I T H E A

International Journal

INFORMATION TECHNOLOGIES & KNOWLEDGE

2009 Volume 3 Number 2

International Journal
INFORMATION TECHNOLOGIES & KNOWLEDGE
Volume 3 / 2009, Number 2

IJ ITK is official publisher of the scientific papers of the members of the ITHEA International Scientific Society

IJ ITK rules for preparing the manuscripts are compulsory.
The rules for the papers for IJ ITK as well as the subscription fees are given on *www.ithea.org*
The camera-ready copy of the paper should be received by e-mail: info@foibg.com.
Responsibility for papers published in IJ ITK belongs to authors.

General Sponsor of IJ ITK is the Consortium FOI Bulgaria (www.foibg.com).

International Journal "INFORMATION TECHNOLOGIES & KNOWLEDGE" Vol.3, Number 2, 2009

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

Publisher: ITHEA®
Sofia, 1000, P.O.B. 775, Bulgaria. www.ithea.org, e-mail: info@foibg.com
Printed in Bulgaria

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

HARDWARE IMPLEMENTATIONS OF VIDEO WATERMARKING

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

Abstract: Digital watermarking (WM) is the process that embeds an additional, identifying message called a watermark into a host multimedia object, such as audio, image or video for authentication purpose. Recently, digital WM technique, an information hiding technique has been investigated as one of the key authentication methods to maintain authenticity and security of multimedia content. By adding a transparent watermark to the multimedia content, it can be possible to make any malicious alteration detected to verify the integrity and the ownership of the digital media. During the last several years various WM techniques for still image have been extensively invented for software implementations due to the low data rate of these signals. Although the software approach holds an advantage of flexibility, certain computational restrictions may arise when attempting to operate at video rate or in portable devices. The hardware-level design offers several distinct advantages over the software implementation in terms of lower power consumption, reduced area and reliability. Therefore, there is a strong motivation for a move toward the hardware-based implementation for digital video WM system.

In order to give a help for future related research works, this paper presents an up to date overview of digital video WM techniques and discusses the important considerations involving in designing VLSI architecture for a novel WM system in many ways. First of all, it goes through a brief survey on WM theory, laying out common classification criterions, discussing the properties of video WM techniques including the specific requirements as well as the comparison to image WM schemes. Various applications of video WM in practice are discussed. Since each WM application has its own specific requirements, WM design must take the intended application into account. Furthermore, the features of video WM implementations in software and hardware and comparison on those two approaches are presented from several points of view: major advantages, drawbacks and differences through the description of several examples of previous works. In addition, a versatile development methodology for hardware WM implementation including the general scheme of a proposed digital video WM system and testing using the custom breadboard are described.

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

ACM Classification Keywords: B.0 Hardware

1. Introduction

Over the past decade, 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, integrity and security of digital video content. So far, there have been various such techniques [2], of which digital watermarking (WM), a data hiding technique, is one of the most popular approaches. Digital WM is a technique that embeds a secret, unnoticeable signal (called watermark) into the original multimedia objects, like audio, image and video for their protection and authentication. The watermark can be detected or extracted later to claim the authenticity and the ownership of the digital media.

During the past few years several researchers have investigated digital WM with different contributions, implemented both on software and hardware platforms [3]-[14]. 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.

While the software approaches hold advantages of easy implementation and flexibility, certain computational restrictions may arise when attempting to operate at video rate or in portable devices. The hardware implementation offers several distinct advantages over the software implementation in terms of low power consumption, less area usage and reliability. Furthermore, it features real time capabilities and compact implementations. Therefore, there is a strong motivation for a move toward the hardware-based implementation for digital video WM system since real time WM of video streams is too expensive for software [11]. 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 to minimize any unexpected deficiencies, while still providing the sufficient level of security. In the past few years, a great deal of research efforts has been focused on efficiently implement WM systems using hardware platforms. Those include implementations in custom-designed circuitry or application specific integrated circuits (ASICs), as well as field programmable gate arrays (FPGAs) implementations [10]-[14]. 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.

In this paper, we aim to achieve two main goals. The first goal is to provide an in-depth overview of previous works on the field of hardware-based video WM through discussing different watermark classifications, new applications and specific requirements. Following that, existing WM software-based and hardware-based implementations and their comparisons are also described. Secondly, to give a help for future related research works, a development methodology for designing and testing hardware-based video WM system has been discussed. In the proposed methodology, VLSI architecture of a prototype chip for real-time video WM design is briefly described.

The remainder of this paper is structured as follows. Section II tries to classify and discuss the digital video WM techniques in various ways in order to give a thorough overview of conventional WM techniques. The software and

hardware implementations of WM algorithms proposed so far are presented in section III. Finally, a design methodology for hardware video WM implementation and conclusions are presented in section IV and V respectively.

2. Background on Video WM

2.1 Watermarks Classification

WM techniques can be divided into different categories according to various criterions [15]. The general classification of the currently available watermarks is shown in Figure 1. In [16] 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 the field of video WM if the individual frames are treated as images [17]. However, contradictory to still image WM techniques, the video WM methods usually require that the WM encoding and decoding are processed in real time.

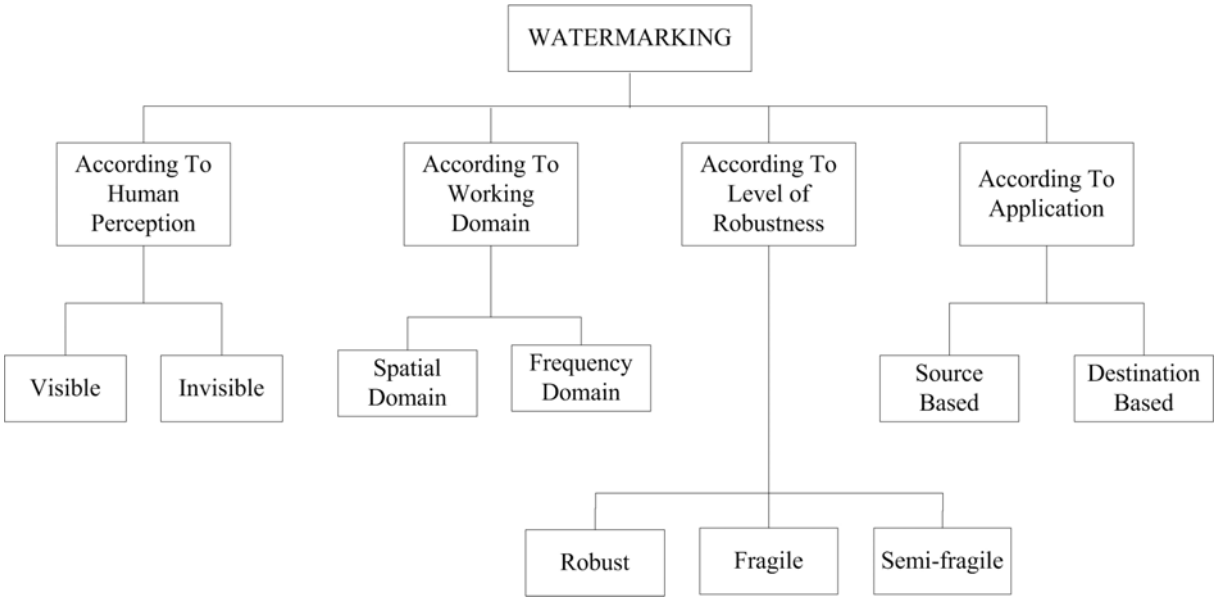


Figure 1. General classification of existing watermarking.

In general, most digital WM techniques proposed can be divided into two different types: visible and invisible watermarks according to human perception [7]. Each of the schemes is equally important due to its unique applications. Sometimes a certain application requires a watermark to be visible, so that the embedded watermark appears visible to a casual viewer. However, the most popular digital WM technique recently used for copyright authentication purposes is the invisible watermark, which marks the more significant digital media object, without perceptually changing it. By extracting the watermark data later with the appropriate decoding mechanism, it can be possible to make any malicious alteration detected to verify the integrity and the ownership of the digital media. The

hidden watermark can be meaningful, like a logo or tag or the information representing the customer identity. A simple example of such techniques can be found in Figure 2.

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 simple addition of a small amplitude pseudorandom 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. However, 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 (frequency domain) can minimize computational demanding operations. Moreover, frequency domain WM methods are more robust than the spatial domain techniques [5]. Therefore, working on compressed rather than uncompressed video is important for practical WM applications.

Figure 2. Example of an invisible watermark.

WM techniques proposed so far for media authentication are usually designed to be robust, fragile and semi-fragile watermarks according to the level of robustness. In copyright protection applications, the robust MW would be detectable even after an image or video frame goes through severe modifications and degradation. For image integrity applications it is efficient to apply a fragile WM which is designed to detect even the slightest change in the image. Most of the fragile WM methods perform the embedding in the spatial domain. Unlike the fragile WM techniques, a semi-fragile algorithm, such as the proposed algorithm, is designed to withstand certain innocuous manipulations but to reject malicious ones. Here, the innocuous manipulations mean the legitimate modifications such as lossy compression performed to the media during storage and distribution. The semi-fragile approaches are generally processed on the frequency domain (such as DCT, DWT), which make it substantially more attractive than the other two WM schemes [5].

From the application point of view, digital WM could be source based or destination based [13]. 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 WM applications are considered in the open literatures and as commercial applications [19]. The reader is referred to [19]-[21] for a more detailed investigation. The applications presented have been distilled down 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.

In Figure 3, a sketch of a simple video surveillance (VS) system, in which WM is used to authenticate VS data, is given [20], [21]. 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.

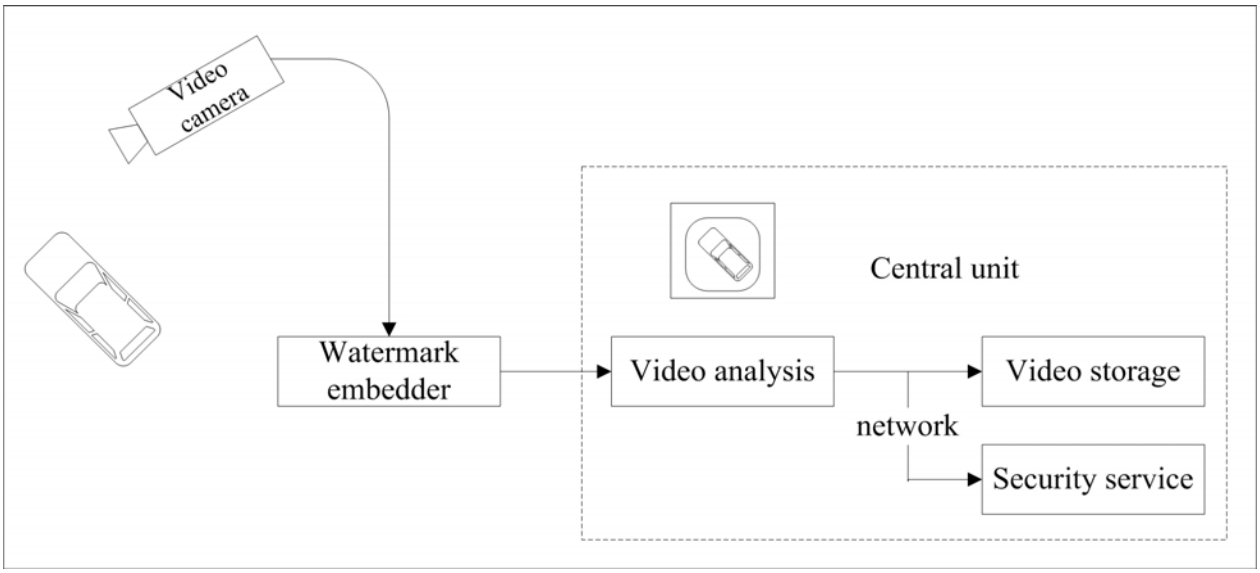


Figure 3. WM-based authentication for automatic VS.

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 4.

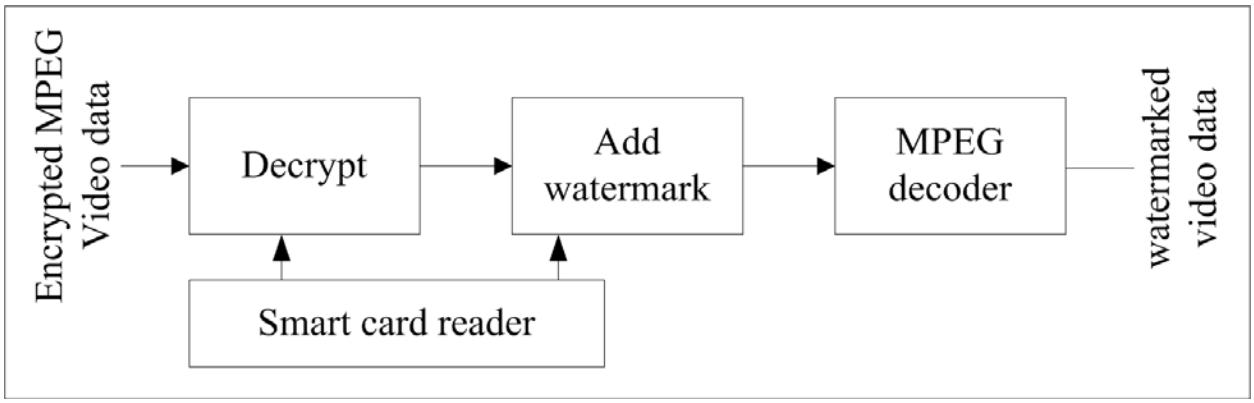


Figure 4. Set-top box with WM capabilities.

