 (Fit2). We remind that evaluated observables {T(λ), R(λ), R'(λ)} with these parameters satisfy the uncertainty conditions.

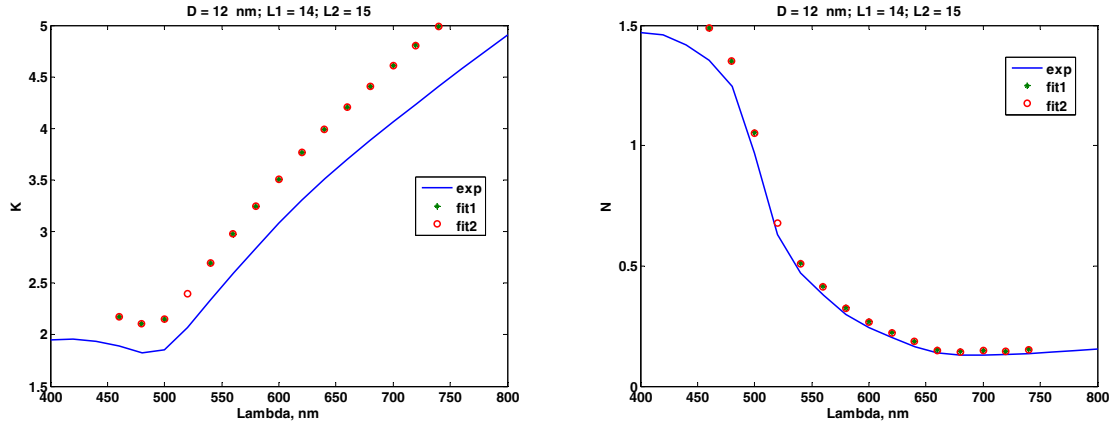


Figure 1. Spectral dependence of the extinction coefficient (a) and refractive index (b), D = 12 nm

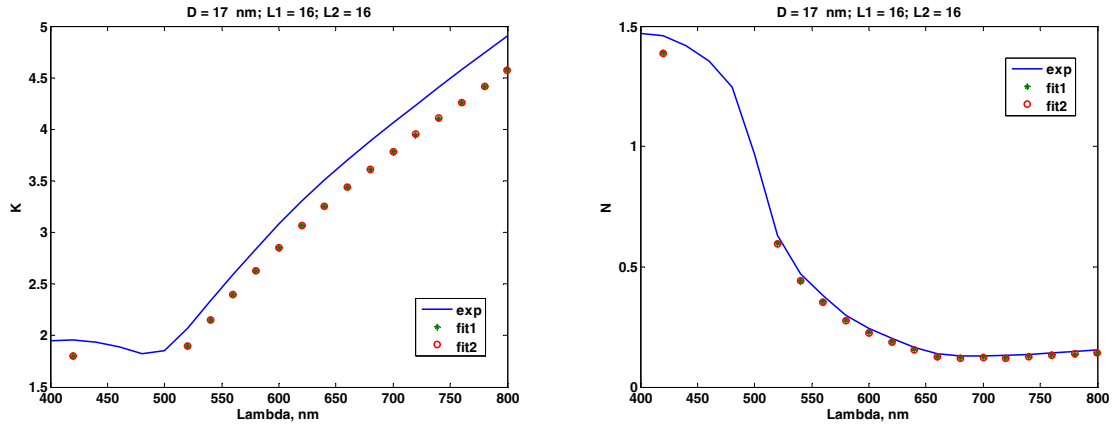


Figure 2. Spectral dependence of the extinction coefficient (a) and refractive index (b), D = 17 nm

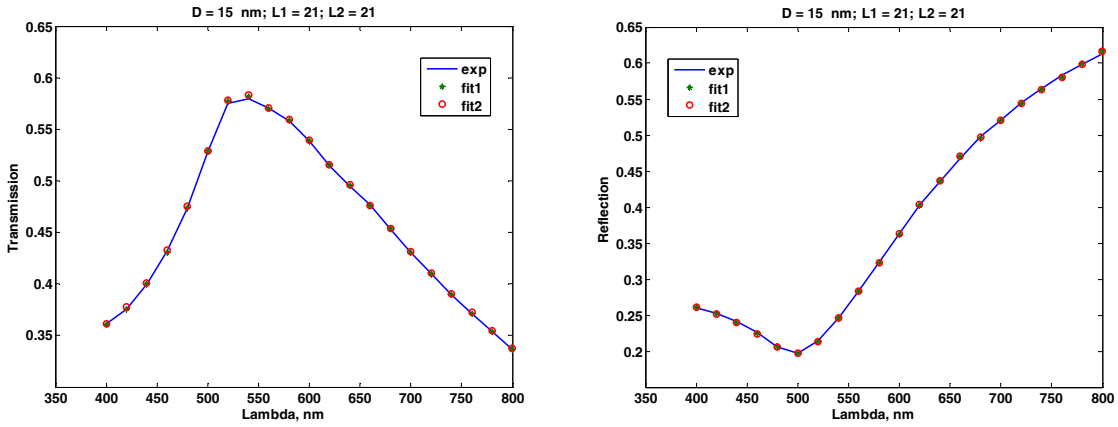


Figure 3. Spectral dependence of the transmission (a) and reflection (b), D = 15 nm

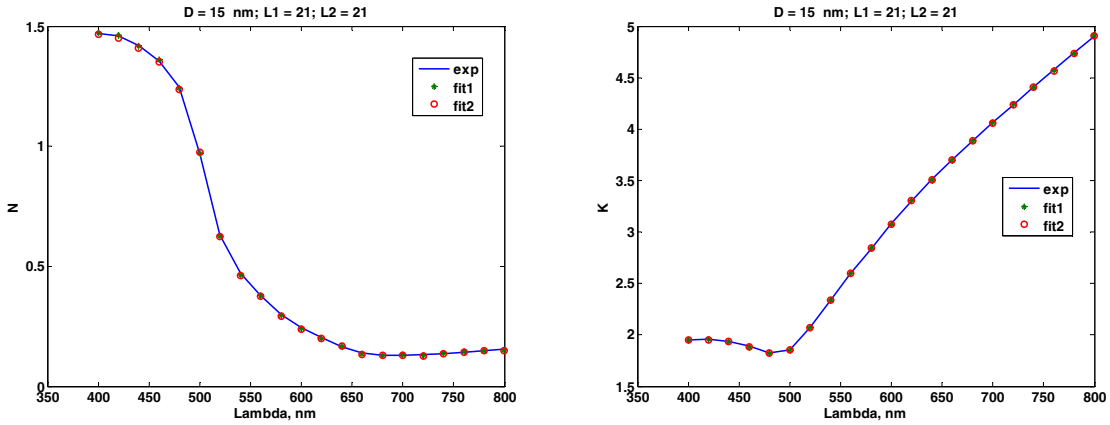


Figure 4. Spectral dependence of the extinction coefficient (a) and refractive index (b), D = 15 nm

In Figure 2 we show the results of a fit for D = 17 nm. The length of realization by Fit1 and Fit2 is 16 estimations, for n and k out of 21 needed realizations on λ. The values of the refractive index are underestimated, in order to catch up with the overestimated physical thickness. The normalized norm of the fit for n(λ) is 0.0059 (Fit1) and 0.0065 (Fit2). The results for k(λ) is 0.06632 (Fit1) and 0.06637 (Fit2).

Figure 3 presents the results for the fit at D = 15 nm. There the length of realization is at maximum and the norms of the estimations are much less than the ones described above. More, the choice of the best physical thickness estimate can be done on considerations on the length of parameter vectors. The normalized norm of the fit for n(λ) is 0.0007 (Fit1) and 0.001 (Fit2). The results for k(λ) is 0.0005 (Fit1) and 0.0007 (Fit2). The norms of the fit on {T(λ), R(λ), R'(λ)} are below 0.03%.

Conclusion

New algorithm is proposed for the solution of inverse optical problems. The unknown parameters evaluation does not need an initial guess and it does not use iterative procedures. Besides, it is derivative free. Thus we overcome the intrinsic deficiency of minimization techniques, such as gradient search methods, Simplex methods, etc. The price of it is a need of more computing power, but our algorithm is easily implemented in structures such as grid clusters. In fact, the simplest form of parallel computing - fractional computing, is neccessary. We have shown with numeric simulations that this approach can be successfully applied to processing of spectrophotometric data. Another advantage is that it has a potential for generalization for other inverse optical problems, related to ellipsometric measurements, FTIR measurements, etc.

Acknowledgments

This work was partially supported by the National Science Foundation at the Ministry of Education of Bulgaria by grant D01-377/2006.

Bibliography

1. [Born, 1957] E. Born and P. Wolf, Principles of Optics, Ed. Princeton, New York, 1957.
2. [Himmelblau, 1970] D. Himmelblau, Process analysis by statistical methods, Ed. John Wiley & Sons, New York, 1970.

Authors' Information

Peter Sharlandjiev – *Central Laboratory of Optical Storage and Processing of Information, BAS, Acad.G.Bontchev St. bl. 101, Sofia-1113, Bulgaria; e-mail: pete@optics.bas.bg .*

Georgi Stoilov – *Institute of Mechanics, BAS, Acad.G.Bontchev St. bl. 4, Sofia-1113, Bulgaria; e-mail: gstoilov@imbm.bas.bg.*

NORMAL ECG RECOGNITION FOR EXPRESS-DIAGNOSTICS BASED ON SCALE-SPACE REPRESENTATION AND DYNAMIC MATCHING

Nataliya Bilous, Michael Bondarenko, Gleb Kobzar, Alexey Krasov, Artyom Rogozyanov

***Abstract**: A novel approach of normal ECG recognition based on scale-space signal representation is proposed. The approach utilizes curvature scale-space signal representation used to match visual objects shapes previously and dynamic programming algorithm for matching CSS representations of ECG signals. Extraction and matching processes are fast and experimental results show that the approach is quite robust for preliminary normal ECG recognition.*

***Keywords**: electrocardiogram, express-diagnostics, curvature scale-space, dynamic programming, dynamic time wrapping.*

***ACM Classification Keywords**: I.5 Pattern Recognition.*

***Conference**: The paper is selected from International Conference "Intelligent Information and Engineering Systems" INFOS 2008, Varna, Bulgaria, June-July 2008*

Introduction

Automatic diagnostics of cardiac diseases is the traditional task of medical cybernetics. At current large experience is accumulated in this area. In particular attempts of complete automation of heart activity diagnostics have failed to be enough robust comparatively to traditional human diagnostics. The most essential stage of traditional ECG analysis is recognition of its major elements, which consist of finding out the QRS-complex, selection of it's characteristic points (tops of Q, R, S indents, scopes of indents and their borders).

The results of element analysis (annotation) and measuring of their parameters are used for ECG interpretation with purpose of correct diagnosis. Two basic categories of algorithms, applied in different systems of automatic diagnostics currently exist. Algorithms that design logic of doctor-diagnostician fall into the first category. The algorithms of the second category as a rule are based on the methods of multidimensional statistical analysis and theory of possibility.

Indisputable advantage of medical algorithms - in possibility of their rapid realization. It is conditioned by the fact that they concentrate experience of diagnostics accumulated in medicine and do not require preliminary teaching. Diagnostic possibilities of such algorithms are limited to the modern level of medicine development and quality of concrete algorithms – to the competence of persons assembling them as technicians and medics.

Advantage of unmedical diagnostic algorithms is that they can utilize any parameters of electrocardiogram representation. Due to it backlogs of information which in clinical practice remain unutilized appear accessible for such algorithms. The lack of these algorithms is complication of teaching. In case substantial difficulties are related to the selection of the well probed patients with diseases which an automat must learn to distinguish. Nevertheless, the algorithms of the second category are considered more perspective, because access to the new information should make diagnostics more effective.

Various automatic algorithms of the second group have been proposed, such as the threshold-crossing intervals X and the auto-correlation function [1] and algorithms based on neural-networks [2]. Time-frequency (t — f) analysis [3] and wavelet analysis [4] have also been used. All these algorithms suppose finding one special disease or detecting that ECG signal could be classified as normal.

This article focuses on building express-diagnostics system for ECG analysis that is to be performed at home or in non-medical environments by professional doctors or even users having no medical knowledge. The diagnostic result s of such system corresponds to three possible binary decisions: normal ECG, disease, not known.

We propose a novel approach of normal ECG recognition which is based on scale-scale signal representation used for geometric object shape recognition previously - curvature scale-space (CSS) and our dynamic programming algorithm for matching ECG signal represented as CSS descriptors. Though the idea of scale-scale signal representation for ECG analysis is far not new [5] including the fact that wavelets are multi-scale by nature the approach proposed may have some valuable advantages over existing scale-scale approaches.

The article is organized in 4 main sections. The first section describes CSS briefly as it's not a well-known technology in medical signal processing. The second section introduces how to apply CSS to ECG signals. The third section focuses on dynamic programming matching algorithm proposed. And the last chapter describes some tests performed.

Curvature Scale Space representation of shape

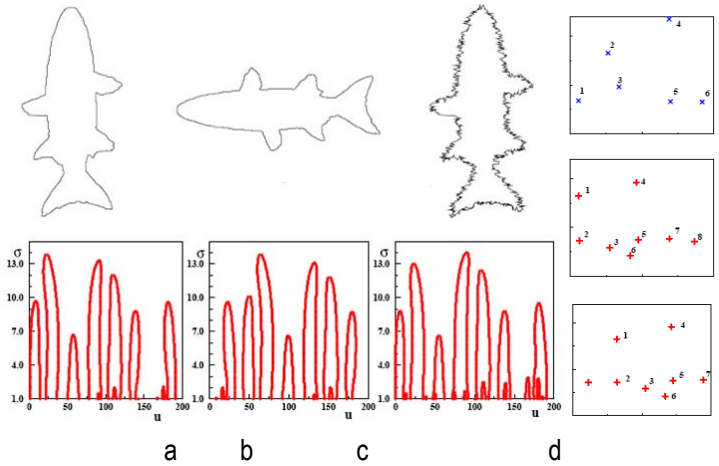
The Theory of Scale-Scale Signal Representation was introduced by Vitkin and Coendric in 1983. The methodology consist in embedding a measured signal into a one-parametric family of derived signals, the scale-space, where the parameter, denotes scale parameter $\sigma \in \Re_+$, is intended to the current level of scale. For a signal $f : \Re \to \Re$, the scale-space representation $L : \Re \times \Re_+ \to \Re$ is defined as [6]:

$$L(x;0) = f(x). \qquad (1)$$

And representation at coarser scales are given by convolution of the given signal with Gaussian kernels of successively increasing width:

$$L(x,\sigma) = g(x,\sigma) * f(x). \qquad (2)$$

Curvature Scale-Space was introduced later by Mokharitan for geometric object shape representation as follows [7]. Having curvature of each closed contour $L(x,\sigma)$ points calculated, curvature zero-crossing points can be found easily. Then Curvature Scale-Space is built by locating zero-crossings in (u,σ) space (u - normalized arc length, σ - Gaussian filter [7]. The resulting CSS can be represented as a binary image of CSS (fig.1). Cross-sections of CSS by horizontal lines define position of zero-crossing points on the corresponding $L(x,\sigma)$ contour curve.



a b c d

Fig. 1. Contour curves and corresponding CSS Images: а – contour, б – it's rotation, в – noise effect

It is obvious that CSS images of normalized curves are invariant to affine transformations and noise. Rotation of an object causes circular shift of it's CSS representation (fig.1a,b). The same effect is caused by the change of

contour starting point. Because of normalization scaling also don't affect the view of CSS. Picture 1 (c) also shows that noise effects in appearance of small arcs at the low levels of σ but don't affect main arcs.

These properties of CSS image are used for effective representation and recognition of object shapes. It should be also noted that successful identification of shapes based on CSS representation don't require CSS Images themselves but arc maximums only [7]. The set of CSS maximas consists of pairs and form well-known CSS description which was selected as one of the main shape descriptors for MPEG-7 standard.

Application of CSS to ECG Representation

Curvature Scale-Space representation and matching process was introduced for closed curve contours only and are not suitable for signal analysis. We will show that it is possible to adapt CSS for ECG representation in this section and ECG matching based on such representation in the next section.

As curvature scale-space methodology has it's root's in multi-scale signal representation area switching back from closed contours to signal segments representation is very easy. Signal curvature zero-crossing points may be acquired in the same way calculating the first and the second discrete differences (fig. 2).

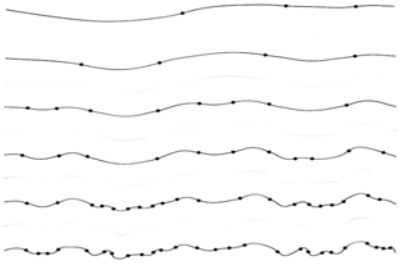


Fig. 2. Evolution of the signal and appropriate curvature zero-crossing points.

CSS images of such zero-crossings found on each scale of ECG signal evolving will look the same as for geometric object shape contour (fig.1). The only difference here is in representation of CSS image maxima which affects matching process also. As closed contour may be started to detect from different points CSS image is also circular. In a difference an ECG signal starts and finishes at certain points. So CSS image for ECG and the set of maxima as well is not be shifted during acquisition and matching process.

Matching CSS Representations of ECG

Let's first analyze CSS matching algorithm for closed contour curves introduced by Mokhtarian [7]. As mentioned before, every object in the database is represented by the locations of the maxima of it's CSS images. Matching algorithm compares two sets of maxima and assigns a matching value to them which represents the similarity between the actual boundaries of objects. The first step in CSS matching is to shift one of the two sets of maxima so that the effect of randomly selected starting points is compensated. Since the exact value of required shift is not available, we choose several values for it and then find the best match among them. The vest choice is a value that shifts one CSS image so that its major maximum covers the major maximum of the other CSS image. Other possible choices are those values which accomplish the same with the second and possibly the third major maxima. The matching value will be the summation of the straight line distances between the matched pairs plus the vertical coordinates of the unmatched maxima.

Obviously as the nature of curves differs for contours and signals matching algorithm of CSS representations introduced by Mokhtarian with it's main principle to shift maximums is not suitable for CSS representations of ECG signals. Moreover the algorithm described will fail to match curves with large dissimilarities of structure distances (heart rhythm change).

A different algorithm of matching CSS representations of ECG based on dynamic programming is proposed. It is organized in the same way as Dynamic Time Warping (DTW).

Having two sets of maxima $Q= q_1,q_2,...,q_i,...,q_n$ ($q_i = (\sigma_i, x_i)$, where σ is a scale parameter and x is position of the maxima) and $C=c_1,c_2,...,c_j,...,c_m$ ($c_i = (\sigma_i, x_i)$) of length n and m respectively matching of these sets using proposed DP matching algorithm requires to construct an n-by-m matrix (DP matrix) where the (i^{th}, j^{th}) element of the matrix contains the distance $d(q_i,c_j)$ between the two points q_i and c_j (in a metric defined below). Each matrix element (i,j) corresponds to the alignment between the points q_i and c_j. This is illustrated in Figure 3a. A warping path W, is a continuous (in the sense stated below) set of matrix elements that defines a mapping between Q and C. The k^{th} element of W is defined as $w_k = (i,j)_k$ so we have:

$$W = w_1, w_2, \ldots, w_k, \ldots, w_K \quad \ldots\ldots \max(m,n) , K < \qquad (3)$$

The warping path is typically subject to several constraints:

- Boundary conditions: $w_1 = (1,1)$ and $w_K = (m,n)$, simply stated, this requires the warping path to start and finish in diagonally opposite corner cells of the matrix.
- Continuity: Given $w_k = (a,b)$ then $w_{k-1} = (a',b')$ where $a–a'<=1$ and $b-b'<=1$. This restricts the allowable steps in the warping path to adjacent cells (including diagonally adjacent cells).
- Monotonicity: Given $w_k = (a,b)$ then $w_{k-1} = (a',b')$ where $a–a'>= 0$ and $b-b'>= 0$. This forces the points in W to be monotonically spaced in time.

There are exponentially many warping paths that satisfy the above conditions, however we are interested only in the path which minimizes the warping cost:

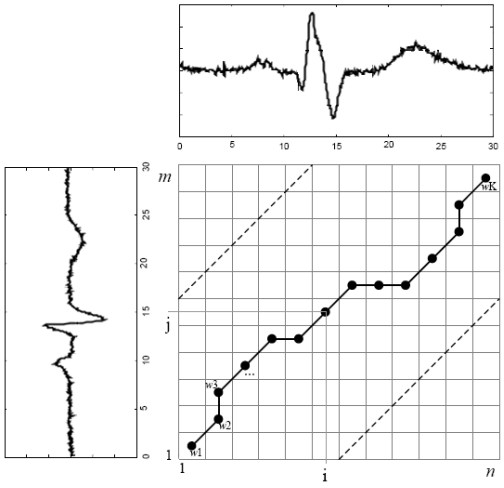


Fig. 3. Dynamic matching of two ECG signals

$$Cost(Q,C) = \min\left\{\sqrt{\sum_{k=1}^{K} w_k} / K \right. \qquad (4)$$

This path can be found very efficiently using dynamic programming to evaluate the following recurrence which defines the cumulative distance $\gamma(i,j)$ as the distance $d(i,j)$ found in the current cell and the minimum of the cumulative distances of the adjacent elements:

$$\gamma(i,j) = d(q_i,c_j) + \min\{ \gamma(i-1,j-1) , \gamma(i-1,j) , \gamma(i,j-1) \}$$

The difference from DTW here is the way to calculate values $d(q_i,c_j)$ which is the basement of any DP algorithm. We propose to calculate $d(q_i,c_j)$ according to matching of two points or skipping the match which corresponds to diagonal, horizontal and vertical movements from w_{k-1} to w_k in DP matrix. Following formulas are used in different cases:

$$\begin{aligned} d_{match}(q_i,c_j) &= \left\| q_i - c_j \right\|, \\ d_{skipi}(q_i,c_j) &= \sigma_i, \\ d_{skipj}(q_i,c_j) &= \sigma_j, \end{aligned} \qquad (5)$$

where $d_{match}(q_i,c_j)$ - distance for matching two points or a diagonal movement,

$d_{skipi}(q_i,c_j)$ - distance for skipping the point q_i or a horizontal movement,

$d_{skipj}(q_i,c_j)$ - distance for skipping the point q_j or a vertical movement,

$\left\| q_i - c_j \right\| = \sqrt{(\sigma_i - \sigma_j)^2 + (x_i - x_j)^2}$.

Such formulation of $d(q_i,c_j)$ has already been proven in practical application of geometric shape matching as a part of Mokhtarian's algorithm.

Finally the cost of matching could be found in $\gamma(n,m)$ using backtracking algorithm while building the optimal path W.

Experimental Results

51 ECG signals was selected for testing from the international ECG database PhysyoNet (Physikalisch-Technische Bundesanstalt – PTB, the National Metrology Institute of Germany). This database was assembled by doctor Michael Oeff, M.D. (the Department of Cardiology of University Clinic Benjamin Franklin in Berlin, Germany) from ECG's of healthy volunteers and patients having different classes of heart diseases.

Table 1. Distribution of classes

Diagnostic class	Number of ECG signals
Myocardial infarction	5
Cardiomyopathy/Heart failure	8
Bundle branch block	5
Dysrhythmia	4
Myocardial hypertrophy	5
Valvular heart disease	5
Myocarditis	5
Miscellaneous	4
Healthy controls	10
Total	51

Techniques of CSS representation and dynamic matching were realized as a computer program which was tested on the assembled database of ECG signals.

As a result of testing the developed program on three takings (average cost for I, II and III taking) the table of likeness for all of ECG signals of the base was acquired against "normal" ECG signal depicted earlier. Setting the threshold of dissimilarity to $TP=119$ no unhealthy man ECG of was treated as "normal". And 1 of 10 healthy ECG signals was treated as "not normal". Following the well known technique of automatic ECG interpretation quality estimation [9] we assume that all database signals are verified (normal forms and pathological forms). Then counting up the general number of the followings events: correct classification of normal ECG (TN), improper classification of normal ECG as pathological (FP), improper classification of pathological ECG as normal (FN), correct classification of pathological ECG (TP).

In total, quality of ECG interpretation is calculated as "probability that classification is correct" according to the following formula:

$$TA = \frac{TP + TN}{TP + FP + TN + FN} \ 100\% \qquad (6)$$

Thus according to experiment results $TP=41,\ TN=9,\ FP=0,\ FN=1$ and resulting $TA=98\%$. After introduction of the third diagnostic result – "not known" $FN=0$ may be achieved. In the case if a user received "not known" result he can repeat the measurements. And if the same result achieved again it would be recommended to consult with professional doctor-cardiologist.

Conclusion

A novel approach of normal ECG recognition based on scale-space signal representation is proposed. The approach uses curvature scale-space signal representation used to match visual objects shapes previously and dynamic programming algorithm for matching CSS representations of ECG signals. The main advantage of the approach over the existing scale-space representations (ex. wavelet based) and correlation methods (including direct dynamic and derivative dynamic time warping considered in the article) is faster extraction and matching process as it is done in feature space.

Experimental results show that the approach is quite robust for preliminary normal ECG recognition. It is planned to conduct larger tests on different ECG databases and comparative tests to compare the quality of diagnostics and the rapidness with other interpretation techniques.

Bibliography

[1] R.H. Clayton, A. Murray, R.W.F. Campbell, Comparison of four techniques for recognition of ventricular fibrillation of the surface ECG, Med. Biol. Eng. Comp. 31 (1993), pp. 111—117.

[2] K. Minami, H. Nakajima, T. Toyoshima, Real-time discrimination of ventricular tachyarrhythmia with Fourier transform neural network, IEEE Trans. Biomed. Eng. 46 (2) (1999), pp. 179—185.

[3] V.X. Afonso, W.J. Tompkins, Detecting ventricular fibrillation, IEEE Eng. Med. Biol. 14 (1995) pp. 152—159.

[4] A.S. Al-Fahoum, I. Howitt, Combined wavelet transformation and radial basis neural networks for classifying life threatening cardiac arrhythmias, Med. Biol. Eng. Comput. 37 (1999), pp. 566—573.

[5] F. Jager; I. Koren; L. Gyergyek Multiresolution representation and analysis of ECG waveforms, Computers in Cardiology Proceedings. Volume 23-26 (1990), PP.:547 – 550

[6] T. Linderberg Scale Space Theory in Computer Vision, Kluwer Academic Publishers. (1994), 440p.

[7] F. Mokhtarian, S. Abbasi, J. Kittler Efficient and Robust Retrieval by Shape Content through Curvature Scale Space, Int. Workshop on Image DataBases and Multimedia Search (1996), pp. 35-42.

[8] H. Strik, L. Boves Averaging physiological signals with the use of a DTW algorithm. In Proceedings of SPEECH'88, 7th FASE Symposium, Edinburgh, Book 3 (1988), pp. 883-890.

[9] Medical electrical equipment. Part 3. Particular requirement for the essential perfomance of recording and analysing electrocardiographs, IEC, Geneva (1996), 75p.

Authors' Information

Nataliya Bilous *– Ph.D., head of ITLCVS laboratory, professor of Kharkov National University of Radio Electronics, Kharkov, Ukraine; e-mail: belous@kture.kharkov.ua*

Michael Bondarenko *– professor of software department, rector of Kharkov National University of Radio Electronics, president of Ukrainian association of distance learning, academician of the Ukrainian academy of science of higher school: belous@kture.kharkov.ua*

Gleb Kobzar *– Researcher of ITLCVS laboratory, Kharkov National University of Radio Electronics, Kharkov, Ukraine; e-mail: glebkobzar@yahoo.com*

Alexey Krasov *– Master of Software Department, Kharkov National University of Radio Electronics, Kharkov, Ukraine; e-mail: belous@kture.kharkov.ua*

Artyom Rogozyanov *– Master of Software Department, Kharkov National University of Radio Electronics, Kharkov, Ukraine; e-mail: belous@kture.kharkov.ua*

APPLICATION OF THE ARTIFICIAL INTELLIGENCE ALGORITHMS WITH A COGNITIVE GRAPHIC AS A TOOL FOR A SYSTEM ANALYSIS OF ELECTROPHYSIOLOGICAL PROCESSES

Nina Dmitrieva

Abstract: *Summarizing the accumulated experience for a long time in the polyparametric cognitive modeling of different physiological processes (electrocardiogram, electroencephalogram, electroreovasogram and others) and the development on this basis some diagnostics methods give ground for formulating a new methodology of the system analysis in biology. The gist of the methodology consists of parametrization of fractals of electrophysiological processes, matrix description of functional state of an object with a unified set of parameters, construction of the polyparametric cognitive geometric model with artificial intelligence algorithms. The geometry model enables to display the parameter relationships are adequate to requirements of the system approach. The objective character of the elements of the models and high degree of formalization which facilitate the use of the mathematical methods are advantages of these models. At the same time the geometric images are easily interpreted in physiological and clinical terms. The polyparametric modeling is an object oriented tool possessed advances functional facilities and some principal features.*

Keywords*: Life and Medical Sciences; Computational Geometry and Object Modeling.*

ACM Classification Keywords*: I.2 Artificial Intelligence I.2.1 Applications and Expert Systems H.4, J Medicine and science*

Conference*: The paper is selected from XIVth International Conference "Knowledge-Dialogue-Solution" KDS 2008, Varna, Bulgaria, June-July 2008*

Introduction

It is known, the existent methods of analysis of electrophysiological processes give little information about early dysfunctions and development of a principally new approach to an analysis of complicate oscillation processes is an actual task [Fedotchev and others.2003]. To decide this task the most perspective approach is a system approach with the use intelligence tools [Pospelov D.A., 1982; Zenkin A.A., 1991; Gorodeckiy V.I.,2002; Workshop on Autonomous Intelligent Systems,2007]. In this aspect the description of analog processes and their analysis are considered as a cell of classification and decision-making with management elements [Fominach 1998; Yachno and others, 2004].

Case-Based Reasoning

From our point of view the relationships of parameters of system have a more significant expression because they are determined in geometric model as the information connections [Dmitrieva N.V., 1990, 2005].WE have to say that the class such phenomena as information connections of physiological processes can not be observed by modern means of physiological experiments. The decision way consists of a polyparametric cognitive modeling of oscillation processes. On this methodological basis we have developed earlier some cognitive models for system analysis of electrocardiogram, electroencephalogram, electroreovasogram and also of functional state of the whole organism for medical diagnostics [Dmitrieva N.V., 1990,2003, 2005]. Such diagnostics have deal with a multitude of parameters of a system and it is polyparametric diagnostics. There are some summaries of the accumulated experience in development of cognitive models for the system analysis of electrophysiological processes.

Short description of a general structure algorithm of intelligence models for system analysis of electrophysiological processes.

The principal structure and function of cognitive models of different electrophysiological processes are unified therefore it is rational considering it as a general structure. For a model construction it is used a set of amplitude and time parameters of electrophysiological processes, which are equidimensional but have different scale. Similarity and nesting of parameters are observed in all electrophysiological systems. The objective parameters are only used in the absolute value, the models of separate electrophysiological processes are constructed on a different scale. This generates model variations, which are some set of models having a unified algorithm of the structure of an artificial intelligence system. Quality difference of such models consists not only of there scale but also of forming of a standard ("ideal") functional state of physiological system which is served for a relative estimation of real processes. For the construction of such a standard there are used parameter modes of every electrophysiological system (ECG, EEG, and others) and maximal and minimal values of parameters without pathognomonic meaning. The structure of model is very simple: parameter values (Mo, maximal and minimal) of individual oscillation processes are represented in two-axis system. Every leg has its own dimension. The totality of parameters is enclosed in the counter of right triangle. The legs of a triangle are vectors of time and amplitude parameters.

Thus it is created a range of parameter changing which have not meaning for clinic but have significance for an evaluation of functional state of physiological system and organism as a whole. The general counter limited maximal and minimal values (external and internal triangle counters) is an intelligence transformer, its work is based on a structure-algorithm of artificial intelligence system. The transformer makes permanent analysis of a character of behavior in the absolute values and classification (for example transfer of a patient in clinical diagnosis). The result is a visualized pattern and quantitative expression of intersystem connection of parameters, which are defined by geometry and symmetry of the model structure. Absolute values of a person's health are inter square limited by external and internal triangles, i.e. in significant range, as at the same time the relationships of parameters are rested near the invariant value [Dmitrieva N.V., 2004].

Summarizing the accumulated experience in the polyparametric cognitive modeling of different physiological processes (electrocardiogram, electroencephalogram, electroreovasogram and others) and developing on this basis of some diagnostic methods (9 patents in Russia) led us to formulating a methodology of polyparametric cognitive modeling. The gist of the methodology is in marking of fractals of electrophysiological processes, the paramerization of its, description of functional state of an object with a unified set of parameters; construction of the cognitive geometric model of fractal with artificial intelligence system algorithms, using of methods of image recognitions for the analysis of multi-dimension data. Such models present the instrument to analyze multidimensional data, performing knowledge engineering and data mining. The geometry model enables displaying relationships adequacy to require of the system approach. Principal singularities of this methodology: full objective elements of models (physical values of parameters); strict geometry structure of model high degree of formalization easing use of mathematical methods. At the same time the geometric images are easy interpreted in physiological and clinical terms making this model preferable [17, 18].

As an examples the polyparametric analysis of ECG and EEG on a basis of polyparametric cognitive modeling are presented.

The polyparametric intelligence model of electrocardiogram **(ECG)** (fig,1) is intended for a system analysis and preclinical diagnostics of early cardiodysfuntions. Elements of the model are commonly accepted amplitude and time parameters of ECG-cycle such as P, T, R, QRS, QT, RR, which are represented in two-axis system Methods of projective geometry and general principles of symmetry are used to define relationships between elements of the model (know-how of the method). The model gives possibility to get meaningful and useful relationships of parameters, which are new diagnostics signs, classify multidimensional data, and to interpret data in commonly used clinical and physiological terms. The model was implemented in a software package.

The model makes it possible to forecast preclinical changes in heart activity. The model was verified in clinical study with participation of 101 volunteers. Simultaneously independent diagnostics of cardiovascular system of the volunteers were performed by using clinical physiological analysis of ECG in 12 deviations [Dmitrieva, 1989]. This method of system analysis of ECG can be particularly useful for preventive medicine.

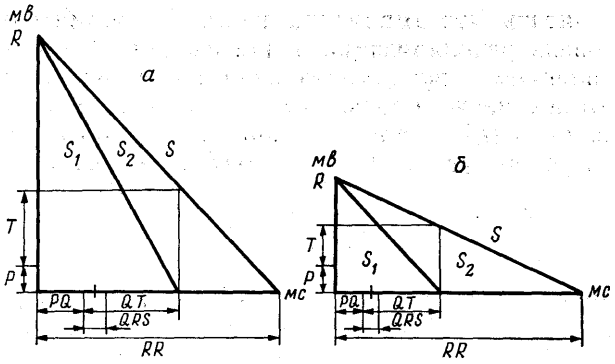


Fig.1.

The model of ECG of the health people (a) and patient with the hypertonic diseases (b). (R, T, P- amplitude, QRS, PQ, QT, time parameters, S_1 and S_2 –Systolic and Diastolic parts of cardio circle).

Short description of cognitive model of system analysis of rhythms of a human electroencephalogram.

Elements of model are commonly accepted amplitude and time parameters of a human electroencephalogram according to recommendation of the International Federation of Encephalography and Clinical Neurophysiology [1984]. Amplitude and time parameters are presented on legs of two right triangulates presenting left and right cerebral hemispheres. The totality of parameters of the whole brain are enclosed in counter of isosceles triangle with superposition of legs of the right triangles. Hypotenuses of every triangles are commonly accepted in clinic indexes of rhythms of each hemisphere expressed in per cent for 10 sec epoch. Each hemisphere has its own intelligence transformer of system realizing algorithm of functions of model and making classification. Relationships between elements of the model are defined with projective geometry and general principles of symmetry (know-how of the method). For standard model ("ideal functional state) and for health people these relationships are invariant (W≈1.309).

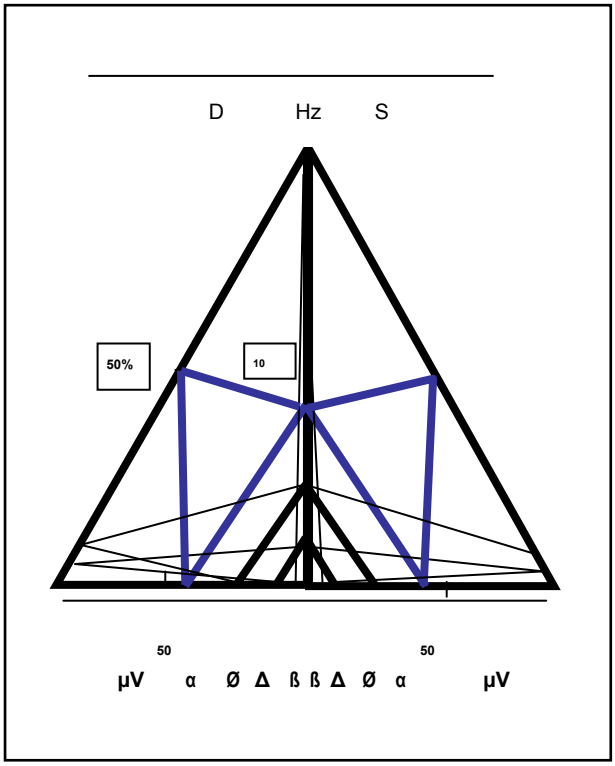


Fig. 2 The model of a structure of rhythms of EEG. (S- left hemisphere, D-right hemisphere of brain. Method of construction is in the text.).

This visualized model presents the strict characteristic of electrical processes and gives possibility to the following math development and at the same time it may be very easily interpreted in clinical- physiological terms.

Discussion

Evidently electrophysiological processes possess a fractal auto organization by principals of scale-invariant similarity and proportion of parameters is an order of ones. On the basis of such model the totality of bioelectrical generators can be rationality presented as a level-based model of hierarchy systems of connected nonlinear endogen electrical generators with different frequency of order presented in evolutionary coordinated code. It is new knowledge allowing presenting of a set of oscillation processes as a unified complicated system.

Development of the theory of informatics for the last years drew the attention researchers of living systems to comprehension of code connections of elements of systems [System. Harmony. Symmetry. 1988]. The special interest in research of regularity of code relationships in different systems presents a conception computer' cognitive graphic based on general principals of symmetry and methodology of artificial intelligence systems [Zenkin A.A., 1991]. Interpretation of our results on the basis of symmetry allows characterizing the maintenance of object on the whole and its changes, i.e. optimality and functional (code) balance of object [System. Harmony. Symmetry.1988]. Without doubt it is shown on a principal basis of functional order as a time organization of physiological processes.

Uniform schemes of description of patterns are open possibilities for using of an algebraically approach to their analysis [Gurevich and others, 1999], which is perspective for supporting diagnostics decisions.

Conclusion

The new polyparametric methodology for a system analysis of electrophysiological processes by modelling ones with application of artificial intelligence algorithms with general principals of geometry and symmetry is proposed. The polyparametric models possess system attribute, data mining and knowledge discovery exhibiting information connection between parameters of processes as elements of physiological systems. Models offered composition, classification properties, characterizing quality and quantity as the whole object as parts one.

Bibliography

[Gorodeckiy, 2002] Gorodeckiy V.I.., Samoylov V.V., Malov A.O. Modern state of technology of knowledge engineering from data base. J. News of artificial intellect. 2002. №4. P.3-9.

[Dmitrieva,1998] Dmitrieva N.V. The polyparametric functional-diagnostics method on basis of intelligence systems. // Proceeding of VI nationality conf. artificial intellect (CII-98).Pushino-on-Oka, 1998. P. 456–462

[Dmitrieva,2000] Dmitrieva N.V. The syndrome analisis of polyparametric patterns of functional states. J. News of artificial intellect 2000.№1.P.120-129

[Dmitrieva, 2004] Dmitrieva N.V. The electrophysiological mechanisms of adational processes :J. Human Physiology. 2004. № 3 P. 47-56.

[Dmitrieva, 2003] Dmitrieva N.V. The polyparametric systems as tool of a system analysis of electrophysiological processes. J. News of artificial intellect. 2003. № 1 P20.

[Jirmunskay, 1991] Jirmunskay E.A. The clinical electroencephalography. M. 1991.

[Zenkin,1991] Zenkin A.A. Cognitive computer graphic. M Science. 1991.

[Livanov,1984] Livanov M.N. Rithms of electroencephalogram// J.Visshey nervnoy deaytelnosty. 1984.V.34.4. P. 613-626.

[Mandelbrot, 2002] Mandelbrot B. Fractal geometry of nature. M. 2002.

[Pospelov, 1982] Pospelov D.A. On the way to artificial intellect. M.: Science. 1982

[System. Harmony. Symmetry. 1988] System. Harmony. Symmetry. /Ed. Tuchtin V.S. and Urmancev U.A. M.: Misl, 1988. 315p

[Fedotchev at al. 2003] Fedotchev A.I., Bondar A.T., Akoev I.G. Rhythmic structure of human EEG: modern state and a tendency researches..J. Uspechi physiological sciences. 2000. v.31.№3.P.39-53.

[Yachno and others. 2004] Yachno V.G., Nuydel I.V., Ivanov A.E., and others Investigation of dynamic regimes of neurosimilar systems, // Information Technology and computational systems. 2004. 1. P 126-148.

Author's Information

Nina V. Dmitrieva *- N.D. Prof. Head of the System Analysis Laboratory Research Insitute of Normal Physiology. Russian Academy of Medical Sciences.Mochovay str. Moscow, 125009 Russia. E-mail: nvdmitrieva@inbox.ru*

MINIMUM MEAN SQUARE ERROR APPROACH IN IMAGES PROCESSING FOR FULL-FIELD DISPLACEMENTS AND DEFORMATION MEASUREMENTS

Georgi Stoilov, Vasil Kavardzhikov, Dessislava Pashkouleva

***Abstract**: A vision system is applied to full-field displacements and deformation measurements in solid mechanics. A speckle like pattern is preliminary formed on the surface under investigation. To determine displacements field of one speckle image with respect to a reference speckle image, sub-images, referred to Zones Of Interest (ZOI) are considered. The field is obtained by matching a ZOI in the reference image with the respective ZOI in the moved image. Two image processing techniques are used for implementing the matching procedure: – cross correlation function and minimum mean square error (MMSE) of the ZOI intensity distribution. The two algorithms are compared and the influence of the ZOI size on the accuracy of measurements is studied.*

***Keywords**: Full-field displacement measurement, Image correlation, Minimum mean square error of intensity distribution.*

***ACM Classification Keywords**: J.2 Physical Sciences and Engineering (Physics)*

***Conference**: The paper is selected from International Conference "Intelligent Information and Engineering Systems" INFOS 2008, Varna, Bulgaria, June-July 2008*

Introduction

The development of reliable, affordable, easy to use full-field measurement techniques is crucial when one wants to characterize the mechanical effects on solids and structures. Materials which are homogeneous at the scale of observation but subject to complex load histories or heterogeneous lead to kinematics fields that need multi-scale analysis to relate the microstructure to the mechanical response. The full-field measurements allow a synergism of two areas that for a long time lived in separate environments – experiments and simulations, to be established.

The fast development of CCD or CMOS cameras that can be used at various scales enables for rapid dissemination of contactless techniques for full-field measurements. Usually these measuring techniques are based on image processing algorithms using correlation principles that were initially introduced in fluid mechanics since the development of Particle Image Velocimetry (PIV) [1, 2] in the 1970's. It is also adapted in solid mechanics since early 1980's [3]. The texture heterogeneities at a surface (or in a volume) of observation are utilized for determination of displacements fields using pattern matching algorithms. With the generalization of very powerful and affordable computers, these techniques start to be commonly used under various conditions [4].

A speckle like pattern is preliminary formed on the surface under investigation. To determine displacements field of one speckle image with respect to a reference speckle image, sub-images, referred to Zones Of Interest (ZOI) are considered. The field is obtained by matching a ZOI in the reference image with the respective ZOI in the moved image.

In cases when the mean values of visibility (contrast) of regions possessing a ZOI dimension is significantly different, the usage of cross- correlation algorithms leads to increasing the probability for a wrong identification the corresponding ZOI of the second image. Such a situation may be a consequence of a non uniform illumination, as well as of non equal reflective properties of observant surface. It is appropriately in these cases an algorithm, based upon minimum mean square error (MMSE) calculation to be used.

Algorithm

The deformation field of the investigated surface can be obtained when two surface images are compared. The first image is stored in computer before the object deformation. It is considered as reference image. A second image is registered after deformation of the observed surface. For every pixel from the reference image an area surrounded that pixel is called zone of interest (ZOI). To determine the vector field of deformation one has to match the corresponding ZOI-s in the two images. If the function of the deformation is a smooth function it is possible to skip calculations in a reasonable distance. A vector field of deformation, derived by the above described way is shown in fig. 1

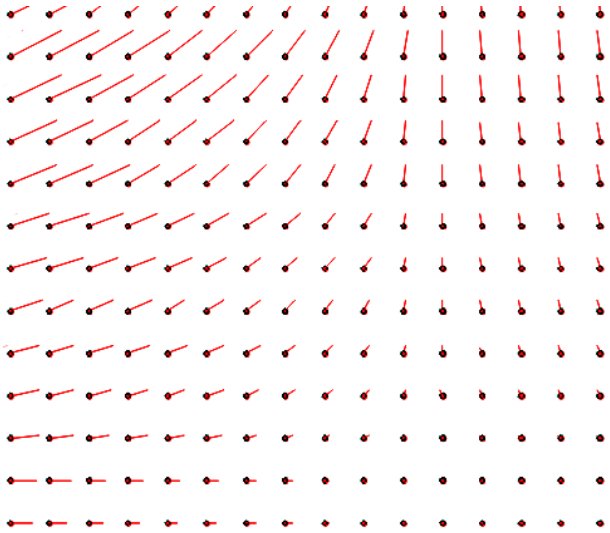
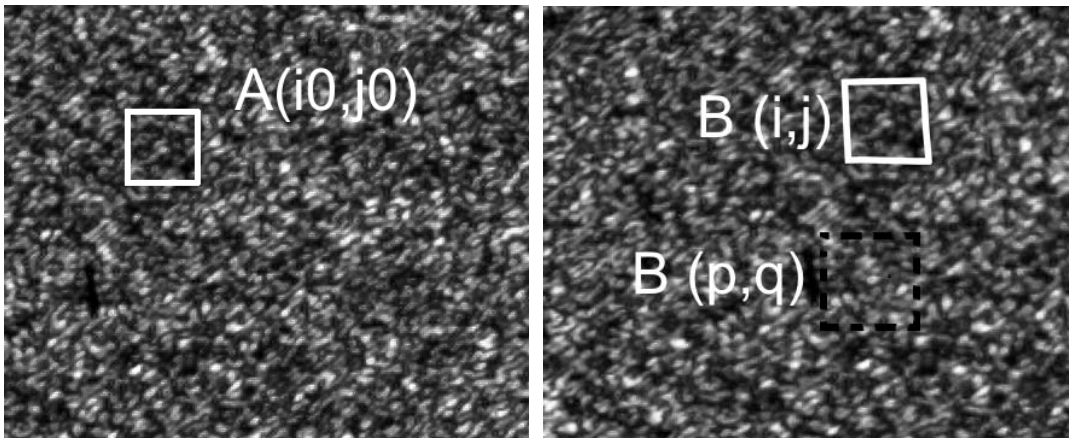


Fig. 1 A vector field of deformation

Let us consider a ZOI $A(i_0,j_0)$ in the reference image (the square zone in Fig. 2-a).



a b

Fig. 2. (a, b) Two consecutive images of the investigated surfaces, registered in the process of its deformation, a- reference image, b- deformed image

One can solve the correlation problem for finding the location displaced ZOI A(i0,j0) to a ZOI B(i,j) position (Fig. 2b – the deformed square zone), accepting an assumption, that the whole surface of interest is illuminated and reflects the light so that no ZOI-s of the image with different mean intensity value. Under this condition ZOI B(i,j) can be searched for in a near vicinity of ZOI A(i0,j0), having a radius equal to the maximum expected reallocation. Calculating the 2D correlation function the ZOI B(i,j) is found with a big degree of a probability at this location at

which maximum value of correlation (a correlation peak) is obtained during the above procedure performance. There exists however situation in which the assumption of uniform, mean image intensity is not satisfied. A wrong correlation maximum can be obtained at position of ZOI B(p,q) (the circular zone in Fig. 2b), due to the fact that the mean intensity of the second image at this position exceeds the mean intensity value of the entire image. As a result the degree of probability of true correlation peak location decreases.

The algorithm of ZOI matching using MMSE calculation instead of correlation function is an acceptable alternative for overcoming this difficulty. By usage of the algorithm a "correlation" peak is generated only when the ZOI $A(i_0,j_0)$ and ZOI B(i,j) values are identical and as an amplitude as positions of pixels corresponding to the zones.

Theory

According to correlation theory a narrow autocorrelation peak can be obtained by processing of a large ZOI. This is a guarantee of precise localization of the peak. At the same time it is well known that the number of calculations increases by power 2 of the ZOI size. Looking for a numerical method, ensuring good enough accuracy in cases when small ZOI size is processed we have found that application of the standard correlation method leads to unsatisfied results in case of small ZOI size having low contrast of the intensity variations. The correlation formula (1) presents a linear function of the amplitude of discrete input signals.

$$C_{(i0,j0,i,j)} = \sum_m \sum_n A_{(m+i0,n+j0)}.B^*_{(m+i,n+j)} \quad (1)$$

Where *A* is a ZOI before deformation and *B* is the image after deformation, *C* is the signal cross- correlation function, *i0* and *j0* are indexes of the reference ZOI A images when a Cartesian coordinate system is used, *i* and *j* are the corresponding indexes in the same coordinate system of the corresponding ZOI in B, m and n are indexes of the digital integration. B* is a conjugation value of *B*. The multipliers which are independent of the processed information are ignored.

$$C_{(i0,j0,p,q)} = \sum_m \sum_n A_{(m+i0,n+j0)}.B^*_{(m+p,n+q)} \quad (2)$$

If B is presented as multiplication (3) of a normalized function N and the mean contrast for this ZOI the two correlation functions can be compared as function of the contrast (4).

$$B_{(m+i,n+j)} = k_{(i,j)}.N_{(m+i,n+j)} \quad (3)$$

$$k(i,j)C^N_{(i0,j0,i,j)} < k(p,q)C^N_{(i0,j0,p,q)} \quad (4)$$

One can see from equation (4) that it is possible such a value of the quotient of the image contrasts K(i,j)/K(p,q) to be found that satisfy the inequality. This means that the correlation function value C(i0,j0,p,q) of ZOI A and B will have its maximum at an other ZOI B(p,q) with coordinates p and q that is not corresponding to the "true" solution C(i0,j0,i,j) in Fig. 2b.

In this example we have found ZOI B(p,q), similar to ZOI B(i,j). The wrong zone has a larger K (higher contrast of 2D intensity distribution) in comparison to the ZOI B(i,j). Since we are searching for the highest value of the correlation function, a "wrong" location of ZOI B with coordinates p and q could be determined. The utilization of MMSE approach leads to comparison of the signal according to both: its shape and its amplitude (equations (5) and (6)).

$$E_{(i,j)} = \sum_m \sum_n [A_{(m,n)} - B_{(m+i,n+j)}]^2 = \sum_m \sum_n e^2_{(m+i,n+j)} \quad (5)$$

The characteristic of the function of MMSE *E* can be seen easily If the contrast is presented as *1+k* that is a increasing function too.

$$E_{(i,j)} = \sum_m \sum_n [A_{(m,n)} - (1+k)N_{(m+i,n+j)}]^2 = \sum_m \sum_n ke^2_{(m+i,n+j)} = kE_{N(i,j)} \quad (6)$$

where *e* is the error between reference ZOI A and target ZOI B.

To obtain a function similar to the cross correlation function operations normalization and inversion are used.

$$C = 1 - \frac{E}{Max(E)} \tag{7}$$

or

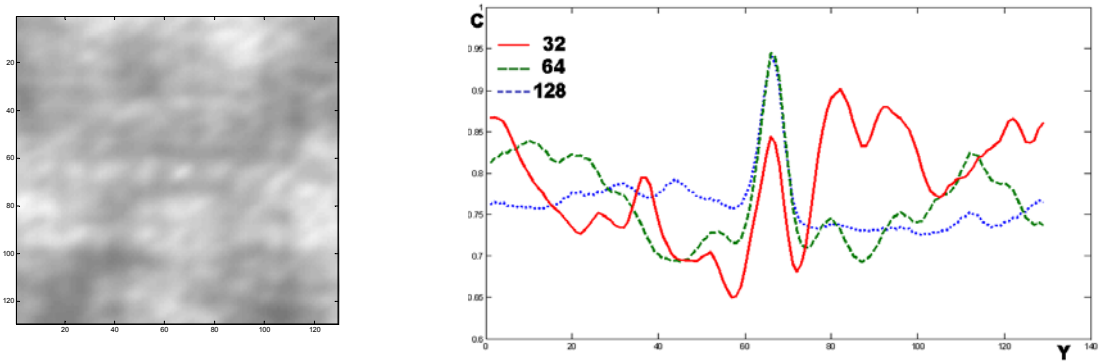
$$C = 1 - \frac{kE_N}{Max(Ek)} = 1 - k\frac{E_N}{Max(Ek)} \tag{8}$$

It can be seen from equation (2) that at the correlation algorithm, if one of the input signals increases the correlation value increases as a result. When MMSE algorithm is implemented the increasing one of the input signals results to increasing of the error(6) and decreasing of the correlation peak value (8).

Experimental results

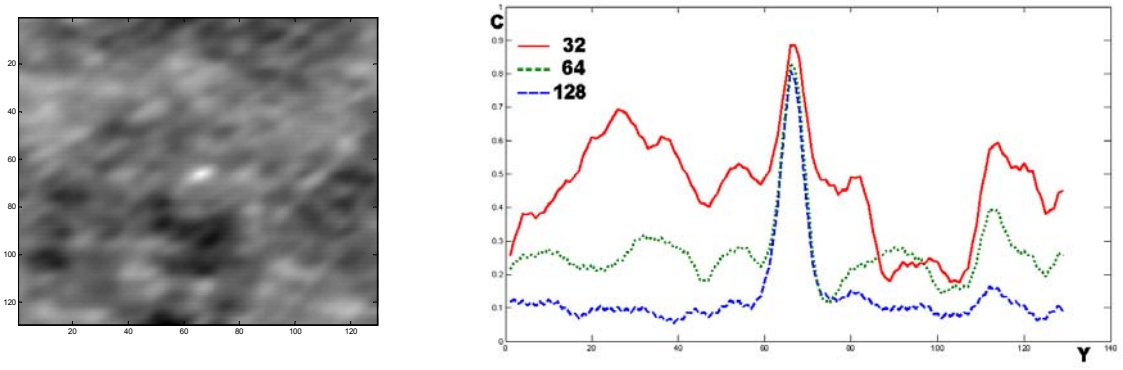
The above described algorithms for ZOI matching were optimized according to its parameters and were applied to the image shown in Fig. 2. The correlation maximum is calculated using the cross-correlation function for a square region of search for matching region 128x128 pixels and with ZOI size 32x32 pixels. Three ZOI size - 128, 64 and 32 pixels, are shown and results obtained are depicted in Fig 3-b. It is seen that the decreasing of ZOI size leads to increasing of the value of the correlation function at wrong position. At the ZOI size 32 pixels, the "true" correlation peak value is smaller than the value of the peak obtained in the wrong location.

MMSE method is used for carrying out the similar calculations and the results are presented in Fig 4.



a b

Fig. 3. a - Correlation function calculated in a 2D region, b - a distribution of correlation values along to the mean vertical cross section of the region at 128, 64 and 32 pixel size of ZOI



a b

Fig. 4. a – Mean Square Error for a 2D region, b - a distribution of correlation values along to the mean vertical cross section of the region at 128, 64 and 32 pixel size of ZOI

Computer programs for realizing the correlation and MMSE algorithms are developed in C++ environment. The computing of cross-correlation function is carried out by a standard optimized algorithm, and of MMSE – by a gradient algorithm. The processing time of performing the MMSE calculations is about 50% longer than the time of correlation calculations.

Conclusion

The solution of ZOI correspondence problem by a cross-correlation approach is a fast and accurate way for determining of deformation vector fields in mechanics of solids. In case when relatively big deformations of second image , related to the reference image are observed, a reduction of the images area or decreasing of ZOI size have to be introduced for obtaining acceptable solution. The usage of MMSE approach leads to increasing the reliability of ZOI searching, which is paid by an increasing computing time. A combination of the two algorithms that would unit their advantages can be realized to reach optimal full-field measurements. Routine calculations my be carried out by the correlation algorithm and a switch to MMSE algorithm can be programmed for the image regions in which decreasing of correlation algorithm reliability is expected or estimated.

Acknowledgments

This work was partially supported by the National Science Foundation at the Ministry of Education of Bulgaria by grant *№ BY-TH-923.*

Bibliography

1. T. D. Dudderar and P.G. Simpkins, Laser T. D. Dudderar and P.G. Simpkins, Laser speckle photography in a fluid medium, Nature, v. 270, 45-47, 1977.
2. C.J.D. Pickering and N.A. Halliwell, Speckle laser in fluid flows: signal recovery with two step processing, Appl. Opt., v. 23, 1128-1129, 1984.
3. M.A. Sutton, W.J. Wolters, W.H. Peters, W.F. Ranson and S.R. McNeill, Determination of displacement using improved digital correlation method, Im. Vis. Comp., v. 1, 133-139, 1983.
4. [Hild, 2007] F. Hild and St. Roux, Lecture Notes in Knowledge-based Multifunctional Materials: Full-Field Measurements and Identification in Solid Mechanics, Ed. KMM-NoE Network of Excellence, Warsaw, 2007.

Authors' Information

Georgi Stoilov *– Institute of Mechanics, BAS, Acad.G.Bontchev St. bl. 4, Sofia-1113, Bulgaria;*
e-mail: gstoilov@imbm.bas.bg.

Vasil Kavardjikov *– Institute of Mechanics, BAS, Acad.G.Bontchev St. bl. 4, Sofia-1113, Bulgaria;*
e-mail: kavarj@imbm.bas.bg.

Dessislava Pashkouleva *– Institute of Mechanics, BAS, Acad.G.Bontchev St. bl. 4, Sofia-1113, Bulgaria;*
e-mail: dessip@imbm.bas.bg.

WLAN / WMAN INTEGRATION ARCHITECTURE AND TRAFFIC CONTROL FOR VOICE TRANSMISSION

Boris Tsankov, Kiril Kassev, Radostin Pachamanov

***Abstract**: The popular technologies Wi-Fi and WiMAX for realization of WLAN and WMAN respectively are much different, but they could compliment each other providing competitive wireless access for voice traffic. The article develops the idea of WLAN/WMAN (Wi-Fi/WiMAX) integration. WiMAX is offering a backup for the traffic overflowing from Wi-Fi cells located into the WiMAX cell. Overflow process is improved by proposed rearrangement control algorithm applied to the Wi-Fi voice calls. There are also proposed analytical models for system throughput evaluation and verification of the effectiveness using WMAN as a backup for WLAN overflow traffic and the proposed call rearrangement algorithm as well.*

***Keywords**: WLAN/WMAN integration, Wi-Fi, WiMAX, voice traffic.*

***ACM Classification Keywords**: C.2.5. Local and Wide Area Network, C.4. Performance of Systems – Voice Traffic*

***Conference**: The paper is selected from Sixth International Conference on Information Research and Applications – i.Tech 2008, Varna, Bulgaria, June-July 2008*

Introduction

Two different air interface standards for wireless access to telecommunication core network became recently very popular. The first standard is for wireless LAN (WLAN), called Wi-Fi (Wireless Fidelity) by Wi-Fi Alliance [1] that certifies network devices to comply with the IEEE 802.11 standard (last version [2]). Today Wi-Fi is widely used and now almost all laptops and other handheld devices as well came with Wi-Fi build-in. The second standard is for wireless MAN (WMAN), called WiMAX (Worldwide Interoperability for Microwave Access) by the WiMAX Forum [3] that promotes conformance and interoperability of the IEEE 802.16 standard [4]. Both standards offer user mobility, and are used for voice traffic although, it is not in the same circumstances.

The differences between Wi-Fi and WiMAX are well known [5], but we are looking for an opportunity to complement each other and to integrate them into one wireless access system for voice traffic.

Wi-Fi is unbeatable to install fast, easy and cheap A*ccess Point* (AP). The cell radius is optimized around 100 m. The standard and its contention-based MAC can not guarantee the required Quality of Service (QoS) and especially the very important QoS parameter - voice latency. The standard does not allow for differentiated service level for each user. A possible solution of QoS guarantee is to substitute the widespread use of *Distributed Coordination Function* (DCF) with *Point Coordination Function* (PCF) although with some loss of simplicity.

WiMAX equipments are considerably more expensive and the range of covering is somewhat 5 to 50 km distance from the *Base Station* (BS). The 802.16 standard is with well developed centralized system of *Polling Services* of a *Request-Grant* MAC and specifically designed to support voice. The WiMAX system throughput (hundreds of voice connections) is much more than Wi-Fi throughput (hardly exceeding 10 -20 voice connections). The WiMAX call range and throughput can be increased introducing provided by the standard mesh mode (see [6] and the references therein).

There are proposals for integration of Wi-Fi with 3G cellular networks [7], [8]. Integration between WiMAX and space based telecommunication infrastructure is considered in [9]. Finally, there are some attempts to integrate

Wi-Fi and WiMAX [10], [11], [12]. In all these cases of heterogeneous networks the subscriber devices have to be compatible with different standards employed. One approach is to associate the existing separated systems in a device. Another, more advanced approach is proposed in [13] for Wi-Fi/WiMAX integration instead of a simple duplication of the radio systems.

In this article WLAN/WMAN (Wi-Fi and WiMAX) network architecture is considered and a particular mechanism for traffic control (rearrangement of calls) is proposed. Analytical methods proposed by authors for capacity analysis are applied. Numerical results are presented with an evaluation of the served traffic gain due to integration and due to proposed call rearrangements.

The terms WLAN – Wi-Fi and WMAN – WiMAX as well will be used in the article text interchangeable.

System Description

The proposed WLAN/WMAN integrated access network is depicted on Fig. 1. The Wi-Fi *hot spots* do not cover the entire WiMAX cell, but only the areas with expected high density of traffic sources. There is separate backhauling organized for BS and Aps to the edge router of the core network, but there are other possible arrangements as well. Special controls tasks are assigned to *Resource Manager* (RM on Fig. 1).To fulfill its traffic control functions the RM is connected to BS and Aps with signaling channels. RM might be incorporated with WiMAX BS. A subscriber located in the coverage of a Wi-Fi cell can be served by both – WLAN and WMAN, but always at first an attempt is made the arriving call to be served by the WLAN. A subscriber located in the coverage of WMAN only is served by WMAN. We will consider two possible algorithms of traffic control for subscriber located in the coverage of both systems:

- ***Overflowing.*** If there are enough resources in the WLAN cell, the arriving call is put into service by the WLAN. In case there are not enough resources in the WLAN cell, the arriving call is directed to the WMAN network. If there are not enough resources in WMAN either, the call is lost. Such algorithm of traffic control is proposed in [12] and in an author's previous work [14], where the cell corresponding to WMAN is called *umbrella-cell.* This algorithm we will call here *Simple Overflowing* (SO).
- ***Overflowing with Rearrangement.*** This algorithm is an extension of the previous one. If during the service in WMAN of an overflow call, in the WLAN cell of that call one resource became free (i.e. the service of one call in the WLAN considered has been completed), the overflow call is directed back (*rearranged*) to his WLAN to complete its service there. The algorithm is proposed by the authors in [14] and it is called here correspondingly O*verflowing with Rearrangement* (OR).

When we say "number of resources" of certain device, we have in mind the number of voice calls can be served through that device. This numbers are fixed by means of a Call Admission Control (CAC) mechanism acting in BS and each AP (Fig. 1).

System Throughput Evaluation

We are proposing analytical methods for system capacity analysis for both algorithms separately.

For **SO** we will apply the well known *equivalent random theory* [15], [16, p. 156] proposed by R. Wilkinson to solve the problem that the overflowing traffic is not Poisson traffic any more.

Suppose there are *k* WLANs; the traffic arriving at WLAN *i* is A_i $(i = 1,...,k)$; WLAN *i* has n_i resources forming a full-availability group. The traffic overflowing from WLAN *i* is with mean value:

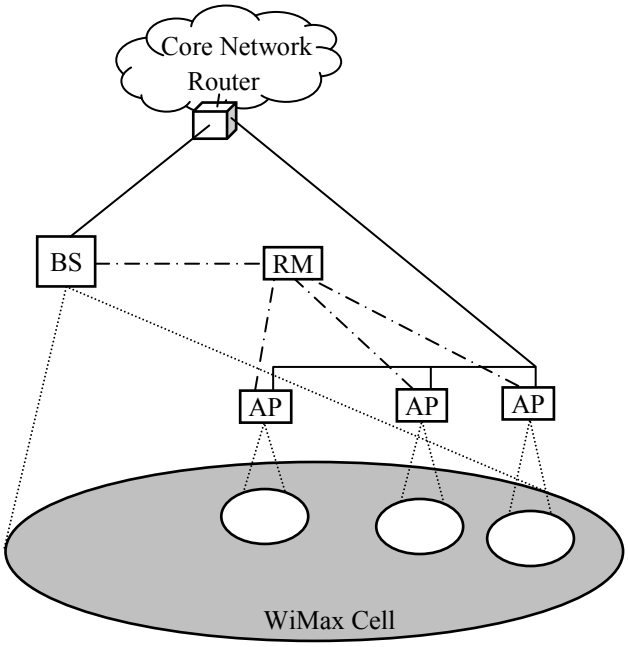


Fig. 1. WLAN/WMAN integration architecture

$$M_i = A_i E_B(n_i, A_i) \tag{1}$$

and variance

$$V_i = M_i\left[\frac{A_i}{n_i + 1 - A_i + M_i} + 1 - M_i\right] \tag{2}$$

Notation $E_B(n, A)$ corresponds to Erlang's B formula for traffic *A* arriving at *n* devices.

The traffic arriving at WMAN has the values

$$M = \sum_{i=1}^{k+1} M_i \qquad V = \sum_{i=1}^{k+1} V_i \tag{3}$$

The values indexed with $k+1$ correspond to "fresh" (Poisson) traffic $A_{i+1} = V_{i+1} = A_{WMAN}$, arriving directly to WMAN from the WiMAX subscriber station SS. The parameters n_e and A_e of the equivalent group are chosen in such a way that the traffic overflowing from it has the same *M* and *V* as the obtained by (3). The values of n_e and A_e are derived from the following equations

$$\begin{cases} M = A_e E_1(n_e, A_e) \\ V = M\left[\dfrac{A_e}{n_e + 1 - A_e + M} + 1 - M\right] \end{cases} \tag{4}$$

Effective approximate solution of (4) is given by Y. Rapp [15]:

$$\begin{aligned} A_e &\approx V + 3z(z-1) \\ n_e &\approx A_e \frac{M+z}{M+z-1} - 1 - M \end{aligned} \tag{5}$$

where $z = V/M$.

Traffic loss probability is given by

$$B = E_B(n_e + n_{WMAN}, A_e) \tag{6}$$

n_{WMAN} is resource number in WMAN. Eq. (7) can be used to derive the necessary WMAN resources.

For **OR** although the call control algorithm is more sophisticated it is possible to derive a relatively simple analytical method for calculating the losses and the throughput afterwards. We will consider the system on Fig. 1 as a pool of

$$n = n_{WMAN} + \sum_{i=1}^{k} n_i$$

resources used by all $k+1$ traffic flows with the following restrictions: each WLAN traffic flow A_i can use only $n_i + n_{WMAN}$ resources; flow A_{WMAN} can use only n_{WMAN} resources. We propose to apply the multidimensional Erlang's-B formula [16], [17]:

$$p(j_1, j_2, ..., j_k, j_{WMAN}) = Q \frac{A_{WMAN}^{j_{WMAN}}}{j_{WMAN}!} \prod_{i=1}^{k} \frac{A_i^{j_i}}{j_i!}, \; j_i = 0,...,n_i; j_{WMAN} = 0,...,n_{WMAN}$$
$$Q^{-1} = p(0,...,0) \tag{7}$$

Calculation of state probabilities (7) can be simplified using the convolution method, first time proposed by V. Iversen [16, p. 180]. The convolution algorithm used we will describe with following steps. Notations with index *k+1* stand for WMAN.

1) Calculate the state probabilities for each flow assuming there are not other flows in the system:

$$P_i = \{p_i(0), p_i(1), ..., p_i(n_i)\}, \quad i = 1,...,k+1 \tag{8}$$

It is important only the relative value of probabilities $p_i(x)$, therefore start with $q(0) = 1$ and calculate the values of $q_i(x)$ as ratio to $q(0)$. Normalize the relative state probabilities:

$$p_i(j) = \frac{q_i(j)}{Q_i}, \quad j = 0,1,...,n_i$$
$$Q_i = \sum_{j=0}^{n_i} q_i(j), \quad i = 1,...,k+1 \tag{9}$$

2) By consecutive convolutions calculate the aggregate state probabilities for the total system excluding *i*-th traffic flow:

$$Q_{(k+1)/i} = P_1 * P_2 * ... * P_{i-1} * P_{i+1} * ... * P_{k+1} \tag{10}$$

Calculations are done consecutively from left to right.

3) Calculation the traffic losses. First calculate the convolution:

$$Q_{k+1} = Q_{(k+1)/i} * P_i \tag{11}$$

The result from this convolution is:

$$Q_{k+1}(j) = \sum_{x=0}^{j} Q_{(k+1)/i}(j-x).p_i(x) = \sum_{x=0}^{j} p_x^i(j) \tag{12}$$

where in the term $p_x^i(j)$, *i* is traffic flow number, *j* is the total number of busy resources and *x* is number of busy resources of *i*-th flow.

Repeating both steps 2) and 3) for every traffic flow, the traffic losses for *i*-th flow are:

$$B_i = \sum_{j \in \Omega_i} p_x^i(j) / Q \qquad (13)$$

where Ω_i the set of states for which the calls of *i*-th traffic flow are rejected.

Q is the normalization constant:

$$Q = \sum_{j=0}^{n} Q_{k+1}(j) \qquad (14)$$

The system from Fig. 1 with application of OR resembles a multidimensional teletraffic system called "shearing with call limitations" [18], where all traffic flows have full access to the common resources, but some or all of the traffic flows can occupy only limited number of these resources. For system with call limitations the convolution method is giving an exact solution. Unfortunately in our system every Wi-Fi flow have reserved "own" resources that can not be used by the neighboring Wi-Fi subscribers leading to not exact results by the convolution method. Only for the case *k* = 1 the convolution method is exact one. The simulation in [14] for a similar teletraffic system (k = 6) shows the error of convolution method application is acceptable for practical network design purposes.

Numerical Results

The current section is dealing with an analytical evaluation on the effect of accommodating the WiFi overflow traffic by a WiMAX cell, applying both the SO and OR traffic control algorithms. The system we are considering (depicted on Fig. 1) consists of a single WiMAX cell with a wide coverage area, which covers a total of 10 WiFi cells with much smaller coverage, each having a resource size of simultaneously serving 6 calls, and an equal offered traffic load A_i in range of 0.5 to 12 erl. The traffic flows to the WiFi AP resources and the direct flow to the WiMAX resources are all Poisson and independent of each other. The overall offered traffic load of the system is a summation of these independent flows.

Since the WiFi bandwidth resources are cheaper compared to the WiMAX ones, it is preferred the user to choose the WiFi for network access and when a call to a WiFi cell is blocked due to occupation of the cell resources, the call to be rerouted to the WiMAX cell. Fig. 2 depicts the overall carried traffic of the system as a function of applying an appropriate traffic control algorithm, considering different values of WiMAX resources n_{WMAN}. In this case, we assume no direct traffic load A_{WMAN} to the WiMAX cell. The same figure also depicts the impact of having the WiMAX resources as a backup of WiFi in terms of the overall carried traffic.

The second point of the analysis, presented at Fig. 3, is dealing with the calls from users, situated outside the WiFi coverage, who are directly served by the WiMAX cell. In this case, there is a direct traffic load A_{WMAN} to the WiMAX cell. For the purpose of the analysis, we assume A_{WMAN} to be three times greater than the traffic load offered to each WiFi cell. As one can see from the results, applying the OR algorithm can significantly improve the overall network performance.

Conclusion

We have presented an integrated WLAN/WMAN architecture based on IEEE 802.11 and IEEE 802.16 standards and specifically proposed an algorithm for voice traffic call control. We have proposed analytical models for system throughput evaluation and verified the effectiveness of using WMAN as a backup for WLAN overflow traffic and the proposed call rearrangement algorithm.

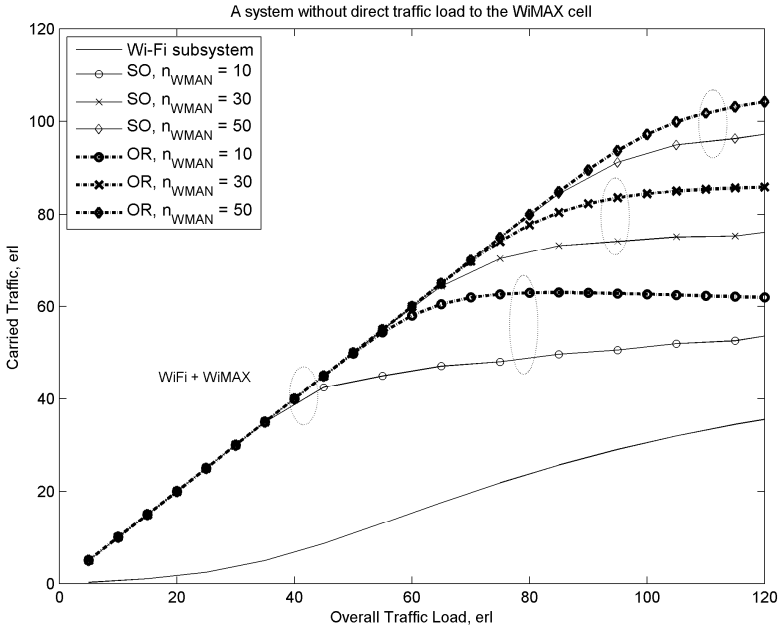


Fig. 2 Overall carried traffic, considering a system without direct traffic load to the WiMAX cell

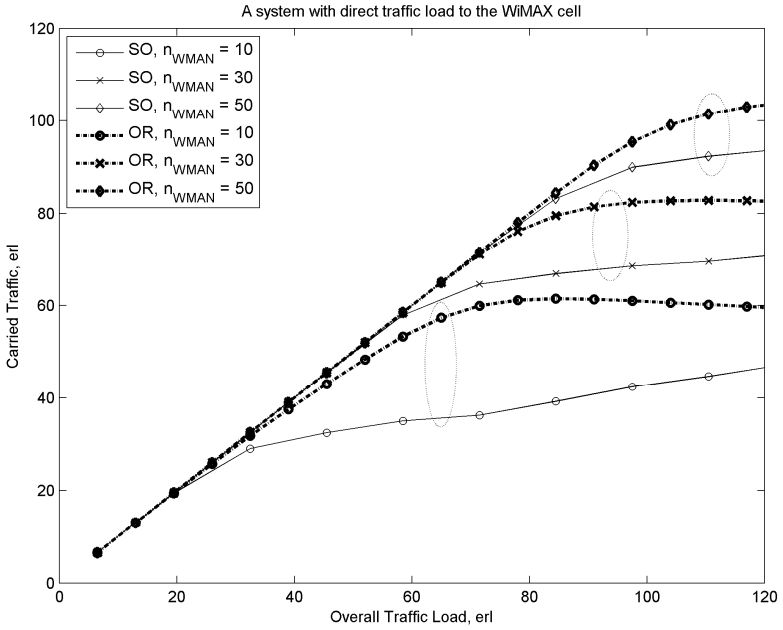


Fig. 3 Overall carried traffic, considering a system with direct traffic load to the WiMAX cell

Acknowledgement

This work was supported by the Bulgarian Ministry of Education and Science with grant ВУ-ТН-105/2005 (project "Multimedia Packet Switching Networks Planning with Quality of Service and Traffic Management" and by the Technical University of Sofia – grant 080910dni-7 (project "Optimal Telecommunications Resource Allocation with Cross-Layer Design")

Bibliography

[1] www.wi-fi.org

[2] IEEE 802.11 Working Group, IEEE 802.1-2007: LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications, 12.06.2007.

[3] www.wimaxforum.org

[4] IEEE 802.16e-2005, IEEE standard for local and metropolitan area networks – Part 16: Air interface for fixed and mobile broad band wireless access systems – Amendment for physical and medium access control layers for combined fixed and mobile operation in licensed bands, May 20, 2005, http://ieee802.org/16/pubs/80216e.html.

[5] N. Fourty, T. Val, P. Fraisse and J.-J. Mercier, Comparatice analysis of new high data rate wireless communication technologies "From Wi-Fi to WiMAX", Proc. of Joint Int. Conf. on Automatic and Autonomous Systems and Int. Conf. on Networking and Services (ICAS/ICNS 2005), Oct. 2005, pp. 68-68.

[6] B. Tsankov, Routing tree branch evaluation in WiMAX multi-hop network for voice traffic, Proc. of Nat. Conf. with Int. Particip. ELECTRONICS 2008, Sofia, 29 – 30 May, (submitted for presentation).

[7] H.-Y. Wei and R. D. Gitlin, Two-hop-relay architecture for next-generation WWAN/LAN integration, IEEE Wireless Communications, vol. 11, n. 2, Apr. 2004, pp.24 – 30.

[8] H.-Y. Wei and R. D. Gitlin, Incentive scheduling for cooperative relay in WWAN/LAN two-hop-relay network, Proc. Of IEEE Wireless Communications and Networking Conf., WCNC 2005, vol. 3, pp. 1696 – 1701.

[9] R. Giuliano, M. Luglio and F. Mazzenda, Interoperability between WiMAX and broadband mobile space networks, Communications Magazine, vol. 46, n. 3, March 2008, pp. 50-57.

[10] S Jindal, A. Jindal and N. Gupta, Grouping Wi-MAX, 3G and Wi-Fi for wireless broadband, Proc. of First IEEE and IFIP Int. Conf. in Central Asia on Internet, Sept. 2005, pp. 1-5.

[11] P. Neves, S. Sargento and R. L. Aguiar, Support of real-time services over integrated 802.16 metropolitan and local area networks, Proc. of 11th IEEE Symposium on Computers and Communications (ISCC'06), Oct. 2006, pp. 15-20.

[12] H. Pan, S. Yamamoto, D. Kawasaki and H. Tani, Capacity analysis and resource management of an integrated WiMAX and WiFi access system for VoIP, Proc. of 2nd Int.Conf. on Access Networks $ Workshops 2007, AccessNet'07, pp.1-8

[13] S. T. Gul, C. Moy and J. Palicot, Two scenarios of flexible multi-standard architecture designs using a multi-granularity exploration, Proc. of 18th IEEE Int. Symp. On Personal, Indoor and Mobile Radio Communications (PIMRC'07), 2007, pp. 1-5.

[14] B. Tsankov and R. Pachamanov, Performance evaluation of an umbrella-cell in cell networks. Proc. of TELECOM 2006, St. Konstantin and Helena - Varna, Oct. 2003, (in Bulgarian)

[15] B. Tsankov and Ts. Slavova, Traffic in telecommunications networks and systems, „Tehnika", Sofia, 1979, (in Bulgarian).

[16] V. B. Iversen, Teletraffic Engineering Handbook, ITU-D, Geneva, 2003

[17] B. Tsankov, Telecommunications – fixed, mobile and IP, "Novi znania", Sofia, 2006, (in Bulgarian)

[18] B. Tsankov, R. Shishkova, Teletraffic Models of a Common Trunk Group Serving ISDN-Traffic. Proc. of 2nd Int. Sem. on Teletraffic Theory and Computer Modeling, Moscow, Oct., 1989, vol. 3, pp. 22-41.

Authors' Information

Boris Tsankov *– Associate Professor, Technical University - Sofia, 8th, Kl. Ohridski Blvd, 1000 SOFIA, Bulgaria; e-mail: bpt@tu-sofia.bg*

Kiril Kassev *– Assistant Professor, Technical University - Sofia, 8th, Kl. Ohridski Blvd, 1000 SOFIA, Bulgaria; e-mail: kmk@tu-sofia.bg*

Radostin Pachamanov *– Solution Engineer, Ericsson Telecommunications Bulgaria, 55, N. Vaptsarov Blvd., 1407 SOFIA, Bulgaria; e-mail: radostin.pachamanov@ericsson.com*

IMPROVED HYBRID MODEL OF HMM/GMM FOR SPEECH RECOGNITION

Poonam Bansal, Anuj Kant, Sumit Kumar, Akash Sharda, Shitij Gupta

***Abstract**: In this paper, we propose a speech recognition engine using hybrid model of Hidden Markov Model (HMM) and Gaussian Mixture Model (GMM). Both the models have been trained independently and the respective likelihood values have been considered jointly and input to a decision logic which provides net likelihood as the output. This hybrid model has been compared with the HMM model. Training and testing has been done by using a database of 20 Hindi words spoken by 80 different speakers. Recognition rates achieved by normal HMM are 83.5% and it gets increased to 85% by using the hybrid approach of HMM and GMM.*

***Keywords**: Speech Recognition, GMM, HMM*

***ACM Classification Keywords**: I5.4 Pattern Recognition- Applications- Conference proceedings*

***Conference**: The paper is selected from International Conference "Intelligent Information and Engineering Systems" INFOS 2008, Varna, Bulgaria, June-July 2008*

Introduction

Speech signal primarily conveys the words or message being spoken. Area of speech recognition is concerned with determining the underlying meaning in the utterance. Success in speech recognition depends on extracting and modeling the speech dependent characteristics which can effectively distinguish one word from another. A general pattern recognition system consists of 4 parts, the feature extractor, pattern trainer, pattern classifier and decision logic. One of the most common and successful feature extraction processes is MFCC [Kotnik, 2002] which has been implemented in this model. Of the various speech/ speaker recognition models available, the most commonly used are Artificial Neural Networks (ANN), Hidden Markov Model (HMM) and Gaussian Mixture Model (GMM). Presently HMM is widely used as one of the successful speech recognition process. HMM considers the speech signal as quasi- static for short durations and models these frames for recognition. It breaks the feature vector of the signal into a number of states and finds the probability of a signal to transit from one state to another [Rabiner, Juang, 1993]. Viterbi search, forward-backward and Baum-Welch algorithms are used for parameter estimation and optimization [Rabiner, 1989], [Rabiner, Juang, 1991]. GMM on the other hand considers a signal to contain different components that are independent of each other. These components represent the broad acoustic classes that represent certain vocal tract configurations [Reynolds, Rose, 1995]. Thus it is more inclined towards modeling features concerning to words having specific characteristics. Each component is optimized using EM algorithm [Dempster, 1977]. Various ways of combining these models have been proposed. One way is to use a 3 state HMM and implement GMM on the middle state to obtain the observation probabilities for that state [Rodriguez, 2003]. Here we have proposed a novel model combining GMM and HMM to provide more enhanced speech recognition engine, which we call as the hybrid model. We have proposed the extraction of likelihood value by combining the likelihood values of both the models using decision logic. In the rest of the paper we discuss individually about the HMM, the GMM and how the hybrid model was developed. At the end, we compare the HMM model with the hybrid model using a 20 word, 80 utterances data set.

Hidden Markov Model

Hidden Markov model (HMM) is a statistical model in which the system being modeled is assumed to be a Markov process with unknown parameters, and the challenge is to determine the hidden parameters from the observable parameters. The extracted model parameters can then be used to perform further analysis, for example for pattern/speech recognition applications. In a hidden Markov model, the states are not directly visible, but variables influenced by the state are visible. Each state has a probability distribution over the possible output tokens. Also, the state transitions are probabilistic in nature. Therefore, the sequence of tokens generated by an HMM gives some information about the sequence of states. The complete HMM model is denoted as $\lambda = (A, B, \pi)$.

HMM Model Parameters

1. A, the state probability distribution, A = {aij}.
2. B, the observation symbol probability density, B = {bj(k)}.
3. π, the initial state distribution, π = { πi}.

The HMM training procedure requires codebook to estimate model parameters. In codebook, the large number of observational vectors of the training data is clustered into M observational vectors clusters using K-means iterative procedure. Based on this clustered observational vectors, estimates of the model parameters are generated during HMM training. The HMM training procedure tries to estimate the value of state probability distribution (A), observation symbol probability density (B), and initial state distribution (π). The observational vectors of training sequences are segmented for each of the N states, a maximum – likelihood estimates of the set of the observations that occur within each state j each of the observation vectors within a state is coded using the M – code – word codebook. Before proper estimation of the model parameters they are initialized to good initial estimates that are essential for rapid and proper convergence of the re-estimation formulas. The parameters are re-estimated using Viterbi Algorithm, Forward- Backward Algorithm and Baum/Welch Algorithm.

Forward- Backward Algorithm

Forward Algorithm

The forward variable αt(i) is defined as αt(i) = P(o1,o2,… ,ot,qT = i|λ) i.e. the probability of the partial observation sequence (until time t) and state i at time t, given the model λ. αt(i) is inductively computed by following steps:

Initialization:

$$\alpha_1(\imath) = \pi_i B_i(o_1), 1 \le \imath \le N \qquad (1)$$

Induction:

$$\alpha_{t+1}(j) = [\sum \alpha_t(i) a_{ij}] b_j(o_{t+1}), 1 \le t \le T-1 \qquad (2)$$

Termination:

$$\mathrm{P(O}|\lambda) \;=\; \sum \alpha_t(\mathrm{i}) \qquad (3)$$

Backward Algorithm

The backward variable βt(i) is defined as βt(i) = P(ot+1,ot+2,…,oT,qT = i|λ) i.e. the probability of the partial observation sequence from t+1 to the end, given the state i at time t and the model λ. βt(i) is inductively solved as follows:

Initialization:

$$\beta_t(i) = 1, 1 \le i \le N \qquad (4)$$

Induction:

$$\beta_{\mathrm{t}}(i) = \sum a_{ij} b_j b_j(o_{t+1})\ \beta_{t+1}(j),\ t = \mathrm{T}-1, \mathrm{T}-2, ..., 1, \quad 1 \le i \le \mathrm{N} \qquad (5)$$

Combining Forward and Backward variables, we get:

$$\mathrm{P(O}|\lambda) = \sum \alpha_t(i) \beta_t(i), 1 \le t \le \mathrm{T} \qquad (6)$$

The Viterbi Algorithm

To find the single best state sequence, for the given observation sequence, δt(i) = max P[q1q2…qt-1, qt = i, o1o2…ot|λ], δt(i) is the highest probability and ψt(i) is the track for path t.

Initialization:

$$\delta_1(i) = \pi_i b_i(o_1), 1 \le i \le N$$
$$\psi_1(i) = 0 \qquad (7)$$

Recursion:

$$\delta_t(j) = \max[\delta_{t-1}(i) a_{ij}] b_j(o_t), 2 \le t \le T, 1 \le j \le N \qquad (8)$$

$$\psi_t(t) = \arg\max[\delta_{t-1}(i)a_{ij}], 2 \le t \le T, 1 \le j \le N$$

Termination:

$$P^* = max\,[\delta_t(i)]\,1 \le i \le N$$
$$q_T^* = \arg\max\,[\delta_t(i)]\,1 \le i \le N \quad (9)$$

Path (state sequence) backtracking

$$qT^* = \psi_{t+1}(q^*_{t+1}), t = T-1, T-2, \ldots 1 \quad (10)$$

The Baum/Welch Algorithm

To adjust the model parameters to satisfy a certain optimization criteria is done by the Baum/Welch algorithm.

$$\xi_t(i,j) = \frac{\alpha_t(i)a_{ij}b_j(o_{t+1})\beta_{t+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_t(i)a_{ij}b_j(o_{t+1})\beta_{t+1}(j)} \quad (11)$$

$$\gamma_t(i) = \sum_{j=1}^{N}\xi_t(i,j) \quad (12)$$

$$a_{ij} = \frac{\sum_{t=1}^{T-1}\xi_t(i,j)}{\sum_{t=1}^{T-1}\gamma_t(i)} \quad (13)$$

$$b_{ij} = \frac{\sum_{t=1}^{T}\gamma_t(j)}{\sum_{t=1}^{T}\gamma_t(j)} \quad (14)$$

Using the final re-estimated A, B and π, the value of LIHMM is calculated with respect to all the word models available with the recognition engine by using Viterbi algorithm. The Viterbi algorithm takes model parameters and the observational vectors of the word as input and returns the value of matching with all particular word models. This is the likelihood values of the word (LIHMM) passed to hybrid training model.

The Gaussian Mixture Model

The GMM can be viewed as a hybrid between parametric and non- parametric density models. Like a parametric model, it has structure and parameters that control the behavior of density in known ways. Like non-parametric model it has many degrees of freedom to allow arbitrary density modeling. The GMM density is defined as weighted sum of Gaussian densities:

$$p_{GM}(x) = \sum_{m=1}^{M} w_m\, g(x, \mu_m, C_m) \quad (15)$$

Here m is the Gaussian component (m=1…M), and M is the total number of Gaussian components. wm are the component probabilities (∑wm = 1), also called weights. We consider K-dimensional densities so the argument is a vector x = (x1, ... , xK)T . The component pdf, g(x, μm, Cm), is a K-dimensional Gaussian probability density function (pdf).

$$g(x, \mu_m, C_m) = \frac{1}{(2\pi)^{K/2}|C_m|^{1/2}} e^{-\frac{1}{2}(x-\mu_m)^T C_m^{-1}(x-\mu_m)} \quad (16)$$

where μm is the mean vector, and Cm is the covariance matrix.

Now, a Gaussian mixture model probability density function is completely defined by a parameter list given by

θ = {w1, μ1, C1... w1, μ1, C1}m m=1…M

Organizing the data for input to the GMM is important since the components of GMM play a vital role in the making of word models. For this purpose, we use K- Means clustering technique to break the data into 256 cluster centroids. These centroids are then grouped into sets of 32 and then passed into each component of GMM. As a result we obtain a set of 8 components for GMM. Once the component inputs are decided, the GMM modeling can be implemented.

EM Algorithm

The expectation maximization (EM) algorithm is an iterative method for calculating maximum likelihood distribution parameter estimates from incomplete data (elements missing in feature vectors). The EM update equations are used which gives a procedure to iteratively maximize the log-likelihood of the training data given the model. The EM algorithm is a two step process:

Estimation Step in which current iteration values of the mixture are utilized to determine the values for the next iteration

$$\gamma(m,t) = \frac{w_m^{(i)} g(X_t, \mu_m^{(i)}, C_m^{(i)})}{\sum_{j=1}^{M} w_j^{(i)} g(X_t, \mu_j^{(i)}, C_j^{(i)})} \quad (17)$$

Maximization step in which the predicted values are then maximized to obtain the real values for the next iteration.

$$\mu_m^{(i+1)} = \frac{\sum_{t=1}^{T} \gamma_{m,t} X_t}{\sum_{t=1}^{T} \gamma_{m,t}} \quad (18)$$

$$W_m^{(i+1)} = \sum_{t=1}^{T} \gamma_{m,t} \quad (19)$$

$$\lambda_{m,j}^{(i+1)} = \frac{\sum_{t=1}^{T} \gamma_{m,t} (x_{t,j} - \mu_{m,j}^{(i+1)})^2}{\sum_{t=1}^{T} \gamma_{m,t}} \quad (20)$$

EM algorithm is well known and highly appreciated for its numerical stabilities under a threshold values of λmin. Using the final re-estimated w, μ and C, the value of LIGMM is calculated with respect to all the word models available with the recognition engine as:

$$L_{GMM} = \frac{1}{T} \sum_{t=1}^{T} \log P_{GM}(x_t) \quad (21)$$

The Hybrid Model

The GMM/HMM hybrid model has the ability to find the joint maximum probability among all possible reference words W given the observation sequence O. In real case, the combination of the GMMs and the HMMs with a weighted coefficient may be a good scheme because of the difference in training methods. The i^{th} speaker independent GMM produces likelihood L^iGMM, I = 1, 2,…, W, where W is the number of words. The i^{th} speaker independent HMM also produces likelihood L^iHMM, I = 1, 2,…, W. All these likelihood values are passed to the so – called likelihood decision block, where they are transformed into the new combined likelihood L' (W):

$$L'(W) = (1 - x(W)) L^{i}_{GMM} + x(W) L^{i}_{HMM} \quad (22)$$

where x(W) denotes a weighting coefficient.

The value of x is calculated during training of the Hybrid Model. In Hybrid Testing, the subset of training data is used and it's HMM & GMM likelihood values are calculated which are combined using weighing coefficient. Static values of weighted coefficient are also used in order to get higher recognition rate.