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 a copy protection scheme using WM techniques shown in Figure 5, consumers can make copies of any original source, but they cannot make copies of copies.

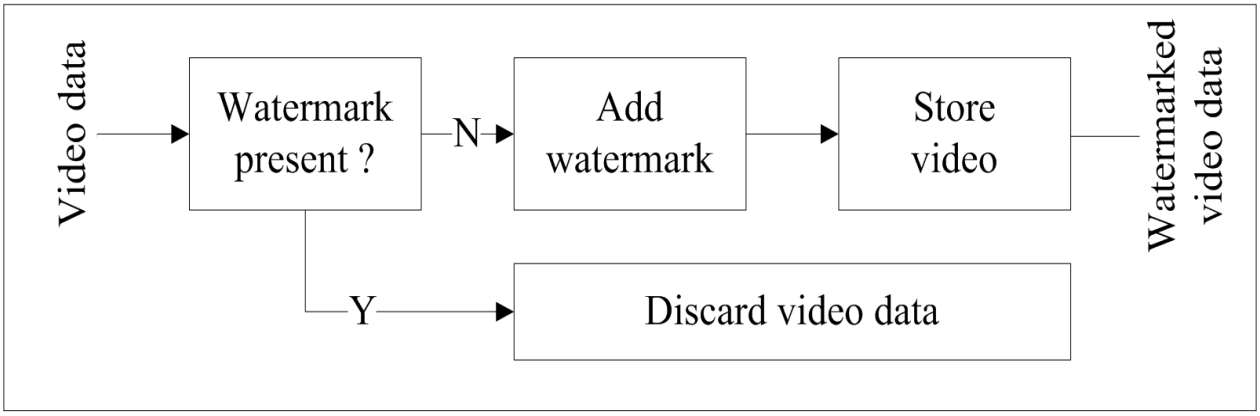


Figure 5. Video recorder with copy protection.

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.

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

In practical video storage and distribution systems, digital 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 [18]. Therefore, working on compressed rather than uncompressed video is important for practical WM applications. Before we describe the video WM techniques, a briefly description of the video compression standards will be presented in the next flowing subsection.

3.1 Compression Standards

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 [22]-[24]. 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. In the following, we refer particularly to description of MPEG-2 video compression technique.

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.
MJPEG	• Low bit rate video compression format based on JPEG compression. • Relatively low computational complexity.
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 general a video sequence can be divided into multiple group of pictures (GOP), representing sets of video frames which are contiguous in display order as illustrated in Figure 6(a). Each video frame is separated into slices and macro blocks. The block layer is formed by the luminance and chrominance blocks of a macro block. An encoded MPEG video sequence is made up of two frame-encoded pictures: intra-coded frames (I frames) and Inter-coded frames (P or B frames). P-frames are forward prediction frames and B-frames are bidirectional prediction frames. Within a typical sequence of an encoded GOP consisting of ten frames, as shown in Figure 6 (b), P-frames may be 10% the size of I-frames and B-frames 2%.

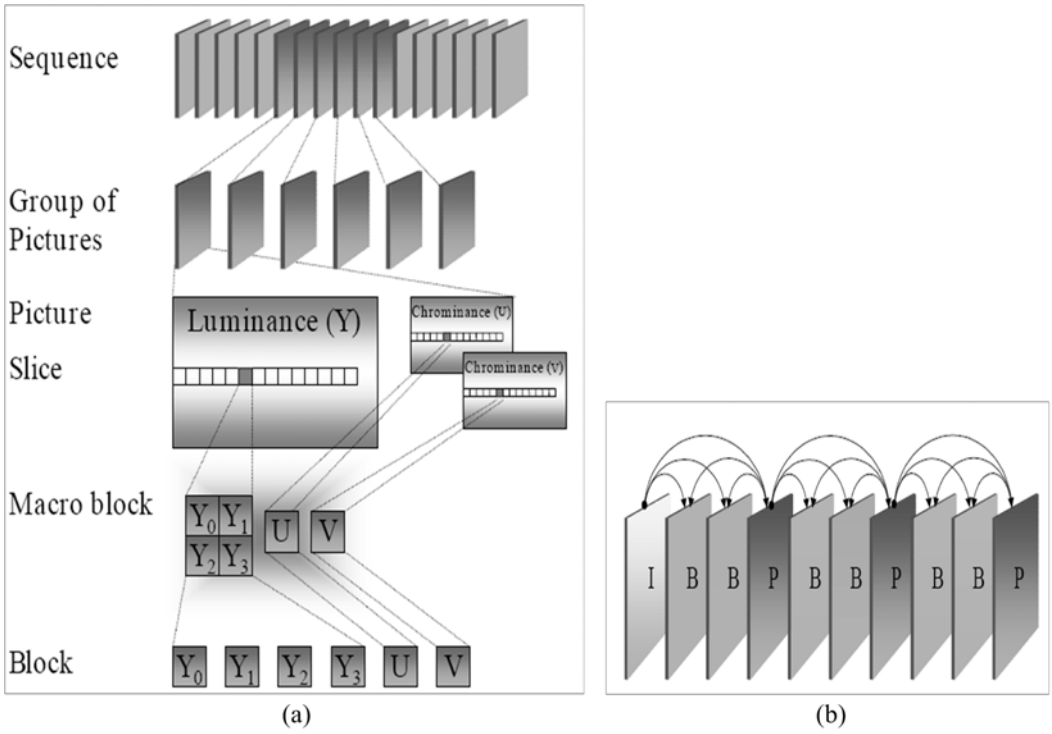


Figure 6. MPEG-2 GOP.

The MPEG-2 video compression algorithm is based on the basic hybrid coding scheme. As can be seen in Figure 7 this scheme combines inter-frame and intra-frame coding to compress the video data [23].

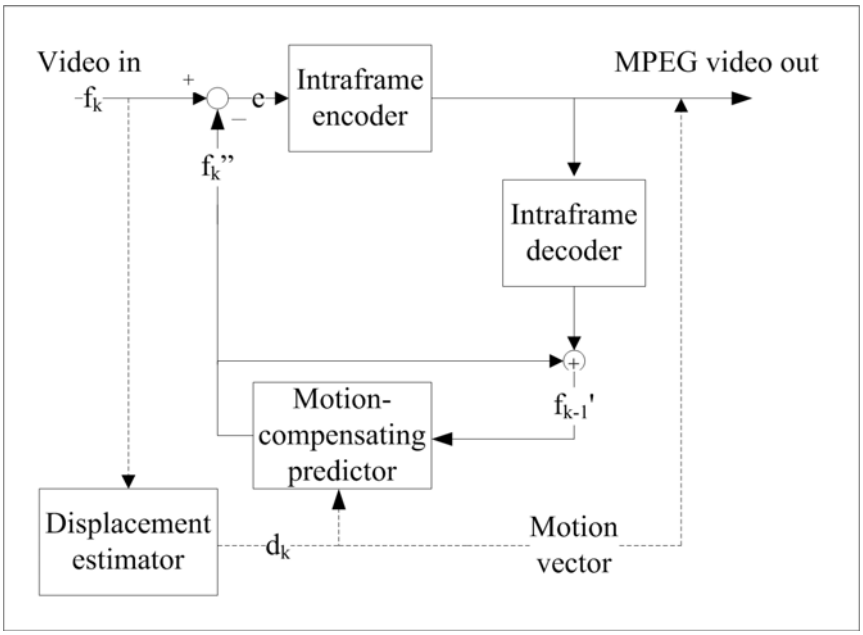


Figure 7. Block diagram of MPEG-2 encoder.

Within a GOP the inter-frames are temporally predicted by other motion compensated frames to reduce the temporal redundancy among the video frames. For each new inter-frame f_k, the motion compensating predictor will generate a prediction frame f_k' based upon the reconstructed previous frame f_{k-1}'' and a displacement estimate d_k that is obtained by an analysis of f_k. Since the original video is not available at the decoder d_k, also called motion vector, has to be transmitted. The prediction error (e), which is called the displaced frame difference, is encoded by the intra-frame encoder. The data flow of the MPEG-2 video encoder is shown in Figure 8.

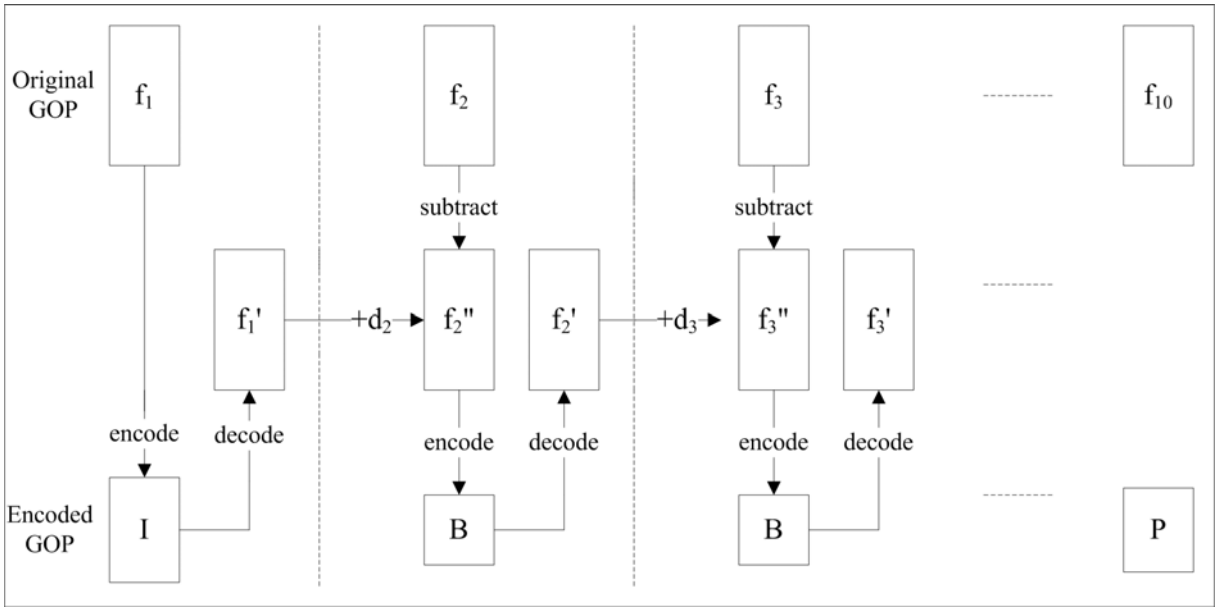


Figure 8. Data flow of MPEG-2 encoder.

The spatial redundancy in the prediction error (e) of the predicted frames and the intra-frames (I-frames) is reduced by the intra-frame encoder using the following operations(just like a JPEG image compression): they are split into blocks of size N × M (8 × 8) pixels which are compressed using the DCT, quantization (Q), zig-zag-scan, run-level-coding (Tuple coding) and entropy coding (VLC). Figure 9 depicts the procedure for the encoding of a single 8×8 block which is, in the bit-stream, represented as a series of Hunffman codewords.