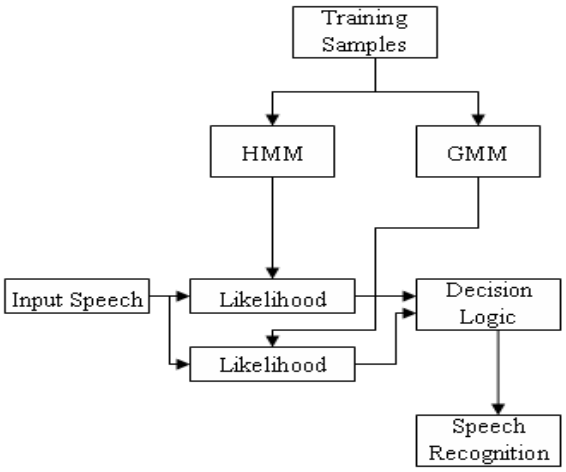


Figure 1: Hybrid Model Block Diagram

Results and Conclusions

The HMM model developed had been tested for different combinations of cluster sizes and number of states. The success rate (%) with different combinations of clusters and states are shown in figure 2. It was observed that as the cluster size increases success rate also increases, because of the reduction in quantization distortion , while increasing the number of clusters. Secondly the time complexity of the recognition engine is also computed with various cluster sizes, and it was observed that 3- state HMM model having 128 cluster size has the least time complexity as comparing with other combinations. Since its time complexity is small it can be useful in applications where speed is the deciding factor. The GMM/HMM Hybrid model created has 8 components and 256 clustering size in GMM and 3 states and 128 clustering size in HMM . Success rate with fixed as well as varying values of weighed coefficient was tested and is shown in figure 4. It was observed, that with the varying value of the weighted coefficient x ,the success rate of word recognition by using hybrid model has increased to 84.38% , which earlier was 83.6% with HMM alone. And by fixing the value of x to 0.9 success rate further increased to 85%. The complexity curve for different cluster size is shown in figure 3.

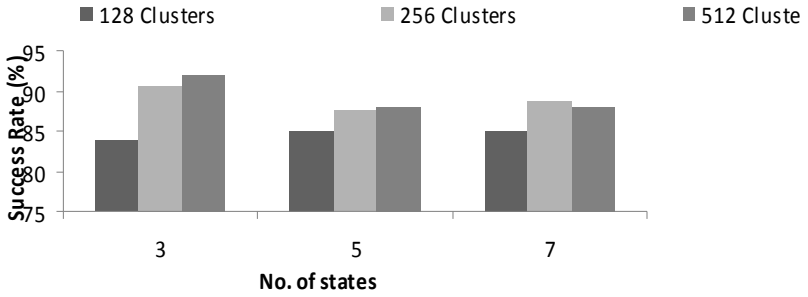


Figure 2: Success Rate verses different combinations of clusters and states

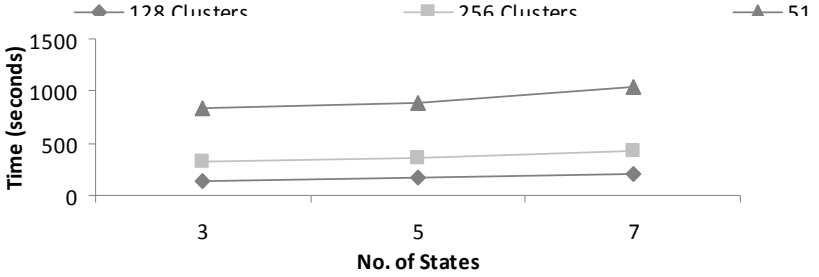


Figure 3: Complexity v/s clusters and states

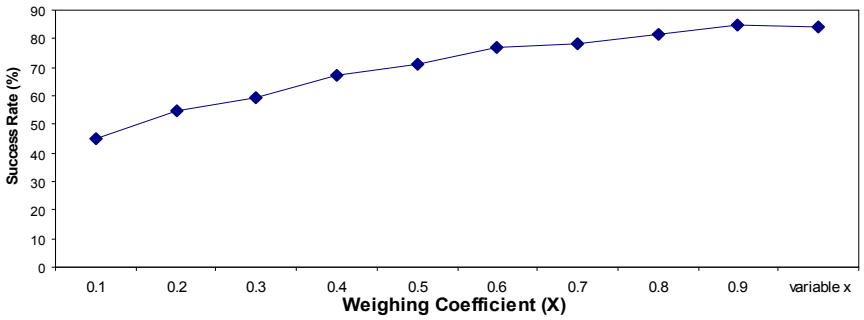


Figure 4: Success Rate v/s weighing coefficient

Bibliography

[Dempster, 1977] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. In: Journal of the Royal Statistical Society: Series B, 39(1) pp. 1–38, November 1977.

[Kotnik, 2002] B. Kotnik, D. Vlaj, Z. Kačič, B. Horvat. Robust MFCC Feature Extraction Algorithm Using Efficient Additive and Convolutional Noise Reduction procedures .In: ICSLP'02 Proceedings, Denver, Colorado, USA pp 445- 448, 2002.

[Rabiner, Juang, 1993] L. Rabiner and B. H. Juang. Fundamentals of Speech Recognition. Prentice Hall, New Jersey, 1993.

[Rabiner, Juang, 1991] B. H. Juang, L. R. Rabiner. Hidden Markov Models for Speech Recognition. In: Technometrics, Vol. 33, No. 3, pp. 251-272, August 1991.

[Rabiner, 1989] L. Rabiner. A tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, In: Proceedings of IEEE Volume 77 No. 2 pp 257-286, February 1989.

[Reynolds, Rose, 1995] D. A. Reynolds and R.C. Rose. Robust text-independent speaker identification using Gaussian mixture speaker models. In: IEEE Transactions on Speech and Audio Processing, 3(1), pp. 72–83, 1995

[Rodriguez, 2003] E. Rodriguez, B. Ruiz, A. G. Crespo, F. Garcia. Speech/Speaker Recognition Using a HMM/GMM Hybrid Model. In: Proceedings of the First International Conference on Audio- and Video-Based Biometric Person Authentication, pp. 227- 234, April 2003

Authors' Information

Poonam Bansal – *Assistant Professor , Department of Computer Science and Engineering, Amity School Of Engineering and Technology, 580, Delhi Palam Vihar Road, Bijwasan, New Delhi 110061, India ; Member ITHEA. e-mail: pbansal89@yahoo.co.in*

Anuj Kant – *Department of Computer Science and Engineering, Amity School Of Engineering and Technology, 580, Delhi Palam Vihar Road, Bijwasan, New Delhi 110061, India.*

Akash Sharda – *Department of Computer Science and Engineering, Amity School Of Engineering and Technology, 580, Delhi Palam Vihar Road, Bijwasan, New Delhi 110061, India.*

Sumit Kumar – *Department of Computer Science and Engineering, Amity School Of Engineering and Technology, 580, Delhi Palam Vihar Road, Bijwasan, New Delhi 110061, India.*

Shitij Gupta – *Department of Computer Science and Engineering, Amity School Of Engineering and Technology, 580, Delhi Palam Vihar Road, Bijwasan, New Delhi 110061, India.*

RESEARCH OF INFLUENCE OF MEDICATION PREPARATIONS ON THE PROCESS OF RENEWAL OF THE BROKEN EQUILIBRIUM OF MAN ORGANISM ON THE DOCTOR OF PHYTOTHERAPY WORKSTATION

Anatoliy Bykh, Elena Visotska, Boris Samura, Olga Kozina, Andrei Porvan, Elena Chernikh, Sergey Parhomenko

***Abstract**: In the given work by authors new approach to the exposure of degree of influencing of medications of vegetable origin in a time of renewal of broken equilibrium of man organism is offered. During realization of the given approach it is suggested to use the mathematical vehicle of.*

***Keywords**: Imitation design, Markov processes, phitotherapy*

***ACM Classification Keywords**: Decision Making*

***Conference**: The paper is selected from Sixth International Conference on Information Research and Applications – i.Tech 2008, Varna, Bulgaria, June-July 2008*

Introduction

During last years enhanceable interest is noted to the alternative methods of medical treatment of man and, especially, to phytotherapy. It is method of medical treatment of man organism by medications of vegetable origin. It can be used as independent or additional type of medical treatment and prophylaxis of different diseases and also rehabilitation of patients with the chronic diseases. A new coil in development of this type of medical treatment demanded the revision of attitude toward one, and also developments of modern methods of leadthrough and estimation of its efficiency. The phytotherapy foresees setting of medical plants. It allows individualizing the process of medical treatment taking into account classification of illnesses, their etiopathogenetic essence, and also to get most and as much possible rapid clinical effect with the minimum of sides actions, that is not always arrived at the use of synthetic medicinal preparations [Самура *2003*].

One of important stages of prophylaxis and medical treatment of different diseases is the ground of expedience of application of one or another type of therapy. As a result the real research directed on the exposure of degree of influencing of medications of vegetable origin on a time of renewal of the broken homoeostasis of man organism is of great scientific and practical interest.

Methods and materials

For research of influencing of synthetic and phytotherapeutic preparations on the process of renewal of the broken equilibrium of man organism the two groups of patients were inspected. Each of which was broken on sub-groups depending on the initial state of patients. The division on sub-groups was conducted according to general classification of the state of man organism [3]:

E1 as the conditionally healthy;

E2 as the initial changes (i.e. the such boundary state of organism, when it isn't impossible to talk about the presence of declining or some concrete disease, but the changes in the usual healthy state are already noted);

E3 as the easy degree of deviation from a standard or easy stage of disease;

E4 as the middle degree of deviation from a standard or middle stage of disease;

E5 as the heavy degree of deviation from a standard or heavy stage of disease;

E6 as the extremely heavy degree of deviation from a standard or extremely heavy stage of disease;

E7 as the fatal outcome.

At primary examination the all inspected were certain in some sub-group from initial five, namely sub-group with the states of E2 – E6. In other words, on the inspection conditionally healthy or dying people did not act. However the medical treatment was conducted to the moment of transition of patient in the state of E1 – conditionally healthy. Also in one group the accident of fatal outcome was noted at providing of medical treatment. In other words, the state of E7 was not initial, but passage is possible to E7. The first group was made by 150 patients who in the moment of primary examination passed medical treatment only by synthetic preparations. The second group consists of 150 patients. This group was conducted medical treatment by vegetable medications. In the depending from state on weight, patients together with the offered medical treatment by vegetable medications adopted synthetic preparations. The dynamic supervision after patients included the clinical, laboratory-biochemical and bacteriologic examinations. Control examinations were conducted one time per a month. On every examination concerned current value of the state of organism. From the chosen current status, the adjustment or establishment of diagnostic and medical strategy was carried out.

For the ground of application efficiency of vegetable medications at renewal of the broken equilibrium of organism the modeling of influence of synthetic and phytotherapeutic preparations on the process of renewal of homoeostasis of man was conducted. On the basis of data about probability of that a patient is in one of five initial states and also data about probability of change of its state on neighboring to the moment of next examination by mathematical one vehicle of theory of Markov processes the numeral values of probabilities of that a patient on a next step will be is in each of seven states were received. Values of probabilities of that a patient in the moment of current examination is in one of seven states, came forward as basic data for the calculation of probability of the state of patient on the next step of design [Медик 2007, Гармоткина 2005].

Human organism as most complex organized biological systems can be represented by the difficult system with the great number of entrances and outputs and this system in different moments of time can be in one k-th of the N possible distinguished standings.

There is also discrete kind of temporal intervals between the separate steps (by stages) of process of diagnostics and treatment.

As is generally known, the system is determined if the vector of probabilities of the initial state of the system $P_k(0)$, and also conditional probabilities of transitions P_{ki} from current state *k* into the *i*-th state during f chosen interval of time Δt, i.e. during one step, are known.

It means that if present probability of a patient was on start point in one of five initial states, and also probability of change of patient's states till the next checking point are known, it is possible to calculate probabilities of being of the system at any *k*-th state on a next step, i.e. at the moment of time $t + \Delta t$ [Токмачев 2003].

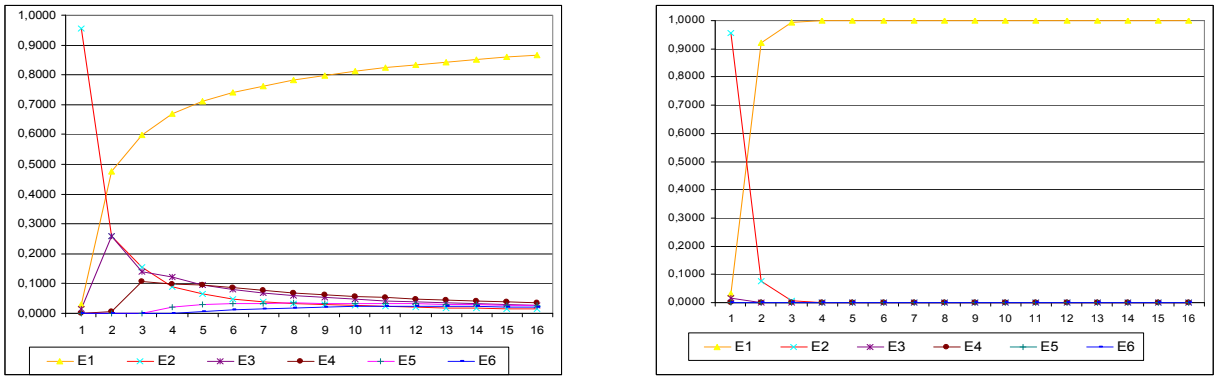
Mathematically this probability will be determined on the formula of complete probability:

$$P_k(t+\Delta t) = P_1(t)\cdot P_{1k} + P_2(t)\cdot P_{2k} + P_3(t)\cdot P_{3k} + P_4(t)\cdot P_{4k} + P_5(t)\cdot P_{5k} + P_6(t)\cdot P_{6k} + P_7(t)\cdot P_{7k}$$

Thus also there is the requirement, that the sum of probabilities of all states is equal 1.

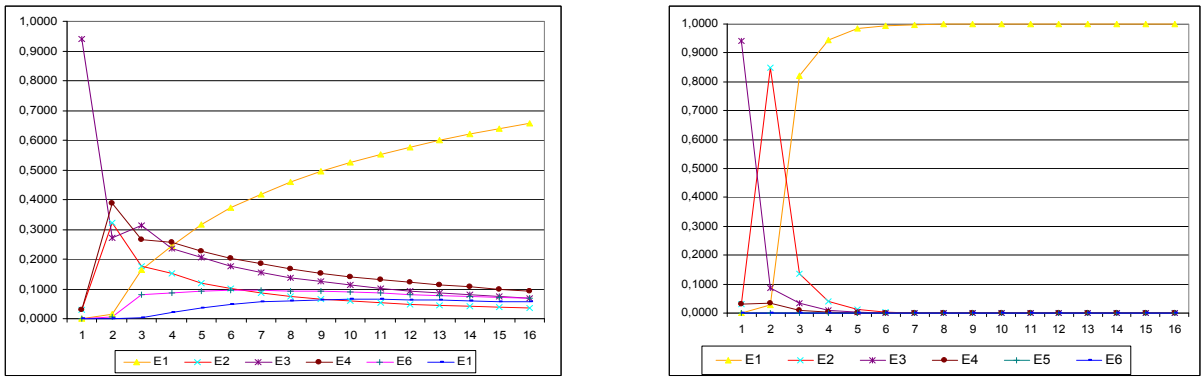
Values of probabilities of a patient in the moment of current checking point is in one of seven states, base as beginning data for the calculation of probability of the patient's state on a next step of imitation.

The results of imitation in form of changes dependences of these probabilities from the initial patient's state in a sub-group and from the got type of treatment are shown on the fig. 1 – 5.



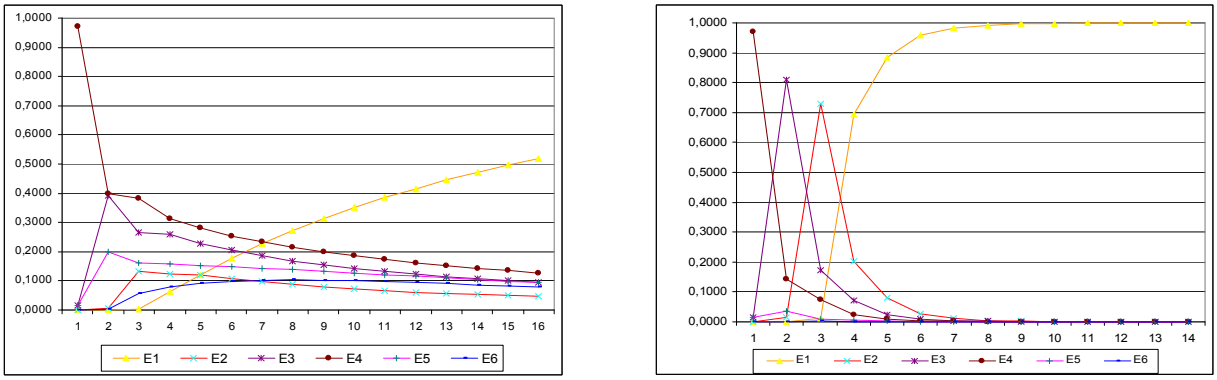
a) b)

Fig.1. Probability of presence of the state of E1–E7 in the first (a) and second (b) groups for sub-groups with the initial state of E2



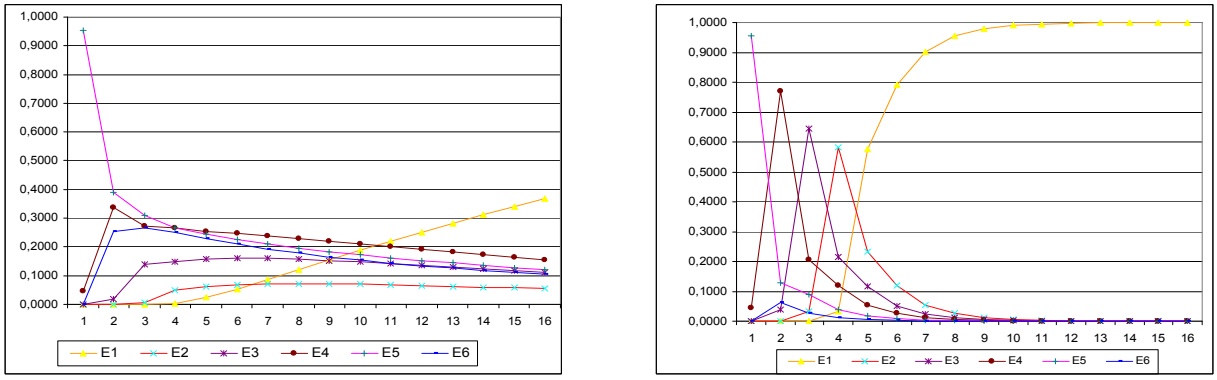
a) b)

Fig.2 Probabilities of presence of the state of E1–E7 in the first (a) and second (b) groups for sub-groups with the initial state of E3



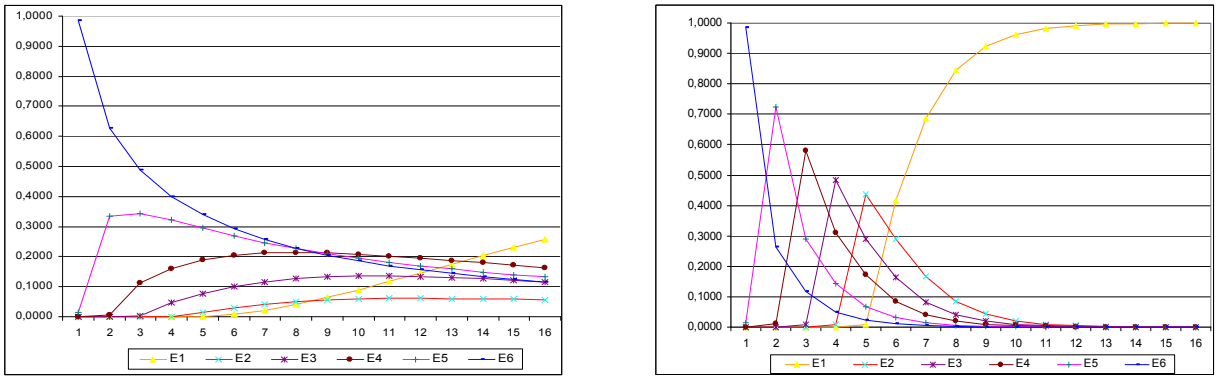
a) b)

Fig.3 Probabilities of presence of the state of E1–E7 in the first (a) and second (b) groups for sub-groups with the initial state of E4



a) b)

Fig.4 Probabilities of presence of the state of E1–E7 in the first (a) and second (b) groups for sub-groups with the initial state of E5



a) b)

Fig.5 Probability of presence of the state of E1–E7 in the first (a) and second (b) groups for sub-groups with the initial state of E6

Analysis and discussion

General review of the received graphs a) from every figure shows on the increase of quantity of steps in medical treatment of patients only by synthetic preparations from the first group, at condition that a patient came on the first examination with more heavy form of disease. Similar situation and with the graphs on figures б) for the second group of patients which got phytotherapeutic preparations. Thus, to talk about a critical moment in the dynamics of the state of patient it is possible on a step, when probability of the «conditionally healthy» state higher than probabilities of all the other states. For example, on Fig.2a this step is the number 5, and on a Fig. 4a this step is the number 12. The number of such step shows minimum duration of medical treatment for every group of patients.

Pair analysis of the graphs of probabilities of the states for two groups of patients at identical initial states of patients shows that minimum duration of medical treatment at patients from the second group shorter already since a Fig 2. I.e. even at presence of easy stage of disease: for the initial state of E3 difference in minimum duration of medical treatment between groups is equal 2 steps. I.e. 2 months, at the initial state of middle degree of declining is difference in 4 steps or in 4 months, and at the heavy degree of declining is already in 9 steps or in 9 months. That indicates on efficiency of the use of phytotherapeutic preparations.

Similarly note should be taken, that for the second group of patients the minimum duration of medical treatment gives more high absolute values of probabilities of «conditionally healthy» state. So, at the initial state of E3 the probability of the state of E1 in the first group at fifth control examination is 31%, and for the second group on the third control examination is 82%. Similar situation for the initial state of E5: for the first group of the state of E1 achieves the value 22% over a year, and for the second group is 58% over 4 months of medical treatment with the phytotherapeutic preparations application.

Fact that for all the time of treatment of patients from first group with, for example, the middle degree of deviation (fig. 3a) till the moment into which probability of state «de been ease healthy» becomes maximal from all possible states, most value of probability at state E4, i.e. of the initial state, is very interesting. In contrary to the second group of patients (рис.3b) already on the second step of treatment with using of phytotherapeutic facilities, probability of more easy degree of deviations becomes most. This tendency is saved on every subsequent step up to a «critical moment», when from all states maximally probably becomes «de been ease healthy».

That is why all graphs of dependences of probability of the states for the second group have a peaks kind, but the graphs of probabilities for the first group of patients are declivous and in the separate places of probability of the «unhealthy» states, i.e. of all except for the state E1, even increase during one-two the steps, for example, steps 2-4 on fig.3a, steps 2-5 on fig.4a and steps 2-6 on fig.5a.

It marks that especially during these stages the results of treatment are extremely unsteady.

Comparison of the models and actual results of states probabilities during renewal of equilibrium disturbance of human organism with using phytotherapeutic facilities showed that exactness of their coincidence on the interval t=3 made 91% (table1) and 86% at treatment by synthetic facilities.

Conclusion

The increased interest to the use of phytomedications in treatment requires additional researches of their efficiency. The results of the conducted research allow getting numeral confirmation of effective application of phytotherapy facilities. The comparative analysis of numeral results of probabilities of the patient's states showed reduction of duration of treatment, more rapid transition from the grave conditions into lights, and also receipt of more credible, i.e. steady positive changes at renewal of equilibrium disturbance of human organism with the additional use of phytomedications. From other side, as well as all other results of modeling, offered models can be efficiently used on prognostication of results of medical tactic.

Bibliography

[Самура *2003*] Самура Б.А., Черных В.Ф., Банный И.П. Фитотерапия в клинике внутренних болезней. Х.: Изд-во НФаУ: Золотые страницы, 2003. – 416

[Медик 2007] Медик В.А., Токмачев М.С. Математическая статистика в медицине: учеб. пособие. М.: Финансы и статистика, 2007. – 800 с.

[Токмачев 2003] Токмачев М.С. Цепи Маркова в прогнозировании медико-социальных показателей // Обозрение прикладной и промышленной математики. Т.10. Вып. 2. М., 2003. – С. 517-518.

[Гармоткина 2005] Гармоткина О.В. Марковская модель заболевания населения // Искусственный интеллект. Вып. 2. Х.: ХНУРЭ, 2005. – С. 22-26.

Authors' Information

Anatoly Bykh *– Doctor of Physics and Mathematics, professor, Head of Biomedical Electronics sub-faculty of Kharkov National University of Radio Electronics*

Elena Visotska *– PhD, lecturer of Biomedical Electronic Devices and Systems Department of Kharkov National University of Radio Electronics*

Boris Samura *– Doctor of medicine, professor of National Pharmaceutical University (Kharkov)*

Olga Kozina *– PhD, lecturer of Computers and Programing Department of National Technical University 'KPI'*

Andrey Porvan *– engineer of Biomedical Electronic Devices and Systems Department of Kharkov National University of Radio Electronics*

Elena Chernikh *– PhD, lecturer of Computers and Programing Department of National Technical University 'KPI'*

Sergey Parhomenko *– PhD student of Computer Software epartment of Kharkov National University of Radio Electronics*

SMART PORTABLE TESTER FOR BIRD FLUE EXPRESS-DIAGNOSTICS: PRINCIPLES OF DESIGN

Oleksandr Palagin, Volodymyr Romanov, Mykola Starodub, Igor Galelyuka, Oleksandra Skrypnyk, Kyryl Skyba

Abstract: *In the V.M. Glushov Institute of Cybernetics of National Academy of Sciences of Ukraine in collaboration with O.V. Palladin's Institute of Biochemistry of National Academy of Sciences of Ukraine the smart portable device for express-diagnostics of acute viral infections, including bird flu, is designed. The device is based on the effect of surface plasmon resonance. The principles of device are described in the article.*

Keywords: *bird flu, portable device, biosensor, surface plasmon resonance, acute viral infection.*

ACM Classification Keywords: *J.3 Life and Medical Sciences – Medical information systems*

Conference*: The paper is selected from Sixth International Conference on Information Research and Applications – i.Tech 2008, Varna, Bulgaria, June-July 2008*

Introduction

Ukraine and others countries suffer from spreading of retroviral infection and others dangerous infectious diseases, including Newcastle disease, bird flu etc. At the same time laboratories for express-diagnostics of listed diseases are absent in the areas of pandemics. This diagnostics is made to order in well-found laboratories, witch are situated in several large cities or others countries. As a result of untimely detection of epidemic every country incurs large losses, and not only animals' lives, but also people's lives are under thread in areas of epidemic.

Taking into account the necessity of medicine in the infectious diseases express-diagnostics of people and animals, such as bird flu, anthrax, tularemia, Newcastle disease and others, the urgent problem now is development, creating and experimental tests of portable biosensor device with GSM-connection and GPS-system for express-diagnostics of acute infectious diseases (including ones of retroviral origin) in the field conditions with transmission of infectious disease place and character by radio-channel.

Because of modern achievements of microelectronics and informatics means it is possible now for the first time in Ukraine to create enough cheap portable biosensor devices on the base of surface plasmon resonance (SPR) for monitoring of especially dangerous infection agents in the field conditions of exploitations. These devices don't need qualified personnel for exploitation. There are no portable industrial devices for express-diagnostics of bird flu and others similar acute viral infections are absent in Ukraine and in the world (see table 1).

Problem description

Taking into account the necessity of wide using of biosensor devices for express-diagnostics of animals and people illness for acute viral infection, such as bird flu, in the places of possible pandemic spreading, in the V.M. Glushkov Institute of Cybernetics of National Academy of Sciences of Ukraine (NASU) in collaboration with O.V. Palladin's Institute of Biochemistry of NASU there is designed of enough simple and cheap biosensor device on the base of SPR for this goal.

Table 1. Comparative parameters of biosensor devices for bird flu diagnostics

Device name	*Vereflu*	*Biosensor for bird flu diagnostics*	*Interference biosensor*	*Biosuplar*	*Biacore 3000*	*Biosensor on the base of SPR*
Developer	*Laboratory "Veredus" (Singapore), company "ST Microelectro-nics"*	*Institute of bioengineering and nanotechnology, Institute of genome, Institute of biomolecular and cellular biology (Singapore), company "ST Microelectronics"*	*Biophysical engineering group from University of Twente city, Holland*	*Germany, www.micro-system.de*	*Sweden, www. biacore.com*	*Institute of Cybernetics of NASU, Institute of Biochemistry of NASU, Ukraine*
Portable	*Yes*	*Yes*	*Yes*	*No*	*No*	*Yes*
Development stage	*Experimental*	*Experimental*	*Experimental*	*Laborato-rial*	*Laboratorial*	*Experimental*
Test time length	*2 hours*	*30 minutes*	*2 minutes*	*1 hour*	*1 hour*	*2-5 minutes*
Cost	*150 USD*	*Not determined*	*Not determined*	*10 000 USD*	*30 000 USD*	*300 USD*

As we can see from table, portable biosensor devices for express-diagnostics of bird flu or other acute infectious diseases viruses are until now on the development stage. It is evidence of urgency of chosen direction.

The features of new biosensor device are the following:

- direct analysis without using reagents;
- express-diagnostics during several minutes;
- analysis with using only 1-2 drops of blood;
- price of analysis no more than 1-2 USD;
- resolution is enough to detect the presence of diagnostic-significant immune components, for example ones typical for bird flu, in threatening quantity in the people or animals blood;
- small sizes and weight (less 0,5 kilogram);
- battery supply;
- field application, for example in the places of gathering of migrant birds or in the places of dearth of poultry;
- processing of analysis data, to show analysis results on build-in display and transfer them by radio-channel to regional medical institutions;
- simple use;
- high productivity – tens of analysis per hour;
- low cost.

Principles of design

The device belongs to area of biological substances researches by analysis of SPR curve shift, size of which is proportional to mass or concentration of biological substance in the research liquid. The device is intended for the express-diagnostics of acute infectious diseases, including bird flu, in the field conditions and data transferring by radio-channel from places of epidemic to medical institutions. In the device it is used SPR effect, according to which the size of SPR curve shift or, it is the same, the change of refractive index is proportional to the antibodies concentration in the ill bird blood, which is placed on the sensitive surface of the device. The device sensitive surface is covered by layer of analyte, which contains special proteins of bird flu (antigens). Ill bird antibodies interact with these proteins.

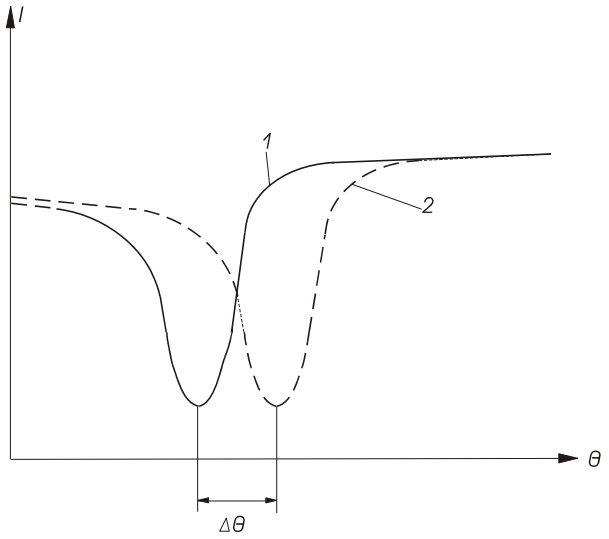


Figure 1. Resonant characteristics of SPR sensor:
1 – with absence of analyte molecules on receptor layer;
2 – with presence of analyte molecules on receptor layer.

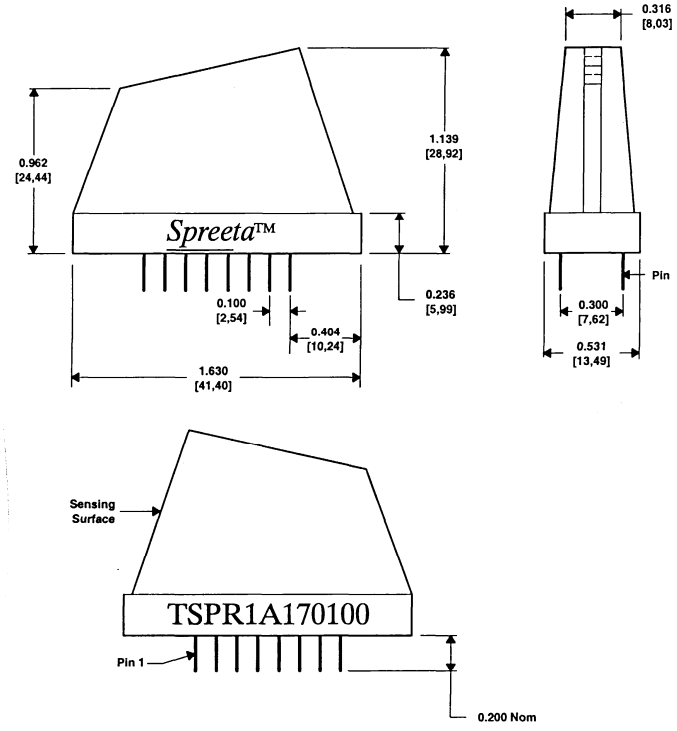


Figure 2. Surface appearance of laboratory-on-crystal *Spreeta*

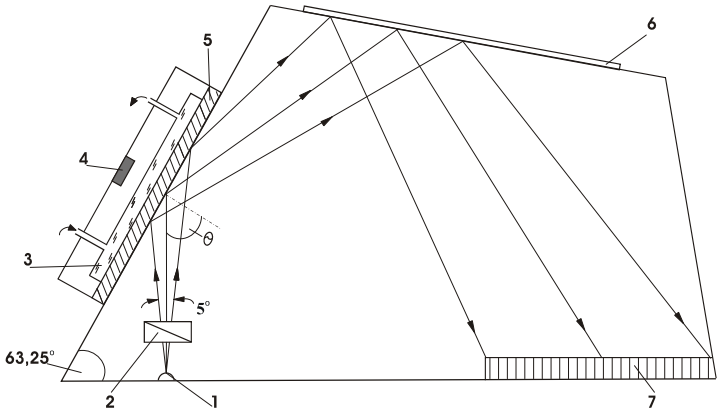
Because of "antigen-antibody" reaction antibodies, if they are present in the bird blood, combine with device sensitive surface, what causes shift of SPR curve (see fig. 1). Date about presence or absence of antibodies, which are got by means of device in the field conditions, are transferred by radio-channel to the medical center.

In the offered device there is used, such called, laboratory-on-crystal Spreeta TSPR 1A170100 (see fig. 2), which is manufactured by company Nomadics, Inc on the base of SPR [Nomadics, Ti]. Spreeta contains prism with sensitive surface. The prism has build-in LED, mirror, photodetectors and logic units. LED exit is connected by means of optical line to sensitive surface of prism, sensitive surface of prism is connected by means of optical line to the mirror, which has optical connection with ruler of photodetectors, exit of which is exit of Spreeta. LED lights up sensitive surface and light beam beats off to mirror. Then reflected beam falls on photodetectors of Spreeta (see fig. 3).

Laboratory-on-crystal Spreeta converts the SPR curve shift to discrete pulses. These pulses are digitizing, processing in the portable device and then data are transferred by radio-channel.

The functional diagram of designed portable device is shown on the fig. 4, which includes Spreeta sensors, analog-digital converter, microprocessor unit for electronic processing, display and radio-channel.

Analysis data are transferred from device to the medical center or laboratory by means of radio-channel. As radio-transmitter it is used the original unit, which is developed by company "VD MAIS" [vdmais]. Besides data analysis also information about place of analysis is transferred by means of radio-channel. For this goal it is used GPS-system, which is built in the radio-transmitter. Principle of operation is based on the transferring all data through radio-channel (by means of GSM-technology) directly to the Internet, and then to the server of medical organization.



1 – LED; 2 – polarizer; 3 – flow-through cell; 4 – thermistor; 5 – sensitive surface; 6 – mirror; 7 – photodetectors

Figure 3. Optical scheme of laboratory-on-crystal *Spreeta*

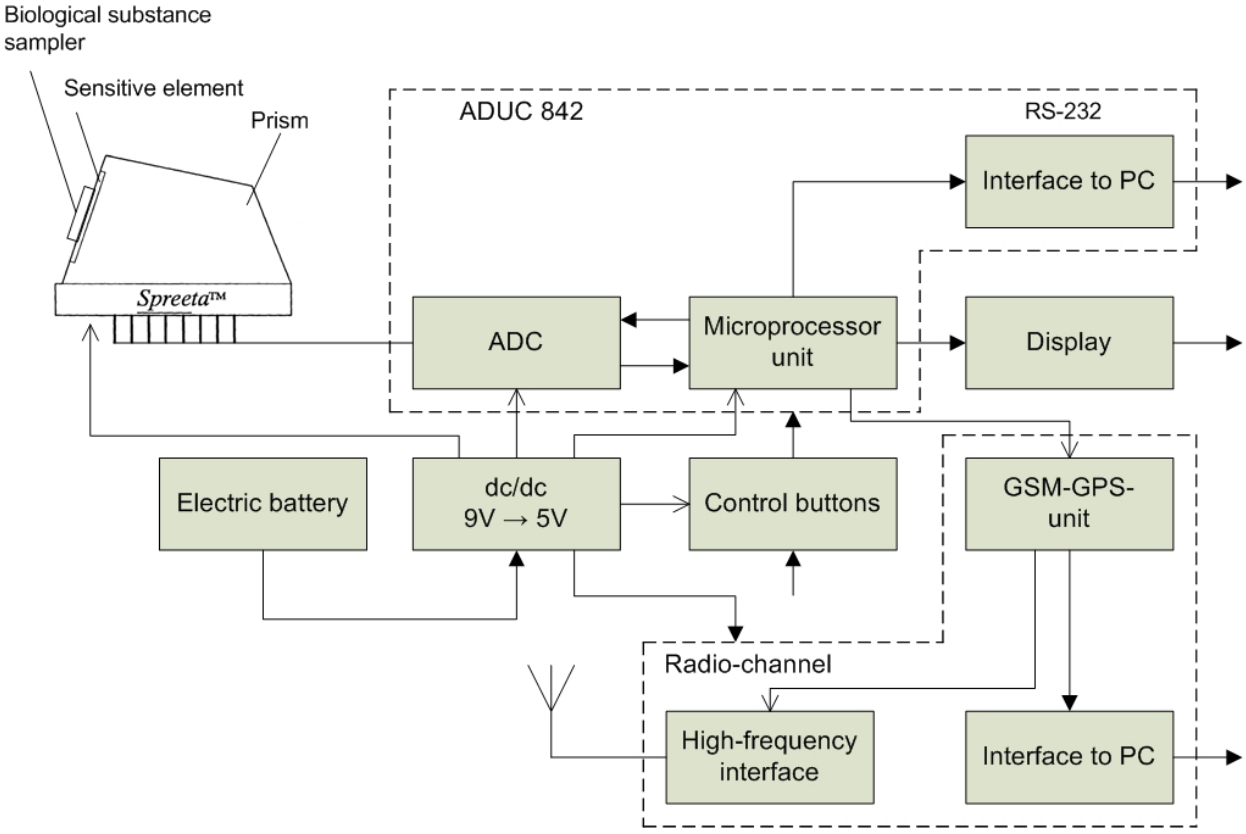


Figure 4. Functional diagram of the device

Conclusions

- as a result of made review of scientific research materials and documentation on mass electronic components and systems we can make next conclusion – off-the-shelf portable devices for express-diagnostics of acute viral infection, including bird flu, are absent on the world market. Now there are fulfilled several enough prospective researches in the some leading universities of the world and conducted tests of experimental samples;

- present off-the-shelf laboratorial devices are not suitable for mass using in the places of pandemic and, in addition, they are very expensive and need high qualified personnel for exploitation;
- specialist of V.M. Glushkov Institute of Cybernetics of NASU together with specialists of O.V. Palladin's Institute of Biochemistry of NASU developed original portable biosensor device with radio-channel for express-diagnostics of acute viral infections [Palagin, 2007; Patent, 2006], including bird flu [Romanov, 2007; Starodub, 2007];
- this time experimental sample of device is under investigation.

Bibliography

[Nomadics] http://www.nomadics.com/.

[Ti] http://www.ti.com.

[vdmais] http://vdmais.kiev.ua.

[Palagin, 2007] Palagin O., Romanov V., Starodub M., Brayko Yu., Galelyuka I., Imamutdinova R., Sarahan Ye. Portable Devices for Express-Diagnostics of Photosynthesis, Viral Infections and Mycotocsicosis // Proceeding of the IV International scientific-practical conference "Urgent question and organizational-lawful aspects of Ukraine-China collaboration in the high technologies field" – Kiev, Ukraine. – 2007, 10 october. – P. 135–138.

[Patent, 2006] Patent of Ukraine № 18099. Device for express-diagnostics of infectious diseases // Romanov V.O., Starodub M.F., Galelyuka I.B. Romanova O.V., Starodub V.M. Bul. № 10 from 16.10.2006.

[Romanov, 2007] O. Palagin, V. Romanov, I. Galelyuka, M. Kachanovska. Virtual Laboratory for Computer-Aided Design of Biosensors // Computing. – 2007, Vol. 6, Issue 2. – P. 68–76.

[Starodub, 2007] Starodub M.F., Romanov V.O., Kochan R.V., Sachenko A.O., Kochan O.V. Implementation of SPR-Biosensors for Express-Diagnostics of Acute Viral Infection and Mycotocsicosis // Proceedings of the IEEE International Workshop on Medical Measurements and Applications MeMeA 2007, Warsaw – Poland, May 4–5, 2007.

Authors' Information

Oleksandr Palagin *– depute-director of V.M. Glushkov's Institute of Cybernetics of National Academy of Sciences of Ukraine, Academician of National Academy of Sciences of Ukraine, Doctor of technical sciences, professor; Prospect Akademika Glushkova 40, Kiev–187, 03680, Ukraine; e-mail: palagin_a@ukr.net*

Volodymyr Romanov *– head of department of V.M. Glushkov's Institute of Cybernetics of National Academy of Sciences of Ukraine, Doctor of technical sciences, professor; Prospect Akademika Glushkova 40, Kiev–187, 03680, Ukraine; e-mail: dept230@insyg.kiev.ua*

Mykola Starodub *– head of department of O.V. Palladin's Institute of Biochemistry of National Academy of Sciences of Ukraine; Leontovycha str., 40, Kiev–30, 01601, Ukraine; e-mail: nstarodub@yahoo.com*

Igor Galelyuka *– research fellow of V.M. Glushkov's Institute of Cybernetics of National Academy of Sciences of Ukraine; Prospect Akademika Glushkova 40, Kiev–187, 03680, Ukraine; e-mail: dept230@insyg.kiev.ua*

Oleksandra Skrypnyk *– software engineer of V.M. Glushkov's Institute of Cybernetics of National Academy of Sciences of Ukraine; Prospect Akademika Glushkova 40, Kiev–187, 03680, Ukraine; e-mail: dept230@insyg.kiev.ua*

Kyryl Skyba *– M2M expert of company "VD-MAIS", Gylyanska str., 29, Kiev, 01033, Ukraine; e-mail: m2m@vdmais.kiev.ua*

KFTKGA: KNOWLEDGE FLOW TRACER AND GROWTH ANALYZER FOR COMMUNITY LEARNING

S.M.F.D. Syed Mustapha, Charles A.H. Chong

***Abstract**: The advances in building learning technology now have to emphasize on the aspect of the individual learning besides the popular focus on the technology per se. Unlike the common research where a great deal has been on finding ways to build, manage, classify, categorize and search knowledge on the server, there is an interest in our work to look at the knowledge development at the individual's learning. We build the technology that resides behind the knowledge sharing platform where learning and sharing activities of an individual take place. The system that we built, KFTGA (Knowledge Flow Tracer and Growth Analyzer), demonstrates the capability of identifying the topics and subjects that an individual is engaged with during the knowledge sharing session and measuring the knowledge growth of the individual learning on a specific subject on a given time space.*

***Keywords**: Knowledge tracing, Knowledge measuring, SECI model*

***ACM Classification Keywords**: H.1.2 User-Machine System*

***Conference**: The paper is selected from International Conference "Intelligent Information and Engineering Systems" INFOS 2008, Varna, Bulgaria, June-July 2008*

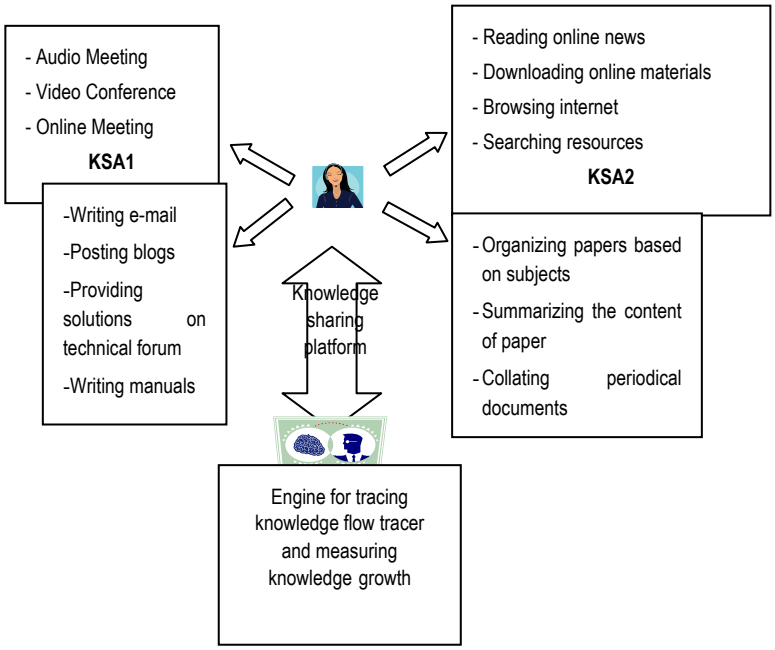
Introduction

In the past decade, there is a continuous growth of interest in knowledge management system as reflected through the research activities reported by the industrialists and academicians. The research in this field is driven by the compelling factors of building knowledge worker, generating knowledge economy and developing knowledge-based society. Organizations are enthusiastic to observe these changes and the impacts to the entire organizational settings, cultures and behaviours. Recently, the hopes for the knowledge management successes have been argued and presumed as hypes or failures. The analyses for these failures have spawned many hypothetical solutions, methodologies and approaches to overcome the shortfall. The implication of these studies still emphasizes the importance of knowledge as the essential element to the modern practices of knowledge-intensive organization. The drawbacks lie within the human factors more than the technology. As such, knowledge management approaches have gradually evolved from technology driven to technology-and-community driven. Our approach, so-called Intelligent Community Informatics (thereafter, ICI) focuses on the three knowledge management components which are the community, knowledge and learning in equal balances. This paper discusses our work where the three components are integrated and operated symbiotically to create the learning community environment by sensing the learning flow of the individual as well as the community and measures the growth of knowledge of different domain. This is demonstrated in our system called KFTGA (Knowledge Flow Tracer and Growth Analyzer) as discussed in section 2. Section 3 describes the application of KFTGA on SECI model environment; Section 4 gives the conclusion of the work.

KFTGA (Knowledge Flow Tracer and Growth Analyzer)

The engine for tracing the knowledge flow and measuring the knowledge growth resides as the backend system incorporated in the knowledge sharing platform. Figure 1 shows the four types of KSAs that an individual like Allen can do in any knowledge sharing applications (such as blogs, forums, e-mails, instant messenger, video calls). There are many knowledge sharing applications that are available in the market that are currently used in

the research laboratory which supports different modes of communication (synchronous or asynchronous, text, audio or video).



KFTGA role is to monitor the knowledge that transpired around the KSA application in terms of the topics and the subject domain that an individual or a community is engaging in. For example, KFTGA can sense whether Allen and her friends share their knowledge through various stages of KSA on a specific subjects or topics. Another way, the system is also able to identify what are the various subjects or topics that linger around the community during the knowledge sharing activities. In addition to that, the KFTGA system can determine which topics or subjects are more dominant than the others. To the best of our knowledge, there is no technology that has been built that integrates the content of the knowledge transpired among various knowledge sharing applications such that the content can be analysed for the purpose of tracing knowledge flow and measuring knowledge growth. The three aspects of content analysis are given below:

1. Relevancy – given the knowledge domain, the system will inspect whether the topics/subjects covered in the knowledge sharing activities are within the ambit of the knowledge domain.
2. Growth – the system periodically examines the size of the knowledge repository enlarging based on the two different point of time on a specific set of knowledge domain.
3. Variation – the system detects whether there are variations of side topics being discussed within the main topics/subjects of the knowledge sharing activities. For example, within the main topic of "knowledge management" there are variations such as "Analytic Hierarchy Process", "collaborative learning" or "ontology".

The content that is captured from various knowledge sharing applications comes in different format needs to be transcribed into text format to be processed and analysed with respect to the Relevancy, Growth and Variation. The challenging parts of the work are

i. to extract the conceptual meaning for each transcription generated from the informal discourse (discussion, blogs, emails or chat) as well as from the formal knowledge sources such as printed materials like manuals, scientific documents, reports. The challenge is that these sources are not predetermined in terms of the format, language style, document structure and presentation.

ii. to cluster the groups of the keywords which belong to the specific topics/subject such that these group can be the determinant factor to detect the relevancy of a given transcription. This approach is not similar to the traditional approach where the word dictionary such as WordNet or Ontology is used to classify the words.

iii. to measure the growth size of the transcriptions on specific subject between two progressive time points from the transcripts which are purely textual information. The measurement of the growth size needs to be translated into numerical value.

iv. to find the conceptual differences between a transcript and the domain knowledge where the difficulties would be to find the subtopics which are included in the main stream of the domain knowledge.

KFTGA technology

This paper addresses the methods of tracing the knowledge flow and measuring the growth of knowledge for community knowledge. The community knowledge is built on the formal and informal knowledge that are transcribed from several forms of communications and knowledge sharing activities such as online meeting, video conferencing, face-to-face, documents uploading and downloading, message posting etc. The reposition of these transcriptions is called community knowledge which evolves over a period of time. Therefore, it is essential to have a mechanism to be able to automatically analyze the content of the knowledge that flows within the repository and also measure the progress and development. By doing this, knowledge sharing attributes such as the learning pattern, the community interest and the knowledge building performance can be gauged. This research work provides three aspects of knowledge tracing and knowledge growth analyzing which are the relevancy, growth and variation. The interesting part is that these values are determined mainly from the textual transcripts generated from formal and informal knowledge sources.

Relevancy

Automated detection on relevancy is difficult since it requires one to understand the subject being discussed in the forum, e-mail, scientific document or report in which they may come in a short or long text. For a short text, there may not be enough words to be analyzed in order to determine the context of the discourse. Other difficulties are that the words usages may differ from one author to another even though they are used to refer to the same context. We avoid the usage of dictionary or word ontology as building them can be time consuming, laborious and costly. Our method allows new subject domain to be added into the community knowledge without the need to upgrade manually the new keywords definition as in the static ontology or dictionary. The KFTGA system builds the keywords definition automatically in incremental manner. The relevancy is determined from the two processes described in Figure 2 and Figure 3.

The set of keywords for a subject domain is built by analyzing formal sources that are published officially such as journals, proceedings, operation manuals, technical reports and others. The characteristics of these sources should be that the contents are specific to a certain topic, there exists finite set of discriminatory keywords that populates the document to represent the subject domain of the document and the representative discriminatory keywords are common to other sources of similar topics.

For example, the keywords such as “socialization”, “externalization”, “internalization” and “combination” are the discriminatory keywords for the articles on “SECI model” topic. The process flow in Fig. 2 produces the discriminatory words extracted from each document and these words are considered as the well-representatives for the chosen topic/subjects. That means if the collections of documents are about “Knowledge management”, then those DW words are the acceptable words being commonly used by the professionals, practitioners and researchers in that domain. The collection of words is now being treated as the pools of words under the category of a specific domain. This approach has an advantage that it can build its own recognizable word list without the need to build a dictionary or ontology which is usually built manually. Once the DW words are compiled, they can be used in the Words Mapper in the form of weight to identify whether a given document is Relevant to the domain as shown in Figure 3.

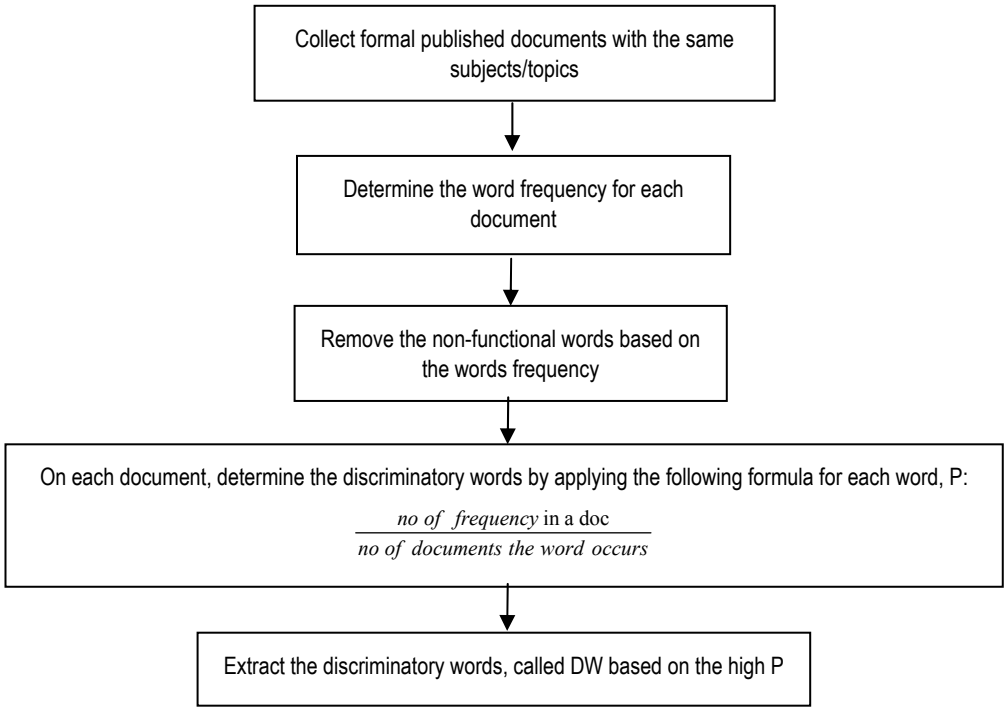
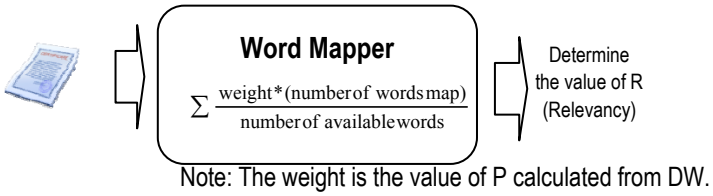


Figure 2. Determine the Discriminatory Words (DW)



Note: The weight is the value of P calculated from DW.

Figure 3. Word Mapper to determine Relevancy

Growth

The engine runs over a period of time that it is possible to measure the development aspect of the resources, so-called growth. There are four dimensions of growth that are The growth is defined as the increment of i) the number of resources that are made available in the knowledge repository (RG) ii) the number of terms that are relevant to the topic/subject (TG) iii) the number of references/usages (RUG) iv) the number of resource maintenance activities (MG).

Resource Growth

$$T_1 = \frac{(R)_{T1} - (\overline{R})_{T1}}{total\ number\ of\ resources}$$

$$T_2 = \frac{(R)_{T2} - (\overline{R})_{T2}}{total\ number\ of\ resources}$$

$$RG = T_2 - T_1$$

R – number of relevant resource

$\overline{R}$ - number of irrelevant resource

where T_1 and T_2 are two different time points and $T_1 < T_2$

Term Growth

$t_i \in (t_1, t_2, ...t_n)$ in T where T is the term in a document D

$$t_{i\delta} = \frac{\left[\left(\sum_1^n 1 \text{ for each } t_i \text{ in D}\right)^{T_1} - \left(\sum_1^n 1 \text{ for each } t_i \text{ in D}\right)^{T_2}\right]^2}{(\text{number of terms in D})^2}$$

$$\text{TG} = \sum_i^n t_{i\delta} \text{ for } \forall \text{i} \in \text{D}$$

where T_1 and T_2 are two different time points and $T_1 < T_2$

Reference Growth

r_{D_j} – the number of references for document D_j

$\bigcup D_j$ – the total number of references for $\forall D$

$$RUG = \left[\sum_{j=1}^{N} \frac{r_{D_j}}{\bigcup D_j}\right]^{T_2} - \left[\frac{r_{D_j}}{\bigcup D_j}\right]^{T_1}$$

where N is number of document and T_1 and T_2 are two different time points and $T_1 < T_2$

Maintenance Growth

$A = \{a_1, a_2 ... a_n\}$ where a is the maintenance activity

$a_i = \bigcup_1^n f(a_i)$ where f is the frequency for activity a

$$\text{MG} = \left[\frac{\sum_1^n a_i}{N_A}\right]^{T_2} - \left[\frac{\sum_1^n a_i}{N_A}\right]^{T_1}$$

where N is the total number of maintenance activity

Variations

For every main concept, there are always sub concepts that are associated to it. For example, if the main concept is VOIP (voice over IP), the possible sub concepts that are associated to it could be “skype”, “economics of VOIP”, “phone adapter”, “Cisco”. The difference between the main concept and the sub-concepts are that the main concept appears as the frequent words in all documents while the sub-concepts appear as the frequent words in some specific documents. This can be determined from the following algorithm.

1. For all $d_i \subset D$ where i = {1,2,...N} and D is a set of documents
2. Calculate f_{d_i} *for* all i on each $t_j \in d_i$, called $f_{d_i}^{t_j}$
3. For each $t_j \in d_i$, *calculate* N where
   N is the number of documents that t_j occur, called N_{t_j}
4. Given P as the total number of documents, then if $\frac{N_{t_j}}{P} > \varepsilon$
   where ε is the threshold, add t_j to list T'
5. Given P as the total number of documents, then if $\frac{N_{t_j}}{P} < \mu$
   where μ is the threshold, add t_j to list T"
6. T' = main concepts and T" = sub concepts

KFTGA application on SECI model

KFTGA uses SECI as the knowledge creation model for its community of learning. SECI model, introduced by Nonaka and Takeuchi [Nonaka and Takeuchi, 1995], has received world-wide attention by many researchers. It proposes four types of knowledge conversion based on tacit knowledge and explicit knowledge as shown in Error! Reference source not found.. Each quadrant represents the conversion type, tacit-to-tacit (socialisation), tacit-to-explicit (externalisation), explicit-to-explicit (combination) and explicit-to-tacit (internalisation). The common understanding about the model is that the knowledge creation will take place as the result of several knowledge conversion spirals.

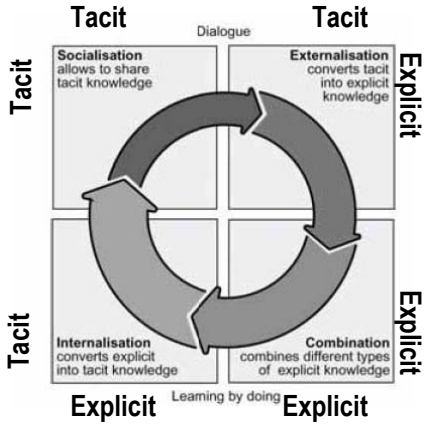


Figure 4 SECI model (adapted from [1]Error! Bookmark not defined.)

The SECI model has received an equal acceptance, rejection and modification from various researchers. Chatti [Chatti, 2007] issues a significant relevancy of SECI model to Web 2.0 by describing the emerging technologies that can be used for different quadrant of SECI model; social media (e.g. Wikipedia (reference), MySpace (social networking), gather.com (social networking), YouTube (video sharing), Second Life (virtual reality), Digg (news sharing), Flickr (photo sharing) and Miniclip (photo sharing)) for *Socialisation*; Discussion channel such as Blogs, Chats, E-mails, IM and Video-conference and any community publishing portals to express opinions such as ratings/voting/feedbacks support *Externalisation*; managing knowledge through a systematic source integration techniques such as RSS/Atom, Pod/vodcasting or any Mashup applications can be considered as *Combination*; and Multi-player gaming and simulations are examples of Web 2.0 applications that fall under *Internalisation*. Hämäläinen [Hämäläinen, 2003] reports the SECI models fits well in the software and research development work where all quadrants are found to be significantly exist in all of the three tested case in the R&D work even though the degree of its applicability may differ. Rice and Rice advocate the possible implementation of SECI model for project in a multi-organisational environment where the employees are not necessarily situated locally [Rice et al, 2005].

KFTGA captures the learning activities that can be characterised by the SECI model in order to trace the knowledge flow evolution of for an individual as well as the group.

Conclusion

KFTGA emphasizes on building the technology where the learning flow of the community can be gauged. The SECI model is used as the reference model for categorizing the four types of knowledge flow phases. The model covers well on every aspects of knowledge sharing activities (KSAs) that could take place on the web environment. It is hypothesized that the KFTGA system that we build could reside behind the knowledge sharing

platform and could perform two essential functions. Firstly, is to trace the knowledge of certain subject or topic that an individual is engaging with during the knowledge sharing session; and secondly, is to measure the growth of knowledge pertaining to the subject or topic. In other words, KFTGA will be able to determine whether an individual has delved a lot on certain topic and the quantity of knowledge the person has invested in.

Bibliography

[Nonaka and Takeuchi, 1995] Nonaka, I and Takeuchi, H (1995) The Knowledge-Creating Company: How Japanese Companies Create the Dynamics of Innovation. New York: Oxford University. http://commerce.concordia.ca/el-sayed/KCreation.pdf

[Chatti, 2007] Chatti, M.A. The Web 2.0 Driven SECI Model Based Learning Process, IEEE ICALT 2007 (to be published) http://en.wikipedia.org/wiki/Social_media

[Hämäläinen, 2003] Hämäläinen, L. Knowledge creation, communication and role of IT in traditional software R&D teams, e-BRF 2003. Abstract: http://www.ebrc.info/kuvat/2020.pdf

[Rice et al, 2005] Rice, J. L. & Rice, B. (2005). The Applicability of the SECI Model to Multi-Organisational Endeavours: An Integrative Review. *International Journal of Organisational Behaviour*, 9, 671-682. http://www.usq.edu.au/resources/vol9no8rice.pdf

Authors' Information

S.M.F.D Syed Mustapha *– Associate Professor in the Faculty of Information Technology, University Tun Abdul Razak, 16-5 Jalan SS6/12 Kelana Jaya, Selangor Darul Ehsan, MALAYSIA; E-mail: syedmalek@unitar.edu.my*

Charles A.H. Chong *– Researcher; BT Multimedia Sdn Bhd, 1B-17, Block 1B, Plaza Sentral Jalan Stesen Sentral 5 Kuala Lumpur Sentral 50470 Kuala Lumpur Malaysia; e-mail: charles.chong@bt.com*

EASY CREATION OF SEMANTICS-ENHANCED DIGITAL ARTWORK COLLECTIONS

**Darina Dicheva, Christo Dichev, Cristan Brown,
Dhanya Veetil, William Thomas**

***Abstract**: In this paper we propose an approach for cost-effective employing of semantic technologies to improve the efficiency of searching and browsing of digital artwork collections. It is based on a semi-automatic creation of a Topic Map-based virtual art gallery portal by using existing Topic Maps tools. Such a 'cheap' solution could enable small art museums or art-related educational programs that lack sufficient funding for software development and publication infrastructure to take advantage of the emerging semantic technologies. The proposed approach has been used for creating the WSSU Diggs Gallery Portal.*

***Keywords**: Cultural Heritage, Semantic Web, Topic Maps, Personalization.*

***ACM Classification Keywords**: H.3 Information Storage and Retrieval - Systems and Software (User profiles and alert services)*

***Conference**: The paper is selected from International Conference "Intelligent Information and Engineering Systems" INFOS 2008, Varna, Bulgaria, June-July 2008*

Introduction

In response to the increasing concern for preserving cultural heritage and making it universally accessible, numerous virtual museum projects have been initiated worldwide. Most of these projects are focused on problems related to using content management systems to enter or convert, catalogue, and index digital content. The content is typically indexed with a mix of local and industry-standard metadata, such as Dublin Core, Categories for the Description of Works of Art, etc.

The potential of the emerging Semantic Web technologies to enhance the efficiency of the search within virtual museum collections and make them interoperable has also attracted the attention of researchers and practitioners in the area. Several projects have been recently investigating how the Semantic Web technologies can be deployed to provide better indexing and search support within large virtual museum collections, including the Dutch Continuous Access to Cultural Heritage (CATCH) Program [CATCH url] that includes the Cultural Heritage Information Personalization Project (CHIP) [CHIP url], Charting the Information Landscape Employing Context Information (CHOICE) Project, etc., the Finish Museum-Finland and Culture-Sampo Projects [Hyv¨onen et al, 2005], the Danish Knowledge Management in Museums Project, etc. These large-scale projects aim to employ semantically richer knowledge representation scheme than traditional metadata schemas in order to support enhanced semantic search and browsing, reasoning, and personalized access to the collections. They typically use W3C Semantic Web languages RDF and OWL to semantically annotate the targeted museum collections, and develop ad hoc frameworks requiring considerable programmers' expertise and resources. It is not easy to reuse the created platforms for new museums/digital art collections.

In this paper we propose an approach for cost-effective employing of semantic technologies to improve the use of digital artwork collections. It highly reduces the involved software development being based on a semi-automatic creation of a Topic Map-based art gallery portal (through the use of our existing Topic Maps tools). Such a 'cheap' solution could enable small art museums or art-related educational programs that lack sufficient funding for software development and publication infrastructure to take advantage of the emerging semantic technologies. We have bed tested our proposed approach in creating the WSSU Diggs Gallery Portal.

The paper is organized as follows. We start with a discussion of the advantages of using the Topic Map technology for developing semantics-navigated Web portals, which made the development of the generic Web portal possible. Section 3 describes our approach to creating a semantics-enhanced virtual art gallery portal. In

Section 4 we discuss the search, navigation, and personalization in the Diggs Gallery Portal, and we conclude in Section 5.

Topic Maps-Driven Semantic Web Portals

The ISO Semantic Web standard Topic Maps (TM) [Biezunski et al, 2000] provides a very suitable technology for semantic structuring and annotation of digital artworks collections and can contribute significantly for improving the efficiency of their searching and browsing. The following backs our proposal:

- Topic Maps offer a standards-based solution for knowledge integration, or Seamless Knowledge [Pepper, 2004]. The Topic Maps data model supports naturally the design of information spaces that offer ontologically rich representations of information, based on heterogeneous information sources, in an integrated fashion.
- The Topic Maps model is designed to be understandable for both machines and humans. This allows for easy and quick creation of Topic Map-based end user interfaces. It also allows the use of some free general purpose topic map tools (such as TM4L) for creating, maintaining, and searching TM-based art collections.
- Topic Maps provide a perfect model for web portals; this allows for quick and easy modification, adaptation and reuse of portals.
    - By implementing the very structure of the portal, a topic map actually drives it: navigating the portal is navigating the topic map.
    - Topic maps support a subject-centric content organization that allows very efficient construction of “topic pages” that contain all the available information in the collection about a particular subject.
- The Topic Maps model supports elegantly both semantic navigation and search:
    - The semantic aspect of Topic Maps is covered with associations. Differently from RDF statements, they are not triples (<subject, predicate, object>), but bind together an arbitrary number of topics. Each of the participating topics plays a specific role. Thus associations support naturally semantic navigation in topic maps.
    - Topic Maps-based collections can offer contextual support to users during their search and query, by helping them to express correctly their information needs so as to find quickly needed information. This is especially important when users don’t know what they are looking for, as well as for small collections where a standard keyword-based search will often fail to find matching objects.
- The fundamental “merge” feature of Topic Maps allows for easy and effective merge of existing heterogeneous information resources while maintaining their meaningful structure. The possibility for merging topics can support:
    - Flexible and efficient re-using and extending of existing collections.
    - Collaborative content authoring.
- Topic Maps can support exchangeability, reusability, and interoperability of information through the identity mechanism built in their model. This is a distributed and democratic mechanism for assigning unique, global identifiers (URIs) known as Published Subject Identifiers, to subjects of common interest, thus making it possible to know when topics subjects are the same.
- The TM concept of scope (context within which a characteristic of a topic may be considered to be true) allows supporting naturally different views or perspectives on a collection. A common use of scope is for providing localized names for topics or defining contexts.

In the context of virtual museums, the Topic Maps model can support subject-centric architecture that is able to accommodate not only digitized images and text, but equally their subjects, themes and movements, their

authors, as well as the people and places, objects etc. mentioned or depicted in those images (or texts). Because of the generality of the Topic Maps paradigm, the conceptual structure can be extended as needed, e.g. to include extra classification schemes such as particular categorization of sculpture artifacts, or to provide more specific types of relationships between images. The subject-centric principles, which are in the center of the Topic Maps paradigm, could have a positive impact on art collections organization. From an authoring perspective, the subject-centric Topic Maps model assumes that the author begins with the subject itself (reified by a particular topic) and extends it with associations and occurrences. Unlike the resource-centric view, in which resources are surrounded by metadata, in the subject-centric view subjects are surrounded by data. A topic map aggregates information to provide binding points from which everything known about a given subject can be reached. This approach affects both the collection creation and information retrieval.

The above features of the Topic Maps technology prove its appropriateness for creating efficient topic map-driven Web portals that support intuitive navigation. Such portals create dynamically the frame structures and content of the rendered webpage from an underlying topic map. Taking an advantage of this semantic technology we have created a generic topic map-driven portal along with a tool for automatic generation of specific portals from the generic one. Our generic portal provides the interface, the topic filtering functionality, and the topic map-directed browsing functionality. The latter supports two different patterns of browsing based on the structure of the underlying topic map. By providing a specific description of the desired presentation categories (indexes) and a specific topic map (an XTM file) to the tool, an author may generate a specific portal from the generic model. Separate portal versions may be generated by specifying different categories that should be explicated by the portal (depending on the application, these could be papers, authors, journals, conferences, or as in this case artworks, artists, exhibitions, etc). The specified categories must be resource or topic types in the corresponding topic map.

Our approach to rapid development of semantics-enhanced artworks portals was to develop such a portal by using our previously developed Topic Maps-related tools: (1) TM4L Editor - a general topic map editor [Dicheva & Dichev, 2006], (2) a generic topic map-driven portal, and (3) a tool for automatic generation of specific portals from the generic one.

As a proof of concept we have created a semantic web portal for the WSSU Diggs gallery [Diggs Gallery url].

The Virtual Diggs Gallery

Diggs Gallery is a small art gallery but offers one of the largest exhibition spaces dedicated to the arts of Africa and the African Diaspora in North Carolina. Exhibitions, publications and programs address a broad range of artistic expression, with special concentration on African-American and regional art. With the creation of the virtual Diggs gallery, our goal was from one side to investigate the appropriateness and easiness of applying the TM technology for creating software for cultural heritage projects, based on Topic Map-based semantic structuring and annotation of digital artworks collections, and from another, to increase the accessibility of the Diggs Gallery artifacts to the general public.

According to our approach, the development of the Diggs Gallery Portal involved:

1. Creating an appropriately structured topic map that contains the semantic annotations of the Diggs Gallery artworks by using our TM4L Editor;
2. Automatic creation of a Web Portal as an instance of our generic Topic Map-based Portal that was slightly adapted to meet the requirements of the Diggs Gallery Portal;
3. Changing the design of the main web page of the portal (optional).

With regard to (1), we designed a topic map model of an artwork collection and using this model created a topic map in the TM4L Editor. Our goal was to organize the collection in a way that helps to highlight the meanings and relationships that exist between the artifacts in the collection and across the collection as a whole. This goal provided the strategic factors forming our design requirement of the model. Fig. 1 displays a screenshot of the Diggs topic map, as created in the TM4L Editor.

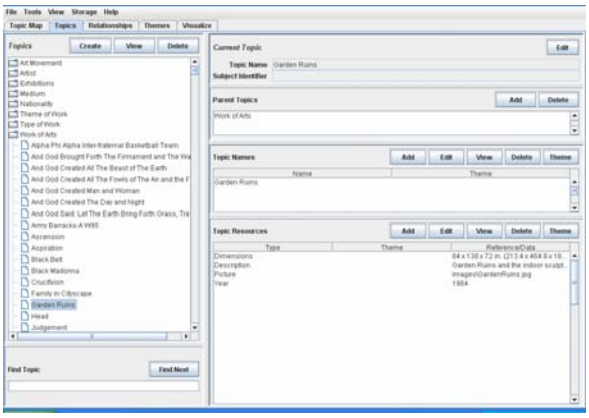


Figure 1. Screenshot from the TM4L Editor used for creating the Diggs topic map.

The topic map model contains the following topics and relationships:

- Topics
  - Art Works
    - Topic name: title of the art work
    - Occurrence types: Year, Dimensions (in inches and centimeters), Narratives / Description, Inscription, Gift from
  - Type of work (e.g. Public Art, Painting, Drawing, Print)
  - Medium (e.g. Acrylic on canvas, Oil on canvas, Granite, Steel, Stone, etc)
  - Theme of Work (e.g. Animals, Still Life, Portrait, Family, Religious, etc.)
  - Collection
  - Exhibition (will be used for constructing the Exhibition History)
  - Art movement (e.g. Renaissance, Impressionism, Graffiti Art, etc.)
  - Artist
    - Topic name: Artist's full name
    - Short (variant name)
    - Occurrence types: Year of Birth, Year of Death, Biography, References (for the biography), Nationality
- Relationships
  - Artist-Work (with roles Artist and Artwork)
  - Type-Work (with roles Type of artwork and Artwork)
  - Medium-Work (with roles Medium and Artwork)
  - Collection-Work (with roles Collection and Artwork)
  - Exhibition-Work (with roles Exhibition and Artwork)
  - ArtMovement-Work (with roles Art movement and Artwork)
  - Theme-Work (with roles Theme of artwork and Artwork)
  - Artist-Nationality (with roles Artist and Nationality)

With regards to (2), we have used our program for automatic generation of a Web portal from our generic Topic Map-based portal. To match the particular art gallery portal requirements this process was followed by minor adaptation involving replacement of the names of the artworks upon rendering with the actual image (as shown in Fig. 2), or with an icon representing the artwork (as shown in Fig. 5).

Figure 2. Diggs Gallery Portal: An artwork view.

Figure 3. Editing functionality in the Diggs Gallery Portal.

Figure 4. Searching Diggs Gallery Portal.

Figure 5. Diggs Gallery Portal: A personalized entry page for a registered user.

Although the generated Diggs Gallery Portal was fully functional and usable, we decided to extend further its functionality so that authorized end users have a publication channel , i.e., can add new artworks or edit the existing information directly from the portal, without a need to download the topic map, edit it in TM4L and upload it again on the site. Thus we implemented an extension to the portal allowing direct authoring of the underlying topic map. The editing menu is displayed only to registered users that are appropriately authorized. Figure 3 displays a screenshot from the Diggs Gallery interface showing the editing menu.

Search, Exploration, and Personalization in the Diggs Gallery Portal

As we already mentioned, the Topic Maps model supports elegantly both semantic navigation and search. Dubbed "the GPS of the information universe", topic maps are destined to provide powerful new ways of navigating large and interconnected corpora – their connectivity can support intuitive navigation and topical search. This is the feature that made possible the proposed topic map-driven generic portal, since it ensured seamless navigation using the topic map connectivity without the need of defining specialized domain-dependent queries to the portal back-end. The resulting portal provides versatile facilities for searching, browsing, and navigating around the art collection. Semantic search based on ontological concepts and semantic browsing allow dynamic presentation of semantic associations between search objects to the end-user as recommendation links with explicit explanations. Thus a portal visitor can navigate and explore the collection in a way dictated by their interests. For example, when seeing an artwork, if the visitor clicks on the Art Movement value for this work, say it is Harlem Renaissance, they will be presented with all artworks of the Harlem Renaissance movement. If they click on the Artist, the artist page will be displayed, etc.

In addition to the topic map-based navigation, for accessing the content of the collection we currently support the two dominant interface types for searching and browsing arts collections: keyword based search, and searching based on grouping and semantic relationships to the other objects in the collection. We have implemented multi-faceted search over type of work, medium, and theme of work. Figure 4 displays a screen shot from the Diggs Gallery Portal illustrating this functionality.

In line with the best practices in the field, we have also extended the Diggs Gallery Portal with personalization functionality. Personalization targets improving museum websites usability by supporting user navigation and assisting users in finding information of interest. To this end, virtual museums automatically adapt the content presentation using user data stored in a user profile. This data typically include user interests, however, personal characteristics, such as age, gender, education, previous knowledge, etc. could support even better comprehension of the collection. Studies have shown that understanding is stimulated when the system uses concepts familiar to the users (considering their interests and knowledge level. User data is typically collected in two ways: (1) users fill in online forms to provide data, and (2) the system monitors users' activities to infer and record their preferences. In addition, [Rutledge et al, 2006] propose an interactive approach for determining user interests in a museum collection with the help of a dialog, which uses artefacts from the semantically annotated collection of the Rijksmuseum Amsterdam to elicit specific user's interests in artists, periods, genres and themes and uses these values to recommend relevant artefacts and related concepts from the museum collection.

Similarly to the leading large-scale projects in the area, the visitors of the Diggs Gallery Portal get personalized view on the collection depending of their user profiles. We have employed similar approaches for collecting user data. Upon registration a new user fills in a registration form, where they can provide some personal data (if they wish so). Our methods are non-intrusive - the system further builds users' profiles by tracking user behavior and mining their logs. As the visitors review art works in the portal (after logging into the system), their activities are tracked and the tracking data are stored in the system's database. Tracking allows finding out how much time a user has spent on reviewing a specific art work and how often he/she has visited the same art work. In addition, we are currently implementing an artwork rating feature that allows visitors to rate their interest in specific facets of artworks, which enhance system knowledge about the user. The user data is used for generating a personalized view of the art work collection for the user (or recommendation of artifacts from the collection). Once

a visitor logs into the system, it analyzes user's data and a personalized list of art works is generated to be displayed to that particular user (see Fig. 5).

The envisaged personalization will improve the Diggs Gallery website usability by supporting user navigation and assisting users in finding appropriate and interesting information.

Conclusion

This paper sketches our experience involving the efforts to improve user access and navigation to a (small-to-medium size) arts collections, by providing flexible exploration model and by personalizing the user view on the collection. The reported work has a twofold objective: On one hand, to investigate how the technologies and tools developed previously from us can be further expanded and repurposed to support semantically enhanced digital art work collection, and on the other hand, to explore the potentials of Topic Maps-driven digital art collections compared to other technologies in terms of low cost systems. As a conventional small art gallery the "proof of concept" repository is composed of a logically related collection of digital objects that can be used to provide various points of information access, giving the visitors a multi-way and multi-purpose navigation path through the collection. The presented experience confirms that Topic Maps provide a suitable model for building portals and other forms of Web-based information delivery.

Future work includes enhancements to content annotation, concept-based organization, and user interface, and use of the architecture as a platform for more powerful personalization and accelerated multipurpose classification.

Acknowledgement

This work is partly supported by the WSSU RIP 2007 Project "Semantic Annotation of Digital Cultural Heritage Collections and LSAMP.

Bibliography

[CATCH url] Continuous Access to Cultural Heritage Program, http://nwo.nl/nwohome.nsf/pages/NWOP

[CHIP url] Cultural Heritage Information Personalization Project (CHIP) http://www.chip-project.org/demo/

[Hyv¨onen et al, 2005].Hyv¨onen, E., M¨akela, E., Salminen, M., Valo, A., Viljanen, K., Saarela, S., Junnila, M, Kettula, S.: MuseumFinland – Finnish Museums on the Semantic Web. J. Web Semantics, 3(2): 224–241 (2005)

[Dicheva & Dichev, 2006] Dicheva, D., Dichev, C.: TM4L: Creating and Browsing Educational Topic Maps, British Journal of Educational Technology – BJET, 37(3): 391-404 (2006)

[Biezunski et al, 2000] Biezunski, M., Bryan, M., Newcomb, S.: ISO/IEC 13250:2000 Topic Maps: Information Technology, www.y12.doe.gov/sgml/sc34/document/0129.pdf

[Pepper, 2004] Pepper, S.: Seamless Knowledge: Spontaneous Knowledge Federation using Topic Maps, In: Proceedings of Extreme Markup Languages, Montréal, Québec (2004)

[Diggs Gallery url] WSSU DIGS Gallery, http://www.wssu.edu/WSSU/About/Administration/Information+Resources/Diggs+Gallery

Authors' Information

Darina Dicheva *– Paul Fulton/Delta Sigma Theta Distinguished Professor of Computer Science, Winston-Salem State University, 601 Martin Luther King, Jr. Drive, Winston Salem, N.C. 27110; e-mail: dichevad@wssu.edu*

Christo Dichev *– Professor of Computer Science, Winston-Salem State University, 601 Martin Luther King, Jr. Drive, Winston Salem, N.C. 27110; e-mail: dichevc@wssu.edu*

Cristan Brown, Dhanya Veetil, William Thomas *– Students, Winston-Salem State University, 601 Martin Luther King, Jr. Drive, Winston Salem, N.C. 27110*

ONE APPROACH TO KNOWLEDGE MAPPING FOR INTERNATIONAL STUDENT PORTAL[1]

Tatiana Gavrilova, HaiNan Jin

Abstract: *Knowledge portal is an approach used to provide view of domain-specific information on the World Wide Web [13]. In this paper, we present one approach by using ontology engineering as a conceptual backbone and relationships for knowledge extracting, structuring and formalizing as a comprehensive way for building knowledge portal. For illustration of a practical ontology development of knowledge portal, the described ideas are implemented in a system design for international student service.*

Keywords: *Knowledge portal, ontology engineering, international student service*

ACM Classification Keywords: *I.2.4 Knowledge Representation Formalism and Method — Semantic Network*

Conference*: The paper is selected from XIV[th] International Conference "Knowledge-Dialogue-Solution" KDS 2008, Varna, Bulgaria, June-July 2008*

Introduction

Today, data and information can be retrieved from many different sources, such as databases, World Wide Web, knowledge bases, and other specific information systems. Using web-based technologies, knowledge portals are developed by providing a single point of access to various types of information [5].

In this paper, we attempt to develop a knowledge portal, which aims to integrate and organize the data/information resources dispersed across web resources [2] by using ontology technique that makes them useful, and a framework of user-centric design for accessing the requested information.

Knowledge Portal

The key to a successful knowledge dissemination strategy is to channel the knowledge to the communities of practice and at the same time provide means for information exchange and peer-to-peer collaboration [9]. One of the models for a virtual collaborative research environment that provides means for both, knowledge sharing and collaboration is the “Knowledge Portal” model. Or we can say that the aim of knowledge portals is to make knowledge accessible to users and to allow users the exchange of knowledge [1]. Knowledge portals specialize in a certain topic in order to offer deep coverage of the domain of interest and, thus, address a community of users.

Ontology Methodology

Traditionally, basic philosophical definition and its further development are pointing that term ontology stands for study of “being” [14]. In information science, today, ontology is a set of distinctions, explicitly made in order to understand and view the world (see Fig.1). There are some of varieties of definitions of this milestone term [7, 8]:

1) Ontology defines the basic terms and relations comprising the structured vocabulary of a topic area, as well as the rules for combining terms and relations to define extensions to the vocabulary.
2) Ontology is an explicit specification of a conceptualization or a hierarchically structured set of terms for describing a domain that can be used as a skeletal foundation for a knowledge base.

This definition clarifies the ontological approach to knowledge structuring while providing sufficient freedom for open-ended, creative thinking [11]. Ontology as a useful structuring tool may greatly enrich modeling process, providing users of KM-systems as an organizing axis to help them mentally mark their vision of the domain

[1] The work is partly supported by RFFI grant 08-07-00062-a

knowledge [4]. For example, ontological engineering can provide a clear representation of a company's structure, human resources, physical assets, and products, and their inter-relationships.

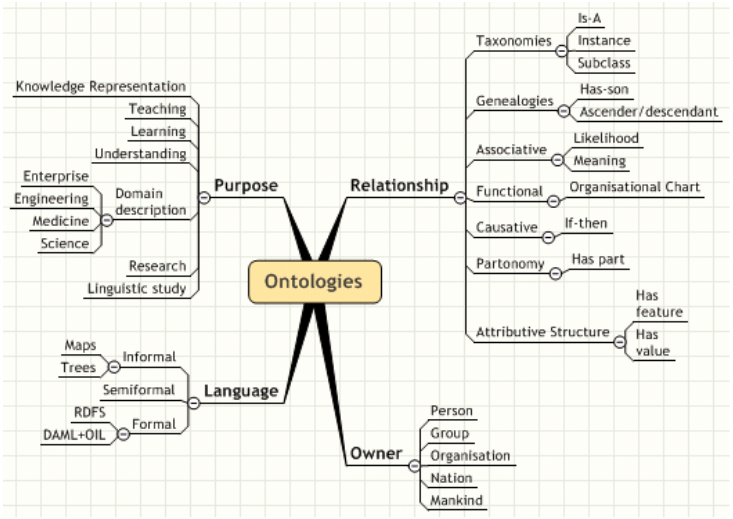


Figure 1. Classification of ontologies

Here we present a method which is called *Four-step Algorithm* [5] to create an ontology:

Step1. Goals, strategy and boundary identification: Identifying the purpose of the ontology and the needs for the domain knowledge acquisition.

Step2. Glossary development or meta-concept identification: This time consuming step is devoted to gathering all the information relevant to the described domain.

Step3. Laddering, including categorization and specification: Having all the essential objects and concepts of the domain in hand, the next step is to define the main levels of abstraction, property and relationship of each concept.

Step4. Refinement: The final step is devoted to updating the visual structure by excluding any excessiveness, synonymy, and contradictions. Meanwhile, we also follow Gestalt (good form) principles by M. Wertheimer [16] to achieve the harmony. As mentioned before, the main goal of the final step is try to create a beautiful ontology.

Knowledge Portal Construction for International Student Service

In this section, for illustration of a practical ontology development for knowledge portal, the described ideas are implemented in a system design for international student service.

International students are continually facing the problem of searching and applying through today's information system which is needed to analyze massive volumes and varieties of data. This situation is not only limited to students, but also took place to organizations and educators. Thus, based on web-based technology, knowledge portal is required and developed to facilitate users finding relevant, domain-specific information by using university database. The portal should provide access to information related to a wide variety of activities [12]. In the first place it will probably concentrate on teaching and learning and student administration, other areas will also needed to be considered: student's life service, financial assistant, passport and visa service, etc.

Members of students and visitors will have different system requirements. But in general, there are some similar parts could be regarded, the following table summarizes the most highly ranked (see Table 1):

Table 1. General Student Service Requirement

Study Information	
	New Student
	Current Student
	Graduating Students
Life Service Information	
	Student Services
	Employment
	Student Union
	Entertainment
Passport and Visa Service	
	Passport Service
	Visa Service
Financial Service	
	Scholarships
	Graduate Assistantships
	Loans

Ontology Design

Following the aforementioned 4-step algorithm, we try to describe the exact practical procedures on each step by representing all the visual structures.

Step 1. Purpose and Goals Identification. It is important to analyze the purpose and proposed usage of the ontology at the beginning of development of process. User requirements analysis is a key part of the user-centered design process, which increases the likelihood that an implemented system matches users' needs and behaviors [6]. If adopted, it could help to overcome some of obstacles to a successful portal implementation of the university information system.

Goals analyze: A university is seeking to provide useful information for international student. Each parts should be integrate, tactic and available. The data will help student to:

- Providing useful and up-to-date information about university and program.
- Helping and guiding news to apply the program and make consultation.
- Information for current and graduating student.
- On arrival service, daily service and activity consultation.
- Passport and Visa service
- Financial Assistant

Step 2. Glossary development or meta-concept identification. The second step is devoted to gathering all the information relevant to the described domain. To achieve this goal, we collected the terms of two sides which are from the point of view of student and point of view of university, expecting to cover all the possible situations that will be referred. The terms and concepts from these sources are combined to build a single glossary (see Table 2).

Table 2. Glossary of terms and concepts

Applications	*Student Union*	*Transportation*	*Russian*
Applicants		*Address*	*Certification*
Applications Flow	*Medical Service*	*Uni. Program*	*Orientation*
Applicants Requirement	*Equipment Network*	*International Interchange*	*International Education*
On Arrival	*Dormitory Service*	*Passport Service*	*Entrance Check-up*
Housing	*Security Service*	*Visa Service*	*FAQ Contact*
Health Insurance	*Activities*	*Scholarships*	*Full-Time Service*
Expenses	*Off-campus Employers*	*Academic Attestation*	*Education Certification*
Finances			

Step 3. Laddering, categorizing and specifying. After creating all essential objects and concepts, we start to build practical ontology of the system. First we built an initial visual structure of the glossary terms based on a set of preliminary high level concepts and the categorization of the glossary terms.

Utilizing MindGenius [10] we build initial ontological categorization or meta-concept architecture from point of view of university that includes all faculties, departments and staffs, it's a complete description of the university framework (see Fig.2).

Figure 2. Main meta-concepts for knowledge structuring

However, the portal needs to be much more attended to the practical problem, composed more precise concepts and hierarchies by analyzing the glossary and previously visual structure [15]. Students from the first visual map can not easily gain the information or knowledge about their university and campus life. Thus, it is should not only from the view of university but more concerned with students.

Meanwhile, there still seems to be some disadvantages that may need to be adjusted further. One of them is too many branches used in the map which may lead visitors to disordered corner. Users sometimes doesn't have enough time and endurance to navigate from one place to another in internet to find information they want, therefore, a specific, ordered and well-structure configuration is more significant [17]. We modify the previous map to make it more effective and humanizing (see Fig.3).

Figure 3. Adjusted portal's structure

Furthermore, we have to think over all the possibilities that may occur when the foreign students are coming to the university, "What's the real help they want to gain from university, not what's the help university wants to give?" [3] – that's the only way to solve the problem.

The following diagram shows us the solution (see Fig.4). The output of this step is a user-centered ontological map, which covers the domain hierarchically.

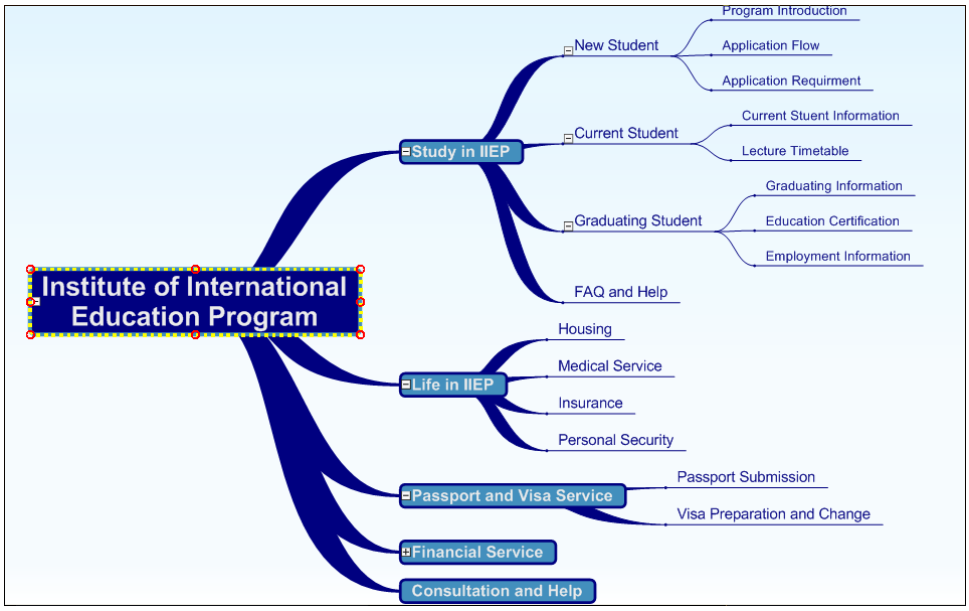


Figure 4. User-centered portal's structure design

Based on the detailed concept map, here we use *Top-Down* structuring strategy to create basic relationships between concepts via filling with term of glossary. The output of is a large and detailed map, which covers the domain hierarchically. The implemented result is shown in Protege 3.3 (see Fig.5 and Fig.6).

Step 4. Refinement and Harmony. To access refinement and harmony, we should update the ontology by taking into consideration of balance and clarity. We removed all the excessiveness, synonymy, and contradictions, then use standard, consistent relationships to simplify understanding.