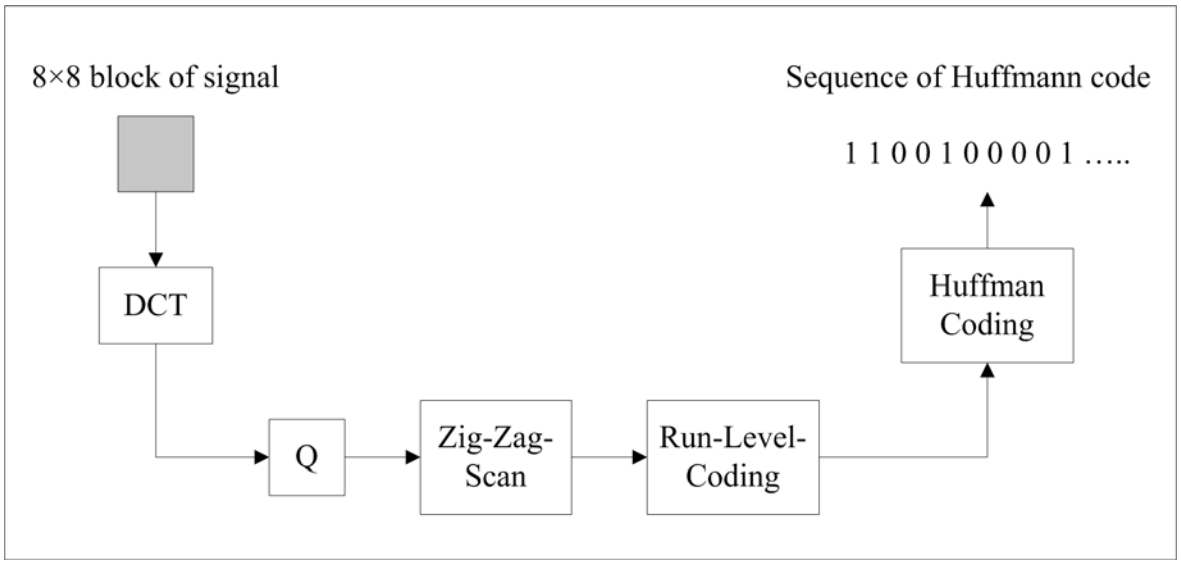


Figure 9. DCT encoding of 8×8 pixel block.

3.2 Software Implementations

Similar to image WM implementations, there exist tow kinds of video WM implementations: software and hardware, each having advantages and drawbacks. In software, the WM approaches hold advantages in terms of easy implementation and flexibility since 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. Moreover, 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.

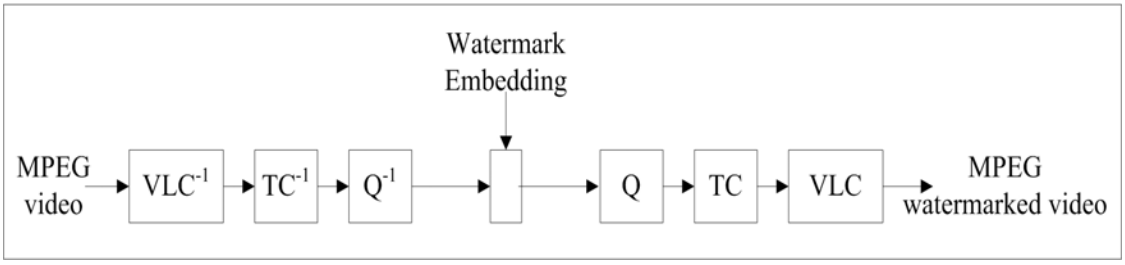


Figure 10. Block diagram of DCT-coefficient domain video WM.

A real-time WM algorithm for MPEG compressed video should closely follow the MPEG compression standard to avoid computationally demanding operations, like DCT and inverse DCT transforms or motion vector calculation.

Therefore, a MPEG compressed video WM algorithm, as shown in Figure 10, which operates on the DCT coefficient domain level only needs to perform VLC coding, tuple coding (TC) and quantization (Q) steps.

The basic idea for this WM algorithm is described as following several steps.

1. Generating a watermark message with the same manner and size as the video frame (to be watermarked).
2. Dividing watermark data into blocks of size N×M (such as 8 × 8) and computing the DCT coefficients for each watermark block.
3. The MPEG video frame is decoded (VLC^{-1} and TC^{-1}) and the resulting quantized DCT coefficients for all blocks then inversely quantized (Q^{-1}) in order to obtain the DCT coefficients.
4. The watermark is embedded into the video frame in DCT domain block-by-block according to certain algorithms and watermarked frame block is obtained.
5. The watermarked DCT coefficients for all blocks are re-encoded and the final result is a compressed watermarked video stream.

A major problem of directly modifying DCT-coefficients in an MPEG encoded video stream is drift or error accumulation. In an MPEG encoded video stream predictions from previous frames are used to reconstruct the actual frame, which itself may serve as a reference for future predictions. The degradations caused by the watermarking process may propagate in time, and may even spatially spread. Since all video frames are watermarked, watermarks from previous frames and from the current frame may accumulate and result in visual artifacts when decoding the MPEG watermarked video. By adding drift compensation signal during watermarking can solve this issue [4].

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 intra 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.

Wu proposed a method that adds a DCT transformed pseudorandom pattern directly to the DC-DCT coefficients of an MPEG compressed video stream [5]. The WM process only takes the luminance values of the I-frames into account. A spread spectrum method is used to watermark video frames, described by Shan in [6]. In a color frame, the mid-frequency DCT coefficients of a green component of the frame are selected to embed the watermark since it is the most robust after compression. Another research work on DCT coefficient domain WM for MPEG-2 compressed video has been presented in [8]. The proposed WM scheme was designed to be undeletable, perceptually invisible, statistically undetectable, and robust to lossy compression and survives to video manipulation and processing.

3.3 Hardware Implementations

Over the last decade, numerous software-based WM algorithms have been invented [15]. However, WM implementation in hardware, especially for video stream, is a recent interest in the area. Up to 1999, no work on video WM implementation in hardware had been shown [11]. However, the watermarking of video streams in real-time applications is mostly suitable for hardware implementations, thus motivating research efforts to that direction.

The hardware WM implementation is usually implemented in custom-designed circuitry, application specific integrated circuits (ASICs) or field programmable gate arrays (FPGAs). As shown in Table 3, we provide a comparative view of most hardware-based video WM designs developed so far. The overall advantage of this scheme over the software implementation is in terms of lower power consumption, reduced area and reliability. It can 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.

Table 3. Hardware-based implementations of Video WM.

Author	Design type	WM	Multimedia	Domain	Chip features
Maes	FPGA/IC	Invisible-robust	Video	Spatial	17/14 kG logics
Mathai	Custom IC	invisible	Video	Wavelet	1.8V
Tsai	Custom IC	Invisible-robust	Video	Spatial	NA

For example, in 2000, Strycker et al. proposed a real time video WM scheme, called Just Another Watermarking System (JAWS), for television broadcast monitoring [10]. JAWS is a well-known video WM algorithm and because it works on uncompressed real time video data, the author is allowed to concentrate on the watermark process and not on the compression issues. Therefore it is more suitable for hardware implementation. 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 [11], [12]. 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 spatial and transform domain is presented by Tsai and Wu [13]. In this scheme, the concepts of spread spectrum from the field of communication 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 stream 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. Hardware-based Video WM Design – A Development Methodology

In this section, we present a development methodology to design a hardware-based video WM system. Although our end goal is to implement the whole system monolithically on a single chip, it is expected that more than one prototype will be designed before a final version is issued. Therefore, it is worthwhile to first focus on determining the core elements of the system which include six functionality modules, such as video camera, watermark compressor, watermark generator, watermark embedder, control unit and memory, as well as the interface between them. A top view of a general scheme for the developed solution is depicted in Figure 11.

To improve the overall performance of the design, the expected system architecture should be designed to make most computational operations performed with temporal parallelism (using pipelining) and spatial parallelism (using parallel hardware).

Field-programmable gate array (FPGA) devices can be used to implement any logical function that an application-specific integrated circuit (ASIC) could perform [25]. The basically building blocks of many FPGA architectures are the programmable logic components that can be configured to perform logical functions and memory elements for data storage. Currently, the high density techniques of FPGA devices have been successfully used to build entire system on a single chip [26]. As they provide significant features in terms of high performance, efficient implementation, lower development cost and flexibility, it make them to be a highly attractive solution for hardware implementation of real-time video WM system. Here, a FPGA will be chosen to implement a prototype chip of the proposed video WM system, of which the FPGA implementation parts are shown in shady blocks as shown in Figure 8. Finally, the overall performances of the hardware implementation using FPGA can be evaluated by

performing experiments (offline and real-time with a CMOS image sensor) on the custom versatile breadboard described below.

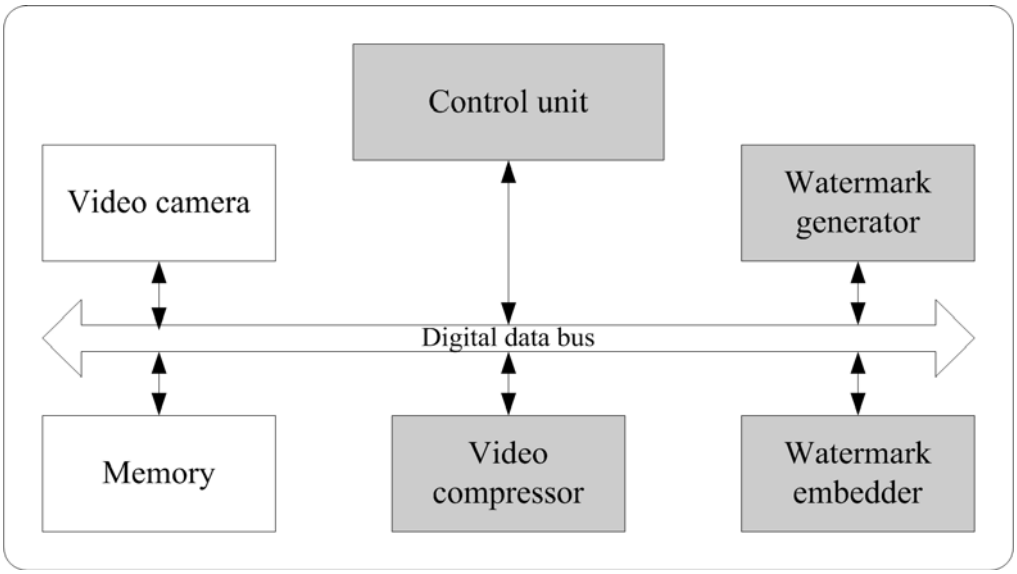


Figure 11. A general scheme of the video WM system.

The specifically designed board emulates a System-On-a-Chip (SoC) platform, allowing the incorporation of custom VLSI designs (such as the image sensor) with peripheral elements and digital logic implemented on an FPGA device. It features low-noise, separated digital and analog power supplies, 12 bit analog voltage and current biasing, 12 and 18 bit A/D converters, an SRAM memory and several I/O ports including LVDS, RS-232 and direct test points for maximum testing flexibility. The designer can choose what part of the system he wants to implement in VLSI and what elements he would use of those available on board. For the discussed digital video WM system implementation a basic imager is first designed, and then the WM modules including video compressor, watermark generator and embedder and control unit are implemented on the FPGA, together with all other required control logic, making use of the A/D converter and SRAM memory to aid the implementation of more complex algorithms. Therefore, the captured video stream can be watermarked at the origin such that the system security is improved as it is certain that the video data entering the system is untouched by any external party. Finally, the compressed watermarked video stream can be sent to SRAM memory for storage or transmitted to the host PC through the frame grabber for performance analysis.

5. Conclusions

In this paper, an in-depth overview of previous works on the field of digital video WM techniques was provided in order to provide help for further research works. Common WM classification criterions and requirements, including

general properties and specific constrains for video WM scheme, has been analyzed. Furthermore, various applications of video WM in practice were discussed, as well as the comparisons between software-based and hardware-based implementations from several points of view: major advantages, drawbacks and differences. Four examples of previous software and hardware WM implementations were also shown. In addition, a development methodology for hardware WM implementations including the general scheme of a proposed digital video WM system and its testing using the custom versatile breadboard were described.

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] F. Mintzer, G. Braudaway, and M. Yeung, “Effective and ineffective digital watermarks,” in proc. IEEE Int. Conf. Image Process., vol. 3, 1997, pp. 9-12.

[8] Chiou-Ting Hsu, and Ja-Ling Wu, “Digital Watermarking for Video,” 13th International Conference on Digital Signal Processing Proceedings, DSP 97. Vol. 1, July 1997 pp. 217 – 220.

[9] Christoph Busch, Wolfgang Funk, and Stephen Wolthusen, “Digital Watermarking: From Concepts to Real-Time Video Applications”. IEEE Computer Graphics and Applications. Vol. 19, Issue 1, Jan.-Feb. 1999. pp. 25 – 35.

[10] 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.

[11] 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.

[12] 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.

[13] 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

[14] 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.

[15] 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.

[16] 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.

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

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] ISO/IEC 13818-2:1996(E), "Information Technology – Generic Coding of Moving Pictures and Associated Audio Information" , Video International Standard, 1996.

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

[24] Andrey Filippov, "Encoding High-Resolution Ogg/Theora Video with Reconfigurable FPGAs," in Xcell Journal. Second Quarter 2005.

[25] B. Shackleford, G. Snider, R. J. Carter, E. Okushi, M. Yasuda, K. Seo, and H. Yasuura, "A high-performance, pipelined, FPGAbased genetic algorithm machine," Genetic Programming and Evolvable Machines, vol. 2, no. 1, pp. 33–60, March 2001.

[26] S. O. Memik, A. K. Katsaggelos, and M. Sarrafzadeh. "Analysis and FFGA Implementation of Image Restoration Under Resource Constrain," IEEE Trans. on Computers, VoL52. N0.3, Mar. 2003.

Authors' Information

Xin Li – 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: fishi@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, Isreal; e-mail: orly@atips.ca

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

Billy 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 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: capsule endoscopy, gastrointestinal disorders, aided magnetic levitation.

ACM Classification Keywords: J.3 Life and Medical Sciences

Introduction

Gastrointestinal (GI) disorders, including cancers, are common medical problems which affect up to 35% of the world population [1, 2]. 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].

Endoscopy

An endoscope (Figure 1.) is a fiberoptic instrument that can be inserted in the GI tract through the mouth or the rectum.

Endoscopy is a common diagnostic technique which allows direct viewing of the internal wall of the organ [2]. Photographs of different sections of the internal wall can be obtained, thus helping the medical diagnosis of GI disorders. Endoscopy is important for the early detection of cancers. People who are at low risk for cancer, or have no symptoms, are recommended to schedule an endoscopy every 3 – 5 years after turning 50 years old. Those who are at high risk should take regular endoscopic examination prior to age of 40 [1].

The endoscope is a long flexible tube having a camera and light source at the end. This camera is connected to an eyepiece or a video screen in order to display the images on a color TV set. Modern endoscopes can diagnose GI disorders, as well as perform treatment [2]. They can view inside the body of a patient without any significant pain or discomfort, and include two different optical fibers: an image fiber and a light fiber. The light fiber delivers illumination into the body cavity, while the image fiber transmits an image of the illuminated body cavity back to a lens system. Depending on the medical application, the fiberoptic endoscopes are divided into two categories: regular endoscopes such as gastroscopes and colonscopes, and ultra-thin endoscopes, utilized in needlescopes, ophthalmic endoscopes, and angioscopes.

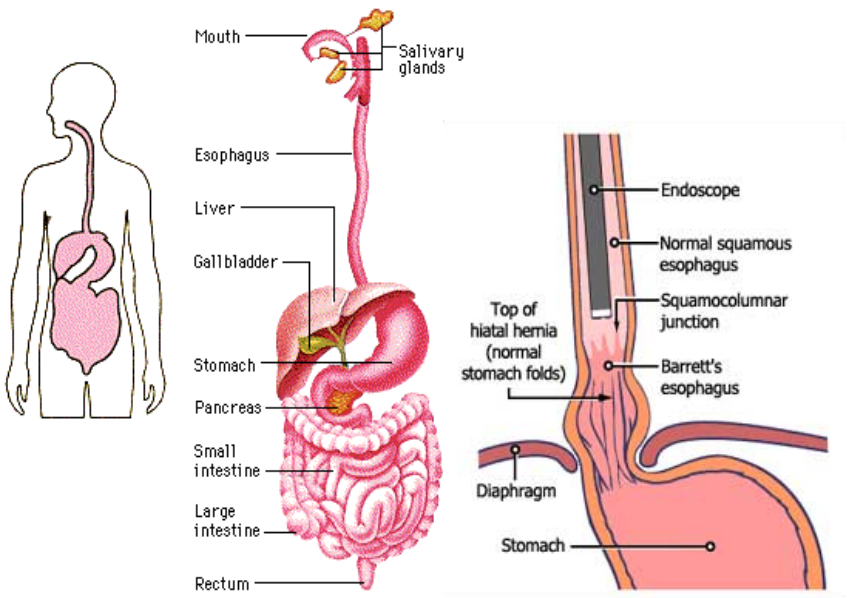


Figure 1: GI tract and endoscope

Capsule Endoscopy

A capsule endoscope is a small swallowable microelectronic pill that contains light-emitting diodes (LEDs), a microelectronic imager, a battery, and a radio-frequency (RF) transmitter. Upon illumination by the LEDs, the imager can periodically take pictures and transmit them to a data recorder wirelessly, while the capsule moves throughout the GI tract as a result of natural peristalsis. Capsule endoscopy (CE) has been utilized for diagnosing small intestinal bleeding and has been recently suggested for diagnosing abnormalities of the esophageal wall [4].

The beginning of CE can be traced to 1997, when a video capsule with wireless endoscopic system was invented [3]. The capsule had an innovative lens and could be readily swallowed. Since it was propelled by natural peristalsis, there was no external steering control on the capsule to position it. In 2001, a system to monitor physiological parameters in the GI tract was proposed [5]. This was followed by the invention of a system to monitor physiological parameters in human lumen in 2004 [6]. For both inventions, the capsule was affixed to the inner wall of the lumen using invasive objects such as attachment band or pin. In 2004, magnetically controlled capsule endoscopy was suggested [7]. Magnetic force was used to propel the capsule and to control its direction, but magnetic levitation was not employed. There was a disclosure of an energy saving method, using a motion detector to sense the movement of the capsule and shut down the circuit when not in use [8]. A magnetic navigation system for imaging in the gut was also disclosed [9]. It was mechanically coupled with a permanent magnet which orients the capsule according to the force produced by the magnetic field. None of the aforementioned techniques for capsule movement control utilized magnetic levitation to precisely position and steer the capsule. In 2003, diamagnetic facilitation was used to levitate a mass, but not for the purpose of stabilizing the position of an endoscopic capsule [10].

Problems with Capsule Endoscopy

The dependence of the capsule movement on natural peristalsis and the inability to achieve positional stabilization preclude broader applicability of this diagnostic technique. Thus, control over the position of the capsule is a major unresolved issue in CE. Since the capsule cannot be positioned close to the inner walls of larger-lumen organs, it cannot be affixed at a particular location for local physiological monitoring or imaging.

Microsystem Diagnostic Capsules

In the recent years, swallowable capsule endoscopy (CE) has evolved as a serious 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]. The advent of microsystem design has made it possible for a capsule endoscope of the size of a medical pill to 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 and the available literature quantifying its peristaltic forces [11]), 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 [11].

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 [10]. 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 [9]. 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) [10].

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 [12] 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 (Fig. 2).

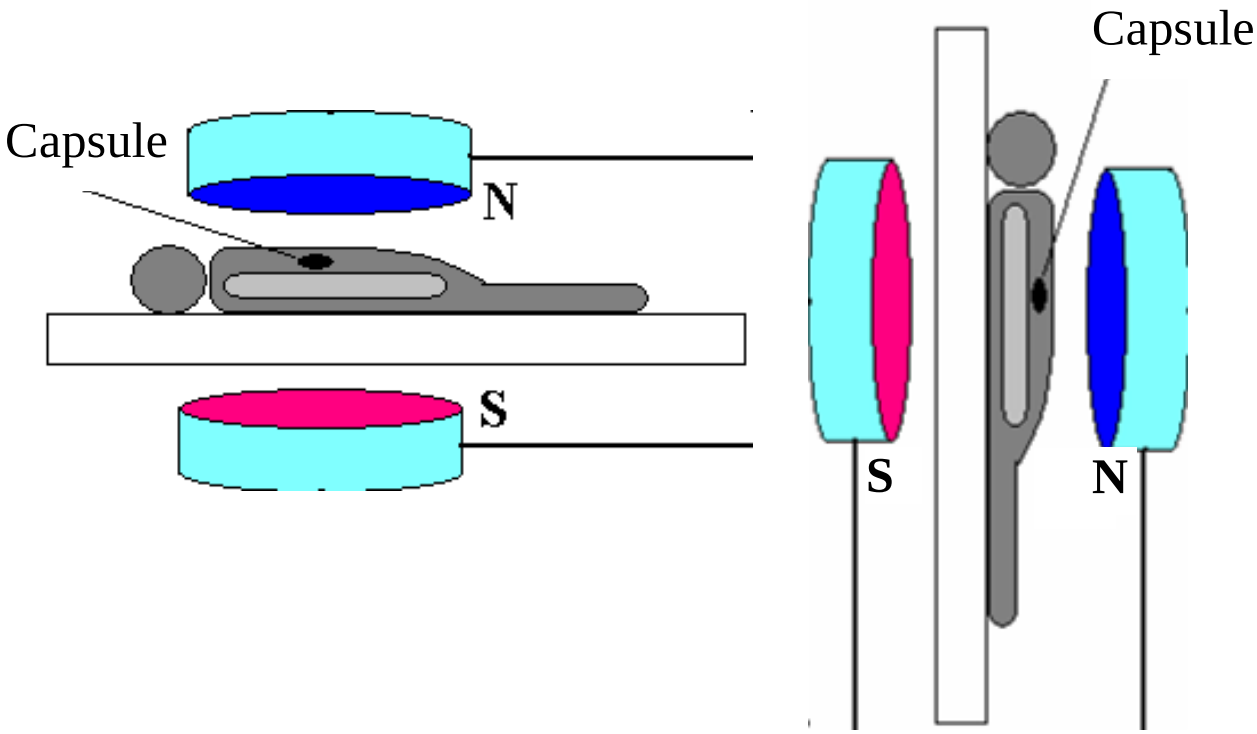


Fig. 2. Positioning the patient during diagnosis, supine (left) and upright (right).

Modeling Setup

The force between two permanent magnets or between a permanent magnet and an electromagnet is the result of non-uniform magnetic field [13]. 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.

Experimental Setup

A levitation kit [12], 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.

Laboratory Setup for Magnetic Levitation

To understand better the requirements for magnetic levitation, the levitation kit mentioned above was purchased and built. The setup (Figure 3) was analogous to a patient in the supine position. The initial goal was to establish stable levitation and explore the features of Maxwell 3D. In this case, the forces on the levitating magnet were gravitational and magnetic. The peristaltic force was not considered at the time. The setup was gradually refined. A larger hand-wound coil was used as the electromagnet, with roughly 5000 turns of AWG 20 copper wire. The core of the coil was a 1018- stainless-steel screw. A spherical NdFeB permanent magnet was placed in a clay-filled capsule to give a combined mass of 10.9g.

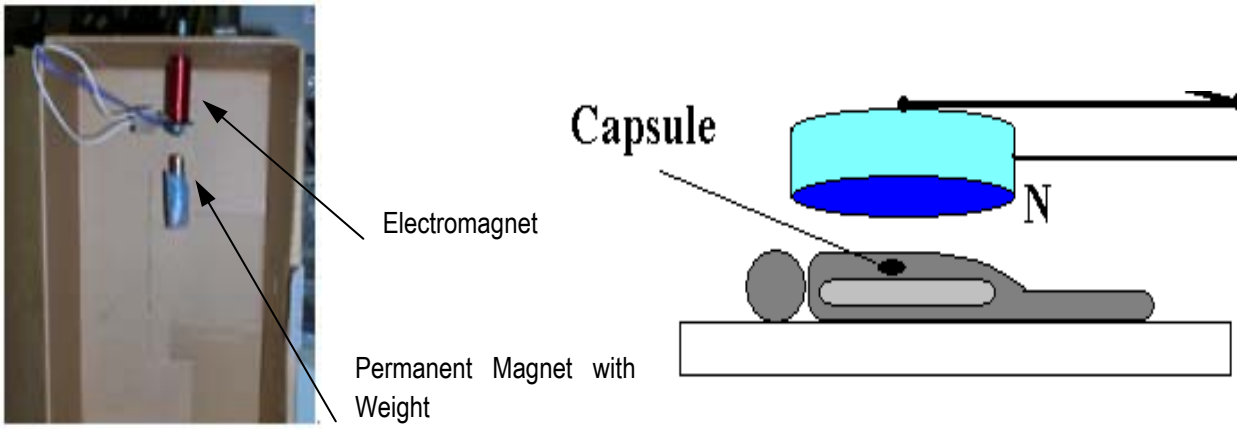


Figure 3: Laboratory setup modeling a patient in the supine position

A simulation model was built using Maxwell 3D. The Hall Effect sensor was placed directly below the coil along the z-axis (the centre axis of the coil). 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.

The simulation agreed with the physical setup, as the upward force on the magnet given by Maxwell 3D was 0.106 N, which was the combined weight of the capsule and the magnet.

Distribution and Direction of the Magnetic Field

The magnetic field was plotted for analyzing its distribution and direction. Due to magnetic torque, the magnet oriented itself so that its magnetic dipole was in the direction of the coil magnetic field (the z-axis). Thus, the magnetic field along the z-axis consisted of the z-component only. The magnetic field was the greatest around the magnet.

To understand the effect of varying the vertical position of the magnet (in order to imitate the stabilization of the magnet position), simulations were run with the magnet slightly below and slightly above the equilibrium position. When the magnet was above the equilibrium position, the control circuit decreased the coil current and, in effect, the magnetic field of the coil. The attractive magnetic force between the coil and the magnet decreased as a result, and the magnet was brought back to the equilibrium position by gravity. Similarly, when the magnet dropped below the equilibrium position, the control circuit increased the coil current to lift the magnet back up.

The biggest challenge in running the simulations was finding the corresponding coil current for each non-equilibrium position. To verify the simulation results, the coil current of the physical setup was measured for each position. Since the examined variation of the magnet position was very small (in the range of +/- 1 mm), light projection was used to magnify the separation distance between the Hall Effect sensor and the magnet (Figure 4). The magnet was physically forced to a non-equilibrium position and the corresponding coil current was recorded.