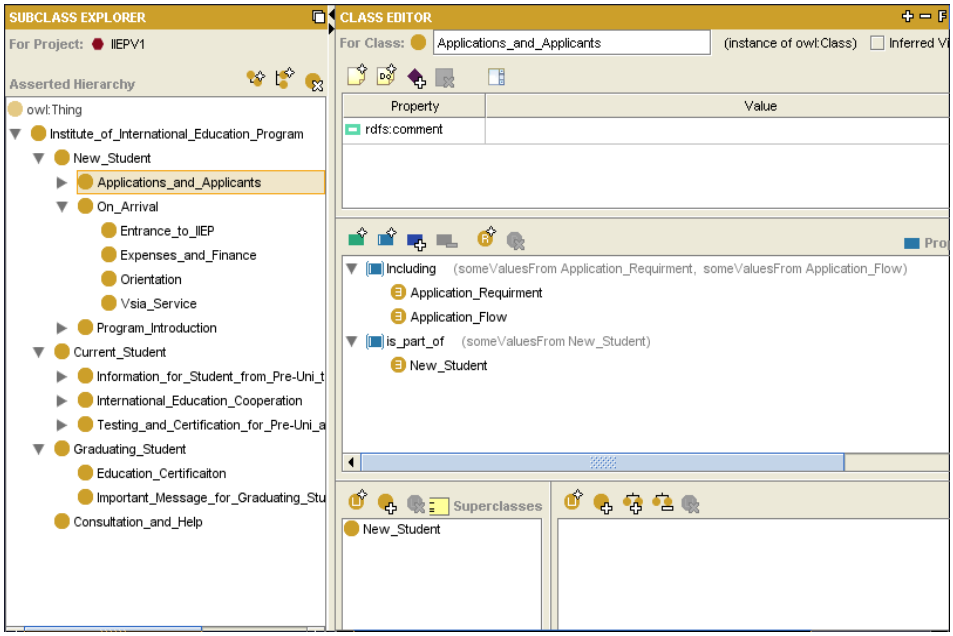


Figure 5. Relationship construction in Protege

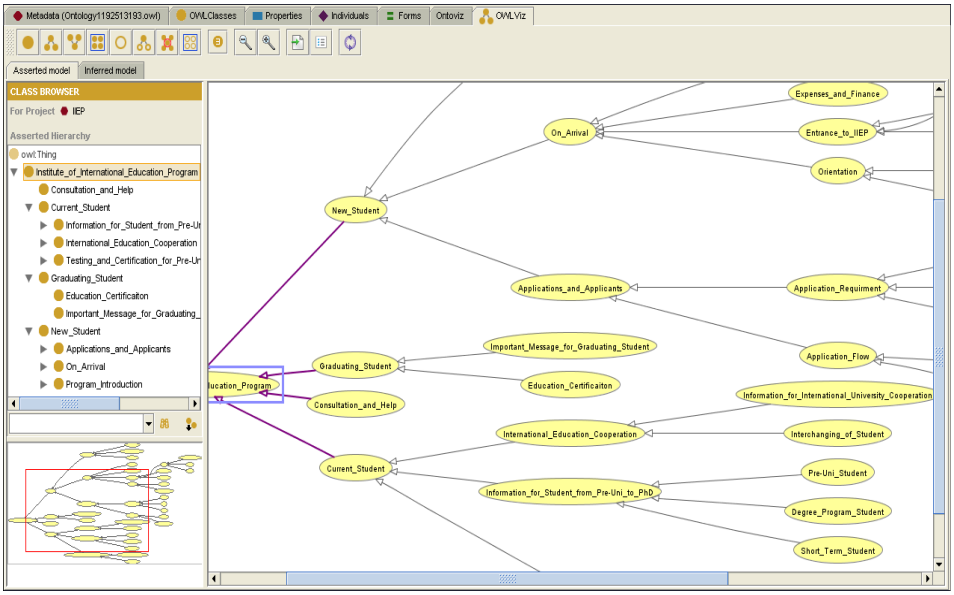


Figure 6. Hierarchical structure construction in Protege

Conclusion

The knowledge portal is a user-centered environment which a user could gain access to information and knowledge from a single internet location. To achieve this goal, we have demonstrated the strategy for designing a system for international student service by using ontology technique which may lay a conceptual foundation and supports for building knowledge portals. The future work is to use web-based technology to implement this design in the university information system.

Acknowledgements

The work was partly supported by grant of Russian Foundation of basic studies N **08-07-00062-a**.

Bibliography

[1] Altmann, R., Bada, M., Chai, X., Carillo, M. W., Chen, R., Abernethy, N. (1999). RiboWeb: *An Ontology-based System for Collaborative Molecular Biology*. *IEEE Intelligent Systems.*

[2] Ankolenkar A., et al (2000). DAML-S: web service description for the Semantic Web. In *The Semantic Web - ISWC 2002*. Springer

[3] Brusilovsky, P., and Rizzo, R.,. (2002), Map-Based Horizontal Navigation in Educational Hypertext. In Proceedings of Hypertext, University Of Maryland, College Park, USA

[4] Gavrilova, T., Brusilovsky, P., Yudelson, M., Puuronen, S. (2006). Creating Ontology for User Modelling Research, Workshop "Ubiquitus User modeling" on European Conference on Artificial Intelligence ECAI 2006, Riva del Garda, Italy.

[5] Gavrilova, T., Gorovoy, V. (2003). Ontological Engineering for Corporate Knowledge Portal Design, In "*Processes and Foundations for Virtual Organisations",* Eds. L. Camarinha–Matos and H. Afsarmanesh, Kluwer Academic Publishers.

[6] Gavrilova, T., Laird D. (2005). Practical Design of Business Enterprise Ontologies, In "*Industrial Applications of Semantic Web" Eds. Bramer M. and Terzyan V*., Springer.

[7] Gómez-Pérez, A., Fernández-López, M., Corcho, O. (2004). Ontological Engineering with examples from the area of Knowledge Management, e-Commerce and the Semantic WebSpringer.

[8] Guarino, N., Giaretta, P. (1998). Ontologies and Knowledge Bases: Towards a Terminological Clarification, In *Towards Very Large Knowledge Bases: Knowledge Building & Knowledge Sharing* . IOS Press.

[9] Martin, P., Eklund. P. (1999). Embedding Knowledge in Web Documents, In *Proceedings of the 8th Int. World Wide Web Conf. (WWW'8), Toronto,* Elsevier Science B.V.

[10] MindGenius. http://www.mindgenius.com/.

[11] Noy N. F., and McGuinness D. L.,. (2001). *Ontology development 101: a guide to creating your first ontology. Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880.* Stanford University Publication, CA.

[12] Staab, S. et al. (2000). Semantic community web portals. *WWW9*. Amsterdam.

[13] Staab, S., Maedche A. (2003). *Knowledge Portals— Ontologies at Work,* Thesis, Karlsruhe, Germany, University of Karlsruhe.

[14] Stamper, R. (2000). *New directions for System Analysis and Design*. In J.Filipe (Ed.), *Enterprise Information Systems*, Kluwer Academic Publishers, Dordrecht.

[15] Turban, E., Aronson, J. E., and Liang, T. P. (2005). Decision Support Systems and Intelligent Systems, Pearson Prentice Hall.

[16] Wertheimer M. *Productive Thinking*. Enl.ed. L.: Ass. Book Publ., 1966.

[17] Wiederhold, G., and Genesereth, M. (1997). The Conceptual Basis for Mediation Services. *IEEE Expert / Intelligent Systems.*

Authors' Information

Tatiana Gavrilova *– Head of Information Technologies in Management Dpt., St. Petersburg State University. Room 618, Grazhdansky pr.28, St. Petersburg, Russia. E-mail: tgavrilova@gmail.com*

HaiNan Jin *– Postgraduate student of St. Petersburg Polytechnic University, Room 713, Grazhdansky pr.30, St. Petersburg, Russia. E-mail: jinhainan@gmail.com*

EMOTIONS AND A PRIOR KNOWLEDGE REPRESENTATION IN ARTIFICIAL GENERAL INTELLIGENCE

Andrey Gavrilov

***Abstract**: In this paper a prior knowledge representation for Artificial General Intelligence is proposed based on fuzzy rules using linguistic variables. These linguistic variables may be produced by neural network. Rules may be used for generation of basic emotions – positive and negative, which influence on planning and execution of behavior. The representation of Three Laws of Robotics as such prior knowledge is suggested as highest level of motivation in AGI.*

***Keywords**: Emotions, neural networks, knowledge representation, hybrid intelligent systems.*

***ACM Classification Keywords**: I.2 Artificial Intelligence – General – Cognitive simulation*

***Conference**: The paper is selected from International Conference "Intelligent Information and Engineering Systems" INFOS 2008, Varna, Bulgaria, June-July 2008*

Introduction

Most sufficient problem in Artificial Intelligence is development of AI functionally similar to human mind. In last time this problem is especially actual in accordance with growth of research and development in intelligent robotics, in particular, humanoid robots. Artificial Intelligence oriented on solving of all tasks by human-like way is named as Artificial General Intelligence [Goertzel and Pennachin, 2007].

Most important issues for development of Artificial General Intelligence are:

- to figure out role and mechanism of introducing of emotions in process of thinking,
- to figure out is it needed to use a prior knowledge and if yes then how to implement one,
- how to implement confabulation, i.e. thinking about future, and connection it with planning and actions,
- how to realize consciousness and thinking about itself,
- how to provide friendly thoughts and behavior of AGI with respect to human beings.

In last time there were published many hypothesis and theories about these [Arbib, 1972], [Pribram, 1971], [Wermter and Sun, 2000], [Hawkins and Blakeslee, 2004], [Hoya, 2005], [Wang, 2006], [Wang and Wang, 2006], [Minsky, 2006], [Hecht-Nielsen, 2006]. Unlike these works we try to connect in one system concepts of emotions, a prior knowledge and friendly to human being behavior.

In this paper we propose the model of AGI based on performance about basic emotions (positive and negative), hybrid architecture consisting of rule-based a prior knowledge representation with linguistic variable (LV) and neural network producing values of LV. In section 1 we are talking about role of emotions in AGI. In section 2 we propose language for representation of a prior knowledge. In section 3 we suggest implementation of Three Laws of Robotics of A. Asimov in proposed language as highest level of motivation in AGI. This a prior knowledge may be expanded by adding rules for logical recognition of “good” or “bad” situations with inputs as result of preprocessing and classification by neural network.

In contrast to another attempts to use the emotions for simulation of mind (for example, in [Goerke, 2006]) we suppose that we have just two main emotions (positive and negative). Many different more specific emotions known from psychology are definition of combination of basic emotions with expression of ones by mimics and other elements of behavior oriented on communication between human beings or with state of organism. Also we suppose that these basic emotions must be produced by a priory determined rules on high level and preprocessing by neural networks on low level. Unlike to these

Role of emotions

A role of emotions in our mind is very wide. There is not definite performance about emotions. It is explained probably by different view on emotions in different sciences. For example, psychologists basically are interesting external expression of emotions in communications between human beings, but it is not enough for developers of AI. For them it is most interesting implementation of emotions as internal states and its influence on behavior. It is possible to look on emotions from different points of view, e.g. influence of emotions on attention, acceleration of decision making, connection between emotions and metabolism, usage of emotion for communications and so on. In this paper we will focus on connection of emotions with motivation or planning.

We suppose that emotions and motivations are very close. Moreover motivation is based on emotions and most sufficient reason of any activity is attraction to positive emotions and avoidance of negative emotions. Thus we have just two general emotions – positive and negative. All other emotions are kinds of these basic emotions with any nuances as result of influence of state of organism (system) and features of interaction with another person.

We believe emotions influence on both selection of goal and achievement of it. The successful process of achievement of goal is reason of positive emotions and present of any unexpected obstacles is the reason of negative emotions.

In figure 1 our performance about connection between perception and generation of emotion is shown. We especially did not write in details unit “Perception and decision making” because this one may be implemented by different ways with hybridization of different paradigms, e.g. neural networks, reinforcement learning, fuzzy logics and so on, and this problem exceeds the bounds of this paper.

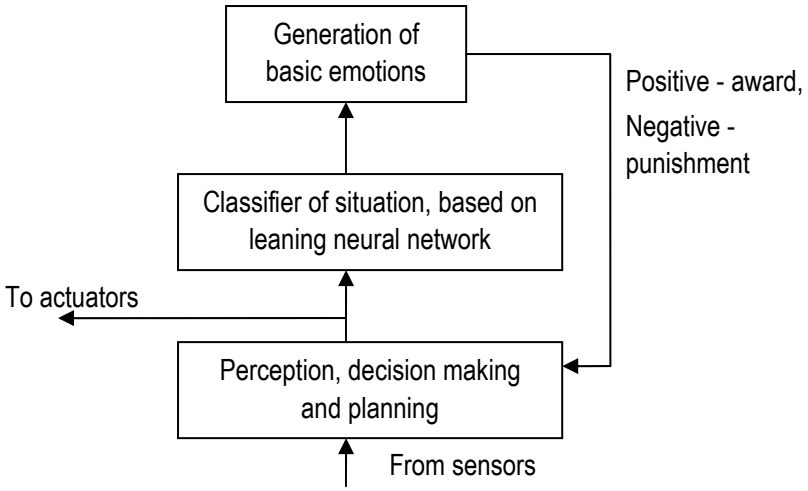


Figure 1. The scheme of connection between perception and emotions

Generation of basic emotions (positive or negative) is based on recognition of “good” and “bad” situations, which may be realized by learning neural network. We suppose that the negative emotions are stimulus for changing of behavior or searching of new plan, whereas the positive emotions are not such crucial for execution of planned behavior. The positive emotions are used as award in any kind of reinforcement learning for storing successful behavior. Produced by such way basic emotions may be stored in associative memory together with any pattern describing an image or situation and may be used for building of plan to avoid or attract to this pattern.

Thus in our performance the highest level of motivation for behavior and control of behavior is the generation of basic emotions - negative and positive.

Representation of a prior knowledge in AGI

It is interesting question how are related in mind learning and a prior knowledge. In beginning of AI this question was not enough actual because developers are dealing with simulation of symbolic intelligence as a prior knowledge and usage of this knowledge for solving of tasks. But later when it was clear that learning by interaction with environment is general property of real intelligence, another extreme approach sometimes is occurred that the mind has not a prior knowledge. But we think that together with development of learning with producing of categories, concepts, rules from learnt neural networks it is needed to provide an opportunity to introduce in AGI a prior knowledge as rules for control of robot at determined situations, for example, like in experiments with simulated mobile robot controlled by hybrid neural network [Gavrilov and Lee, 2007]. In this model of robot the behavior is determined by rules in confusion with obstacle, whereas the movement was controlled by neural network in situations when robot is enough far from obstacles in front. However this is enough primitive example of combination of prior knowledge and learning. For implementation of AGI it is needed more complex representation of a prior knowledge with uncertainty. And a measure of uncertainty of concepts used in rules may be produced by neural networks as result of learning.

We propose follow simple language for representation of a prior knowledge. It is described lower by grammar in BNF notation.

< Rule > ::= If < Antecedent > then < Consequent >

< Antecedent > ::= < Fuzzy symbolic value > | < Function > (<Fuzzy symbolic value >) | < Condition >
| < Antecedent > and | not < Antecedent >

< Condition > ::= < Fuzzy symbolic value > < Relation > < Operand >

< Relation > ::= = | < | > | $\leq$ | $\geq$ | $\neq$

< Operand > ::= < Fuzzy symbolic value > | < Constant >

< Consequent > ::= < Action >

< Function > ::= Increase | Decrease

Here “Fuzzy symbolic value” means value of linguistic variable [Zadeh, 1975]. In this grammar we did not describe the terminal symbols < Fuzzy symbolic value > and < Action > (names of fuzzy variables and actions) depending on concrete implementation of intelligent system. In next section we suggest example with concrete terminal symbols.

Semantics of this grammar is partially determined by fuzzy logics. It means that every fuzzy variable is determined by confidence factor (member function) between 0 and 1, consequent inherits confidence factor of antecedent. But confidence factor of conjunction is calculated as sigmoid function from sum of confidence factors of members of conjunction. Such model of rule is similar to model of neuron in feed forward neural network. Thus if we will not use <condition> as antecedents we can obtain rules very easy extracted from trained neural network where any symbolic value of fuzzy variable is corresponding to output of any neuron and value of output is equal the value of membership function for this symbolic value of fuzzy variable.

We suppose that AGI works in discrete time with any step. Functions “increase” and “decrease” means corresponding updating of value of membership function in compare with previous step.

Three laws of Robotics

There is well known problem to provide friendly to human beings behavior of AGI in robotics. We propose to overcome this problem using implementation of Three Laws of Robotics of A. Asimov as prior knowledge producing positive and negative emotions based on recognition of fuzzy concepts “what is bad” and “what is good” for human beings and robot. These original laws are following:

1. A robot may not injure a human being, or, through inaction, allow a human being to come to harm.
2. A robot must obey orders given it by human beings, except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

Let any classifier of situations provides recognition of follow classes: *Danger_for_man, Danger_for_myself, Command_executed.*

Every of these classes may be viewed as variable with confidence factor or value of linguistic variable with any corresponding value of membership function. It means that if we use, for example, multilayer perceptron as classifier then it has three outputs and value of these outputs may be used as confidence factors.

Variable *Command_executed* shows success of execution of last command got from human being. May be supposed that confidence factor of it is rising during execution of plan.

Let use as actions (< Action > in grammar) the generation of positive and negative emotions “Negative_emotion” and “Positive_emotion”. Then Three Laws of Robotics may be written by rules as:

If *Danger_for_man* then Negative_emotion;

If *Danger_for_myself* and not *Danger_for_man* then Negative_emotion;

If not *Command_executed* and not *Danger_for_man* and not *Danger_for_myself* then

Negative_emotion;

If *Command_executed* and not *Danger_for_man* and not *Danger_for_myself* then

Positive_emotion;

Note that same outputs of these rules may be accumulated in accordance with rules of fuzzy logics.

If we want to get more strong statements of these laws we can append the rules simulating processing of derivative of input variables:

If Decrease(*Danger_for_man*) then Positive_emotion;

If Decrease(*Danger_for_myself*) and not *Danger_for_man* then Positive_emotion;

If Increase(*Command_executed*) and not *Danger_for_man* and not *Danger_for_myself* then Positive_emotion;

Of course this mechanism of implementation of Laws of Robotics does not guarantee friendly behavior of AGI because it is possible incorrect learning of robot to classify situations. However, it is impossible to implement these Laws appropriately and surely for every case in life. For more reliability it is possible to formulate by experts a prior high level knowledge for classification in proposed language. In this case rules based representation of Laws is just minimal knowledge base which may be expanded by rules for estimation of different situations and cases. In this case we can speak about implementation of rules-based logical level of thinking and implementation of associative thinking by neural networks.

Basic positive and negative emotions produced by rules as described above may be used for decision making (selection of action) and planning as award and punishment respectively.

Such implementation of Laws can provide human-like behavior unlike fully determined rules in which it is impossible to clear definite such concepts as “harm”, “danger” and so on. It means that we have not guarantee that robot always will be demonstrate friendly behavior like we see for human beings. His behavior basically depends on careful training.

Conclusion

In this paper we suggest novel performance about implementation of Laws of Robotics as representation of a prior knowledge producing positive and negative emotions influencing on planning and execution of behavior. The language for this representation based on rules and linguistic variables is proposed. The source of information (values of linguistic variables) for rules describing of Laws is classification of situations as result of perception by neural network. The feature of proposed language is an easiness to connect rules with neural networks and to extract rules from trained neural network. It provides the opportunity to grow the rule-based part of AI if it is needed by introducing of expert knowledge or fixing of knowledge obtained by previous training as a prior unchangeable knowledge, in particular, highest knowledge about classification "what is good and what is bad" needed for Laws of Robotics.

Positive and negative emotions produced by rules may be used for planning as awards and punishments respectively for reinforcement learning [Sutton and Barto, 1998].

In future work we plan to implement the architecture of AGI in simulated mobile robots and in characters in game based on proposed in this paper mechanism of representation of prior knowledge and generation of basic emotions.

Bibliography

[Arbib, 1972] M.A. Arbib. The Metaphorical Brain. Wiley-Interscience, 1972.

[Gavrilov and Lee, 2007] A.V. Gavrilov, S,-Y. Lee. Usage of Hybrid Neural Network Model MLP-ART for Navigation of Mobile Robot. Proceedings of International Conference on Intelligent Computing ICIC'07, China, August, 2007, LNAI 4682. Springer-Verlag, Berlin, Heiderberg, 2007, 182-191.

[Goerke, 2006] N. Goerke. EMOBOT: A robot Control Architecture based on Emotion-like Internal Values. In book: Mobile Robots, Moving Intelligence (eds. J. Buchli). ARS/pIV, Germany, 75-94.

[Goertzel, 2006] Ben Goertzel. Patterns, hypergraphs and embodied general intelligence. Proceedings of International Joint Conference on Neural Networks (IJCNN 2006), July, 16-21, 2006, Vancouver, BC, Canada., (2006), 451-458.

[Goertzel and Pennachin, 2007] Ben Goertzel and Cassio Pennachin. Artificial General Intelligence. Springer-Verlag, Berlin, Heidelberg, 2007.

[Hawkins and Blakeslee, 2004] J. Hawkins and S. Blakeslee, On Intelligence, Times Books, 2004.

[Hecht-Nielsen, 2006] R. Hecht-Nielsen. The mechanism of thought. Proceedings of International Joint Conference on Neural Networks (IJCNN 2006), July, 16-21, 2006, Vancouver, BC, Canada., (2006), 1146-1153.

[Hoya, 2005] Tetsuya Hoya. Artificial Mind System. Kernel Memory Approach. Springer, Berlin, 2005.

[Minsky, 2006] M. Minsky. The emotion machine. Simon&Shuster, New York, 2006.

[Pribram, 1971] K.H. Pribram Languages of the Brain: Experimental Paradoxes and Principles in Neuropsychology: Prentice Hall/Brandon House, N.Y., 1971.

[Sutton and Barto, 1998] Sutton R.S., Barto A.G. Reinforcement learning: An introduction. – MIT Press, Cambridge, MA, 1998.

[Wang, 2006] P. Wang, Rigid Flexibility: The Logic of Intelligence, Springer, 2006.

[Wang and Wang, 2006] Yingxu Wang, Ying Wang. Cognitive Informatics Models of the Brain. IEEE Trans. on Systems, Man, and Cybernetics — Part C: Applications and Reviews, 36(2), 2006, 203-207.

[Wermter and Sun, 2000] Stefan Wermter, Ron Sun, Hybrid Neural Systems. Springer-Verlag, Heidelberg, Germany. 2000.

[Zadeh, 1975] L.A. Zadeh, The concept of linguistic variable and its application to approximate reasoning, Inform. Sci. I: 8, (1975), 199-249

Authors' Information

Andrey Gavrilov *– PhD, Associate Professor, Novosibirsk State Technical University, Karl Marx av., 20, Novosibirsk, 630092, Russia; e-mail: andr_gavrilov@yahoo.com*

ONTOLOGY-BASED MODEL OF REPRESENTATION OF KNOWLEDGE ABOUT LANGUAGE MAPPINGS

Margarita Knyazeva, Vadim Timchenko

Abstract: *The paper presents a short review of some systems for program transformations performed on the basis of the internal intermediate representations of these programs. Many systems try to support several languages of representation of the source texts of programs and solve the task of their translation into the internal representation. This task is still a challenge as it is effort-consuming. To reduce the effort, different systems of translator construction, ready compilers with ready grammars of outside designers are used. Though this approach saves the effort, it has its drawbacks and constraints. The paper presents the general idea of using the mapping approach to solve the task within the framework of program transformations and overcome the disadvantages of the existing systems. The paper demonstrates a fragment of the ontology model of high-level languages mappings onto the single representation and gives the example of how the description of (a fragment) a particular mapping is represented in accordance with the ontology model.*

Keywords: *Ontology; Language mappings base; Programming language mappings; Language mappings editor; single representation of program.*

ACM Classification Keywords*: I.2.5 Artificial intelligence: programming languages and software*

Conference*: The paper is selected from XIVth International Conference "Knowledge-Dialogue-Solution" KDS 2008, Varna, Bulgaria, June-July 2008*

Introduction

Traditionally, program analysis, parallelizing and optimizations are applied to programs represented not in high-level languages but by means of different schemata and models that are suitable and convenient representations to work with source programs (e.g. Martynyuk schemata, Lavrov schemata, program representations in the form of various graphs) and are described in detail for example in the works [Voevodin, 2002] [Voevodin, 1992] [Kasyanov, 1988]. Therefore, systems designed for optimizations and parallelizing, especially those that work with several languages of source program texts, have their own single internal representations into which the analyzed program can be transformed and which is used for its further processing.

As a rule, internal representations in systems of program optimization and parallelizing are implemented in such structures as lists, trees, graphs. In some systems, e.g. Polaris [Blume, 1992], SUIF/SUIF2 [Wilson, 1994], Open Parallelizing System (OPS) [Shteinberg, 2004], the internal program representation is realized in the form of class hierarchy in an object-oriented language. The major advantages of this representation are simplicity of designing, modifiability and extensibility.

The internal program representation in OPS is a universal program information data structure and the foundation for design and analysis of information dependencies in the program, program transformations and algorithms to facilitate their execution. The system is supposed to be open to potential changes in the language of source texts due to adding or replacing relatively small programs.

Many systems for designing internal program representation, including OPS, SUIF/SUIF2, use external programs or libraries (Translator Construction Systems (TCS) ANTLR, SableCC, Sage++, Bison, YACC, compilers of Portland Group Inc.) with ready language grammars that makes them dependent on outside designers. Besides, while there is a tendency to improvement of such characteristics of TCS as (i) usability and simplicity of the translator interface that make it possible to integrate it into a software tool; (ii) usability of the description of programming language grammar; (iii) readability of the generated code, yet one may experience difficulty in using them. For example, when one describes own grammars, such tools normally impose constraints on their class and form of specifying that are conditioned by the used parsing method. Moreover, one has to make the effort to

integrate program representations with the help of foreground compilers based on such tools and own internal representations.

Thus, in many systems support of new languages is constrained by the absence of own parser (SUIF/SUIF2, Cetus) which causes regular difficulties in debugging, modifying or integrating into the system of the translator generated by the external TCS or set of classes for its implementation; or by their internal program representation being oriented to the particular language (Polaris, Parafrase).

System of construction of optimizing and parallelizing compilers (SCOPC) [Tapkinov, 2006], a structural predicate system, is an attempt to avoid those constraints. It is a program complex based on the implementation of structural predicate grammars and tools of the structural graph that is an internal program representation used in this system. The main purpose of this system is tools for developing optimizing and parallelizing compilers. Besides, the system can be used for teaching translation methods and carrying research and experiments on developing algorithms and translation methods.

The internal representation is developed for subsets of languages (C, Pascal, Fortran) and is here the result of syntactic and context analysis of the program with the help of structural predicate grammars (SP Grammar).

SP Grammar is a logic grammar in which rules are described in terms of terms and logical formulae. This language can turn out to be difficult for users of this system who are not specialists in mathematical logic. Moreover, the user has to know the internal program representation structure.

The Program Transformation System (PTS) is being developed in the Intellectual Systems Department, the Institute for Automation & Control Processes, the Far Eastern Branch of the Russian Academy of Sciences, to conduct research into program transformations. The support of an extensible set of programming languages of program source texts which must then be translated into the single intermediate representation [Artemieva, 2002] [Artemieva, 2003] is one of the requirements to this system. In order to comply with it the concept of a subsystem of generation of the single internal program representation [Kleshchev, 2007] [Knyazeva, 2008] is introduced as a solution of the problem of multilingualism in program transformations. This approach seeks to avoid the above drawbacks and restrictions in program transformations and parallelizing.

Designing a multilingual system open to new source languages requires especially flexible ways of formalizing information about them. So, to achieve this goal it was decided to use the mapping approach.

The mapping approach to translation started to be developed in early 1970s in the Research Computing Center, Lomonosov Moscow State University. The basic concepts of the approach and expressive means of appropriate description were developed by V. Sh. Kaufman [Kaufman, 1978] [Bunimova, 1978]. By mapping of the language L1 onto the language L2, Kaufman understands a mapping (in its common mathematical sense) $p : I1 \rightarrow I2$, where I1 and I2 are sets of texts acceptable in these languages. The value of the mapping is that it records the way of interpretation of L1 constructions by L2 means and is the most important part of the translator construction task and gives the freedom to choose translation algorithms as it depends on nothing but L1 and L2. Three relations must be fixed to assign the mapping from L1 onto L2: (i) relation that describes the structure of L1 texts; (ii) relation that describes the structure of L2; (iii) relation that describes the connection between those two relations.

Each relation is a system of elementary relations that are expressed in a declarative way by V-language means [Kaufman, 1977]. The significant part of this formalism is based on the apparatus of mathematical logic (predicates) and expressive means of Backus-Naur form.

The practical value of the mapping approach as a technology of constructing translators or interpreters is the design of automated TCS that must deal with a problem of how to construct a translator that implements this mapping according to a formally assigned mapping (translator specifications). The translator implements the mapping if it produces the mapped text according to the projection as a result for each text in the source language. This is the so-called problem of the solvability of mappings. Kaufman's approach did not solve this problem.

This paper presents a way of describing language mappings that can be easily used by specialists who lack mathematical training. On the other hand, it is oriented to the implementability of the above mappings, i.e. possibility in principle of constructing an interpreter that would generate this program in the single representation on the basis of the description of the mapping of a programming language onto the single internal representation

and of the program in this programming language. Thus, in this case there is no problem of mapping solvability as all the described mappings are known-solvable due to the way of representing their description.

The paper was written with financial support from the Far Eastern Branch of the Russian Academy of Sciences (initiative-based research project "Internet system for controlling information about program transformations" – 06-III-A-01-007) and by the Russian Foundation for Basic Research (project 'Research into collective control in the semantic web of information resources of different levels of commonality' – 06-07-89071-a).

Implementation-oriented approach to description of language mappings. The main ideas.

The first idea of the implementation-oriented approach to the description of language mappings is to develop and describe the ontology model of a programming language for which its mapping onto the single internal representation is written.

The programming language ontology is a set of information that describes a set of concepts of this language and interrelations between these concepts, i.e. the way of uniting them into language constructions. It can be said that the language ontology defines the abstract syntax of this language. The model of the language ontology is represented by the semantic network of concepts connected with each other by directed arcs and provided with a special markup.

The second idea is to describe the connection between the language ontology and concrete syntax of this language that defines the content of the correct language constructions from the point of view of the syntax of this language. The connection between the language ontology and the concrete syntax of this language contains such language elements as the punctuation content, defines the lexeme order and so on. This kind of information restricts the way of expressing the sense which is in the abstract program representation in the form of a text.

These models are necessary for the parsing of program texts and for representations of these programs in the abstract syntax. The program in the abstract syntax is a program represented in terms of the programming language ontology that does not contain elements of the concrete syntax.

The third idea is to develop and describe the model of ontology of programming language mappings onto the single representation in accordance with which each mapping is described.

The above formalization makes it possible to develop the conception of subsystem of generation of the single internal program representation that consists of easily customizable to the correspondent model of the ontology of program components and ontologies controlled by the correspondent models that can be changed if necessary [Kleshchev, 2007] [Knyazeva, 2008].

Model of the ontology of language mappings onto the single representation

By the projection, the following mapping is understood:

$$\mathbf{P}: C \rightarrow E, \text{ where}$$

C = {concepts} – a set of concepts of the programming language in terms of which the program is represented;

E = {elements} – a set of elements of the single representation language.

Below is the fragment of the ontology model of programming language mappings onto the single representation. The model is a specification of the abstract syntax of the language of the mapping description. The description of the operational language semantics which is oriented to an interpreter is not represented here, but its informal description for the user is given.

To describe we used the following symbols of the language of specifications which is used in the work [Yershov, 1977]: **unit** – uniting; [a] – a is not necessary (can be absent); = – concept definition; **:** – position opening, * – possibly indefinite attribute; **ser** – serial component.

Mapping description language = (**ser** programming language **:** Mapping description)

Description: Mapping description language specifies the rules of description of mapping of the assigned programming language onto the single program representation.

Semantics:

Mapping description = (**ser** correspondence **:** Correspondence description)

Description: Mapping description specifies a set of correspondences of the concepts of the assigned programming language with the constructions of the single program representation language.

Semantics:

Correspondence description = (programming language concept **:** STRING, construction of the single representation language **:** Elements of the single representation language)

Description: In the correspondence the structure of the construction of the single representation language that must be correspondent to the concept of the source programming language, i.e. the concept of the programming language is correspondent to a set of elements of the single representation language structured in a specific way.

Semantics: To interpret the description of the structure of the construction of the single representation language in this correspondence for the specified concept of the programming language.

Elements of the single representation language = ([fragments **:** Set of fragments], [control arcs **:** Set of control arcs], [attributes **:** Set of attributes])

Description: The construction of the single representation language consists of sets of fragments, control arcs and attributes that can be empty empty.

Semantics: To develop the construction of the single representation language with the specified structure.

Set of fragments = (**ser** fragment **:** Fragment class)

Fragment class = **unit** (*Expression*, *Program block*, *Conditional statement*, *Loop_with_step, Loop_ with_precondition, Loop_with_postcondition, Procedure_call, Dynamic_variable_elimination*, *Assignment, Input, Output*, *Description_of_one_variable*,

Description_ of_one_function,

Description_of_one_parameter, *Block_of_descriptions_of_variables, Block_of_descriptions_of_functions, Block_of_descriptions_of_parameters*, Name of other fragment class)

Name of other fragment class = (name **:** STRING)

Description: *Set of fragments* is a set of names of fragment classes (they are represented by a component *Fragment class*) specified in the single representation language. If during the mapping description there are not enough names of fragment classes specified in the core of the language of names of fragment classes, one can define a new fragment class (by specifying its name).

Thus defined fragment classes are added to the specialized (for the particular programming language) extension of the single representation language which, if necessary, is specified during the mapping description of the programming language onto the single representation.

Semantics: To gradually create fragments a set of names of which is defined by a component *Set of fragments*.

Set of control arcs = (**ser** control arc : Control arc)

Description: *Set of control arcs* is a set of control arcs.

Semantics:

Control arc = (arc name : **unit** (*If, Then, Else, Condition_of_loop, Starting_boundary_of_loop, End_boundary_of_loop, Step, Body, Description_of_parameters*, *Description_of_variables*, *Description_of_functions*, *Right_expression, Left_expression, Parameter_list, Fragment_ancestor,* Other name of control arc), fragment_arc start point : **unit** *(Program_block, Conditional_statement, Loop_with_step, Loop_with_precondition, Loop_with_postcondition, Procedure_call, Dynamic_variable_elimination, Assignment, Input, Output*, *Description_ of_one_function*,

Block_of_descriptions_of_variables, Block_of_descriptions_of_functions, Block_of_descriptions_of_parameters, Name of other fragment class),

fragment-arc end point : **unit** (*Expression, Program_block, Conditional_statement, Loop_with_step, Loop_with_precondition, Loop_with_postcondition, Procedure_call, Dynamic_variable_elimination, Assignment, Input, Output*, *Description_of_one_variable*, *Description_ of_one_function*,

Block_of_descriptions_of_variables, Block_of_descriptions_of_functions, Block_of_descriptions_of_parameters, Name of other fragment class))

Another name of control arc = (name **:** STRING)

Description: *Control arc* is a directed arc that is characterized by its name and connects two fragments. The arc start point is a fragment defined by the selector “fragment_arc start point”. The arc end point is is a fragment defined by the selector “fragment_arc end point”.

If during the mappping description there are not enough control arcs specified in the core of the language of control arcs, one can define a new control arc by specifying its name – a component *Name of control arc* defined by the selector “arc name”, and initial and finite fragments which should be connected by it.

Thus defined control arcs are added to the specialized (for the particular programming language) extension of the single representation language which, if necessary, is specified during the mapping description of the programming language onto the single representation.

The component defined by the selector “fragment_arc start point” can be one of the mentioned fragments. The component defined by the selector “fragment_arc end point” can be one of the mentioned fragments.

Semantics: To create a control arc with the name defined by the selector *arc name*, make a fragment defined by the selector *fragment_arc start point* the initial fragment of the control arc and a fragment defined by the selector *fragment_arc end point* – the finite fragment of the control arc.

If a fragment defined by the selector *fragment-arc end point* does not exist (has not been created) on this interpretation step of correspondence description, it will be created on the basis of description of the next interpreted correspondence and assigned as the finite argument of this control arc.

Set of attributes = (**ser** attribute : Attribute)

Description: *Set of attributes* is a set of attributes.

Semantics:

Attribute = (attribute name : **unit** (*Identifier, Declaration_statements, Loop_counter, Left_part_type, Left_part_of_expression, Right_part_type, Variable_not_parameter, Right_part_of_expression, Is_function, Operation_symbol, By_value, Is_array, Is_pointer, Address_expression, Variable_result_of_function, Original_string_name, Type, Lower_bound_of_array, Upper_bound_of_array,* Another name of attribute),

argument of attribute : **unit** (*Expression, Program_block, Conditional_statement, Loop_with_step, Loop_with_precondition, Loop_with_postcondition, Procedure_call, Dynamic_variable_elimination, Assignment, Input, Output*, *Description_of_one_variable*, *Description_ of_one_function*, *Description_of_one_parameter*

Name of other fragment class),

computable : LOG,

[attribute value : Value])

Another name of attribute = (name : STRING)

Value = **unit** (STRING, INTEGER, REAL, LOG, Fragment class)

Description: *Attribute* is characterized by its name, argument for which it is defined and value it can get. Besides, it can be computable or non-computable.

The argument of the attribute can be one of the mentioned fragments. The attribute value is a value of string, integer, real, logical type or it can be represented by the component *Fragment class.* “Computable” means either that it is necessary to compute the value of an attribute in the process of program scanning in the programming language or it is not necessary because the attribute will assume the value specified directly during the mapping description.

If it is necessary to compute the value of an attribute, it is supposed that this information is always directly accessible (without using any program flow analysis methods) during the lexico-syntactical analysis of the source program.

If during the mapping description there are not enough attributes specified in the core of the language, one can define a new attribute by specifying its name – a component *Name of attribute* defined by the selector “attribute name”, argument defined by the selector “attribute argument”, meaning of indicator about the necessity of computing the value defined by the selector “computable”, and the value defined by the selector “attribute value”.

Thus defined attributes are added to the specialized (for the particular programming language) extension of the single representation language which, if necessary, is specified during the mapping description of the programming language onto the single representation.

Semantics: To create an attribute with the name defined by the selector *attribute name* the argument of which is defined by the selector *attribute argument*, to assign the component *Value* defined by the selector *attribute value*.

If the value if an attribute defined by the selector computable is true, the value defined by the selector *attribute value* must be computed at the stage of program scanning in the programming language.

Fig.1, fig. 2 and fig. 3 demonstrate the examples of correspondence description between Pascal concepts, such as “Expression”, “Assignment statement”, “Variable name”, and constructions of the single representation language corresponding to them.

These descriptions are a fragment of description of the mapping of the Pascal programming language onto the single internal representation and are made in compliance with the above ontology model.

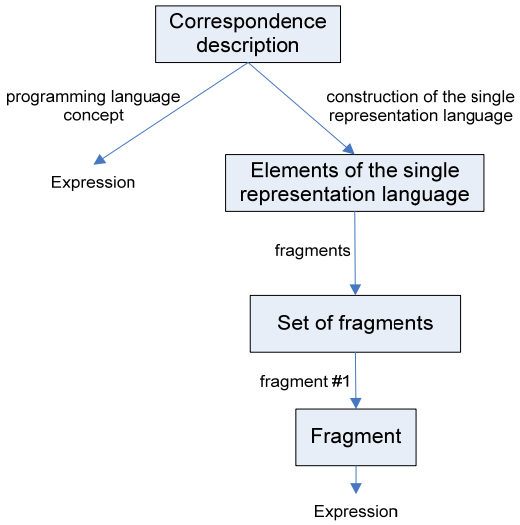


Fig. 1. Description of correspondence between Pascal concept “Expression” and construction of the single representation language.

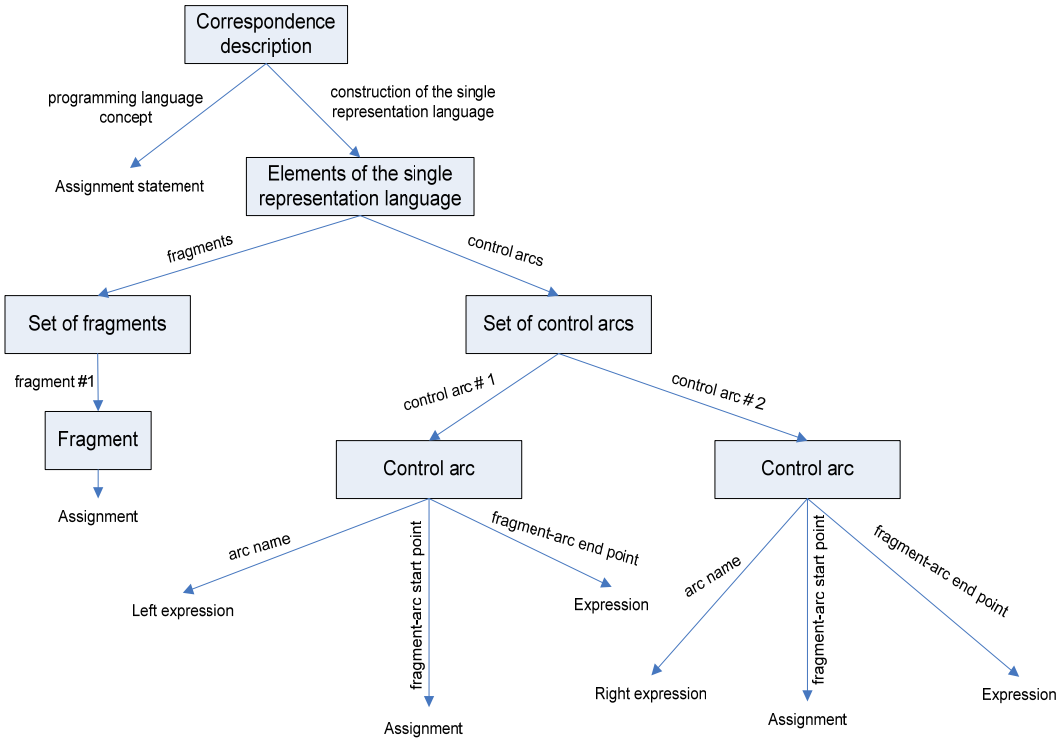


Fig. 2. Description of correspondence between Pascal concept “Assignment statement” and construction of the single representation language.

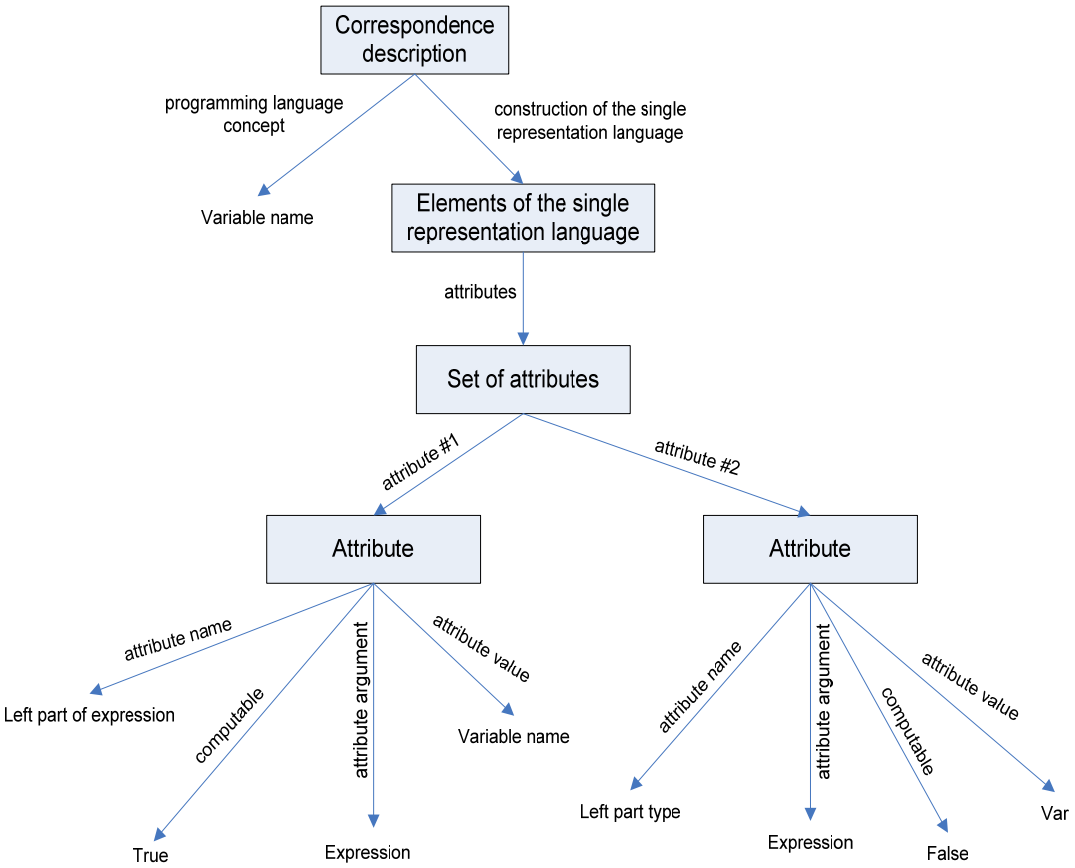


Fig. 3. Description of correspondence between Pascal concept "Variable name" and construction of the single representation language.

Conclusion

The paper presents a short review of some systems in which analysis, optimization and parallelizing are performed on the basis of the internal intermediate representation of these programs. Many such up-to-date systems support several languages of representation of source program texts and, thus, solve the task of their translation into the internal representation using some approaches. Being quite a challenge, this task is complicated by the fact that to make the internal representation as usable and efficient as possible for the flow analysis, optimization and parallelizing becomes more and more important and the translation itself is considered as a subordinate task that can be done with the help of ready tools of outside designers. Thus, solving this problem is still connected with some difficulties and the approaches, being closer and closer to the solution, still have their drawbacks and constraints. The paper also deals with the general idea of using the mapping approach to solve the problem of multilingualism in PTS that can help overcome disadvantages of the existing systems. The paper demonstrates a fragment of the ontology model of high-level languages mappings onto the single representation and gives the example of how the description of (a fragment) a concrete mapping is represented in accordance with the ontology model.

Bibliography

[Voevodin, 2002] Voevodin V.V., Voevodin Vl.V. Parallel computing. Saint Petersburg: BHV-Peterburg, 2002. (In Russian).

[Voevodin, 1992] Voevodin Vl.V. Theory and practice of research on parallelism of sequential programs // Programmirovanie. 1992. #3. P. 38-54. (In Russian).

[Kasyanov, 1988] Kasyanov V. N. Optimizing transformations of the programs. Moskow: Nauka, 1988. (In Russian).

[Blume, 1992] Blume W., Doallo R., Eigenmann R. a. o. Parallel programming with Polaris // Computer. – 1992. – Vol.29, N 12. – P. 78 – 82.

[Wilson, 1994] Wilson R.P., French R.S., Wilson C.S., a.o. SUIF: An infrastructure for research on parallelizing and optimizing compilers // SIGPLAN Not. – 1994. – Vol.29, N 12. – P. 31-37.

[Shteinberg, 2004] Shteinberg B.Ya. Open parallelizing system // Mathematical methods of parallelizing of recurrent loops for upercomputers with parallel memory - Rostov-on-Don: Rostov University, 2004.- P. 166-182. (In Russian).

[Tapkinov, 2006] Tapkinov B.Yu. Internal program representation in System of construction of optimizing and parallelizing compilers // Proceedings of All-Russian Scientific Conference "Scientific service in the Internet: parallel programming technologies", Novorossiysk, 18-23 September 2006. P. 88-91. (In Russian).

[Artemieva, 2002] Artemieva I.L., Knyazeva M.A., Kupnevich O.A. A model of a domain ontology for "Optimization of sequential computer programs". Terms for optimization process description. In 3 parts. // Scientific & Technical Information. Part 1: 2002. №12: 23-28. (In Russian).

[Artemieva, 2003] Artemieva I.L., Knyazeva M.A., Kupnevich O.A. A model of a domain ontology for "Optimization of sequential computer programs". Terms for optimization process description. In 3 parts. // Scientific & Technical Information. Part 2: 2003. №1: 22-29. (In Russian).

[Kleshchev, 2007] Kleshchev A.S., Knyazeva M.A., Timchenko V.A. Smart system of generation of single internal representation in system of program transformations. Second International Conference "System analysis and information technologies" : Conference proceedings. In 2 vol. Vol. 1. – M.: Editorial URSS, 2007. – 288 p. (In Russian).

[Knyazeva, 2008] Knyazeva M.A., Timchenko V.A. Subsystem of generation of single internal representation in system of program transformations. // Software & Systems. - 2008.- №1. (In Russian).

[Kaufman, 1978] Kaufman V.Sh. On technology of translator construction (mapping approach). // "Programmirovanie". 1978. № 5. – P. 36 – 44. (In Russian).

[Bunimova, 1978] Bunimova E.O., Kaufman V.Sh., Levin V.A. On description of language mappings. – "Vestnik MGU. Computational Mathematics and Cybernetics". 1978. №4. – P. 68 – 73. (In Russian).

[Kaufman, 1977] Kaufman V.Sh., Levin V.A. Natural approach to the problem of description of context conditions. – "Vestnik MGU. Computational Mathematics and Cybernetics". 1977. №2. – P. 67 – 77. (In Russian).

[Yershov, 1977] Yershov A.P., Grushetsky V.V. Realization-oriented method of description of algorithmic languages. Preprint, Computational Center of Siberian Branch of Academy of Sciences of the USSR, Novosibirsk, 1977. (In Russian).

Authors' Information

Margarita A. Knyazeva *– Ph.D. Senior Researcher of the Intellectual System Department, Institute for Automation & Control Processes, Far Eastern Branch of the Russian Academy of Sciences: 5 Radio Street, Vladivostok, Russia; e-mail: rita_knyazeva@mail.ru*

Vadim A. Timchenko *- Researcher; Institute for Automation & Control Processes, Far Eastern Branch of the Russian Academy of Sciences: 5 Radio Street, Vladivostok, Russia; e-mail: rakot2k@mail.ru*

UPGRADABLE TREE LEVELS EDITOR OF METAONTOLOGIES, ONTOLOGIES AND KNOWLEDGE FOR A CHEMISTRY[1]

Irene Artemieva, Natalia Reshtanenko, Vadim Tsvetnikov

***Abstract**: Development of upgradable tree levels editor of metaontologies, ontologies and knowledge for a chemistry intellectual system is described. A fragment of chemistry ontology of the third level is given. A dialogue scenario for editing ontologies of the second level is described. Data base schemes for representing ontologies and knowledge are defined. A way for adding graphical components to the editor is described.*

***Keywords**: Upgradable multi-levels editor of ontologies and knowledge, domain ontology, chemistry ontology, domain knowledge*

***ACM Classification Keywords**: I 2.5 – Expert system tools and techniques.*

***Conference**: The paper is selected from XIV[th] International Conference "Knowledge-Dialogue-Solution" KDS 2008, Varna, Bulgaria, June-July 2008*

Introduction

While solving applied chemical tasks researchers have to use ontologies and knowledge of different chemical domains and in turn to solve nested tasks (as the subtasks) of these domains. So the computer systems integrating these ontologies and different domains knowledge for solving chemical tasks are needed. As a scientific domain is being developed so the computer systems must be upgradable. In other words, from one hand it must allow user to add new ontologies and knowledge of new chemical domains, and from the other hand it must allow to add the new program components for solving applied tasks.

One of such kind of systems is the specialized computer knowledge bank for chemistry [Artemieva, Reshtanenko, 2006] – the expandable intellectual Internet-oriented program system for solving the diverse tasks from this professional domain, supporting the mechanisms for collective ontologies and data bases development and for adding new program components for solving applied tasks of this domain. To allow ontologies and knowledge bases to be expanded and developed, this knowledge bank is based on the 3-level chemical ontology [Artemieva, 2007]. The upper level – called chemical metaontology – describes the structure of several 2-nd level chemical ontologies, also known as the meta-ontologies of chemical domains. Each meta-ontology of chemical domain describes the structure of several representations of nested sub-domain ontologies. In its turn, sub-domain ontology describes the structure of information representation in the sub-domain knowledge base.

Chemist ordinary deals with the specialized objects as “compound structured formula”, “spectrum” and so on. The knowledge of such objects is represented in the traditional for chemistry graphical symbols. That’s why the knowledge editors must allow using specialized graphical editors, which may be called by ontology. For example, if some property of an object is the structured formula, than the structured formula editor for assigning this property must be called. The set of possible graphical object types may be expanded in the future; it requires new corresponding graphical editors. So the editor imbedded into the specialized chemical knowledge bank must be patchable with such components.

There are editors [Corcho et al, 2003], [Denny, 2002] allowing to create the domain ontologies by defining the concepts (classes) and their hierarchy. An ontology created is used for editing the domain knowledge. Knowledge elements are represented in these editors as the elements belonging to classes described in the ontology. Another approach to creation of the knowledge editors controlled by the metainformation (ontology of information) is described in the paper [Kleschev, Orlov, 2006]. Ontology is represented as the semantic net. Knowledge is another semantic net and its structure is defined by the ontology. But the methods of creating the

1 This paper has supported by Far Eastern Branch of the Russian Academy of Sciences, the project 06-III-A-01-005 and Russian Fund of Fundamental Investigation, the project 06-07-89071-a

specialized multilevel editors which allow addition of the special components for special objects (including graphical objects) editing are still not described in the literature.

The purpose of this paper is to describe the method of developing the expandable specialized 3-level editor for chemical metaontologies, ontologies and knowledge, based on 3-level chemical ontology [Artemieva, 2007].

Database structure for storing ontologies

Second level ontology is stored by means of database control system. Each domain corresponds to its particular database with the same name. The structure of 2-nd level ontology is fixed by the 3-d level ontology [Artemieva, 2007]. Let us describe the structure of several database tables for 2-nd level ontology representation. Also, let's demonstrate how this structure is correlated with the terms defined in the 3-d level ontology. Examples are written by means of the applied logic language [Kleshchev, Artemieva, 2005].

"Types of objects" – defines what types of objects ("Subname" field) form the current domain described in the 2-nd level ontology, and how they are represented ("SubsType" field). Value of "Subname" field is the string with the name of an object. "SubType" field can be one of : {}R, {}I, {}N. Representation of information is defined by the 3-d level ontology parameter as the name of the set of terms: sort Types of objects: {}N \ ∅.

The view of information of "SubType" field is taken from the description of each element of the set of types as the name of the set : (Type: Types of objects) sort Type: {}(R ∪ I ∪ N).

"Types of objects components" is the table containing the definitions of the types for the objects of the 2-nd level ontology to be created (for each type of object ("Subname" field) it defines what kind of objects will become the components of the current one ("SubsComponent"). This table is linked with the "Object types" table, the values of "Subname" and "Subs Component" fields may be only the types defined in the "SubName" field of "Object types" table. The table may contain several rows with the same "Subname" field value, but the values of corresponding "SubsComponent" field must differ. The representation of information is described with the 3d level ontology parameter as the function, with the input of object type and the output of the set of types:

sort Types of object components: Types of objects → {}Types of objects.

"Own properties of objects" is the table containing the names of the sets of own properties of objects for the 2-nd level ontology to be created. This table is linked with the "Object types" table. The value of the "SubName" field may be only the types defined in the "Object types" table. Value of "SubPrivateProp" field is entered by user. In the 3-d level ontology the term "Own properties of objects" is defined as the constructor for the set of functions [Artemieva, 2007], and the parameter of this constructor is the type of objects: Own properties of objects ≡ (λ(Type: Types of objects) (λ(Area of possible values: {}(Value sets ∪ {}Value corteges)) (j(Type) → Area of possible values)).

"Properties of components" is the table containing the names of the sets of the components of the same type. This table is linked to "Component types" table, the values of "SubName" and "SubsComponent" fields may be only the pairs of values defined in the "SubName" and "SubComponent" fields of the table "Component types"; value of the "SubsComponentsSubsProp" field is the name entered by the user. The term "Properties of components" is also defined in the 3-d level ontology as the constructor for the set of functions [Artemieva, 2007], and the first parameter of this constructor is the object type, second – the set of the components: Properties of components ≡ (λ(Type1: Types of objects) (Type2: Types of objects components(Type1)) (λ(Area of possible values: {}(Value sets ∪ {}Value corteges)) (Object that has type 1 → j(Type1), Objects that has type 2→ Object components(Type1, Type2)(Object that has type1)) → Area of possible values))

"Types of process objects" is the table defining the level of abstraction for the physicochemical process in the domain being defined. In contains the names of objects that may be the participants of the physicochemical process. The information representation is defined by means of the parameter of the 3-d level ontology: sort Types of process objects: {} Types of objects \ ∅.

So the linkage between the database tables corresponds to the relations between the terms of domain metaontology defined by the chemical metaontology. During the creation of the new metaontology of new domain the new database is automatically created and knowledge engineer fills it with the new information. The 1-st level ontology of each domain has the module structure, and each ontology module corresponds to one subdomain. Each module has its own database. 1-st level ontology terms are stored in the table with the following structure:

(1) the field for the term of 1-st level ontoogy defining the name of the property (function name); (2) the set this term belongs to (term of the 2-nd level ontology); (3) the arguments of the function; (4) value area of the function.

The value of the third field is defined automatically by defining the name of the set-term of the 2-nd level ontology, because the 3-d level ontology already contains the definitions for each function. The value area of the function may be the set of names, set of integers or real numbers from some interval, the set of structured formulas etc. If the value area is an interval then the table contains the bounds for this interval.

Each set of the graphical objects has its name. Each element of the set corresponds to its editor. This correlation is stored in the special table which contains the names of the graphical editors and the names of subroutine component of knowledge editor for editing this type of objects. Addition of the new editor component is the responsibility of the attendant programmer.

Information representation structure in the knowledge base module is defined by means of ontology module. Database containing the set of linked tables is automatically created by database control system. The schema of this database is defined in the ontology as the set of terms and their interconnections. If the term is defined in the ontology model as the set, it will be represented in the database as the table containing two fields: unique ID (key field) and the value. If the term is defined as the function, it will become the table where the number of the fields is by one greater (key field) then the sum of arguments number plus the number of elements in the result representation (if the result is not the single value but the Cartesian product then each element of this product corresponds to one table filed). If the result is the predicate then it's regarded as the functions with Boolean result.

The type of each field is defined by means of value restrictions from the ontology module.

Dialogue scenario of the Editor

Process of creation of the 2-nd level ontology (metaontology of the chemical domain) includes the user-definition of the values of all parameters of the 3-d level ontology. Every constructor of the set defines the scheme of terms belonging to this set. Let's describe the fragment of the dialogue scenario for creating 2-nd level ontology by means of the editor.

1. Define the name of new 2-nd level ontology. In this case the empty database is created, with the same name as the 2-nd level ontology. The scheme of database is based on 3-d level ontology. For example, the name of domain can be "Physical chemistry".
2. Define the names of the sets of objects belonging to this subdomain. The representation of each type of objects must be described as one of available: {}R, {}I, {}N (on other words, there can be three variants of object representations - float or integer numbers, or names). For example, for the "Physical chemistry" domain there can be such types of objects as "chemical substances", "chemical elements", "chemical reactions"
3. Define the structure of objects of each type. On other words, confront to the type of objects the set of other types of objects. For example, the components of the objects of the type "chemical substances" are the objects of the type "chemical elements". The components of the objects of the type "chemical reactions" are the objects of the type "chemical substances" etc. The definition of the components for each type is done by means of choosing the set of available types from the list (as defines at the step 2).
4. Define the terms for labeling the sets of own properties of each object type. In this case editor automatically generates the names for the sets by default; afterwards user can edit these names. For example editor forms the names such as "Own properties of objects that have type <chemical elements>" and "Own properties of objects that have type <chemical substances>". User changes them into the "Own properties of chemical elements", "Own properties of chemical substances". Editor forms all names according to the object type definition of step 2.
5. Define the terms for labeling the sets of properties of object components. In this case editor also automatically generates the default names; and user can edit them. For example editor forms the names such as "Properties of components <chemical elements> for objects that have type <chemical substances>", "Properties of components <chemical substances> for objects that have type <chemical reactions>". User changes them into the "Properties of chemical elements of substances", "Properties of substances of

reactions". Editor forms all names according to the pair definitions <type of object, type of component> of the step 3.

6. The same term can have two or more schemes of definition. This set of such schemes is accepted by the scientist society. During this step user defines for each set or terms its name and the set of possible schemes. This scheme definition step uses the information entered earlier during the editing.
7. Define the terms for labeling the sets of names of relations between the objects of different types, so called the sets of common object properties. In this case editor allows to choose the several object types and to enter the term name. For example user enters term name “temperature-dependent material properties” and defines (by choosing the elements from the list of all object types) that the objects (which are the arguments of this property) belong to two sets “chemical substances” and “tabular values of temperature”.
8. Define the level of abstraction for the physicochemical process. In other words, definition of what object types participate in the chemical processes, are their properties taken into consideration in the chemical processes or not. The user is step by step asked about every object type and its components defined by him/her before. For example, participants of the process in the physical chemistry are chemical materials and reactions between them.
9. Define the terms for common properties of the process and its components. The user has defined the level of abstraction (on step 8), so the editor automatically creates the names for the sets of ontology terms. User also can change these names.
10. The definition of relations between objects “object – its component” leads to the fact that each component can include its own components, and in its turn, subcomponent can also include sub-subcomponents etc. During the investigation of chemical process not only the properties of its direct participant are considered but also the properties of participants’ components are also considered. That’s why the purpose of the next step is to define the depth of such nestling. All relations “object – its component” are already defined, so the editor one by one asks the user, which levels of nestling will be considered in this domain. This step uses the information gathered on step 3 and 8.
11. The 2-nd level ontology is used as the base for creating the 1-st level ontology of subdomain. Creation of 1-st level ontology consists of term definition – representatives for the term sets already defined in the 2-nd level ontology. The meta-term of 2-nd level ontology is used for 1-st level ontology term definition. The 1-st level ontology terms are the names of functions – object properties, their components etc. For each function, the definitional domain is defined by the meta-term, value area is defined by user. User chooses the subset of values from available sets; it can be the set of structured formulas, the set of spectrum etc.

The fragment of the second level ontology

As an example of editor’s work let’s see the sample 2-nd level ontology for physical chemistry. Let’s start with the definition of values for 3-d level ontology parameters.

1. Types of objects $\equiv$ {Chemical elements, Chemical substances, Chemical reactions}

The ontology defines the objects of the given types. This set is defined on step 2.

2. Types of object components $\equiv$ (λ(Type: {Chemical elements, Chemical substances, Chemical reactions}) (Type = Chemical substances $\Rightarrow$ {Chemical elements}), (Type = Chemical reactions $\Rightarrow$ {Chemical substances}), (Type = Chemical elements $\Rightarrow \varnothing$)

Chemical elements haven’t components. Components of chemical substances are chemical elements. Components of chemical reactions are chemical substances. This information is defined on step 3.

3. Types of process objects $\equiv$ {Chemical substances, Chemical reactions}

Objects of chemical process are chemical substances and reactions. This set is defined on step 6.

Lets’ define the ontological agreements for 2-nd level ontology. They are defined on step 2 when user chooses the way of representation for each object type.

1. Chemical elements $\subset$ {}N \ $\varnothing$
2. Chemical substances $\subset$ {}N \ $\varnothing$
3. Chemical reactions $\subset$ {}N \ $\varnothing$

Ontological agreements of other types can be defined by means of specialized formula editor.

Now lets define the 2-nd ontology terms, and sensible names for constructors.

1. Own properties of chemical elements ≡ Own properties of objects(Chemical elements)

Term "Own properties of chemical elements" means function, which argument is the set of values or the set of corteges of values (m); the result is the function which argument is the chemical element, and the result is the member of set m. This term is defined on step 4.

2. Properties of substances of reactions ≡ Properties of components(Chemical reactions, Chemical substances)

Term "Properties of substances of reactions" means function, which argument is the set of values or the set of corteges of values (m), the result is the set of functions; where the arguments of each function are chemical reaction or its participant (chemical material), and the result of each function is the member of set m. This term is defined on step 5.

3. Properties of elements of substances ≡ Properties of components(Chemical substances, Chemical elements)

Term "Properties of elements of substances" means function, which argument is the set of values or the set of corteges of values (m), the result is the set of functions; where the arguments of each function are chemical material or chemical element, and the result of each function is the member of set m. This term is defined on step 5.

Let's see the example of definition of the 1-st level ontology term, by means of 2-nd level ontology terms.

sort Atomic weight: Own properties of chemical elements (R(0, ∞))

Term "Atomic weight" means function, which argument is the chemical element and result is positive real number.

Knowledge editor

During knowledge editing user can define only the values allowed by ontology. For example, let's define the term "Current number" as the own property of chemical element. So the definitional domain of this function is the set of chemical elements stored in the table with the same name. Let's define the value area of this function as the integer numbers from 1 to 104. In this case, only the integer number from this range can be assigned as the value of this function for any chemical element.

Another example, let's define the term "Reagents of reaction" as the own property of chemical reaction. So the definitional domain of this function is the set of chemical reactions, and the value area is the set of all subsets of all chemical materials. So for each reaction stored in the table "Chemical reactions" user can choose its reagents from the table "Chemical materials" as he/she defines the values for this function.

If the term is defined as the property of reagent of reaction, then its first argument is the name of reaction, the second one is the name of reagent. In other words, knowledge editor doesn't allow defining the wrong set of arguments. These restrictions are defined in the metaontology of chemistry.

If the information input about structured formulas is needed, then specialized graphical editor is used. The call of this editor is managed by ontology. Entered by user information about structured formula is automatically transformed into the structured description according to the rules of description in the specialized ontology [Artemieva et al, 2006]. Let's show the main terms of this ontology.

The ontology defines the set of possible types of bond: bond types ≡ {simple, double, triple}.

The structured formula describes the structures bonds of chemical elements with each other, and each element has its own number in the structure. Element numbers ≡ I[1, ∞).

The set of mutual relations between the elements is represented with the term "set of bonds" ≡ (∪ (n:I[1, maximum number of bonds])(× chemical elements, element numbers, bond types)⇑ n), which means the set of triple corteges consisting of the chemical element, its number and type of bond. The components of the structured formula are represented as the triple corteges consisting of the chemical element, its number and the set of bonds which this element forms within the structured formula: possible components of the structured formula ≡ (×chemical elements, element numbers, {}set of bonds)). Each structured formula is the sequence of components where the numbers of chemical elements differ for different components: possible structured formulas ≡ {(f: {(n: I[1,∞)) possible components of the structured formula⇑ n})(&(i: I[1, length(f)])(&(j1: {(j: I[1, length(f)]) i ≠ j}) π(2, π(i, f)) ≠ π(2, π(j1, f))))}.

The graphical editor checks all agreements from chemical ontology and knowledge about chemical elements while defining the structured formula [Artemieva et al, 2005], and doesn't allow user defining the values contradicting with the ontology and knowledge. For example, while defining the bond between two chemical elements editor checks, can these two elements with current valencies create this type of bond or cannot.

Conclusion

This paper describes the creation of 3-level ontology and knowledge editor for specialized computer knowledge bank for chemistry. The fragment of chemical meta-ontology is represented. The editor dialogue scenario for creation of chemical domain meta-ontology is shown. The representation structure of ontology and knowledge base by means of database control system is described. The method of adding graphical components to the editor is described. At present time, the prototype of this editor is created; it contains the graphical component for defining the structured formulas of materials.

Bibliography

[Artemieva, Reshtanenko, 2006] Artemieva I.L., Reshtanenko N.V. Specialized computer knowledge bank for organic chemistry and its development based on the ontology. In Artificial Intelligence, Ukraina, 2006, vol. 4. PP. 95-106. ISSN 1561-5359.

[Artemieva et al, 2005] Artemieva I.L., Vysotsky V.I., Reshtanenko N.V. A model for the ontology of organic chemistry. In Scientific & Technical Information, 2005, № . 8. PP. 19-27.

[Artemieva et al, 2006] Artemieva I.L., Vysotsky V.I., Reshtanenko N.V. Description of structural formula of organic compounds in the model for the ontology of organic chemistry. In Scientific & Technical Information, 2006, №2. PP.11-19.

[Artemieva, 2007] Artemieva I.L. Multilevel ontologies for domains with complicated structures // In the Proceedings of the XIII-th International Conference "Knowledge-Dialog-Solution" - KDS 2007, June 18-24, Varna, Bulgaria, Sofia:FOI ITHEA.2007.Vol. 2. PP. 403-410.

[Corcho et al, 2003] Corcho O., Fernandez-Lopez M., Gomez-Perez A. Methodologies, tools and languages for building ontologies. Where is their meeting point? In Data & Knowledge Engineering, 2003. № 46. PP. 41–64.

[Denny, 2002] Denny M. Ontology Building: a Survey of Editing Tools.
URL: http://www.xml.com/pub/a/2002/11/06/ontologies.html

[Kleshchev, Orlov, 2006] Kleshchev A.S., Orlov V.A. Computer knowledge banks. The universal way to solve the problem of information editing // Information technologies. 2006. №5. PP. 25-31.

[Kleshchev, Artemjeva, 2005] Kleshchev A.S., Artemjeva I.L. A mathematical apparatus for domain ontology simulation. An extendable language of applied logic // Int. Journal on Inf. Theories and Appl., 2005, vol. 12, № 2. PP. 149-157.

Authors' Information

Irene L. Artemieva *artemeva@iacp.dvo.ru*

Natalia V. Reshtanenko *reshtanenko@iacp.dvo.ru*

Vadim A. Tsvetnikov

Institute for Automation & Control Processes, Far Eastern Branch of the Russian Academy of Sciences;
5 Radio Street, Vladivostok, Russia

MULTIPLE-VALUED STRUCTURES OF INTELLECTUAL SYSTEMS

Mikhail Bondarenko, Irina Vechirskaya, Grigoriy Chetverikov

Abstract: *The basic construction concepts of many-valued intellectual systems, which are adequate to primal problems of person activity and using hybrid tools with many-valued intellectual systems being two-place, but simulating neuron processes of space toting which are different on a level of actions, inertial and threshold of properties of neuron diaphragms, and also frequency modification of the following transmitted messages are created. All enumerated properties and functions in point of fact are essential not only are discrete on time, but also many-valued.*

Keywords*: many-valued intellectual system, level of actions, parallelism, analyze-coordination processor, processor-supervisor, knowledge base, artificial intelligence hybrid logic, multiple-valued logic, multi-state element.*

ACM Classification Keywords*: I.2.4. Knowledge Represetation Formalisms and Methods.*

Conference*: The paper is selected from XIVth International Conference "Knowledge-Dialogue-Solution" KDS 2008, Varna, Bulgaria, June-July 2008*

Introduction

Construction methods of switching circuits (multiple-valued invertible heterogeneous logic elements) of the first and second kind for language systems of artificial intelligence with invertible properties concerning multiple-valued heterogeneous codes have been modified and they have received a subsequent development effort. Research of hardware-controlled means of obtained models implementation of the language and synthesis methods of multiple-valued invertible heterogeneous logic elements allowed rational ways and their relations to be determined as well as to formulate selection criteria of multiple-valued invertible heterogeneous logic elements for implementation of relations which make up the action base of language systems.

The basic construction concepts of many-valued intellectual systems (MIS), which are adequate to primal problems of person activity and using hybrid tools with many-valued coding are considered. With materialism of a point of view these concepts are agreed with the dialectic laws opened by a man and their manifestations in problems connected with creation of identification systems prediction and recognition of imagery in which the interactive operational mode is a main part of the whole complex of intellectual properties [1, 2].

Those are, for example, the law of unity and struggle of contrasts – as availability in parallel operating in space and time of mechanisms both discrete, and continuous mapping objects of plants; the law of transition from quantitative changes to qualitative-quantitative changes of gradation levels of brightness and the color result in qualitative changes in mapping of objects; the law of negation of negation – as a changes and alternation of coding indications of messages about objects in neurons of a brain – from space to temporal and from two-place to many-valued.

In particular, in works the accent on the concept of neuro-physiologic and neuro-cybernetic aspects of alive brain mechanisms is made. It is connected with the following natural neuron structures from nervous cells – neurons, essentially are highly effective recognizing systems and, for this reason, is of interest not only for doctors physiologists, but also for the experts designing artificial intelligence systems. However direct transfer of research results of neuro-physiologists in engineering practice is now impossible because of a lack of an appropriate bioelectronic technology and an element basis, that has led to development and creation of a set of varieties of artificial neurons realized on the elements of the impulse technology [3, 4].

But also here there were complications because of non-adequate neuron models to a set of the demands made of MIS. Creation of neuro-like models on the basis of multiprocessor in inputting systems technology with programmed architecture, in particular, on the basis of digital integrating structures is offered as the alternative in works. Thus, retaining Neumann structure a MIS are created, being essentially two-place, but simulating neuron processes of space toting different on a level of actions, inertial and threshold properties of neuron diaphragms,

as well as variation of recurrence frequency of transmitted messages. Though it is obvious that all enumerated properties and functions in point of fact, are, essential, not only discrete on time, but also many-valued (are discrete on a level).

1. Structurally Functional Cell Model of a Many-Valued Intellectual System

As the corollary, non-adequacy of used principles of coding and element basis to simulated processes entails a redundancy, complication and non evidence of used mathematical and engineering means of transformations [5], loss of a micro level of parallelism in handling expected fast acting and flexibility of restructuring without essential modifications of architecture and connections.

The originating complications [1], in creation of a many-valued intellectual system promote moving out of the adequacy concept of many-valued logic and structures of MIS creation problems with desirable properties and possibilities.

Therefore, for disclosure of use paths of a knowledge backlog in the field of many-valued coding and structures in MIS creation the conceptual structurally functional model of a MIS cell (Fig.1) is offered.

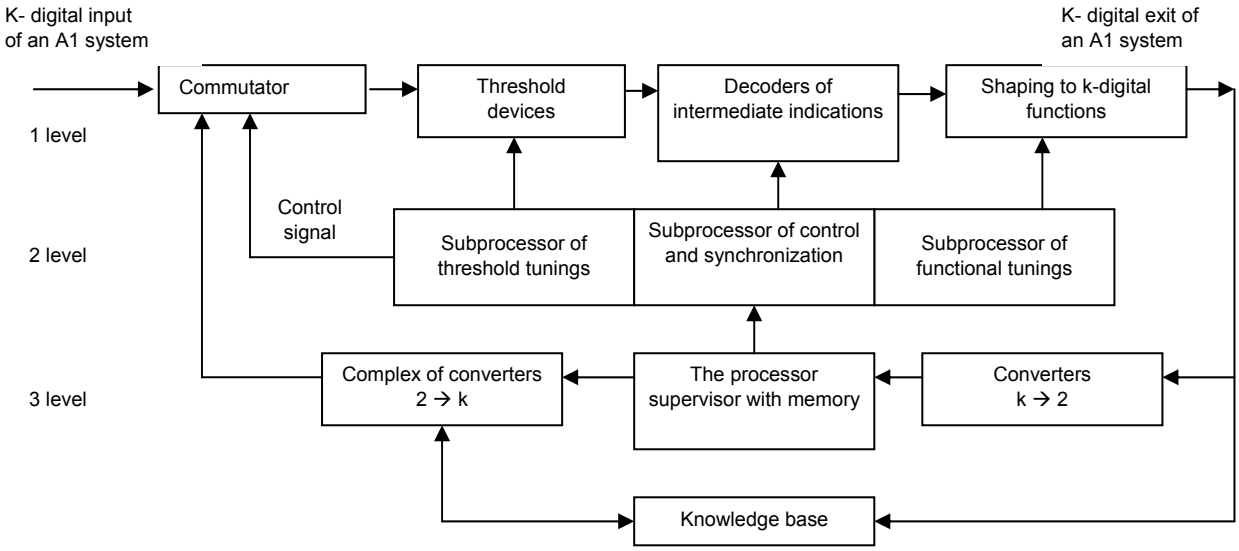


Fig. 1. A conceptual structurally functional model of a MIS cell

Each MIS is characterized by a set of functions fulfilled by blocks, which realize functions and information interchanges. In accordance with solved problems, the structurally functional cell breaks up to three hierarchical levels: functional (analytic-synthetic) – level 1; tactical (analyses-coordination) – level 2; strategic (coordination) – level 3.

The MIS cell increases on a function level both on inputs, and on outputs, and it is integrated with other meshes on inputs of decoders of intermediate indications; at a tactical level – through the analyze-coordination processor; at a strategic level – through the processor-supervisor and knowledge base. The conceptual model of a MIS cell is based on the concept of symbiosis of two- and many-valued tools of data processing, therefore at a strategic level it contain complexes of converters of the data representation form – converters from a two-place code to many-valued 2⟶ k and back 2⟵ k. Obviously, that their use in MIS determines, at what level the problems, are solved in what logic and with what speed (what channel capacity of MIS). Besides the application of these tools excludes necessity of an operator work with two-place translators in input – output of data.

The new principle of the computers construction is offered, in which the principle of organization of brainwork simultaneously with a principle of programmed control assumes as a basis. The principle of organization of brainwork assumes as a basis of operation of such computers, in classical element basis it will be for more to Hilbert machines than for nowadays existing Neumann machines, the basis of which is the principle of programmed control realized rather slowly.

2. Formalization of Construction Principles of Many-Valued Spatial Structures

In the generalized from the two-input universal k-valued structure of a spatial type contains two recognition elements (RE), the control unit (CU), the matrix selector (MS), commutator (C), and keys (K) or the digital-to-analog converter (DAC) (Fig.2).

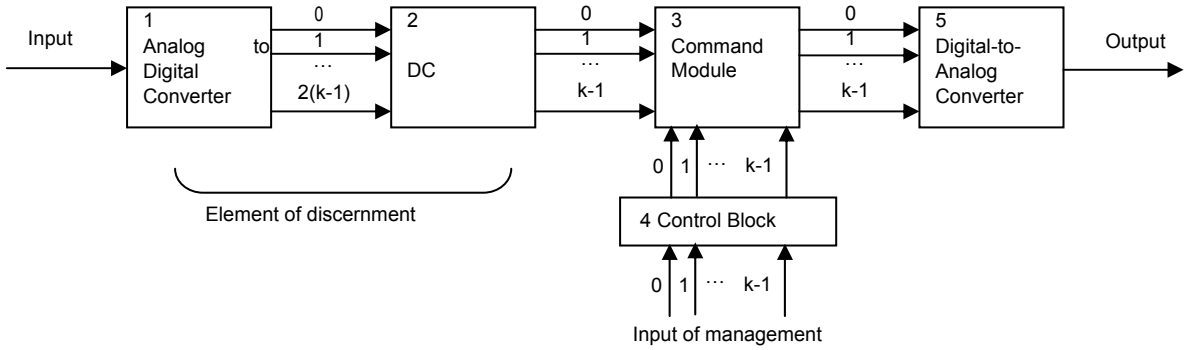


Fig. 2. Universal Multiple-Valued Functional Converter

The logic of the decoders operation in recognition elements 1, 2 is described by the fallowing equation system:

$$f_0 = (x_0, x_1, ..., x_{k-1}) = y^0,$$
$$f_1 = (x_0, x_1, ..., x_{k-1}) = y^1,$$
$$.....,$$
$$f_{k-1} = (x_0, x_1, ..., x_{k-1}) = y^{k-1}.$$

Or in the explicit form at the algebra language of finite predicates [1]:

$$y^0_{1,2} = \overline{x_1},$$
$$y^1_{1,2} = x_1 \cup \overline{x_2},$$
$$y^2_{1,2} = x_2 \cup \overline{x_3},$$
$$.....$$
$$y^{k-1}_{1,2} = x_{k-1}.$$

where x_i and $\overline{x_i}$ $\left(i = \overline{0, k-1}\right)$ - signals of direct and inversion outputs of the ADC units in recognition elements 1, 2. The logic of the matrix selector is described by the following equation system:

$$b_{00} = y^0_1 \cup y^0_2, b_{01} = y^0_1 \cup y^1_2, ..., b_{0(k-1)} = y^0_1 \cup y^{k-1}_2,$$
$$b_{10} = y^1_1 \cup y^0_2, b_{11} = y^1_1 \cup y^1_2, ..., b_{1(k-1)} = y^1_1 \cup y^{k-1}_2,$$
$$.....$$
$$b_{(k-1)0} = y^{k-1}_1 \cup y^0_2, b_{(k-1)1} = y^{k-1}_1 \cup y^1_2, ..., b_{(k-1)(k-1)} = y^{k-1}_1 \cup y^{k-1}_2,$$

where b_{ij} $\left(i, j = \overline{0, k-1}\right)$ - output logical signals of the matrix selector 4. The commutator has two groups by k inputs: the signals from the selector are applied to the first group and control signal values are applied to the second group. In the explicit from the commutator operation is described by the following system:

$$b^{k_0}l^0 \cup b^{k_0}l^1 \cup ... \cup b^{k_0}l^{k-1} = z^{k_0},$$
$$b^{k_1}l^0 \cup b^{k_1}l^1 \cup ... \cup b^{k_1}l^{k-1} = z^{k_1},$$
.....
$$b^{k_{k-1}}l^0 \cup b^{k_{k-1}}l^1 \cup ... \cup b^{k_{k-1}}l^{k-1} = z^{k_{k-1}}.$$

As all k of keys of the output shaper are constantly connected to corresponding k-values of output signals the function values selected by the commutator and the control unit, respectively, will arrive in the converter output (structure) in the course of variations of k-valued functions on the converter inputs. The process control of the logic recommutations is carried out under the action of external control signals.

3. Composition and Decomposition of Spaces

During the last years the problem of understanding people by the computers has become very popular. It arises with some attempts to make work conditions better. The scientists try to formalize natural language with Many-Valued Intellectual Structures. The mechanism of natural language is formally described by facilities of logic mathematics. At the same time in the terms of linguistic algebra the thought is a predicate, the sentence is a formula of predicate operations algebra. Sentence semantics is described by the language of predicate algebra. Grammatical structure of the sentence is described by predicate operations algebra.

We have an algebro-logical language. Using it we can express an algebro-logical structure of natural language. Main task of logic mathematics in the science of language is to know what variant of predicate operations algebra was realized in the natural language. Therefore logic mathematics should strengthen its algebraic tools.

Let's consider the space S

$$S(x_1, x_2, ..., x_n, y) = 1,$$

where y is a point of the space, $(x_1, x_2, ..., x_n, y)$ is its coordinate representation.

Suppose $S(x_1, x_2, ..., x_n, y)$ as a predicate, which is determined on the cartesian product $A \times B$. Let's name the set $A = A_1 \times A_2 \times ... \times A_n$ as a coordinate system of space, B is a space support.

Theorem. Every predicate of disjunktively - conjunctive algebra can be represented as

$$P(x_1, x_2, ..., x_m) = \bigvee_{a_1, a_2, ..., a_m \in U} P(a_1, a_2, ..., a_m) x_1^{a_1} x_2^{a_2} ... x_m^{a_m}.$$

Let's consider a predicate $S(x_1, x_2, ..., x_n, y)$:

$$x_1 : S(x_1, x_2, ..., x_n, y) = \bigvee_{a \in A_1} x_1^a S(a, x_2, ..., x_n, y),$$
$$S_a(x_2, ..., x_n, y) = S(a, x_2, ..., x_n, y).$$

Let's illustrate space exfoliation.

An example. Let's consider the short form of adjective

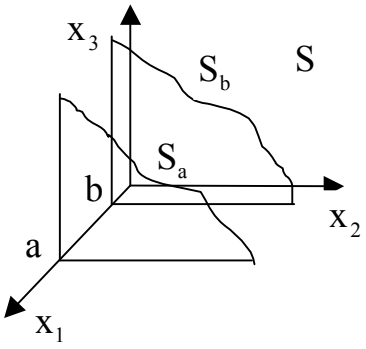


Fig. 3. Graphical interpretation of morphological predicate

$B' = \{мил, мила, мило, милы\},$

$A_1 = \{е, м\}, A_2 = \{у, б\}, A_3 = \{ж, м, с\}, B = \{\_, а, о, ы\}.$

Morphological predicate is represented in the following form:

$$S(x_1, x_2, x_3, y) = x_1^{е} x_2^{у} x_3^{м} y- \vee x_1^{е} x_2^{б} x_3^{ж} y^{а} \vee x_1^{е} x_2^{у} x_3^{с} y^{о} \vee x_1^{м} x_2^{б} y^{ы};$$

$$1) x_1 : S_{е}(x_2, x_3, y) = x_2^{у} x_3^{м} y- \vee x_2^{б} x_3^{ж} y^{а} \vee x_2^{у} x_3^{с} y^{о}, S_{м}(x_2, x_3, y) = x_2^{б} y^{ы};$$

$$2) x_2 : S_{у}(x_1, x_3, y) = x_1^{е} x_3^{м} y- \vee x_1^{е} x_3^{с} y^{о}, S_{б}(x_1, x_3, y) = x_1^{е} x_3^{ж} y^{а} \vee x_1^{м} y^{ы};$$

$$3) x_3 : S_{ж}(x_1, x_2, y) = x_1^{е} x_2^{б} y^{а} \vee x_1^{м} x_2^{б} y^{ы}, S_{м}(x_1, x_2, y) = x_1^{е} x_2^{у} y- \vee x_1^{м} x_2^{б} y^{ы},$$

$$S_{с}(x_1, x_2, y) = x_1^{е} x_2^{у} y^{о} \vee x_1^{м} x_2^{б} y^{ы}.$$

We have 11 predicates of letters recognition in 1), 2) and 18 – in 3). This decomposition shows that the “род” in a certain sense is older than “ударность” and “число”.

Composition is shown in the following form (we used idempotent and true laws):

$$S(x_1, x_2, x_3, y) = x_1^{е}(x_2^{у} x_3^{м} y- \vee x_2^{б} x_3^{ж} y^{а} \vee x_2^{у} x_3^{с} y^{о}) x_1^{м} x_2^{б} y^{ы} \vee x_2^{у}(x_1^{е} x_3^{м} y- \vee x_1^{е} x_3^{с} y^{о}) \vee$$

$$\vee x_2^{б}(x_1^{е} x_3^{ж} y^{а} \vee x_1^{м} y^{ы}) \vee x_3^{ж}(x_1^{е} x_2^{б} y^{а} \vee x_1^{м} x_2^{б} y^{ы}) \vee x_3^{м}(x_1^{е} x_2^{у} y- \vee x_1^{м} x_2^{б} y^{ы}) \vee$$

$$\vee x_3^{с}(x_1^{е} x_2^{у} y^{о} \vee x_1^{м} x_2^{б} y^{ы}) = x_1^{е} x_2^{у} x_3^{м} y- \vee x_1^{е} x_2^{б} x_3^{ж} y^{а} \vee x_1^{е} x_2^{у} x_3^{с} y^{о} \vee x_1^{м} x_2^{б} y^{ы}.$$

Therefore we can analyse the mechanism of language evolution.

Composition and decomposition of space can be used in language formalization.

Thus, main principles of representation of multidimensional relations as a composition of binary relations were discussed. The formal description of adjective’s endings was generated. It was made an attempt to build a relational model of process of ending’s formation as a construction of binary tables.

Conclusion

On the basis of general principles and methods of universality hybridism and parallelism (speed) of k-valued spatial structures a new class of universal functional converters - switching circuits of the third kind which allows to implement analysis, normalization and synthesis of Ukrainian language morphology problems has been created.

The modification of the conventional tables of truth for multiple-valued invertible heterogeneous codes, connected by predicated equations allowing to represent these tables in the compact form is proposed. In so doing, a number of elements of the modified table in 3’ times, where variable p is a depth of decomposition.

It is shown that the algebra and logic method of invertible switching, circuits synthesis of the first kind is propagated to a case of heterogeneous codes, predetermined by predicate equations [6].

The utility of the transfer to switching circuits of the second kind in a number of variables in the initial predicate equation more 8 in a digit number of variables not exceeding 6, which is characteristic for problem solving of morphological analysis and synthesis is justified [7-8].

It is found that switching circuits of direct and back action synthesized on the basis of logic elements and modules possess a number of characteristic properties and properties of full and partial regeneration, contradiction detection in input signals.

The problem solving of principles formalization of the structure organization of computing tools, thus ensures construction of the newest concept for systems of an artificial intelligence; application of space and temporal parallelism at structural and algorithmic levels; creation of procedural an function languages, parallel machines of knowledge bases and the interface. The problem solving of organization principles formalization of universal k -

valued structures of a spatial type by tools of predicate and hybrid logic will ensure construction of a modern concept for artificial intelligence systems, application of spatial parallelism at structured and algorithmic levels; creation of functional languages of parallel machines of knowledge basis; application of symbiosis of two- and many-level heterogeneous coding.

One of circuit for realization of multiple-valued elements is the frequency-harmonic multi-state element which states are coding by amplitude and frequency. This element was made by thin film technology as hybrid integrated circuit.

This paper is devoted to building of formalization methods of the relation. It is a main tools for realization of Many-Valued Intellectual System which focused on parallel information processing and its program realization.

The utilization of developed outlet's instruction strategics is minimized the search time and this is to increase the efficiency of outlet. Developed logical algebra mathematical tool is simplified of given types problem for both knowledge engineer and user.

Bibliography

[1] Bondarenko M., Chetverikov G., Karpukhin A. Analiz problemi sozdaniya novich tekhnicheskich sredstv dlya realizatsii lingvisticheskogo interfeisa. Proc of the 10th International Conference KDS – 2003, Varna, Bulgaria, June 16-26, 2003, pp.78-92.

[2] Bondarenko M., Chetverikov G., Deyneko Zh. Application of a numerically – analytical method for simulation of non-linear resonant circuits. 10th International Conference "Mixed Design Of Integrated Circuits And Systems" (MIXDES 2003), Lodz, Poland, 26-29 June 2003, pp.131-133.

[3] Bondarenko M., Chetverikov G., Leshchinsky V. Synthesis Methods of Multiple-valued Structures of Bilogical Networks // Proc. Of the 12th International Conf. "Mixed design of integrated circuits and systems" (MIXDES 2005), Krakow (Polska), 2005. – Pp. 201-204.

[4] Bondarenko M., Chetverikov G., Karpukhin A. Structural Synthesis of Universal Multiple-Valued Structures of Artificial Intelligence Systems // Proc. Of the 9th World Multi-Conference in Systemics,Cybernetics and Informatics (WMSCI 2005). – Orlando, Florida (USA), 2005. Vol. VII. – Pp. 127-130.

[5] Bondarenko M.F., Konoplyanko Z.D., Chetverikov G.G. Theory fundamentals of multiple-valued structures and coding in artificial intelligence systems.- Kharkiv: Factor-druk, 2003. – 336 p.

[6] Bondarenko M.F., Z.D. Konopljanko, G.G. Chetverikov. Osnovy teorii synteza nadshvydkodiuchikh structur movnykh sistem shtuchnogo intellectu, Monografia. – K.: IZMN, 1997. – 386 s.

[7] Bondarenko M.F., S.V. Lyahovets, A.V. Karpukhin, G.G. Chetverikov. Sintez shvidkodiuchikh structur lingvistichnich ob'ekhtiv., Proc. of the 9th International Conference KDS – 2001, St.Peterburg, Russia, 2001, s. 121–129.

[8] Bondarenko M.F., V.N. Bavykin, I.A. Revenchuk, G.G. Chetverikov. Modeling of universal multiple-valued structures of artificial intelligence systems, Proc. of the 6th International Workshop "MIXDES'99", Krakow, Poland, 17–19 June 1999, pp. 131–133.

Authors' Information

Mikhail Bondarenko *– Professor, Doctor of Technical Sciences; Rector of the State National University of Radio-Electronics P.O. Box 14, Lenin's avenue, Kharkov, 61166; Ukraine.*

Irina Vechirskaya *– Researcher; State National University of Radio-Electronics P.O. Box 14, Lenin's avenue, Kharkov, 61166, Ukraine; e-mail: ira_se@list.ru.*

Grigoriy Chetverikov *- Professor, Doctor of Technical Sciences; State National University of Radio-Electronics P.O. Box 14, Lenin's avenue, Kharkov, 61166, Ukraine; e-mail: chetvergg@kture.kharkov.ua.*

DEVELOPMENT OF GENERALIZED NATURE ENVIRONMENT MODEL FOR EMERGENCY MONITORING

Oleksandr Kuzomin, Illya Klymov, Asanbek Toroev

***Abstract**: In the presented work the problem of generalized natural environment model of emergency monitoring is presented. The approach, based on using CASE-based technologies is proposed for methodology development in solving this problem. Usage of CASE-based technology and knowledge databases allow for quick and interactive monitoring of current natural environment state and allow to develop adequate model for just-in-time possible emergency modeling.*

***Keywords**: emergency, monitoring, model, CASE, classification, clustering*

***ACM Classification Keywords**: H.1.1 Systems and Information Theory*

***Conference**: The paper is selected from Sixth International Conference on Information Research and Applications – i.Tech 2008, Varna, Bulgaria, June-July 2008*

Introduction

The analysis of emergency situations of natural and technogenic character in the different countries of the world for last five years, testifies to growth number of emergency situations average on 5 % per year. According to United Nations experts and large insurance companies for the last century has occurred more than 50 000 natural catastrophes in different countries of the world which has caused death of more than 4 million persons. Necessity of risks handling in extreme natural situations is called by global and national factors which render the negative influence safety of ability to live of world society.

This risks include [1]:

- **rise of spontaneous natural phenomena risks**, caused by global warming, growth of seismic activity, the extension of ozone gaps, etc., and also an intensification of technogenic influence on surrounding environment;
- **rise of probability and scales** of the spontaneous natural phenomena and catastrophes influence on human everyday live;

Managerial process by life safety should include modeling of an environment for risk factors revealing, a risk estimation at all development cycle, development of decision-making subsystem for monitoring, handling of risk and liquidations of emergency situation consequences.

Environment modeling for risk factors revealing provides research of possible threat sources, events initiating emergency situations occurrence, description of the object and existing protection frames, possible scenarios of event course and their ranking. Risk estimation is a process of emergency situation occurrence probability definition throughout the certain period and scale of consequences for health of people, property and a surrounding environment.