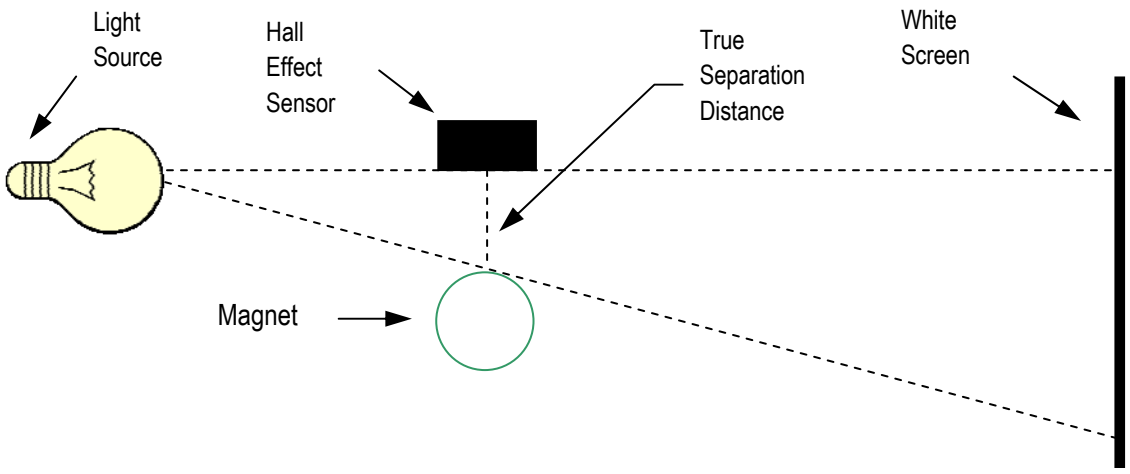


Figure 4: Setup for measuring the magnet position

For the simulations, the magnitude of the coil current was deemed appropriate when it resulted in an upward force on the magnet roughly equivalent to its weight.

The values obtained from the simulations were checked with the measured values, and they coincided.

Deviation of +1 mm from the equilibrium position meant that the magnet was above its equilibrium position (closer to the coil) by 1mm. Deviation of -1 mm meant that the magnet was below (further away) by 1 mm.

Steering in Supine Position

A simulation CAD model (Fig. 5) 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.

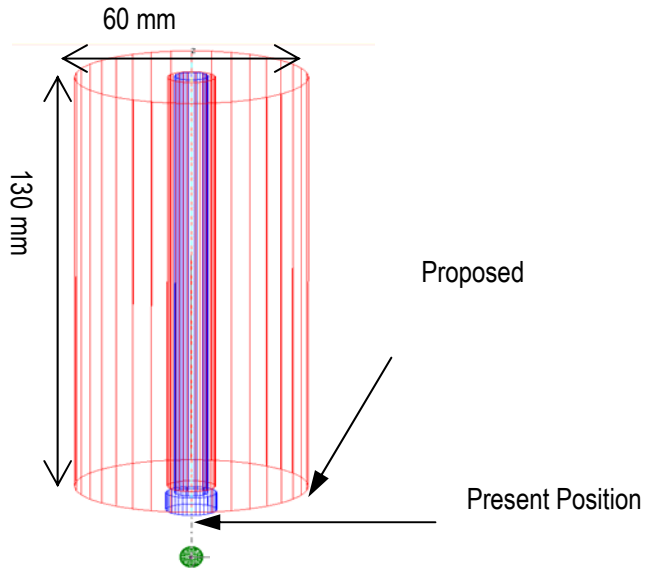


Fig. 5. Experimental setup for supine position
with the present and proposed positions for the Hall Effect sensor.

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. 5) 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. 6 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).

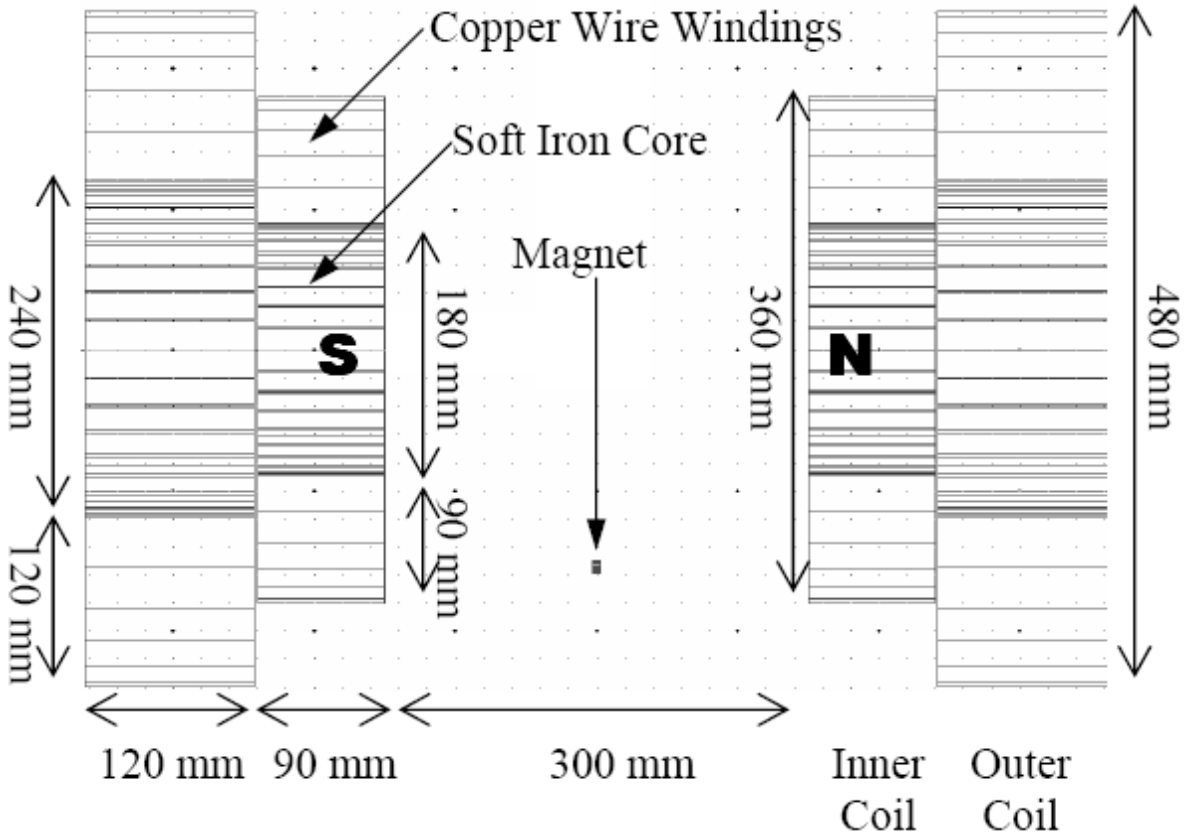


Fig. 6. Cross-sectional view of the simulation model for upright position in an actual application.

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 [10]. Soft iron cores were used to concentrate the magnetic field generated.

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 [10]. This stabilization aid can substantially ease the sensitivity requirements for the Hall Effect sensors.

Results

Steering in Supine Position

Fig. 7 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.

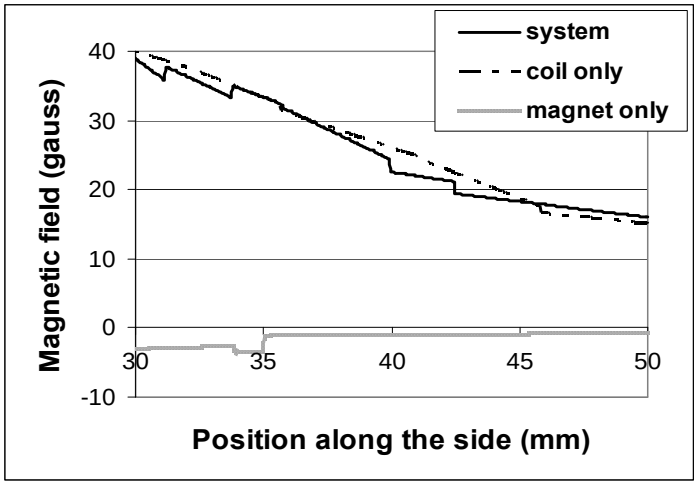


Fig. 7. Breakdown of the field contributions near the proposed position (supine case).

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.8.

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.

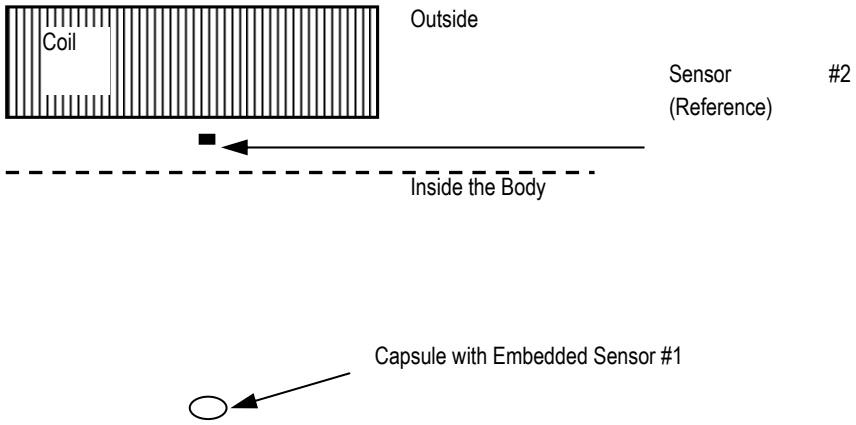


Fig. 8. Two-sensor concept for supine position.

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:

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

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:

$$m = V\rho = 2\pi A_w \rho \sum_{n=1}^{N} a_n \tag{2}$$

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:

$$a_n = a_{n-1} + d_w = a_{n-2} + 2d_w = \ldots = a_1 + (n-1)d_w \tag{3}$$

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:

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

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

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 [11], 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 optimally 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 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] G. J. Iddan and D. Sturlesi, “In Vivo Video Camera System,” U.S. Patent 5,604,531, February 18, 1997

[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] J. T. Kilcoyne, R. Tsukashima, G. M. Johnson, and C. Klecher, "Remote Physiological Monitoring System," U.S. Patent 6,285,897, September 4, 2001.

[6] J. T. Kilcoyne, R. Tsukashima, G. M. Johnson, and C. Klecher, "Implantable Monitoring Probe," U.S. Patent 6,689,056, February 10, 2004

[7] G. M. Wakefield, "Magnetically Propelled Capsule Endoscopy," U.S. Patent Application 20040199054, October 7, 2004

[8] G. J. Iddan and G. Meron, "Energy Management of a Video Capsule," U.S. Patent Application 20040236182, November 25, 2004

[9] S. Jin, "Magnetic Navigation System for Diagnosis, Biopsy and Drug Delivery Vehicles," U.S. Patent 6,776,165, Aug. 17, 2004

[10] 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.

[11] 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.

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

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

[14] 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.

[15] 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. H. Wu – Graduate Student, Department of Electrical and Computer Engineering, University of Calgary, Calgary, Alberta, Canada T2N1N4; e-mail: wu@enel.ucalgary.ca*

Major Fields of Scientific Research: Embedded Systems, Electronic Instrumentation

***Martin** P. Mintchev – Professor, Department of Electrical and Computer Engineering, University of Calgary, Calgary, Alberta, Canada T2N1N4; Fellow, American Institute for Medical and Biological Engineering, Washington, DC, USA; e-mail: mintchev@ucalgary.ca*

Major Fields of Scientific Research: Biomedical Instrumentation, Navigation, Information Systems in Medicine

HIGH-PERFORMANCE INTELLIGENT COMPUTATIONS FOR ENVIRONMENTAL AND DISASTER MONITORING

Nataliia Kussul, Andrii Shelestov, Sergii Skakun, Oleksii Kravchenko

Abstract: In this paper we present different approaches to multi-source data integration for the solution of complex applied problems, in particular flood mapping and vegetation state estimation using satellite, modelling and in-situ data. Since these applications are data- and computation-intensive, we use Grid computing technologies. In such a case computational and informational resources are geographically distributed and may belong to different organisations. For this purpose, we also investigate benefits of different approaches to the integration of satellite-based monitoring systems.

Keywords: data integration, Earth observation, flood mapping, inverse modelling, neural network, Grid technologies, information system, geospatial information.