Development of such modeling system is main tasks of any research for support of natural safety. Let's remind that model **represents** a collection of objects and ratios between them which adequately describes only some properties of an environment. The model is only one of many possible emergency situation interpretation. This interpretation should suit the user in the present state of affairs, at present time. For model four properties are generally characteristic:

- The reduced scale (size) of model, more precisely, its complexity, which degree always is less, than for the original. At model construction simplifications are entered;
- Saving of key relations between different parts;
- Working capacity, i.e. possibility basically to work, as original-modeled emergency situation (anyway, similarly);
- Adequacy to the real properties of the original (a reliability degree)

Analysis of environment risks

The most part of earth surface are slopes. Sites of a surface concern slopes with the angles of slope exceeding 1 degree. They occupy not less than 3/4 of whole earth surface.

The more abruptly the slope is, the more considerably the gravity component, aspiring to overcome cohesive force of particles of breeds and to displace them downwards. To the Gravity help or stir features of slopes structure: durability of breeds, alternation of various structure layers and their inclination, ground waters weakening cohesive forces between particles of breeds. The slope collapse can be called subsidence — separation from a slope of a large-size block of breed. Subsidence is typical for the abrupt slopes added tight breeds (for example, limestones). Depending on a combination of these factors slope processes gain various shape. In order to correctly model such processes a bunch of various factors should be taken into account. Each factor should be estimated if it influence current situation in order to receive simplified model which still meets requirements described above.

Definition of risks estimation of states of natural complexes, etc., the statistical data about emergency situations, the spontaneous phenomena, and also on results of appropriate dangerous events modeling and situations should be grounded on results of the control of availability index of product of the dangerous technogenic objects, the given monitoring of dangerous geological and hydrometeorological processes. [2]

Presence of effective toolkit for quantitative estimation of safety level gives the chance to provide rationing of risks, levels definition of comprehensible risks for the population, a surrounding environment and economy objects, estimation of correspondence degree of current pace to European safety standards.

Handle of emergency situations risks provides the organization of constant danger level observation of natural processes, natural complexes, exogenous geological processes, etc. in a direction of their danger lowering. Regular monitoring of natural risks gives the chance to watch changes of safety level and to receive real estimations of a remainder resource which in the conditions of limited financial resources allows to optimise expenses on repair and a recovery work.

Proceeding from the analysis of existing methods and models for solution formulated above a problem of safety in emergency situations it is necessary to use a system approach which assumes following stages of problem solution:

- subject domain studying (inspection of objects of a surrounding environment);
- revealing and a formulation of emergency situation problem;
- mathematical (formal) setting of emergency situation occurrence problem;
- natural and/or mathematical modeling of researched objects and processes of surrounding environment;
- statistical processing of results of modeling,
- formulation of alternative solutions,
- estimation of alternative solutions,
- formulation of outputs and proposals on solution of emergency situation problem for objects of a surrounding environment.
- estimation of proposals according to knowledge base of experts
- adjustment of proposals and final decision taking

Problem statement

According to presented tasks of research, it is required to develop and investigate object-oriented methodology of specialized information-analytical system development, including components based on precedent knowledge, taking into account different points of view on current situations (modeling of monitoring methods of emergency situations). Such method must provide adaptation to different circumstances of current situation, must be easy extendable and fast enough to obtain result in required (limited) time before emergency situation occurs.

In general, such method can be divided in several parts:

- presentation of situations as logically connected chains;
- detection of most important nature parameters and transforming them to microsituations

- implementation of non-final estimation principle and taking into account expert linguistic estimations of currecnt situations

Such method should improve **quality** of emergency possibility estimation (time of estimation)

In general, the task of minimization risk of emergency appearance should be solved:

$$\Phi = \min_{j,i} \Re\left(X_{const}, x_j^K, x_i^M \right),$$

where X_{const} – set of constant parameters of nature environment (angle of slope, different plants and trees on the slope, type of inner structure and so on);

x_j^K – j-th numeral K – parameter of natural environment (air temperature, humidity, wind speed, etc.)

x_i^M – i-th quality M parameter of nature environment (expert estimation of emergency situations and possible risk of its occurrence, for example, verbal threat estimation – “big”, “average”, “low”, etc.)

Development of nature environment model and system for emergency prediction

As problem situation we will understand still controllable situation which can be classified as possible emergency. Developed method must provide [3]:

- system approach (different levels of situation description, support of decision takin lifecycle, differential point of view on analyzing emergency)
- variety of modeling methods $\{\mathrm{Mod}\}$ and algorithms of solution taking $\{\mathrm{Alg}\}$ both for automated system development and situation analysis for taking best solutions in case of emergency, which influence different aspects of subject (problem situation, taken solutions, additional resources required for estimation or controlling emergency)

In that case solving task of modeling automated system structure $\{Str\}$ and general conception of emergency modeling $\mathrm{M}\left(\mathrm{Sit}_i'\right)$ can be presented in such way:

$$\{Zad, Met, Mod, A\lg, \Pr og, Pk\} \xrightarrow{\{\Pr\}} \{Str\},\ Str_i \in Str,\ i = \overline{1,K}$$

$$\mathrm{M}\left(\mathrm{Sit}_i'\right) = \underset{\mathrm{Zad}_\eta^e \in \mathrm{Zad}}{\overset{e}{\mathrm{extr}}}\ \underset{\mathrm{Zad}_\eta^e \in \mathrm{Zad}}{\overset{e}{\mathrm{extr}}}\ \underset{\mathrm{Met}^e \in \mathrm{Met}}{\overset{e}{\mathrm{extr}}}\ \underset{\mathrm{Mod}_{\tilde{\Pi}\mathrm{N}_i}^e \in \mathrm{Mod}}{\overset{e}{\mathrm{extr}}}\ \underset{\mathrm{Alg}_j^e \in \mathrm{Alg}}{\overset{e}{\mathrm{extr}}}\ \underset{\mathrm{Prog}_\varphi^e \in \mathrm{Prog}}{\overset{e}{\mathrm{extr}}}\ \min_{r \in \Re} \mathrm{M}\begin{pmatrix} \mathrm{Zad}, \mathrm{Met}, \mathrm{Alg}_j, \\ \mathrm{Prog}_j, \mathrm{Mod}_i, \Re \end{pmatrix},$$

where:

- Zad_η^e - η – th task of nature environment control and prognosis for e – th method of system development, $\eta = \overline{1,Z}$;
- Met^e - e – th method of model development, $e = \overline{1,E}$;
- $\mathrm{Mod}_{\Pi\mathrm{C}_i}^e$ - model of nature environment in e – th method of system development;
- $A\lg_j^e$ - j – th algorithm of system development for e – th method of system development; $j = \overline{1,J}$;
- $\Pr og_{\varphi_j}^e$ - φ – th program solution for implementing j – th algorithm in e – th method of system development, $\varphi = \overline{1,\Pi}$;
- Pk^e - estimation quality of result program complex

Structural analysis of emergency monitoring system should be considered as unity of complex system modeling methods and must be developed on the base of powerful informational support systems. Such systems are called

CASE-based (Computer Aided Software Engineering). Architecture of proposed CASE-system is based on paradigm "methodology-model-notation-methods".

Let's review process of development model of emergency situation model $\{Mod\}$ from CASE-based technologies point of view. During model $\{Mod\}$ development we should choose most effective variants of modeling, which provide receiving required adequate emergency risks estimations with minimal time for system development. And configuring. Model $\{Mod\}$ must provide connection of nature situation with search of control solutions $\Im$, which primary aim is prevention or liquidation of emergency consequences with minimal risk for human being R_e and use no more resources Σ, which are possible to use. In other words to solve the problem for situation i for current slope we must develop emergency model in such way:

$$Mod_i = \underset{r_e^i \in R_e}{extr} Mod\left\{X^i, \Sigma^i, \Im^i, r_e^i\right\} \text{ [4],}$$

where $i = \overline{1, N}$ – count of possible states of nature on current slope

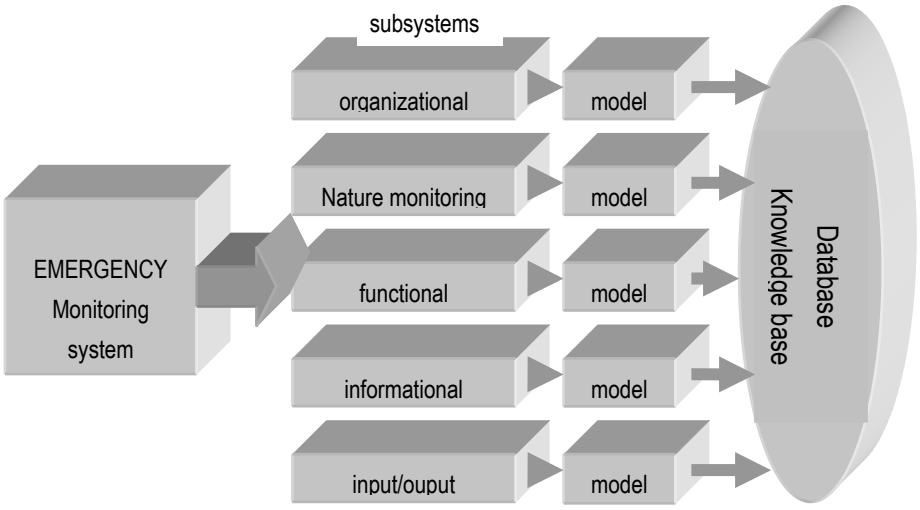
Integratiion in one model this element require to take into account next laws: A – associations; Posl – ordering; K – classification; $\mathrm{K}L$ – clustering.

Association takes place if current problematic nature state Sit_i^t, where t – current time, and situation Sit_i^{t-1}, happened before are connected with each other. This is expressed by equation $\mathrm{A}:\mathrm{Sit}_i^t \to \mathrm{Sit}_i^{t-1}$

For Posl it is obvious that $\mathrm{Posl}:\mathrm{Sit}_i^{t-1} \to \mathrm{Sit}_i^t$, because a chain of interconnected in time situations exists

Using classification $\mathrm{Sit}_i^t = \mathrm{K}\left\{(\mathrm{Sit}_i^t \in \mathrm{K}_{kr}) or (\mathrm{Sit}_i^t \in \mathrm{K}_{\overline{kr}})\right\}$, where K_{kr} – class of emergency and $K_{\overline{kr}}$ is class of non-emergency situations, **group of most important nature parameters are revealed.** In general clustering $\mathrm{Sit}_i^t = \mathrm{KL}\left\{(\mathrm{Sit}_i^t \in \mathrm{KL}_{kr}) or (\mathrm{Sit}_i^t \in \mathrm{KL}_{\overline{kr}})\right\}$ differs from general classification that classes are not determined at step of system creation. Using clustering homogeneous types of data are obtained automatically, reducing time for system development.

After obtaining classification position of current situation Sit_i^t it is possible to build required system using data from closest situations, happened in past which are homogeneous t current one:



Structure of emergency monitoring system

Such computer system of decision-making support grants to expert possibility to handle easily great volumes of the information in realtime a time scale, allowing it to obtain the objective data and make the value judgment. Using effective estimations expert can easily control many parameters during time, along with allowed changes in them:

Monitoring current nature parameters on slope and allowed differences

Conclusion

The offered approach to modeling nature environment allows quickly develop effective system for emergency prediction. Presented method is based on classification and clustering approach as theoretical backend and a strong use of CASE-based technologies as effective solution for quick development of system solving similar tasks (risk estimations in current case). Developed system can be easily implemented by means of any high-level programming language using database/knowledge base backend for storing microsituational data. Effectiveness of such system allow to decrease risks on different slopes and effectively estimate running costs on living/manufacting in different areas.

Bibliography

1. Buser, O., Butler, M. and Good, W. 1987. Avalanche forecast by the nearest neighbors method. IAHS Publ. 162. pp. 557-569.
2. Durand Y., Brun E., Merindol L., Guyomarc'h, Lesaffre B., Martin E. A meteorological estimation of relevant parameters for snow models. Ann. Glaciol.,18, 1993, pp. 65-71.
3. Fuhn P. An overview of avalanche forecasting models and methods. Oslo, NGI, Pub.N 203, 1998, pp. 19-27.
4. Kuzemin A., Toroev A. Mobile means of control and prediction of avalanche climate using information conversion in acoustic range. IDRC-2006. – DAVOS, 2006. – Vol. 2. – pp. 291–294.

Authors' Information

Kuzomin O.Ya.: *Prof. of Information Department, Kharkov National University of Radio Electronics, Head of IMD, (Ukraine),* kuzy@kture.kharkov.ua

Klymov I. M., *student of Kharkiv National University of Radioelectronics, speciality "Intelligent decision support systems",* ilia.klimov@gmail.com

Toroyev A.A.: *director of joint-stock company "Computational techniques and automation systems",* kuzy@kture.kharkov.ua

LEVERAGING EXISTING PLASMA SIMULATION CODES

Anna Malinova, Vasil Yordanov, Jan van Dijk

***Abstract**: This paper describes the process of wrapping existing scientific codes in the domain of plasma physics simulations through the use of the Sun's Java Native Interface. We have created a Java front-end for a particular functionality, offered by legacy native libraries, in order to achieve reusability and interoperability without having to rewrite these libraries. The technique, introduced in this paper, includes two approaches – the one-to-one mapping for wrapping a number of native functions, and using peer classes for wrapping native data structures.*

***Keywords**: wrapping, legacy code, Java Native Interface, peer classes.*

***ACM Classification Keywords**: D.2.13 Reusable Software*

***Conference**: The paper is selected from Sixth International Conference on Information Research and Applications – i.Tech 2008, Varna, Bulgaria, June-July 2008*

Introduction

Contemporary physics simulations evolve over the years and become composite and increasingly complex. A large portion of the system may consist of legacy codes, which are mostly written in languages like FORTRAN, C, and C++. It is ineffective and unreliable to rewrite the entire software system with new design rules, or in new programming languages. Reusing is perhaps the best strategy to handle complexities in software development.

We use the Sun's Java Native Interface (JNI) to leverage existing native libraries. By native code, we mean non-Java code, typically C or C++. As part of the of the Java virtual machine implementation, the JNI is a two-way interface that allows Java applications to invoke native code and vice versa [JNI, 2003]. Thus JNI allows programmers to take advantage of the power of the Java platform, without having to abandon their investments in legacy scientific codes [Malinova, 2006].

In this paper we present creating Java front-ends for some basic functionality of the PLASIMO simulation software. PLASIMO is a framework for modeling low-temperature plasma sources. It has been developed at Eindhoven University of Technology in the department of Applied Physics [Plasimo, 2008]. PLASIMO is written in C++, so in our work we have used the JNI to produce a class library that wraps a set of the PLASIMO's functions and classes. We first discuss the most straightforward way to write wrapper classes – the *one-to-one* mapping. We then introduce how we wrap native data structures using *peer classes*. At the end, we discuss such issues like exception handling and reflection support provided by the JNI.

Methodology

Java applications call native methods in the same way they call methods implemented in the Java programming language. Behind the scenes, however, native methods are implemented in another language and reside in native libraries. We use the JNI to write native methods that allow the Java code to call functions implemented in PLASIMO's native libraries. The JNI allows interaction to occur in both directions: the Java code can invoke native methods; as well the native methods can create, update and inspect Java objects and call their methods (see Figure 1).

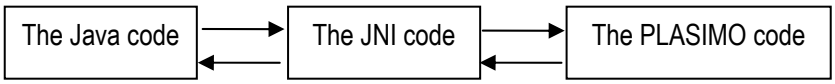


Figure 1. The Java Native Interface as link between Java and PLASIMO

Implementation of a call from Java to a PLASIMO library includes several basic steps, such as [JNI, 2003]:

- Define a Java wrapper class with the native method declaration. This declaration includes the keyword `native` to signify to the Java compiler that it will be implemented externally. The created Java wrapper loads the native library that contains the native code, and invokes the native methods.
- Generate a header file for use by the native (C/C++) code.
- Create the native C/C++ implementation - the native code implements the function definition contained in the generated header file and implements the needed logic as well.
- Compile the native implementation into a native library – create dynamic (.dll) or shared object library (.so) so it can be loaded at runtime.

One-to-one mapping

This section discusses the most straightforward way to write wrapper classes – the one-to-one mapping. This approach requires us to write one stub function for each native function that we want to wrap. Figure 2 presents the one-to-one mapping of the functions belonging to the PLASIMO's `plRuntime` namespace - it contains functions and classes for runtime configuration and library loading/unloading.

```
public class JRuntime {
        private static native void _configure(JNode node);

        public static void configure(JNode node) {
                _configure(node);
                .........
        }
        .........
}
```

a) native function declaration

```
JNIEXPORT void JNICALL Java_JRuntime__1configure
    (JNIEnv *env, jclass clsptr, jobject jnode) {
        ...........
        // invoke plRuntime::Configure
        ...........
}
```

b) native stub function

```
namespace plRuntime
{
        ...........
    void Configure( const plNode & node) {....}
}
```

c) native method definition

Figure 2. One-to-one mapping approach to write wrapper classes.

Each native function (for example, _configure) maps to a single native *stub* function (for example, `Java_JRuntime__1configure`), which in turn maps to a single native method definition (for example, `plRuntime::Configure`). Here the stub function serves two purposes:

- The stub adapts the native function's argument passing convention to what is expected by the Java virtual machine. The virtual machine expects the native method implementation to follow a given naming convention and to accept two additional arguments. The first parameter, the `JNIEnv` interface pointer, points to a location that contains a pointer to a function table. Each entry in the function table points to a JNI function. Native methods always access data structures in the Java virtual machine through one of the JNI functions. The second argument differs depending on whether the native method is a static or an instance method. The second argument to an instance native method is a reference to the object on which the method is invoked, similar to the "this" pointer in C++. The second argument to a static native method is a reference to the class in which the method is defined. Our example, `Java_JRuntime__1configure`, implements a static native method. Thus the `jclass` parameter is a reference to the `JRuntime` class.
- The stub converts between Java programming language types and native types. The JNI defines a set of C and C++ types that correspond to types in the Java programming language. The mapping of primitive types is straightforward. For example, the Java type `int` maps to the C/C++ type `jint` (defined in `jni.h`). The JNI passes objects to native methods as opaque references. Hence, the native code must manipulate the underlying objects via the appropriate JNI functions.

The result is that the created stub function makes calls to the PLASIMO functions and possibly back to the Java methods, and returns results to Java.

Peer classes

One–to–one mapping addresses the problem of wrapping native functions. The examples in the previous section have covered calling standalone C++ functions that return result or modify parameters passed into the function. However, if we create an instance of a C++ class in one native method, another problem encounters: how can C++ classes be used from a Java program and keep objects around while the program is running? One way to handle this situation is to define a Java class called "peer class" that corresponds to the C++ class. Peer classes are classes that directly correspond to native data structures. Each instance of the peer class corresponds to a C++ object, tracking the state of the C++ object.

We have created a number of peer classes that correspond to some basic PLASIMO classes (see Figure 3). The native methods are called within the peer classes, and are the link between the peer classes and the C++ classes.

The JPeer class is an abstract Java class that all peer classes extend, as presented in Figures 4 and Figure 5. The JPeer class provides some common functionality and contains a 64-bit field that refers to the corresponding C++ instance of a PLASIMO class (see Figure 4). Subclasses of JPeer assign specific meaning to that field. If we are on a platform with 32-bit pointers, we can simply store this pointer in an int; if we are on a platform that uses 64-bit pointers, we store it in a long.

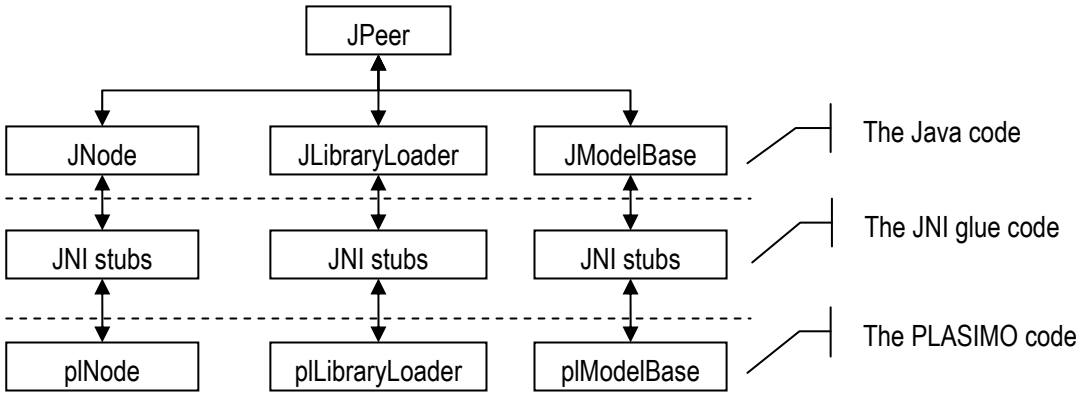


Figure 3. The Java peer classes corresponding to the C++ PLASIMO classes.

The Java peer classes will have the same methods as the C++ classes they represent (it is not necessary all the methods to be wrapped), but the implementation of these methods will be to call the C++ equivalents. For example, take the JNode peer class that wraps the native plNode class (see Figure 5). The Plsimo's plNode is a tree with a variable branching factor. It consists of two parts: a data part of type plLeaf and a variable-sized vector of pointers to children, also of type plNode. Two kinds of plNode-instances occur, sections and data lines: nodes without children are data lines; nodes with one or more children are sections. The children are also plNode's instances, also these can be either sections or data lines. plNode offers a variety of members for accessing its data items, or for section nodes, those of its children.

```
abstract class JPeer {
   private long pointer;

   public long getPointer() {
            return pointer;
   }
   public void setPointer(long ptr) {
            pointer = ptr;
   }

   public abstract void destroy();

   public void finalize(){
            destroy();
   }
}
```

Figure 4. The abstract peer class.

As can be seen in Figure 5, the JNode's constructor calls the native method `create` passing as arguments string values representing a section name and a file name. Then inside the stub function, implementing the JNode's `create` method, an instance of the C++ plNode class is created (see Figure 6).

```
public class JNode extends JPeer{
   private native long create(String secname);
   private native long create(String secname, String filename);
   private native long create(String secname, BufferedReader reader);
   private native void destroy(long p);
   private native void write(String filename, boolean recreate, long p);
   private native void write(PrintWriter writer, boolean recreate, long p);
   private native void read (String filename, long p);
   private native long mount (long tree, long p);
   private native long getSection(String secname, long p);
   ..................
   public JNode(String secname, String filename) throws JException {
        long jnodePtr = create(secname, filename);
        setPointer(jnodePtr);
   }
   public void destroy() {
        long jnodePtr = getPointer();
        if(jnodePtr != 0) {
             destroy(jnodePtr);
             setPointer(0);
        }
   }
   ..................
}
```

Figure 5. The JNode peer class.

The pointer to that instance is saved in the 64-bit field defined in the peer class JNode. In other words, the `create` native method returns this value, and the value is saved in objects of the peer class. When other native methods, such as `destroy`, `write`, etc., are called, this value is retrieved and passed as an argument to the method. The value is then casted into a C++ pointer.

```
JNIEXPORT jlong JNICALL Java_JNode_create__Ljava_lang_String_2Ljava_lang_String_2
  (JNIEnv * env, jobject obj, jstring secname, jstring filename) {
    const char * str;
    str = env->GetStringUTFChars(secname, NULL);
    if(str == NULL) { return 0; }
    const char * c_filename = env->GetStringUTFChars(filename, NULL);
    if(c_filename == NULL) {
        return 0;
    }
    plNode *ptr=0;
    try{
        ptr = new plNode(str, c_filename);
        return (jlong)ptr;
    }
    catch(plParserException& plex){
        std::string msg = "Plasimo exception: " + (std::string)plex.what();
        JNI_throwException(env, "JException", msg);
        return 0;
    }
    .....................
}
```

Figure 6. The stub function implementing a JNode's `create` method.

An important point about peer classes concerns freeing native data structures. Instances of the Java peer classes are garbage collected but not the instances of the C++ classes. C++ has no garbage collection, so it is necessary to think about how objects are destroyed when they are no longer in use. When objects are dynamically created with the `new` operator, the `delete` operator must be explicitly called. In the peer class, this behavior is modeled using the `destroy` method. It is declared as abstract method of the JPeer class. All peer classes that extend the abstract JPeer class provide implementations of the `destroy` method (see Figure 5).

JPeer also defines a `finalize` method that calls `destroy`. When the garbage collector is ready to release the storage used for a Java object, it will first call `finalize` and clean up the memory allocated inside the non-Java code. Since neither garbage collection nor finalization is guaranteed, one cannot rely on `destroy` being called in a timely way. If the Java virtual machine is not close to running out of memory, then it might not waste time recovering memory through garbage collection. That is why we explicitly call `destroy` to invoke the destructor for the C++ class, and avoid memory leaks.

Use the reflection support to call Java from C++

In the previous examples we have discussed calling C++ from the Java code: how can C++ classes be used from a Java program, and keep objects around while the program is running. But since JNI is a two-way interface, we can also call Java from within the native code. JNI allows accessing Java class fields, calling methods, invoking constructors. This involves using the JNI functions that provide reflection support since the reflection allows us to discover at run time the name of arbitrary class objects and the set of fields and methods defined in the class. Although it is possible to call the corresponding Java API to carry out reflective operations, the JNI provides functions to make the frequent reflective operations from native code more efficient, such as: `GetSuperClass`, `GetObjectClass`, `IsInstanceOf`, etc.

In Figure 7 it is shown how to call the JNode's method `getPointer` from a native stub function. Calling a Java method (instance or static) from within the native code involves the following three steps: retrieve the class

reference; retrieve the method identifier; call the method. The `getPointer` method is an instance method defined in JPeer – the JNode's superclass, so we have to obtain the method identifier from a reference to the superclass.

```
JNIEXPORT void JNICALL Java_JRuntime__1configure
    (JNIEnv *env, jclass clsptr, jobject jnode) {

    jclass cls = env->GetObjectClass(jnode);
    jclass super = env->GetSuperclass(cls);
    jmethodID mid = env->GetMethodID(super, "getPointer", "()J");
    if (mid == NULL) {
        return;
    }

    jlong jl = env->CallNonvirtualLongMethod(jnode, super, mid);
    .....
}
```

Figure 7. Use the JNI functions providing reflection support

Exception handling

When the native code detects the exception thrown by the PLASIMO code, it throws a Java exception - in our case an instance of the JException class, as it is shown in Figure 8. This can be seen also in Figure 6.

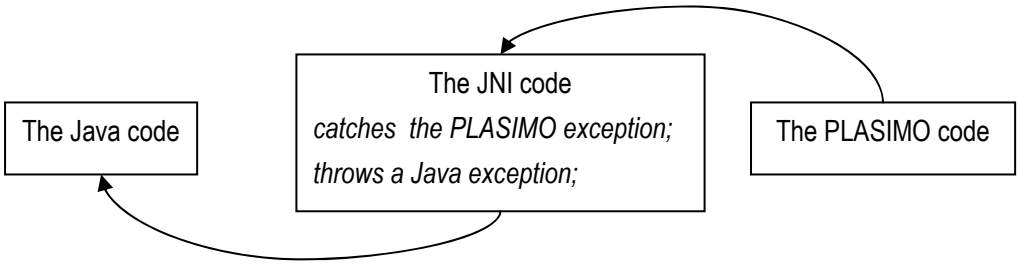


Figure 8. Handling PLASIMO exceptions inside the JNI stub functions.

More interesting is the case when the native code issues a callback to a Java method that itself throws an exception. When the control is returned to the native method, the native code can detect this exception by calling the `ExceptionOccurred` function – the JNI function that performs checks for a pending exception in the current thread. Then the native code can clear the exception by calling `ExceptionClear` and then execute its own exception handling code.

Conclusions

The technique of using the Java Native Interface allows us to leverage code in existing native libraries, particularly the PLASIMO simulation software. Once we have Java-based components to work with, they can easily be transformed into Web services [Mahmoud, 2005]. External client applications will interact with these Web service wrappers, rather than the actual scientific codes. This will alleviate the scientific collaboration with other development teams and will allow us to build connected applications.

Apart from avoiding rewriting PLASIMO in Java, using the existing PLASIMO libraries trough JNI has another advantage. Java has been designed to be portable – compiled programs run on any machine with a Java Virtual

Machine. To accomplish this, the Java compiler compiles to machine-independent byte code that is interpreted at run time by the virtual machine. Although the Java interpreter is efficient, the plasma simulations are supposed to involve cpu-intensive operations on large amounts of data and by leaving these operations out of Java into the PLASIMO compiled libraries, we may expect better performance.

Acknowledgments

This work is supported by the NSF of the Bulgarian Ministry of Education and Science, Project VU-MI-205/2006 and the NPD-07M07 project of the University of Plovdiv "Paisii Hilendarski", Bulgaria.

Bibliography

[Agrawal, 2006] S. Agrawal, V. Chenthamaraksan. Handling events from native objects in Java code, 2006, http://www-128.ibm.com/developerworks/java/library/j-jniobs

[JNI, 2003] Java Native Interface 5.0 Specification (2003). http://java.sun.com/j2se/1.5.0/docs/guide/jni/index.html

[Malinova, 2006] A. Malinova, S. Gocheva-Ilieva, I. Iliev. Wrapping legacy codes for Numerical simulation applications, Proceedings of the III International Bulgarian-Turkish Conference Computer Science, Istanbul, Turkey, October 12-15, Part II, pp. 202-207, 2006.

[Mahmoud, 2005] Q. Mahmoud. Service-Oriented Architecture (SOA) and Web Services: The Road to Enterprise Application Integration (EAI), 2005, http://java.sun.com/developer/technivcalArticles/WebServices/soa/index.html

[Plasimo, 2008] PLASIMO simulation software, http://plasimo.phys.tue.nl .

[Pont, 2004] M. Pont. Calling NAG Library Routines from Java. NAG Technical Report TR 1/04, 2004.

[Pont, 2003] M. Pont. Calling C library routines from Java, Dr.Dobb's Journal, 2003, http://www.ddj.com/architect/184405381

[Smith, 2003] J.Smith. Fast math with JNI, JavaWorld.com, 2003.

[Stearns, 2003] B. Stearns, Java Tutorial: Java Native Interface. Sun Microsystems Inc., 2003. http://java.sun.com/docs/books/tutorial/native1.1 .

Authors' Information

Anna Malinova – *Department of Computer Technologies, University of Plovdiv "Paisii Hilendarski", 24 Tzar Assen St., Plovdiv-4000, Bulgaria; e-mail: malinova@uni-plovdiv.bg*

Vasil Yordanov *– Faculty of Physics, Sofia University, BG-1164 Sofia, Bulgaria; e-mail: v_yordanov78@yahoo.com*

Jan van Dijk *– Department of Applied Physics, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands; email: j.v.dijk@tue.nl*

COGNITIVE APPROACH IN CASTINGS' QUALITY CONTROL

**Jürgen Bast, Valeriy Kamaev, Irina Polyakova,
Andrey Tikhonin, Maxim Zabolotsky, Natalia Kudashov**

***Abstract**: Every year production volume of castings grows, especially grows production volume of non-ferrous metals, thanks to aluminium. As a result, requirements to castings quality also increase. Foundry men from all over the world put all their efforts to manage the problem of casting defects. In this article the authors present an approach based on the use of cognitive models that help to visualize inner cause-and-effect relations leading to casting defects in the foundry process. The cognitive models mentioned comprise a diverse network of factors and their relations, which together thoroughly describe all the details of the foundry process and their influence on the appearance of castings' defects and other aspects.. Moreover, the article contains an example of a simple die casting model and results of simulation. Implementation of the proposed method will help foundry men reveal the mechanism and the main reasons of casting defects formation.*

***Keywords**: castings quality management, casting defects, expert systems, computer diagnostics, cognitive model, modelling, simulation.*

***ACM Classification Keywords**: I. Computing Methodologies - I.6 Simulation and Modelling - I.6.5 Model Development - Modelling methodologies*

***Conference**: The paper is selected from International Conference "Intelligent Information and Engineering Systems" INFOS 2008, Varna, Bulgaria, June-July 2008*

Introduction

By casting defect we understand a technical characteristics mismatch of produced castings and technical requirements which the castings should meet.

Every year production volume of castings grows, especially grows production volume of non-ferrous metals, thanks to aluminium. As a result requirements to castings quality also increase. While we still get castings with defects that make castings more expensive, foundry men from all over the world put all their efforts to manage the problem of casting defects. The issue of reducing casting defects is vital for the foundry industry.

The quality of castings depends on a great amount of factors and parameters. Very often these dependencies are very complex by their nature; they could only be described by numerous factors, which in their turn could have varying interactions with others. Moreover, we experience a lack of knowledge about the collection of factors in general and have limited opportunities of their quality understanding.

The presence of intricate interactions (which can also change dynamically or depend on other factors) and at the same time insufficient information about other relations make the understanding of relations between factors more complicated.

Existing methods and the proposed method of problem solving

Nowadays we see a large number of simulation methods that work with complex dynamic systems and processes. The choice of a method depends on the level of complexity of a system and on the volume of information about it. Today there are four main methods in the sphere of "fighting" with casting imperfections:

1) the first method is based on the use of atlases of casting defects. Many research centres are still working on them. It is difficult to find right reasons with the help of such atlases, especially when there is more than one reason or more than one defect;

2) the second method deals with classical expert systems. Though many scientific groups work on them, such kind of software systems does not give us an opportunity to observe the behaviour of all system parameters when we want to improve values of some of them;

3) the third group works on a so-called black box technology. The black box technologies comprise methods that are based for example on neural networks. Nevertheless, the problem of neural networks lies in the lack of transparency of the process of making inferences. This means that it is impossible to understand the grounds of the conclusion being made;

4) the forth method is simulation. This is a common method, when a system is represented by number of differential equations, which describe energy conservation laws that occur in the system. Moreover it is the most laborious and complicated method, which is also rather costly and time consuming.

Though simulation is the most powerful method very often there is no need to spend time and money to build a model. Frequently the problem lies in understanding of the occurring processes.

In the field of "fighting" with casting defects, it is impossible to build a fully adequate mathematical model, which summarises all the factors and their interrelations. Even now, there are a lot of interferences, which are still not mathematically described and could only have verbal interpretation. Taking into consideration the results of the conducted analysis of existing methods and their disadvantages in the sphere of "fighting" with casting defects, the authors are going to use a new method of informational representation – a cognitive map.

Cognitive modelling makes it possible to conduct fast and more or less exact quantitative virtual experiments with the help of proper software and to get the required information.

Cognitive maps were initially suggested by American psychologist Tolman E. C. to describe behaviour of mice [Tolman, 1948]. Later Axelrod R. used them in politics (model of British politics in Persia) [Axelrod, 1976]. Roberts F. used them in economics (model of energy consumption) [Roberts, 1986].

Commonly a cognitive map is represented as a directed graph G(V, A), where V – factors, A - cause-and-effect relations between factors.

Significant contribution to the development of cognitive maps' theory was made by B. Kosko [Kosko, 1992]. He proposed the most popular modification of cognitive maps - so-called fuzzy cognitive maps (FCM), where values of factors vary from -1 to +1 and some scale is in use.

Cognitive modelling has already been tested with socio-economic systems like regional economy management, industrial safety and so on [Polyakova, 2005]. All such systems are complicated and semi-structured systems, which have a large quantity of interacting factors. These interactions could be changed dynamically. The system of casting defect formation has similar characteristics; therefore the cognitive approach should be rather efficient here.

Like many other scientists the authors see further development of cognitive maps' approach in the direction of joining them with fuzzy logic, where factors are linguistic variables and relations represent data banks of fuzzy rules.

Cognitive simulation is an approach, which is based on the use of cognitive maps in computer simulation. With the help of cognitive maps, it is possible to make visible and transparent all the occurred processes of casting defect formation. Moreover, it is possible to use not only well-known exact interactions, but also interactions, which are only supposed to occur.

The visual representation of a cognitive map gives the possibility to see clearly the whole complex network of reasons of casting defect formation. This visualization helps foundry man to find rapidly all the weighted reasons of defects.

Cognitive modelling example

For example, let us consider a cognitive map of AlSi12-alloy and die casting process. In our research we investigated a great number of factors, which described properties of alloy, mould, work process-related

parameters, work of personnel and machines and also factors, which described casting quality (structure, mechanical properties, measurements and casting surface). We have collected a considerable catalogue of factors. Cognitive approach gives us a unique opportunity to bind all the discovered factors into a single cognitive model and work with them jointly and simultaneously.

In order to arrange reasons of casting defect formation logically we used a so-called Ishikawa diagram (or a fishbone diagram, also known as a cause-and-effect diagram) [Frank, 1993]. This diagram helps to discover several levels of reasons of a problem. We used it in order to find and evaluate system reasons of casting defects.

Afterwards, we developed a cognitive map on the basis of the factors that had been founded with the help of Ishikawa diagrams (Fig. 1).

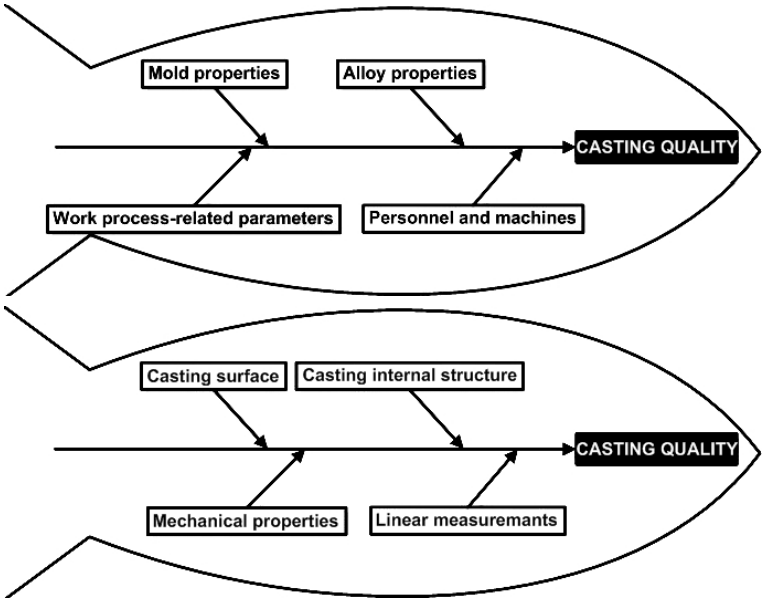


Fig. 1. General view of the Ishikawa-diagrams

Frequently, it is quite a challenge to describe how factors interact with each other. Moreover, we do not have exact information about the character of this interaction. But we know that, for example, “if the die ventilation is insufficient, the probability of porosity defect will increase”. So, we do not have any well-defined relation or mathematical equation. In such case we can take an advantage of cognitive modelling, which supports qualitative interactions and can help us to analyze complex systems with the above mentioned type of relations.

If there are no quantitative data about the interactions, it is possible to use a qualitative value (strong, moderate, weak) to define the interactions together with the Harrington’s scale [Diligensky, 2004] from -1 to +1. An example of possible verbal interpretation sounds: “moderate increase of the die ventilation strongly increases the probability of porosity defect”.

A fragment of classical cognitive map is shown in Fig. 3. The fragment consists of 29 factors and even more interactions. In order to facilitate and simplify the process of creating this cognitive model of casting defects’ formations (and cognitive models in general) we designed a software system. With the help of this software system all the available experience and knowledge of the factory employees and experts can be easily transferred into factors and interactions.

Using qualitative relations, we can also distinguish positive and negative interactions. In other words, a positive interaction shows us that a rise of the factor, which is at the beginning of the arrow, increases the factor value on the arrow end; a negative interaction on the contrary shows us that a rise of the factor, which is at the beginning of an arrow, decreases the factor value on the arrow end.

For example, on the cognitive map in Fig. 3 we can see that the factor "zinc" negatively influences the factor "hot cracking" and consequently decreases it, while the element titanium rises up the probability of "hot cracking".

Among these factors we can choose so-called target factors, which should alter in a desired direction (shown with squares at Fig. 3). Moreover, we can also select so-called control factors (shown with triangles in Fig. 3), which values we should change in order to achieve desired directions of target factors.

The information about factors is represented in the software system as a matrix. This allows us to run a simulation, which virtually shows possible consequences of alterations of one or more factors.

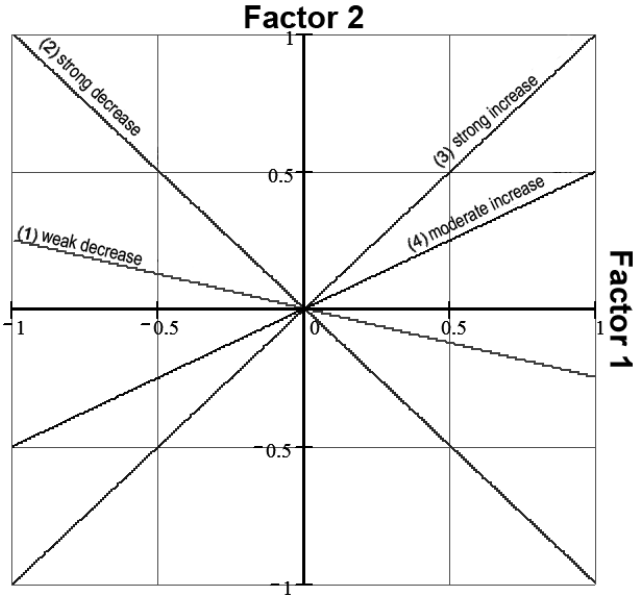


Fig. 2. Qualitative interactions

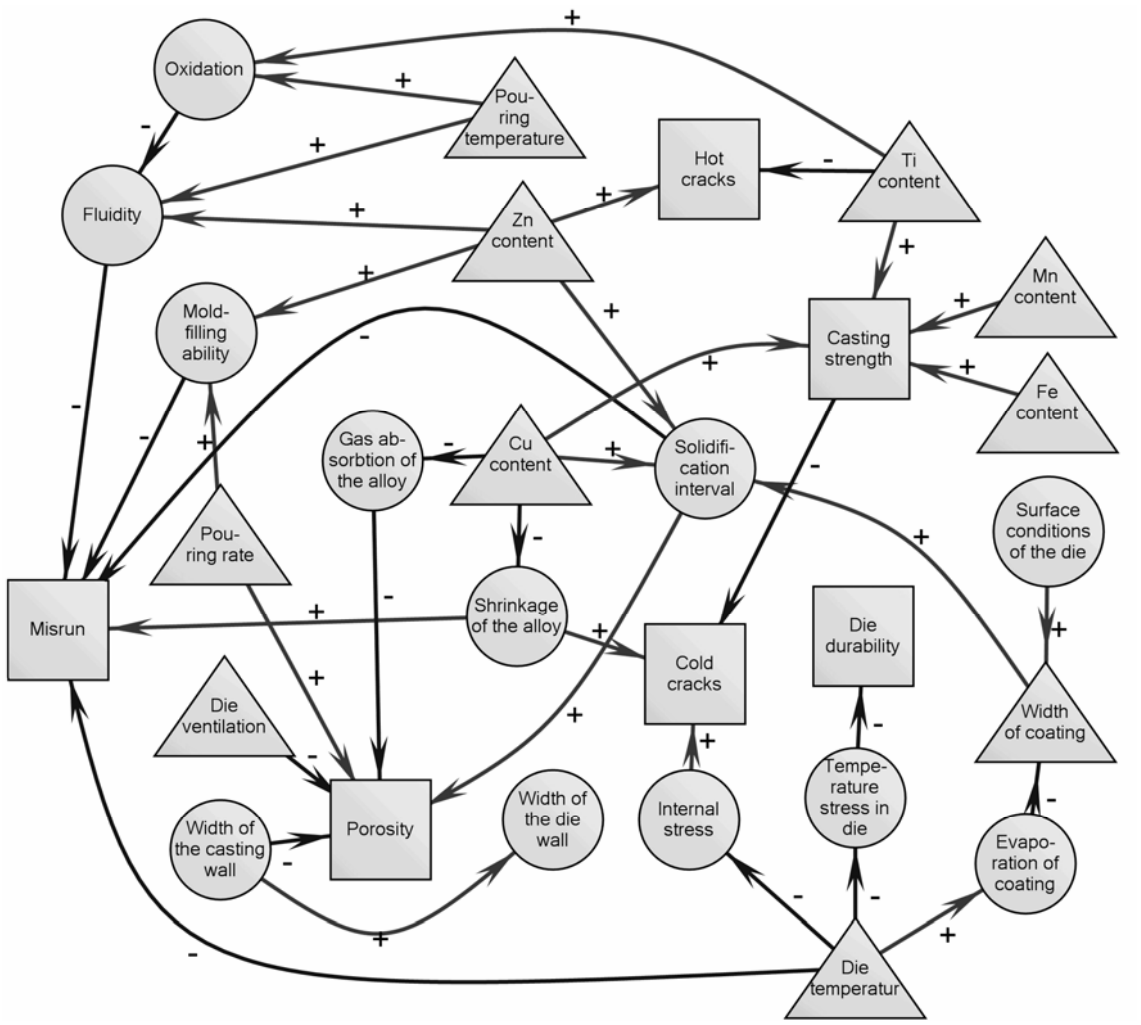


Fig. 3. Fragment of the cognitive map with 29 factors (Δ - control factors, □ - target factors, O- other factors)

Simulation example

To demonstrate a simple simulation example on the above mentioned matrix let us simplify all the interactions in our cognitive model and to make either -0.5 (for all negative) or +0.5 (for all positive).

For example, let us observe the behaviour of some factors when adding zinc into the alloy. We can observe that the addition of zinc at first influences the fluidity and mould-filling ability of the alloy. Alterations of these factors lead to probable reduction of misruns. The results of simulation are shown in Fig. 4.

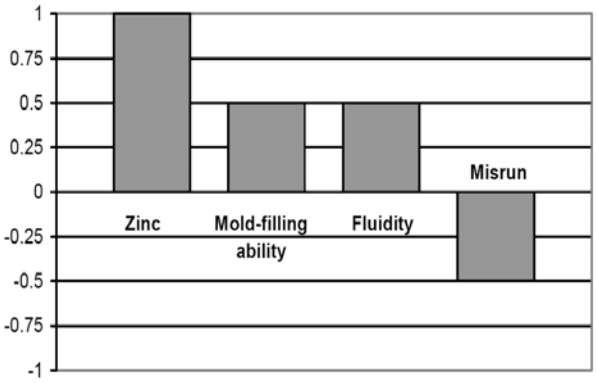


Fig. 4. Simulation example "zinc – misrun"

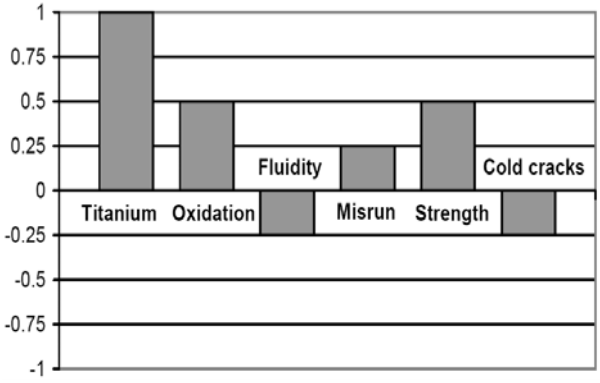


Fig. 5. Example "titanium– misruns and cold cracks"

The next example will demonstrate conflicting goals. This time we choose the factor titanium. If we look at the graph, we can see that the addition of titanium into the alloy will increase oxidation, which in its turn will lead to reduction of fluidity and consequently we could get misruns (see Fig. 5).

Otherwise, the presence of titanium in the alloy increases strength, what in its turn decreases the probability of cold cracks.

This example illustrates how the fight with one casting defect can influence the others. That is why it is highly important to consider the whole network of factors, but not concentrate on a single factor or even a group of factors.

Conclusion

So, the software system, which is being developed and improved, could be used by foundry men in their every day practice and by experts or managers in making fast decisions without a need to conduct long tests and complicated researches. One of the main advantages of the proposed method is that a foundry man can extend a cognitive model and improve results of simulation himself (without any help from the outside), according to his own knowledge and experience about castings formation and their avoidance. The use of the cognitive approach helps us to manage a large quantity of factors. It is highly competitive in the situations, when other methods could either be very expensive or do not show the logic of the occurred processes. This method will help the foundry men to reveal the mechanism and the main reasons of casting defects' formation and take preventive measures in time.

Acknowledgements

Thanks DAAD for financial support.

Bibliography

[Tolman, 1948] E. C.Tolman. Cognitive maps in rats and men. Psychological Review, 55, 189-208, (1948).

[Axelrod, 1976] Axelrod, Roberts (ed.), Structure of Decision (Princeton, NJ: Princeton University Press, 1976).

[Kosko, 1992] Kosko, Bart, Neural networks and fuzzy systems, a dynamical systems approach to machine intelligence (Englewood Cliffs, NJ [u.a.], 1992)

[Roberts, 1986] Roberts, Fred, Discrete mathematical models with application to social, biological and environmental problems (Rutgers University, Prentice-Hall, Inc, Englewood Cliffs, NJ, 1986)

[Brunhuber, 1958] Ernst Brunhuber. Leichtmetall- und Schwermetall- Kokillenguss. Fachverlag Schiele und Schön GmbH, Berlin.1958

[Schneider, 1986] Philipp Schneider. Kokillen für Leichtmetallguß, Gießerei-Verlag.1986.

[Irmann, 1959] Dr.-Ing habil. Roland Irmann. Aluminiumguss in Sand und Kokille. Aluminium-Verlag GmbH Düsseldorf 1959

[Aluminium, 1988] Aluminium-Gußlegierungen, 5. Auflage 1988, VDS Düsseldorf

[Polyakova, 2005] Polyakova I.A., Kamaev V. A., Zabolotsky M. A..,Tikhonin A. V., region's situation analysis and development management with the help of cognitive modeling IT in science, education and business: Works of international conference, Gurzuf, 2005

[Frank, 1993] Gregory, Frank: Cause, Effect, Efficiency and Soft Systems Models, in: The Journal of the Operational Research Society, Vol. 44, No. 4, New Research Directions (Apr. 1993), S. 333-344

[Diligensky, 2004] Diligensky N.V., Dimova L.G., Sevastyanov P.V. Nechetkoe modelirovanie i mnogokriterialnaya optimizaciya proizvodstvennih system v usloviyah neopredelennosti: tehnologiya, ekonomika, ekologiya. Moskva, «Izdatelstvo Maschinostroenie n. 1», 2004.

Authors' Information

Bast, Jürgen *– Professor Dr. Eng., Head of the Department, Institute of Mechanical Engineering Technical University Bergakademie Freiberg, Bernhard-von-Cotta-Strasse 4, D-09596 Freiberg; e-mail: bast@imb.tu-freiberg.de*

Kamaev, Valeriy *– Professor Dr., Head of the CAD Department, Volgograd State Technical University; e-mail: cad@vstu.ru*

Polyakova, Irina *– PhD student; Institute of mechanical engineering Technical University Bergakademie Freiberg, Bernhard-von-Cotta-Strasse 4, D-09596 Freiberg; e-mail: ipolyakova@gmail.com*

Tikhonin, Andrey *– PhD student; Volgograd State Technical University; e-mail: andrey.tikhonin@gmail.com*

Zabolotsky, Maxim *– PhD student; Volgograd State Technical University; e-mail: mxms@rbcmail.ru*

Kudashov, Natalia *– PhD student; Institute of mechanical engineering Technical University Bergakademie Freiberg, natalia@kudaschov.de*

РАСПОЗНАВАНИЕ СЛОЖНЫХ СТЕРЕО И МУЛЬТИ–ИЗОБРАЖЕНИЙ В РЕАЛЬНОМ ВРЕМЕНИ

Адиль Тимофеев, Олег Дерин

***Аннотация**: Рассматриваемые в данной статье методы распознавания сложных стерео- и мульти-изображений в реальном времени анализируют окружающее пространство, выделяют окружающие объекты, классифицируют их и оценивают уровень их важности для решаемой системой задачи. Особенностью данной задачи является как выделение в видеоизображении объектов, так и их классификация, и собственно оценка (рейтинг) их важности.*

***Ключевые слова**: распознавание изображений в реальном времени, виртуальная реальность,3D-сцена, нейросеть.*

***ACM Classification Keywords**: C.2.4 Distributed Systems*

***Conference**: The paper is selected from XIVth International Conference "Knowledge-Dialogue-Solution" KDS 2008, Varna, Bulgaria, June-July 2008*

Введение

Задача автоматического распознавания сложных стерео и мульти – изображений в реальном времени в настоящее время актуальна в медицине, робототехнике, мультимедиа – приложениях, системах обеспечения безопасности. Например, автоматизированные системы позволяют сократить усилия персонала служб безопасности, затрачиваемые на контроль всей охраняемой территории, наблюдаемой сотнями телекамер видеонаблюдения, и сконцентрировать их на наиболее уязвимых участках и подозрительных объектах. Рассматриваемые в данной статье методы анализируют окружающее пространство, выделяют окружающие объекты, классифицируют их и оценивают уровень их важности для решаемой системой задачи. Особенностью данной задачи является как выделение в видеоизображении объектов, так и их классификация, и собственно оценка (рейтинг) их важности.

Методы распознавания сложных стерео- и мульти-изображений в реальном времени

Выделение, сопровождение и классификация недетерминированных объектов в реальной обстановке с перекрытием движущихся объектов препятствиями в условиях изменяющихся фона и условий освещенности, при наличии помех возможно при нейросетевой обработке мульти-изображений. Мульти-изображения (3D-сцены, разнесенные во времени и/или пространстве) с высоким разрешением позволяют измерять статические и динамические параметры объектов (размеры, координаты, скорости, траектории движения) и более надежно выделять подозрительные объекты в 3D-сцене.

Наиболее целесообразно в этом случае использование нейросетевых технологий как наиболее приспособленных к обнаружению недетерминированных объектов в условиях изменяющихся фона и помех.

Следует отметить, что использование «классических» нейросетей (совокупности «равноправных» нейронов) в принципе позволяет решить эту задачу в реальном масштабе времени. Однако «универсализм» такой системы чрезвычайно усложняет аппаратную реализацию и обучение нейросетей и затрудняет её реальное применение. Предложенный ранее логико-аксиоматический метод распознавания сложных изображений основывается на предварительном обучении компьютерной системы понятиям (классам изображений отдельных объектов в разных ракурсах) в форме логических аксиом и на последующем принятии решений путём идентификации и локализации отдельных объектов (например, лица террориста) на сложном изображении средствами адаптивного поиска логического вывода в

исчислении предикатов [1]. Рассматриваемый в данной статье нейросетевой метод распознавания также базируется на предварительном обучении нейронной сети понятиям (классам изображений объектов) с последующей идентификацией и отслеживанием нужного объекта путём параллельной обработки видеоинформации в реальном времени [2–3].

Авторами разработаны методики и математический аппарат, позволяющие редуцировать сложность нейросети путем построения иерархической нейросети («гиперсети») как совокупности адаптивно - связанных нейросетей и ее предобучение [4]. Авторами также разработаны технологии разделения нейросетей на потоковые вычисления, реализуемые на базе FPGA, и сложные вычисления, реализуемые на MCPU, в результате чего аппаратные средства используются оптимально, улучшаются стоимостные, массогабаритные и энергетические характеристики системы [4].

Реализация предлагаемых нейросетевых и мульти-агентных технологий представляет собой программно-аппаратный комплекс, состоящий из сенсоров - телекамер, радио и ИК средств наблюдения, расположенных в зонах контроля, и нейросетевых вычислителей, обрабатывающих сигналы сенсоров, обнаруживающих и оценивающих уровень оценки важности наблюдаемых объектов для решения поставленной целевой задачи. Например, для систем безопасности важным является оценка уровеня террористической угрозы наблюдаемых объектов. При превышении порогового уровня такой оценки информация и изображение подозрительного объекта передается оператору, который принимает решение – игнорировать оценку или принять адекватные меры (закрыть проезд, вызвать группу быстрого реагирования и т.п.). В автоматическом режиме комплекс может принять необходимые меры самостоятельно.

Разработаны, изготовлены и прошли испытания для различных приложений ряд таких комплексов, в том числе [4]:

1. Программно-аппаратный комплекс трехмерного зрения, обеспечивающий автоматическое обнаружение окружающих предметов и обход препятствий подвижной платформой.

Этот комплекс анализирует серию изображений, разлагает 3D-сцену на объекты за счет стереозрения, отделяет движущиеся объекты от неподвижных, с помощью движущихся объектов ориентируется в пространстве и затем сопровождает и классифицирует неподвижные объекты. При движении платформы, на которой установлена данная система, за счет явления зрительного параллакса определяется расстояния до окружающих неподвижных объектов и их размеры, строится (планируется) траектория обхода препятствий в соответствии с выполнением задачи достижения заданной точки. При этом система может преследовать классифицированный движущийся объект (траектория «погони»). Для уточнения расстояния до выбранных объектов могут использоваться дальномеры. Обработка производится бортовым вычислителем каждый кадр (25 раз в секунду).

Рис. 1. Первое мульти-изображение во входной выборке.

Рис. 2. Последнее мульти-изображение во входной выборке.

(телекамера установлена на движущейся платформе)

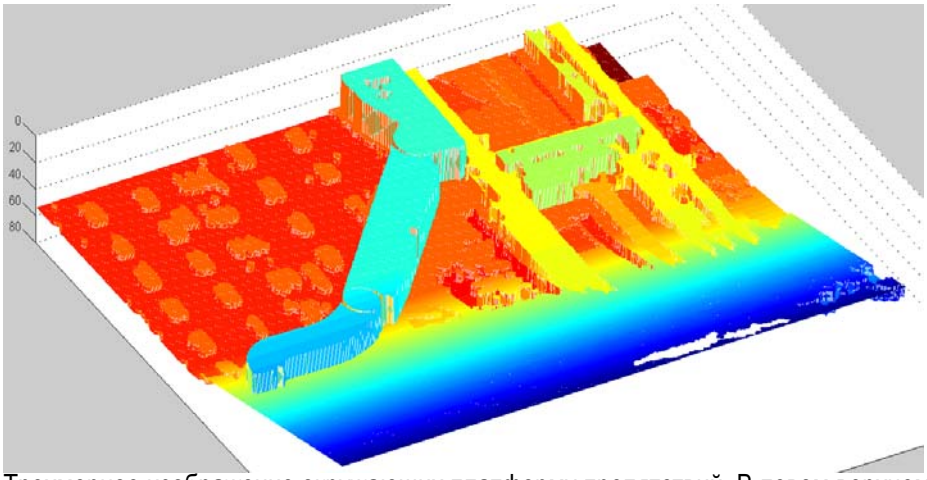


Рис. 3. Трехмерное изображение окружающих платформу препятствий. В левом верхнем углу – расстояние до платформы в метрах

2. Программно – аппаратный комплекс автоматического обнаружения подозрительных лиц и предметов в целях антитеррористической защиты скопления людей в публичных местах.

Этот комплекс предназначен для обнаружения подозрительных лиц и предметов на охраняемой территории, оборудованной многокамерной системой видеонаблюдения, и привлечения к ним внимания оператора, контроля за его действиями и т.п. Изображения со всех телекамер ранжируются по вероятности угрозы (ВУ) и наиболее высокие показатели ВУ поочередно выдаются на экран дополнительного монитора для человека-оператора. Комплекс указывает оператору номер камеры, положение подозрительного объекта в кадре, “историю движения” объекта и параметры, по которым объект признан подозрительным. Одновременно эти данные, а также сведения, переключился ли оператор на данную камеру, его голосовые команды и т.п. записываются в базу данных («черный ящик»), что позволяет отследить принятые меры (аналогично «черному ящику» самолета) и повышает ответственность оператора ввиду того, что зачастую оператор не наблюдает за камерами с должным вниманием и нет возможности проконтролировать его работу.

Обработка производится вычислителем комплекса каждый кадр (25 раз в секунду).

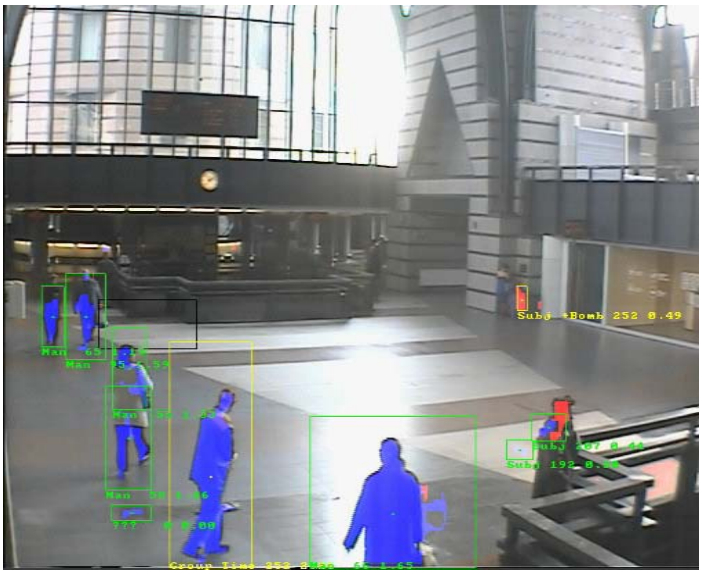


Рис. 4. Результат работы комплекса обнаружения угроз.

3. Программно – аппаратный комплекс контроля подъездных путей стратегических объектов.

Этот комплекс анализирует движение транспортных средств на подъездном пути по данным видеокамер с двух или нескольких точек зрения, телевизионных камер, ИК и радиосредств. По этим данным автоматически обнаруживаются и классифицируются объекты на подъездном пути, сравнивая их характеристики с базой знаний системы, производится их распознавание на классы: «свой», «чужой» или «неизвестный». По результатам распознавания принимается решение об открытии или закрытии проезда с помощью шлагбаума и блокиратора, а также информируется оператор контрольно-пропускного поста.

Обработка производится вычислителем комплекса каждый кадр (25 раз в секунду).

Заключение

Предлагаемые методы и модели позволяют эффективно решать задачи распознавания сложных стерео- и мульти-изображений в реальном времени и обеспечивают безопасность людей и транспортных средств от террористических угроз.

Статья написана при поддержке гранта **РФФИ № 06–08–01612а**.

Bibliography

[1]. *Тимофеев А.В., Косовская Т.М.* Методы адаптивного логического распознавания и его применение. – Труды Международной конференции “Искусственный интеллект – промышленное применение” (Ленинград, 15-19 апреля 1990г.) – Л., ЛВЦ РАН, с. 247–249.

[2]. *Timofeev A.V., Andreev V., Gulenko I.E., Derin O.A., Litvinov M.V.* Design and implementation of multi-agent man-machine interface on the base of virtual reality models. – 9th International Conference SPEECOM’2004 (September 20-22, 2004, St.Petersburg), pp.670-675.

[3]. *Timofeev A.V., Derin O.A., Sova R.U.* Application DSP and ADSP For Artifical Neural Network Control of Dynamic Objects. - Proceedings of the First International Conference Szczecin Poland, December 11-12, 1997 , pp.91-95.

[4]. *Timofeev A.V., Derin O.A., Sova R.U.* Technology of hardware and software designing the neuro-like control for electric drives. - Proceedings of international conference on informatics and control (June 9-13,1997, St.Petersburg), vol.2,u3,pp.632-637..

Информация об авторах

Тимофеев Адиль Васильевич *– Заведующий лабораторией информационных технологий в управлении и робототехнике Санкт-Петербургского института информатики и автоматизации РАН, 199178, Санкт-Петербург, 14-я линия, д. 39; e-mail: tav@iias.spb.su*

Дерин Олег Александрович *– Научный сотрудник лаборатории информационных технологий в управлении и робототехнике Санкт-Петербургского института информатики и автоматизации РАН, 199178, Санкт-Петербург, 14-я линия, д. 39; e-mail: derinoa@mail.ru*

СТРУКТУРНАЯ МОДЕЛЬ ПОЛУТОНОВОГО ИЗОБРАЖЕНИЯ И ЕЕ ИСПОЛЬЗОВАНИЕ В ЗАДАЧЕ СЕГМЕНТАЦИИ ИЗОБРАЖЕНИЙ

Владимир Калмыков, Виталий Вишневский, Татьяна Власова

Аннотация *Предложена структурная модель полутонового изображения. Структурная модель предполагает инвариантное относительно аффинных преобразований описание выделенных в изображении объектов. Форма объекта полностью определяет его описание и представлена его граничным контуром и функцией оптической плотности, которая определена в пределах этого контура. Предложено определение контура полутонового изображения как последовательности, состоящей из отрезков прямых и дуг кривых, причем эти отрезки прямых и дуги кривых являются особыми линиями поверхности, которая соответствует полутоновому изображению. Рассматривается пример использования структурной модели в процессе обработки полутоновых изображений медицинских препаратов, полученных по методу Кирлиан.*

Ключевые слова*: структурный анализ, полутоновое изображение, контур, строковая модель, Кирлиан, сегментация*

ACM Classification Keywords*: I.5.1 Models, I.3.5 Computational Geometry and Object Modeling.*

Conference*: The paper is selected from XIVth International Conference "Knowledge-Dialogue-Solution" KDS 2008, Varna, Bulgaria, June-July 2008*

Введение

Обработка визуальной информации, в частности полутоновых изображений, относится к наиболее сложным задачам искусственного интеллекта и, в то же время, все более актуальным для практического использования в самых различных отраслях науки и технологий. В настоящее время в средствах искусственного интеллекта полутоновые изображения представлены в растровом виде. Такое представление исключает возможность обработки – преобразования, идентификации объектов, отличающихся аффинными преобразованиями – масштабом, положением в поле изображения, поворотом. В современных средствах обработки визуальной информации практически не используются такие понятия как *контуры объектов полутоновых изображений* (за исключением контуров объектов предварительно бинаризованных полутоновых изображений).

В то же время одной из важнейших и наиболее естественных особенностей зрительного восприятия человека является его способность к сегментации поля зрения на объекты, которые отличаются от фона оптической плотностью, цветом, текстурой, др. Основной характеристикой любого объекта является его форма, которая определена контуром – границей между объектом и фоном. Контур, в свою очередь, воспринимается человеком как последовательность отрезков прямых и дуг кривых линий. Форма полутоновых и цветных объектов определяется, кроме того, функцией оптической плотности с учетом цвета, текстуры внутри контура каждого из объектов. Эти особенности зрительного восприятия человека отражены в предлагаемой структурной модели полутонового изображения.

Структурная модель дает возможность однообразного по форме представления произвольных изображений. Задача приведения к структурной модели произвольных изображений, заданных в растровом виде, искаженных помехами в общем случае еще не решена. Однако в отдельных, достаточно многочисленных случаях приведение изображений к структурной модели позволяет существенно повысить скорость и качество обработки визуальной информации, что, в свою очередь, обеспечивает качественное функционирование использующих эти средства информационных технологий. Объекты полутонового изображения, представленные в структурном виде, инвариантном относительно аффинных

преобразований, наилучшим образом подходят в качестве исходных данных для обработки средствами растущих пирамидальных сетей [Гладун,1] и теории распознавания и памяти [Рабинович,2]

Отметим также, что такой структурный анализ формы визуальных объектов, искаженных помехами, хорошо согласовывается с известным стандартом MPEG-7, и может быть к нему адаптирован.

В настоящей работе рассматривается структурная модель полутонового изображения и пример ее использования при разработке информационных технологий для медицинских диагностических систем, в частности, для структурного анализа полутоновых изображений с целью автоматического выделения объектов на примере изображений медицинских препаратов, полученных по методу Кирлиан.

Структурный анализ полутонового изображения

Основой структурного анализа полутонового изображения является модель, которая определяет его структурные элементы. В соответствии с известными представлениями о механизмах зрительного восприятия такими структурными элементами изображения, в частности, есть объекты, расположенные на фоне, который определяется двумерной функцией оптической плотности. Объекты, в свою очередь, определяются контурами, которые их ограничивают, и двумерной функцией оптической плотности в пределах объекта. Контуры являются замкнутыми последовательностями, которые образованы отрезками прямых и дугами кривых линий.

Под изображением понимают часть плоскости, ограниченную некоторой геометрической фигурой, обычно прямоугольником, каждая точка которой характеризуется определенным значением оптической плотности. Другими словами, на части плоскости, ограниченной прямоугольником с размерами *X,У* определена $\rho=f(x,y)$, $(0\le x\le X;0\le y\le У)$. Этой функции можно поставить в соответствие некоторую поверхность $z = f(x,y)$ в пространстве Oxyz.

Предварительно приведем необходимые сведения из области аналитической геометрии в пространстве [Корн,Корн,3].

Множество точек P(x,y,z), координаты которых удовлетворяют системе уравнений

$$x=x(u,v),\ y=y(u,v),\ z=z(u,v) \qquad (1)$$

при подходящих значениях действительных параметров u,v, называется ***непрерывной поверхностью***, если правые части уравнений являются непрерывными функциями параметров. Поверхность может быть также определена уравнением

$$\varphi(x,y,z) = 0 \text{ или } z = f(x,y).$$

Поверхность может иметь более чем одну полость.

Простой поверхностью называется *непрерывная поверхность*, состоящая из одной полости и не имеющая самопересечений (кратных точек). При этом подразумевается, что простые поверхности являются двусторонними (односторонние поверхности, такие как лист Мебиуса, исключаются).

Точка поверхности (1) называется ***регулярной точкой***, если при некотором параметрическом задании поверхности функции (1) имеют в достаточной близости к рассматриваемой точке непрерывные частные производные первого порядка и, по меньшей мере, один из определителей

$$\begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial y}{\partial u} \\ \frac{\partial x}{\partial v} & \frac{\partial y}{\partial v} \end{vmatrix}, \quad \begin{vmatrix} \frac{\partial y}{\partial u} & \frac{\partial z}{\partial u} \\ \frac{\partial y}{\partial v} & \frac{\partial z}{\partial v} \end{vmatrix}, \quad \begin{vmatrix} \frac{\partial z}{\partial u} & \frac{\partial x}{\partial u} \\ \frac{\partial z}{\partial v} & \frac{\partial x}{\partial v} \end{vmatrix}$$

отличен от нуля. Простой кусок поверхности, ограниченный замкнутой кривой, называется ***регулярным***, если все его внутренние точки регулярные. ***Регулярной поверхностью*** называется двусторонняя

простая (замкнутая или незамкнутая) поверхность, составленная из конечного числа регулярных кусков с общими регулярными дугами и точками.

Таким образом, каждому полутоновому изображению можно поставить в соответствие регулярную незамкнутую поверхность в пространстве Oxyρ которая состоит из простых кусков поверхности.

Для поверхности, которая соответствует полутоновому изображению, справедливо следующее ограничение. Каждому значению пары координат (x,y) соответствует одно и только одно значение функции ρ(x,y), то есть перпендикуляр к плоскости изображения в любой точке x,y пересекает воображаемую поверхность один и только один раз.

Контур каждого куска регулярной поверхности является замкнутой последовательностью регулярных дуг кривых и отрезков прямых. Точки контура не являются регулярными точками кусков простых поверхностей. Точки контура – это граничные точки кусков простых поверхностей. Точки контура образуют особые линии поверхности, которые являются граничными, разделяющими разные куски простых поверхностей. В отличие от бинарных изображений, точки которых могут иметь только два значения оптической плотности – черный или белый, области полутоновых изображений, ограниченные контуром, могут иметь различные законы изменения оптической плотности. Соответственно, количество отличающихся друг от друга соседних кусков простых поверхностей, как правило, больше двух. Значит, контуры полутонового изображения могут и не быть, в общем случае, односвязными последовательностями регулярных дуг кривых и отрезков прямых, а состоят из ***ветвей***, соединяющих ***узлы***. Ветви являются особыми линиями изображения. Точки контура, за исключением узлов, являются регулярными точками ветвей. Узлы являются особыми точками контура и изображения. Ветви и узлы вместе с законом изменения оптической плотности каждого куска простой поверхности полностью определяют регулярную поверхность и соответствующую ей область изображения. Во многих практических случаях, однако, когда на фоне расположены простые, не соприкасающиеся друг с другом объекты, контуры полутонового изображения являются односвязными последовательностями регулярных дуг кривых и отрезков прямых, что существенно упрощает задачу структурного анализа.

В полутоновом изображении всегда можно выделить области, для которых значение оптической плотности постоянно, либо меняется по определенному закону. Закон изменения оптической плотности определяется grad ρ – градиентом оптической плотности. Обычно в пределах одной области ρ = const, либо $\partial\rho/\partial x+\partial\rho/\partial y=\text{const}$, либо $\partial^2\rho/\partial x^2+\partial^2\rho/\partial y^2=\text{const}$. В то же время возможны и другие законы изменения оптической плотности.

В соответствии с приведенными определениями полутоновое изображение можно рассматривать как некоторую область регулярной поверхности, состоящую из регулярных кусков простых поверхностей, причем каждый объект изображения соответствует одному или нескольким кускам простых поверхностей.

Цифровая строковая модель произвольного полутонового изображения

С поверхностью в пространстве Oxyρ, которой соответствует полутоновое изображение, совмещена решетка N×M×P, и для каждого пиксела изображения определено среднее в пределах его площади значение оптической плотности p(n,m), принимающее целочисленные значения p(n,m) = (0,P); n = (0,N); m = (0,M). Сторона решетки с N клетками расположена вдоль оси *Ox*, сторона решетки с M клетками расположена вдоль оси *Oy*, сторона решетки с P клетками расположена вдоль оси *Oρ*. Пусть (*ynρ*); *n = (0,N)* – множество параллельных плоскостей, перпендикулярных оси *Ox* в трехмерном пространстве *Oxyρ*. Точно так же (*xmρ*); *m = (0,M)* – множество параллельных плоскостей, перпендикулярных оси *Oy*. Пересечение поверхности изображения с этими плоскостями образует на каждой из плоскостей *ynρ* линию контура $\rho_n(x)$, а на каждой из плоскостей *xmρ* линию контура $\rho_m(y)$, или $\rho_n(m)$ и $\rho_m(n)$ для случая дискретизованного изображения.

Выделение регулярных и особых точек регулярных поверхностей может быть выполнено процессе структурного анализа функций $\rho_n(m)$ и $\rho_m(n)$ дискретизованного полутонового изображения, что дает возможность представить их как последовательности отрезков цифровых прямых и дуг цифровых кривых в плоскостях ρOn при всевозможных значениях m = (0,M) и ρOm при всевозможных значениях n = (0,M) соответственно. Граничные точки отрезков и дуг являются особыми точками линий пересечения, в то время как иные точки являются регулярными точками линий пересечения. Каждая точка *t* дискретизованного изображения принадлежит одновременно двум взаимно перпендикулярным плоскостям $yn_t\rho$ и $xm_t\rho$ и двум пересекающимся линиям контура $\rho_{nt}(m)$ и $\rho_{mt}(n)$ соответственно.

Из определения регулярной поверхности следует, что точка поверхности является регулярной, если она является регулярной точкой горизонтальной и вертикальной линий пересечения.

Если же точка поверхности является особой точкой хотя бы одной из линий – горизонтальной и/или вертикальной линий пересечения, то такая точка является особой – граничной точкой регулярной поверхности – области полутонового изображения.

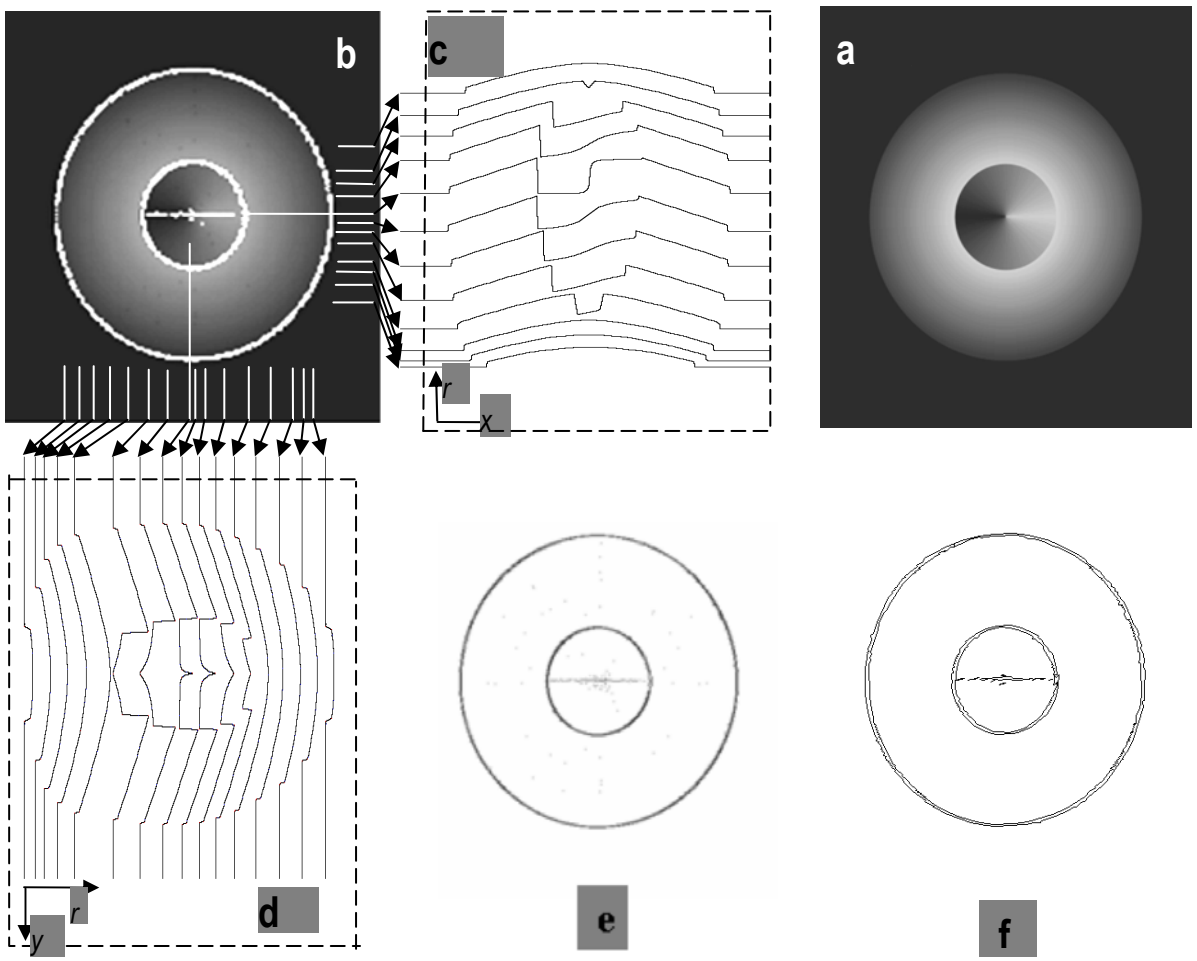


Рис.1. Выделение контуров на полутоновом изображении, выполнено программами, реализующими строковую модель изображения и обработку контуров: a – модельное полутоновое изображение; b – то же самое изображение с выделенными особыми точками; c – кривые оптической плотности – *r* – горизонтальных строк; d – кривые оптической плотности – *r* – вертикальных строк; e – изображения контуров, образованные отдельными особыми точками; f – связные изображения контуров.

Граничные точки областей полутонового изображения (его регулярной поверхности) образуют линии контуров в плоскости *xOy*. Линии контуров, в свою очередь, содержат регулярные и особые точки.

Структурный анализ полутонового изображения, в частности, должен содержать следующие операции.

1. Выделение особых точек регулярных поверхностей (областей изображения).

2. Построение особых линий изображения (контуров), которые ограничивают объекты, по особым точкам регулярных поверхностей.

3. Выделение структурных элементов контуров - отрезков прямых и дуг кривых линий.

На рис. 1 представлен пример структурного анализа с использованием строковой модели, а именно выделение контуров на полутоновом изображении. Программы выполняют над изображением (рис. 1a) следующие операции. Для каждой вертикальной и горизонтальной строк изображения строятся графики функций оптической плотности, примеры которых изображены на рис. 1c,d. Для каждого графика определяется последовательность элементов, из которых он состоит, - отрезков цифровых прямых и дуг цифровых кривых. Граничные точки между элементами графика являются особыми точками графика данной строки и всего полутонового изображения. Особые точки изображения выделены на рис. 1b белым цветом. Особые точки принадлежат линиям контура полутонового изображения. По особым точкам построены контуры полутонового изображения. На рис. 1e представлены контуры в растровом виде, образованные отдельными особыми точками. Им соответствуют контуры в векторном виде - рис. 1f.

Эксперименты по обработке изображений, полученных по методу Кирлиан

Рассмотрим использование структурной строковой модели цифрового изображения на примере обработки изображений медицинских препаратов, полученных по методу Кирлиан.

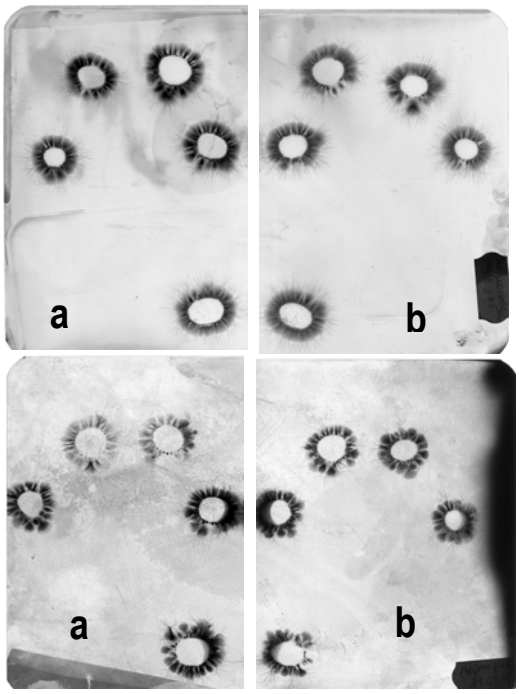


Рис. 2. Примеры изображений Кирлиан: свечения пальцев a)левых, b)правых рук

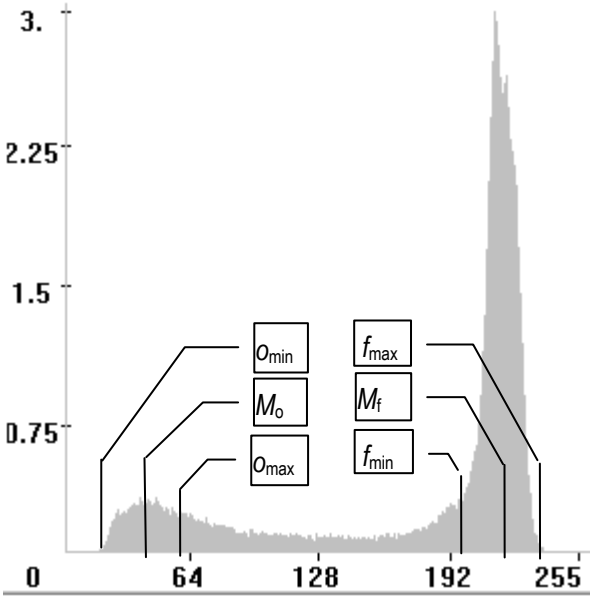


Рис. 3. Гистограмма яркости (оптической плотности) изображения По оси абсцисс определены значения яркости. По оси ординат – условные значения, пропорциональные количеству пикселов, соответствующих данному значению яркости.

M_o,M_f – средние значения яркости объектов и фона; o_{min},o_{max}, f_{min},f_{max} – минимальные и максимальные значения яркости объектов и фона.

Изображения медицинских препаратов вообще содержат объекты, форма которых очень изменчива, но, в то же время, именно в форме содержится диагностическая информация, достаточно уверенно определяемая специалистами при визуальной оценке. Как правило, такие изображения искажены помехами. Изображения медицинских препаратов используются в процессе принятия решений в медицинских диагностических системах. Изображения, которые получены в процессе функционирования таких систем, далеко не всегда могут быть достаточно высокого качества, что значительно снижает возможность их быстрого и полного восприятия экспертами при минимальных расходах времени. В то же время именно в процессе анализа больших количеств таких изображений могут быть полученные новые знания о состоянии здоровья групп населения. Время обработки и принятия решения, так же как и количество экспертов в медицинских диагностических системах, как правило, ограничено. Без автоматизации обработка таких объемов визуальной информации перестает быть эффективной: снижаются качества обработки, растет количество ошибок.

Изображения, полученные по методу Кирлиан (дальше изображение Кирлиан), являются снимками, выполненными на специальной фотопленке, размером А4, на которых зафиксированы свечения от каждого из десяти пальцев (рис. 2).

Для изображений характерны нестабильность и неравномерность фона, значительное количество помех, которые по уровню яркости и величине сравнении с объектами, неустойчивость формы и уровня яркости самих объектов.

Хотя по содержанию обработки в настоящее время эти изображения могли бы считаться бинарными, однако, даже задача бинаризации таких изображений не может считаться тривиальной, не говоря уже о задачах последующей обработки, в частности, задачи выделения и идентификации объектов с целью диагностики. Имеется стандартное программное обеспечение для обработки изображений Кирлиан, полученных на специальных приборах отдельно для каждого пальца [Коротков,4]. Однако использование таких приборов усложняет диагностику, поскольку состояние организма исследуемого за время последовательной фиксации свечений каждого из десяти пальцев может существенно измениться.

Чтобы использовать стандартное программное обеспечение для изображений Кирлиан, на кото-рых одновременно зафиксированы все пальцы рук (рис. 2), необходимо предварительно сегменти-ровать такие изображения и повернуть изображение каждого пальца таким образом, что-бы оно отвечало его вертикальному расположе-нию.

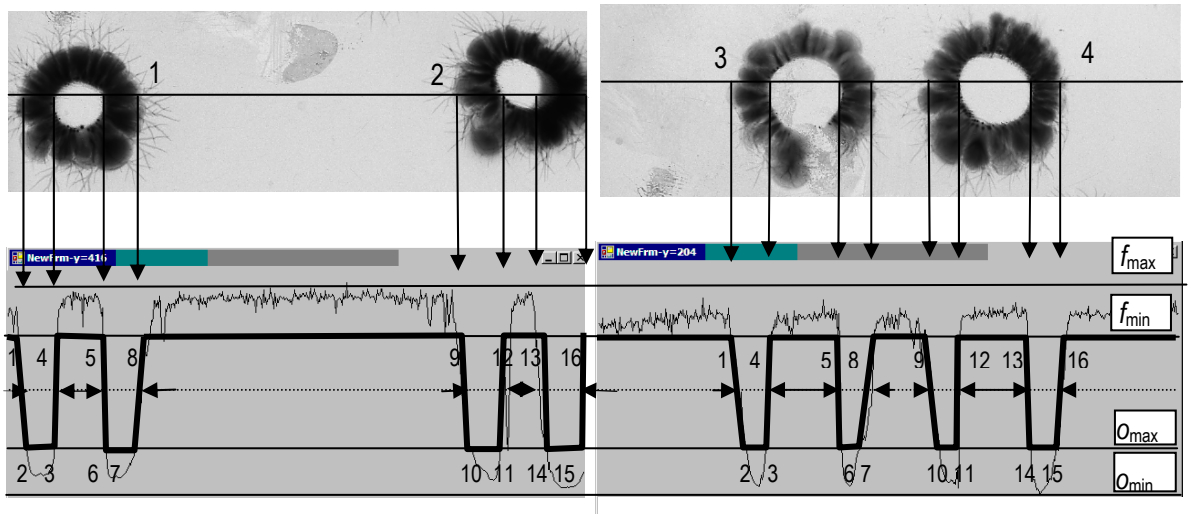


Рис.4. Выделение особых точек в местах пересечения объектов горизонтальными строками.

Предлагаемое программное обеспечение предназначено для автоматической сегментации изображений Кирлиан от пяти пальцев на изображения от каждого пальца отдельно. Работа программы заключается в исполнении следующих операций.

– Вычисляют гистограмму оптической плотности исследуемого изображения Кирлиан (рис.3). По гистограмме, после ее сглаживания, определяют минимальное значение яркости объектов o_{min}, как минимальное значение яркости пикселов изображения, максимальное значение яркости фона f_{max}, как максимальное значение яркости пикселов изображения. Определяют также среднее значение яркости объектов M_o, как первый максимум при росте значений яркости, начиная с нуля, и фона M_f, как первый максимум при убывании значений яркости, начиная с максимального (255).

- На основании предположения о симметричности распределения случайных величин яркости пикселов фона и объектов вычисляют минимальное значение яркости пикселов фона как $f_{min}= M_f -(f_{max}- M_f)$ и максимальное значение яркости пикселов объектов как $o_{max}= M_o +(M_f - o_{min})$.

– Вычисляют функции яркости (оптической плотности) горизонтальной строки $r(m,n)$ m = 1,M для всех горизонтальных строк изображения n = 1,N (рис.4).

– Определяют функцию на изображении

$$v_b(m,n)=\begin{vmatrix} f_{\min}, \text{при } r(m,n) \ge f_{\min}; \\ o_{\max}, \text{при } r(m,n) \le o_{\max}. \end{vmatrix}$$

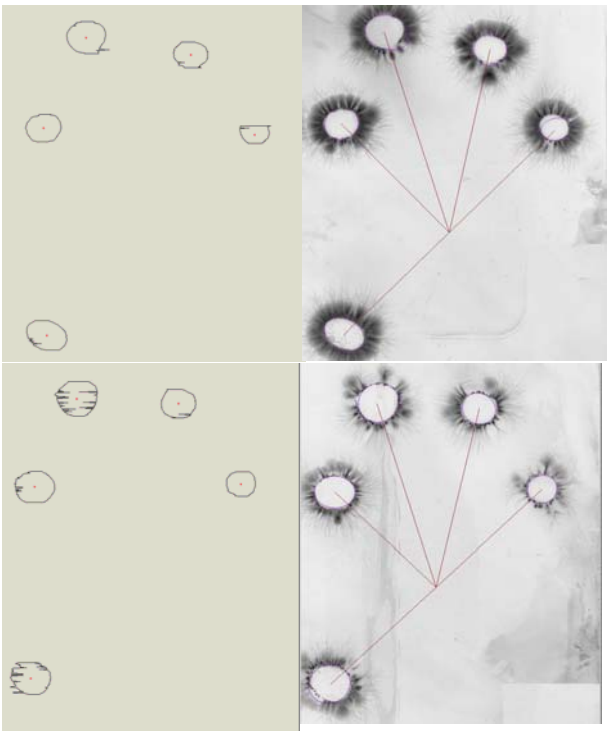


Рис.5. Внутренние контуры объектов (слева). Аппроксимация контуров эллипсами (справа).

Эта функция не является результатом бинаризации, поскольку не рассмотрены значения функции яркости $o_{max}<r(m,n)<f_{min}$. Пикселы с такими значениями функции яркости являются промежуточными между пикселами фона и объекта и не имеют существенного значения для выделения объектов на изображении, по крайней мере, для решения задач по обработке изображений Кирлиан.

Данное преобразование изображения, с учетом динамического диапазона изображения, является нелинейным изменением количества уровней квантования с 256 до 3 и позволяет в значительной мере исключить влияние помех. Примеры функции $v_b(m,n)$ изображены на рис.4 ломаной линией. Числами 1,4,5,8,9,12,13,16, определены особые точки изображения, которые принадлежат фону. Числами 2,3,6,7,10,11,14,15 определены особые точки изображения, которые принадлежат объектам.

Строят внутренний и, при необходимости, внешний контуры объектов, по выделенным особым точкам рис.5. Если контуры определены, то объекты выделены успешно.

Определяют положение объектов свечения пальцев относительно центра ладони. Аппроксимируют внутренние контуры эллипсами, определяют угол поворота каждого пальца, вращают изображение свечения пальца до соответствия вертикальному положению пальца (рис.5) и формируют результирующие файлы изображений.

Заключение

1. Предложенная структурная модель позволяет строить описания объектов полутонового изображения инвариантные к аффинным преобразованиям.
2. Проведенные эксперименты подтвердили большую сложность задач обработки полутоновых изображений и, в то же время, показали возможность и эффективность обработки полутоновых изображений с использованием структурной модели полутоновых изображений.
3. На основе проведенных экспериментов может быть разработана информационная технология, которая может быть использована при создании медицинской диагностической системы.

Литература

[1] Гладун В.П. Планирование решений. – Киев: Наукова думка, 1987. – 167с.

[2] Рабинович З.Л. О естественных механизмах мышления и интеллектуальных ЭВМ // Кибернетика и системный анализ. – 2003. -№5 С.82-88

[3] Корн Г. и Корн Т. Справочник по математике для научных работников и инженеров. - Москва: Наука, 1974.

[4] Коротков К.Г. Основы ГРВ биоэлектрографии. - СПб., 2001. - 134 с.

Информация об авторах

Владимир Калмыков - *старший научный сотрудник, кандидат технических наук, Институт проблем математических машин и систем, просп. акад. Глушкова 42, 03680, Киев 187, Украина; e-mail: kvg@immsp.kiev.ua*

Виталий Вишневский - *заведующий отделением**,** старший научный сотрудник, кандидат технических наук, Институт проблем математических машин и систем, просп. акад. Глушкова 42, 03680, Киев 187, Украина; e-mail: vit@immsp.kiev.ua*

Татьяна Власова *– научный сотрудник, Институт проблем математических машин и систем, просп. акад. Глушкова 42, 03680, Киев 187, Украина; e-mail: chery@immsp.kiev.ua*