ACM Classification Keywords: I.5.1 [Pattern Recognition] Models – Neural nets; G.1.8 [Numerical Analysis] Partial Differential Equations - Inverse problems; D.2.12 [Software Engineering] Interoperability; F.1.2 [Theory of Computation] Modes of Computation - Parallelism and concurrency; F.1.1 Models of Computation - Probabilistic computation; G.4 Mathematical Software - Parallel and vector implementations; H.1.1 [Information Systems] Models and Principles - Systems and Information Theory; H.3.5 [Information Storage and Retrieval] Online Information Services; I.4.6 [Image Processing and Computer Vision] Segmentation - Pixel classification; I.4.8 Scene Analysis - Sensor fusion; J.2 [Computer Applications] Physical Sciences and Engineering - Earth and Atmospheric Sciences

Introduction: Specifics of Earth Observation Problems

Nowadays, due to global climate change the solution of such problems as rational land use, environmental monitoring, prediction of natural disasters and so forth became the task of a great importance. The basis for the solution of these problems lies in the integrated use of data from multiple sources: modelling data, in-situ measurements and remote sensing observations. However, inconsistency of heterogeneous data and measurement techniques, spectral, spatial and temporal disarrangement of data limit the potential of up-to-date technologies for the solution of crucial problems of arising in the different social benefit areas. Therefore, there is a considerable need for the development of methods and technologies for integration of heterogeneous data coming from multiple sources.

Recent advances in satellite and sensor technologies made the Earth Observation (EO) data from space to play a major role in the solutions of applied problems in different domains. Satellite observations enable acquisition of data for large and hard-to-reach territories, and can provide continuous measurements and human-independent information. Such important applications as monitoring and prediction of floods, droughts, vegetation state assessment heavily rely on the use of EO data from space. For example, the satellite-derived flood extent is very important for calibration and validation of hydraulic models to reconstruct what happened during the flood and determine what caused the water to go where it did [Horritt, 2006]. Information on flood extent provided in the near real-time (NRT) can be also used for damage assessment and risk management, and can benefit to rescuers during

the flooding. Both microwave and optical data can provide means to detect drought conditions, estimate drought extent and assess the damages caused by droughts [Kogan et al., 2004; Wagner et al., 2007].

It should be also emphasized that the same EO data sets and derived products can be used for a wide variety of applications. For example, land use/change information, soil properties and meteorological conditions are both important for floods and droughts applications as well as vegetation state assessment. That is, once the corresponding interfaces are developed to enable access to these data and products they can be used in a uniform way for different purposes and applications. Services and models that are common for different EO applications (e.g. flood monitoring and crop yield prediction) are shown in Fig. 1.

The EO domain, in turn, is characterized by large volumes of data that should be processed, catalogued, and archived. For example, GOME instrument onboard Envisat satellite generates nearly 400 Tb data per year [Fusco et al., 2003]. The processing of satellite data is carried out not by the single application with monolithic code, but by distributed applications. This process can be viewed as a complex workflow (DEGREE) that is composed of many tasks: geometric and radiometric calibration, filtration, reprojection, composites construction, classification, products development, post-processing, visualization, etc. [Rees, 2001].

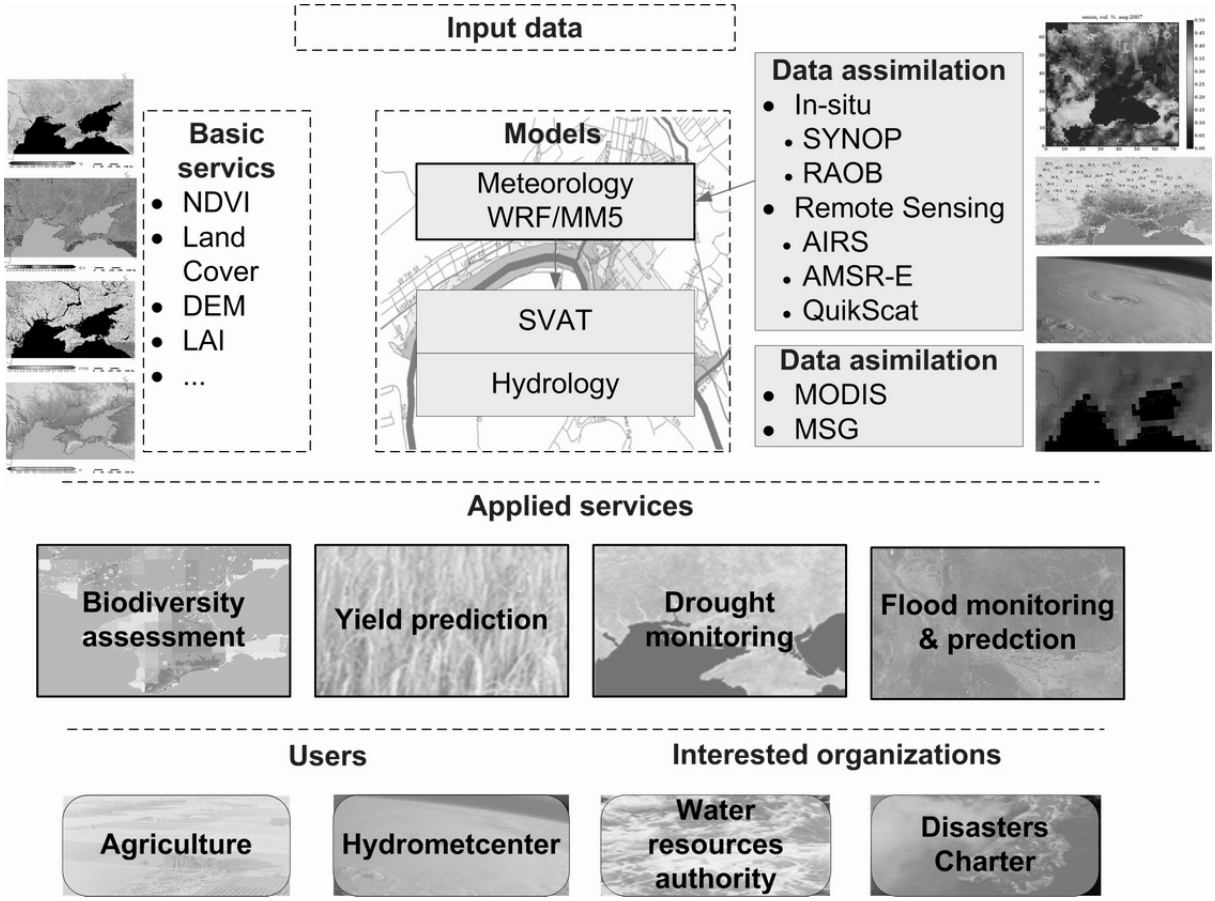


Figure 1. Common services and models for a variety of applications

To enable processing and management of such volumes of data sets and information flows an appropriate infrastructure is needed that will support [Fusco et al., 2003; Shelestov et al., 2006]: access to distributed resources; high flexibility; portal enabling easy and homogeneous accessibility; collaborative work; seamless integration of resources and processes; allow processing of large historical archives; avoid unauthorised access to/use of resources.

Grid can provide appropriate facilities for high-performance computations and efficient data management in EO domain. Grid computing is an emerging paradigm for global computing and a very active research domain for complex, dynamic, distributed and flexible computing and resource sharing [Foster and Kesselman, 2004]. Grid computing belongs to main trends of on-line environment development among with web services, semantic web and peer-to-peer networking. The integration of these technologies is essential for the next generation networks.

Grid systems are recognized to be very efficient for EO and geospatial community for a number of reasons: geospatial data and associated computational resources are naturally distributed; the multi-discipline nature of geospatial research and applications requires the integrated analysis of huge volume of multi-source data from multiple data centres; most geospatial modelling and applications are both data and computational intensive.

In this paper we present different approaches to multi-source data integration for the solution of complex applied problems, in particular flood mapping and vegetation state estimation using satellite, modelling and in-situ data. Since these applications are data- and computation-intensive, we use Grid computing technologies. In such a case computational and informational resources are geographically distributed and may belong to different organisations. For this purpose, we also investigate benefits and approaches to the integration of satellite-based monitoring systems.

The rest of the paper is organised as follows. First, we briefly review the existing Grid-systems for EO data processing. Then we describe in details two applications that are solved using multi-source data integration, and have been ported into Grid platform at the Space Research Institute NASU-NSAU. Then, we focus on the integration of geographically distributed information systems that rely on the use of EO data and implemented using Grid computing technologies.

State of the Art: Grid-based Systems for EO Data Processing

At present, Grid technologies are widely applied in different domains, in particular EO domain. The European DataGrid Project (EDG) was the first large European Commission-funded grid project (www.eu-datagrid.org) [Fusco et al., 2003]. Many of the results of EDG project have been included in the European project Enabling Grids for E-sciencE (EGEE). EGEE aims to develop a service grid infrastructure which is available to scientists 24 hours-a-day (http://www.eu-egee.org). Now EGEE and other existing Grid infrastructures in Europe are transitioned to the European Grid Initiative model (http://web.eu-egi.eu).

Based on the gained experience ESA and the European Space Research Institute (ESRIN) have developed Grid Processing on Demand (G-POD) for Earth Observation Applications (http://gpod.eo.esa.int). Online access to different data is enabled within this project, in particular to data provided by various instruments onboard Envisat satellite (http://envisat.esa.int), SEVIRI instrument onboard MSG (Meteosat Second Generation) satellite, ozone profiles derived from GOME instrument, etc. One of the most important applications is the analysis long-term data. Grid Web Portal provides access to the “Grid-on-demand” resources enabling: personal certification, time/space

selection of data directly from the ESA catalogue, data transfer, job selection, launching and live status, data visualization.

A major challenge for DEGREE (Dissemination and Exploitation of GRids in Earth science, http://www.eu-degree.eu) project was to build a bridge linking the Earth Science and GRID communities throughout Europe, and focusing in particular on the EGEE-II Project. Grid provides appropriate infrastructure enabling international cooperation within GMES and GEOSS. The following problems were within the scope of DEGREE: earthquake analysis, floods modelling and forecasting, influence of climate changes on agriculture

Japan Aerospace eXploration Agency (JAXA) and KEIO University started establishing “Digital Asia” system aimed at semi-real time data processing and analyzing. They use GRID environment to accumulate knowledge and know-how to process remote sensing data. The Digital Asia project is the part of bigger Sentinel Asia project that is targeting on building natural disasters monitoring system Fukui, 2007].

CEOS Wide Area Grid (WAG) project was initiated by the CEOS Working Group on Information Systems and Services (WGISS), and aims at providing horizontal infrastructure enabling efficient integration of resources of different space agencies. WAG testbed infrastructure is currently under development within ESA Cat-1 project “Wide Area Grid Testbed for Flood Monitoring Using Spaceborne SAR and Optical Data” (no. 4181) [Kopp et al., 2007]. Within the WAG project Space Research Institute NASU-NSAU has developed a testbed that integrates resources of Ukrainian Grid segment (Ukrainian Academician Grid) with resources of international organisations (ESA, CEODE-CAS). The tesbed is described in more details in the subsequent sections.

Applications

In this section we focus on the detailed description of the two applications that are solved within Grid environment at the Space Research Institute NASU-NSAU:

- flood mapping from synthetic-aperture radar (SAR) satellite imagery, and
- vegetation state estimation using remote sensing and modelling data.

Flood Mapping from Satellite Imagery. In recent decades the number of hydrological natural disasters has considerably increased. According to [Scheuren et al., 2008], we have witnessed in recent years a strengthening of the upward trend, with an average annual growth rate of 8.4% in the 2000 to 2007 period. Hydrological disasters, such as floods, wet mass movements, represent 55% of the overall disasters reported in 2007, having a tremendously high human impact (177 million victims) and causing high economic damages (24.5 billion USD) [Scheuren et al., 2008].

EO data from space can provide valuable and timely information when one has to respond to and mitigate such emergencies as floods. From satellite imagery we can determine flood areas, since it is impractical to provide such information through field observations. The use of optical imagery (in visible and infra-red range) for flood mapping is limited by severe weather conditions, in particular by the presence of clouds. In turn, synthetic aperture radar (SAR) measurements from space are independent of daytime and weather conditions and can provide valuable information to monitoring of flood events. This is mainly due to the fact that smooth water surface provides no return to antenna in microwave spectrum and appears black in SAR imagery [Rees, 2001].

Flood mapping procedure from SAR imagery represents a complex workflow and consists of the following steps. The first step consists in re-constructing a satellite imagery taking into account the calibration, the terrain distortion using digital elevation model (DEM) and providing exact geographical coordinates. The second step is image segmentation, and the third step consists in the classification to determine the flood extent.

In this subsection we describe a neural network approach to flood mapping from satellite SAR imagery that is based on the application of self-organizing Kohonen's maps (SOMs) [Kohonen, 1995; Haykin, 1999]. The advantage of using SOMs is that they provide effective software tool for the visualization of high-dimensional data, automatically discover of statistically salient features of pattern vectors in data set, and can find clusters in training data pattern space which can be used to classify new patterns [Kohonen, 1995]. We applied our approach to the processing of data acquired from different satellite SAR instruments (ERS-2/SAR, ENVISAT/ASAR, RADARSAT-1 and RADARSAT-2) for different flood events: river Tisza, Ukraine and Hungary (2001); river Huaihe, China (2007); river Mekong, Thailand and Laos (2008); river Koshi, India and Nepal (2008); river Norman, Australia (2009); and river Zambezi, Mozambique (2008) and Zambia (2009).

To this end, different methods and approaches were proposed to flood mapping using satellite imagery:

- multi-temporal technique (http://earth.esa.int/ew/floods);
- threshold segmentation [Cunjian et al., 2001];
- statistical active contour model [Horritt, 1999];
- edge-detection techniques [Niedermeier et al., 2000];
- analysis of time-series of SAR images [Martinez and Le Toan, 2007].

The following shortcomings of the existing approaches can be identified: manual threshold selection and parameters identification; statistical models require a priori knowledge of image statistical properties; application of complex models for noise (speckle) reduction; no spatial neighbourhood between pixel is considered. More detailed description of the existing techniques is given [Kussul et al., 2008a].

Data set description. We applied our approach to the processing of remote-sensing data acquired from different satellite SAR instruments for different flood events:

- ERS-2/SAR: flood on Tisza river (Ukraine), 2001;
- ENVISAT/ASAR Wide Swath Mode (WSM): river Huaihe, China, 2007; river Zambezi, Mozambique, 2008; river Mekong, Thailand and Laos, 2008; river Koshi, India and Nepal, 2008; Ha Noi City, Vietnam, 2008; river Zambezi, Zambia, 2009;
- RADARSAT-1: river Huaihe, China, 2007;
- RADARSAT-2: river Norman, Queensland, Australia, 2009 (see Fig. 2).

Data from European satellites (ERS-2 and ENVISAT) were provided from the ESA Category-1 project "Wide Area Grid Testbed for Flood Monitoring using Spaceborne SAR and Optical Data" (№4181). Data from RADARSAT-1 satellite were provided from the Center of Earth Observation and Digital Earth (China). RADARSAT-2 data were provided by the Canadian Space Agency (CSA) within the GEOSS Architecture Implementation Pilot Phase 2 (AIP-2, http://www.ogcnetwork.net/AIpilot).

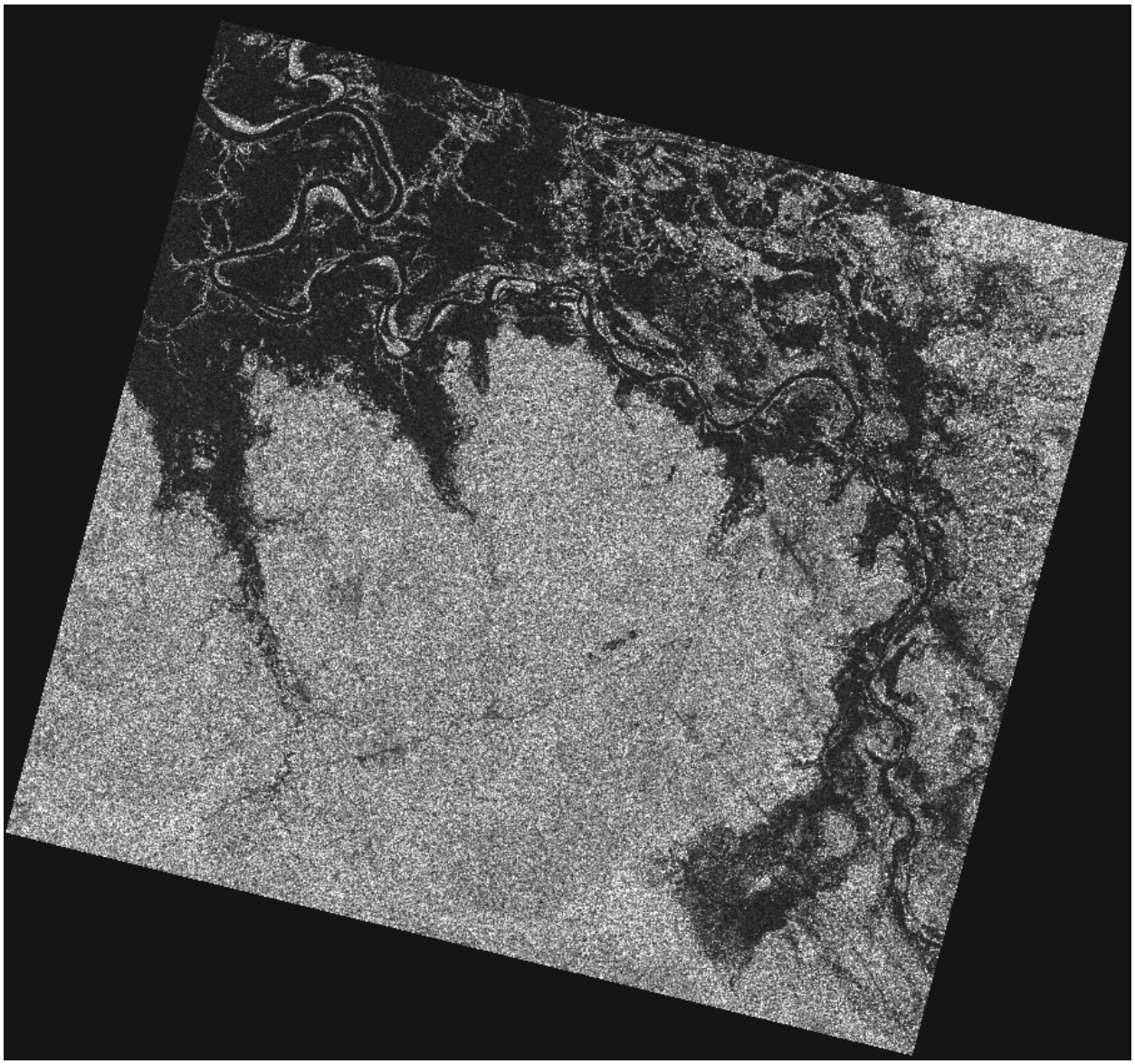


Figure 2. SAR image acquired from RADARSAT-2 satellite (date of acquisition 14.02.2009) during the flood on the river Norman, Australia (RADARSAT-2 Data and Products © MacDONALD, DETTWILER AND ASSOCIATES LTD. 2009 – All Rights Reserved. RADARSAT is an official mark of the Canadian Space Agency)

A pixel size and ground resolution of ERS-2 imagery (in ENVISAT format, SLC — Single Look Complex) were 4 m and 8 m, respectively; for ENVISAT imagery - 75 m and 150 m; and for RADARSAT-1 imagery - 12.5 m and 25 m; for RADARSAT-2 imagery – 3 m both. We used auxiliary data to derive information on water bodies (Landsat-7/ETM+, European Corine Land Cover CLC 2000) and topography (SRTM DEM v.3).

Neural network is built for each SAR instrument separately. In order to train and test neural networks, we manually selected the ground-truth pixels with the use of auxiliary data sets that correspond to both territories with the

presence of water (we denote them as belonging to a class “Water”) and without water (class “No water”). For ENVISAT/ASAR instrument, data from Chinese flood event were used to construct and calibrate the neural network. This neural network, then, was used to produce flood maps for other flood events. Collected ground-truth data were randomly divided into the training set (which constituted 75% of total amount) and the testing set (25%). Data from the training set were used to train the neural networks, and data from the testing set were used to verify the generalization ability of the neural networks, i.e. the ability to operate on independent, previously unseen data sets [Haykin, 1999].

Methodology description. Our flood mapping workflow with input and output data is shown in Fig. 3 [Kussul et al., 2008a].

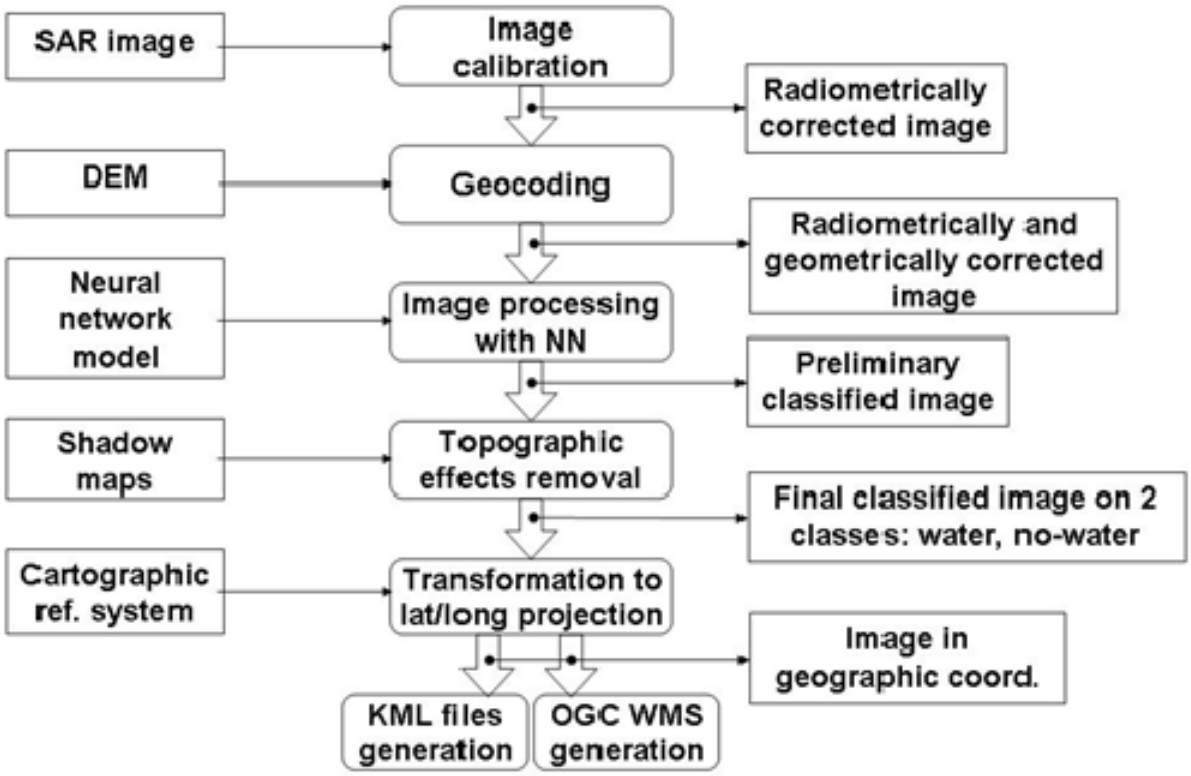


Figure 3. Flood mapping from SAR satellite imagery: workflow

SOM is a type of artificial neural network that is trained using unsupervised learning to produce a low-dimensional (typically two dimensional), discretised representation of the input space of the training samples, called a map [Kohonen, 1995; Haykin, 1999]. The map seeks to preserve the topological properties of the input space. SOM is formed of the neurons located on a regular, usually 1- or 2-dimensional grid. Neurons compete with each other in order to pass to the excited state. The output of the map is a, so called, neuron-winner or best-matching unit (BMU) whose weight vector has the greatest similarity with the input sample x.

The network is trained in the following way: weight vectors $\mathbf{w}_j$ from the topological neighbourhood of BMU vector i are updated according to [Kohonen, 1995; Haykin, 1999]

$$i(\mathbf{x}) = \arg\min_{j=\overline{1,L}} \left\| \mathbf{x} - \mathbf{w}_j \right\|,$$

$$\mathbf{w}_j(n+1) = \mathbf{w}_j(n) + \eta(n) h_{j,i(\mathbf{x})}(n)(\mathbf{x} - \mathbf{w}_j(n)), j = \overline{1,L} \quad (1)$$

where η is learning rate (see Eq. 3), $h_{j,i(\mathbf{x})}(n)$ is a neighbourhood kernel around the winner unit *i*, x is an input vector, $\|\bullet\|$ means Euclidean metric, *L* is a number of neurons in the output grid, *n* denotes a number of iteration in the learning phase.

The neighbourhood kernel function $h_{j,i(\mathbf{x})}(n)$ is taken to be the Gaussian

$$h_{j,i(\mathbf{x})}(n) = \exp\left(-\frac{\left\| r_j - r_{i(\mathbf{x})} \right\|}{2\sigma^2(n)} \right) \quad (2)$$

where $r_j, r_{i(\mathbf{x})}$ are the vectorial locations in the display grid of the SOM, $\sigma(n)$ corresponds to the width of the neighborhood function, which is decreasing monotonically with the regression steps.

For learning rate we used the following expression:

$$\eta(n) = \eta_0 \cdot e^{-\frac{n}{\tau}}, \eta_0 = 0.1 \quad (3)$$

where τ is a constant. The initial value of 0.1 for learning rate was found experimentally.

Kohonen's maps are widely applied to the image processing, in particular image segmentation and classification [Kohonen, 1995; Haykin, 1999]. Prior neural network training, we need to select image features that will be give to the input of neural network. For this purpose, one can choose original pixel values, various filters, Fourier transformation etc. In our approach we used a moving window with backscatter coefficient values for ERS-2 and ENVISAT images and digital numbers (DNs) for RADARSAT-1/2 image as inputs to neural network. The output of neural network, i.e. neuron-winner, corresponds to the central pixel of moving window. In order to choose appropriate size of the moving window for each satellite sensor, we ran experiments for the following windows size: 3-by-3, 5-by-5, 7-by-7, 9-by-9 and 11-by-11.

We, first, used SOM to segment each SAR image where each pixel of the output image was assigned a number of the neuron in the map. Then, we used pixels from the training set to assign each neuron one of two classes ("Water" or "No water") using the following rule. For each neuron, we calculated a number of pixels from the training set that activated this neuron. If maximum number of these pixels belonged to class "Water", then this neuron was assigned "Water" class. If maximum number of these pixels belonged to class "No water", then this neuron was assigned "No water" class. If neuron was activated by neither of the training pixels, then it was assigned "No data" class.

Results of image processing. In order to choose the best neural network architecture, we ran experiments for each image varying the following parameters: (i) size of the moving window for images that define the number of neurons in the input layer of the neural network; (ii) number of neurons in the output layer, i.e. the sizes of 2-dimensional output grid. Other parameters that were used during the image processing are as follows:

- neighbourhood topology is hexagonal;
- neighbourhood kernel around the winner unit is the Gaussian function (see Eq. 2);
- initial learning rate is set to 0.1;
- number of the training epochs is equal to 20.

The initial values for the weight vectors are selected as a regular array of vectorial values that lie on the subspace spanned by the eigenvectors corresponding to the two largest principal components of the input data [Kohonen, 1995].

We applied our approach to determine flood areas from SAR images acquired by the following instruments: ERS-2/SAR, ENVISAT/ASAR and RADARSAT-1. Classification rates for these sensors using independent testing data sets were 85.40%, 98.52% and 95.99%, respectively.

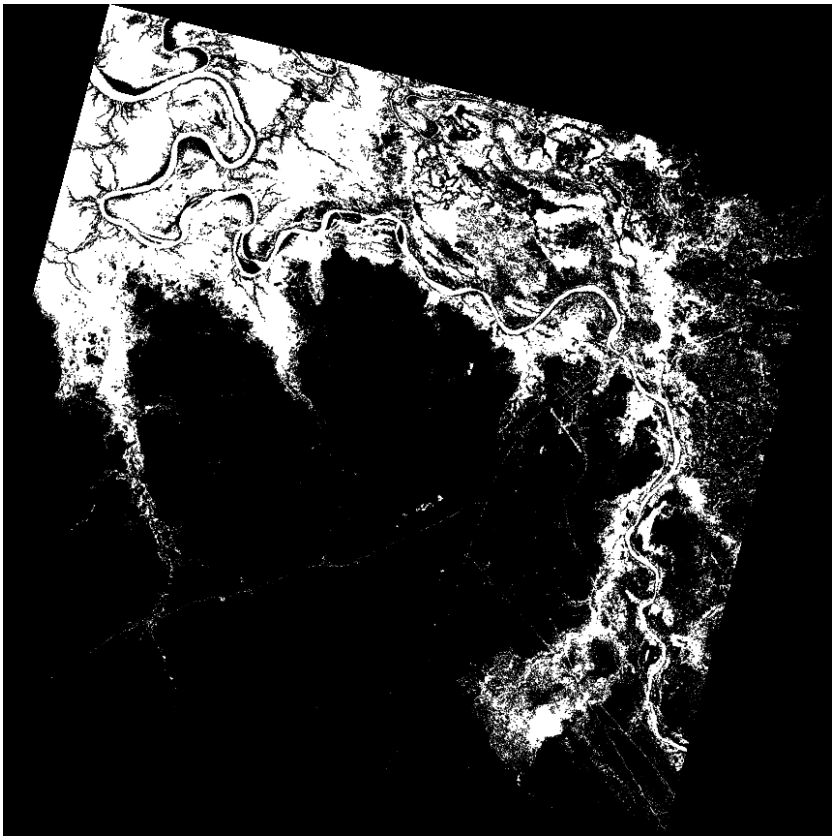


Figure 4. The resulting flood extent shown with white colour for the river Norman, Australia (© Space Research Institute NASU-NSAU 2009; RADARSAT-2 Data and Products © MacDONALD, DETTWILER AND ASSOCIATES LTD. 2009 – All Rights Reserved. RADARSAT is an official mark of the Canadian Space Agency)

For the images with higher spatial resolution (i.e. ERS-2 and RADARSAT-1), the best results were achieved for larger moving window 7-by-7. In turn, for the ENVISAT/ASAR WSM image, we used the moving window of smaller size 3-by-3. The use of higher dimension of input window for the ENVISAT image led to the coarser resolution of the resulting flood extent image and reduced classification rate.

The example of resulting flood extent map derived from RADARSAT-2 data acquired for the river Norman, Australia (see Fig. 2) is shown in Fig. 4.

Implementation. We developed a parallel version of our method and deployed it at the Grid infrastructure. Parallelization of the image processing is performed in the following way: SAR image is split into the uniform parts that are processed on different nodes using the OpenMP Application Program Interface (www.openmp.org). The use of the Grids allowed us to considerably reduce the time required for image processing. In particular, it took approximately 30 min to process a single SAR image on a single workstation. The use of Grid computing resources allowed us to reduce the time to less than 1 min.

Vegetation State Estimation. Estimation of vegetation state from satellite data has proved to be very helpful for agriculture monitoring, climate modelling, natural disasters management [Liang, 2008]. Parameters that can be estimated using optical data include Leaf Area Index (LAI), Fraction of Photosynthetic Active Radiation (FPAR), leaf pigment concentration, water concentration. Here, we will focus on plant moisture estimation from satellite data. This is very important for drought monitoring that becomes one of the major disasters in agricultural countries like Ukraine. For example, drought in Ukraine in 2007 resulted in $100 millions losses.

Water shortage in plants and plant stress in general can be detected by optical satellite data. Vegetation moisture determination is possible mainly due to significant differences in reflectance in Shortwave Infrared band of electromagnetic spectrum (SWIR) of vegetation under water stress and under normal conditions. However, in solar optical domain vegetation reflectance is controlled not only by moisture but by several other factors: leaf structure, pigment concentration, LAI, soil reflectance [Liang, 2004]. Due to this plant moisture estimation is far from trivial.

This estimation task is a massive parallel problem since estimation has to be performed on the per pixel basis. And, even if the problem is not computationally complex for a single pixel, it has to be solved for each pixel of the satellite imagery. For current moderate resolution sensors such as MODIS 1 million pixels has to be processed per day, and new satellite systems such as RapidEye will deliver billions pixels per day. Nevertheless, this problem is highly parallelizable and, thus, is a good candidate to be executed in a Grid environment.

Earlier approaches to vegetation moisture estimation were based on so-called Vegetation Indexes [Ceccato et al., 2002; Gao, 1996]. Index is a simple combination of reflectance in different bands of satellite image which has increased sensitivity to target variable like moisture content and low sensitivity to other factors. For example, one of the popular indexes is a Normalized Difference Water Index (NDWI):

$$NDWI = \frac{\rho_{0,8} - \rho_{1,6}}{\rho_{0,8} + \rho_{1,6}} \quad (4)$$

where $\rho_{0,8}$ and $\rho_{1,6}$ are reflectance value in Near Infrared band (NIR) and SWIR band.

Vegetation Indexes uses only a limited number of spectral bands (2-3) while modern sensors like MODIS, MERIS have 7-15 bands. Also, indexes remain only indirect measures of target variables, and additional regressions have to be used to estimate it. Usually, such regressions require additional calibration using local data which further complicates utilization of Vegetation Indexes. That is why, at present, the modern way to estimate vegetation parameters is based on more sophisticated approach – physical modelling of satellite signal using canopy radiative transfer models [Liang, 2004].

Problem statement. Under modelling approach the estimation problem is considered as inverse to the problem of simulation of satellite signal. For the latter task the wide range of models exists [Liang, 2004], among which several models (like PROSPECT [Feret et al., 2008] and SAIL [Verhoef et al., 2007]) are widely used in remote sensing. For our purpose we will formulate radiative transfer model as a mapping $h:\mathbf{R}^{n_x} \to \mathbf{R}^{n_d}$ that maps state of vegetation $x \in X \subset \mathbf{R}^{n_x}$ into reflectance in different bands $h(x) \in D \subset \mathbf{R}^{n_d}$:

$$d = h(x) + h(x)\varepsilon \tag{5}$$

where d is measurement vector and ε is noise vector. This problem is characterized by multiplicative noise [Bacour et al., 2006].

For instance, for PROSPECT leaf radiative transfer model the dimension of x is four $x = (N, C_{ab}, C_w, C_m)^T$, where N — leaf structure parameter, while C_{ab}, C_w, C_m — concentration of chlorophyll, water and dry matter. Dimension of model output vector $h(x)$ is 2100, however for remote sensing purposes model output has to be aggregated to be comparable with current multispectral sensors. So usually the dimension of observation vector d is much smaller, for instance for MODIS sensor it will be 7.

In this paper the Bayesian approach to inverse problems is considered [Tarantola, 2005]. Within this approach uncertainty in a priory estimate of state vector x and in process of measurement of reflectance vector $h(x)$ has probabilistic nature. Let $\mathbf{x}$, $\mathbf{d}$, $\boldsymbol{\varepsilon}$ — random vectors of a priory estimate of model input, observations and noise in observations, $p(x)$, $p(d)$ and $p(\varepsilon)$ — densities of probability distributions of these vectors. It is assumed that random vectors $\mathbf{x}$ and $\boldsymbol{\varepsilon}$ are independent, while densities $p(x)$, $p(\varepsilon)$ and function h is such, that random vectors $\mathbf{x}$ and $\mathbf{d}$ have common density $p(x,d)$ and components of these vectors have variance.

The solution of inverse problem is conditional density of model input $\mathbf{x}$ with respect of known value of observations vector $\mathbf{d}$ [Tarantola, 2005]:

$$p(x \mid d) \propto p(d \mid x)p(x), \quad x \in \mathbf{R}^{n_x}, d \in \mathbf{R}^{n_d} \tag{6}$$

However, for practical purposes we have to estimate some properties of above conditional density, like mean, standard deviation, median, most probable value etc.

Neural network method to solve inverse problem. There are several methods to estimate properties of (6): Monte-Carlo [Qingyuan et al., 2005], variational [Bacour et al., 2002], lookup tables [Combal et al., 2002] and neural

networks [Bacour et al., 2006]. However, in recent years neural networks gain a lot of attention due to their ability to approximate arbitrary continuous function and computational efficiency [Haykin, 1999].

To solve inverse problem (6) within traditional neural network approach the approximation $f: D \to X$ of inverse mapping to $h: X \to D$ is constructed using neural network, for instance Multilayer Perceptron (MLP). This is performed through minimization of quadratic functional:

$$J(w) = \frac{1}{2}\sum_i \left\| x_i - f(d_i, w) \right\|^2 \tag{7}$$

where function $f(\cdot, w)$ is defined by neural network with weight coefficients w, $\left\{(d_i, x_i), i = \overline{1,n}\right\}$ is learning sample set created via sampling from density $p(x,d)$.

It can be shown (see for instance [Bishop, 1996; Kravchenko, 2009]) that given sufficient number of learning samples neural network with quadratic error criteria will approximate conditional mean $\mathrm{E}[\mathbf{x} \mid \mathbf{d} = d] = \int x p(x \mid d)\, dd$ of network output x given input d. So in the framework traditional neural network approach we can obtain only point estimate of parameters. To overcome this deficiency of traditional neural networks for inverse problem solving we propose to apply neural networks with nonquadratic error criteria, such as Mixture Density Networks (MDN) [Bishop, 1996]. Such networks allow modelling of conditional density $p(x \mid d)$ as a mixture of Gaussian densities.

$$p_{MDN}(x \mid d, w) = \sum_{l=1}^{L} \alpha_l(d,w) \cdot \phi\left(x; m_l(d,w), \sigma_l(d,w)\right) \tag{8}$$

where $\phi(x; m, \sigma) = \dfrac{1}{\left(\sqrt{2\pi}\sigma\right)^{n_x}} \exp\left(-\dfrac{\|x-m\|^2}{2\sigma^2}\right)$ — Gaussian density with mean m and diagonal covariance matrix $\sigma^2 \mathrm{I}$, α_l — mixture coefficients ($\sum_l \alpha_l = 1$), L — number of elements of mixture. Functions $\alpha_l(d,w)$, $m_l(d,w)$ and $\sigma_l(d,w)$ are constructed using MLP with modified output layer. MDN is learned through minimising the following error criteria:

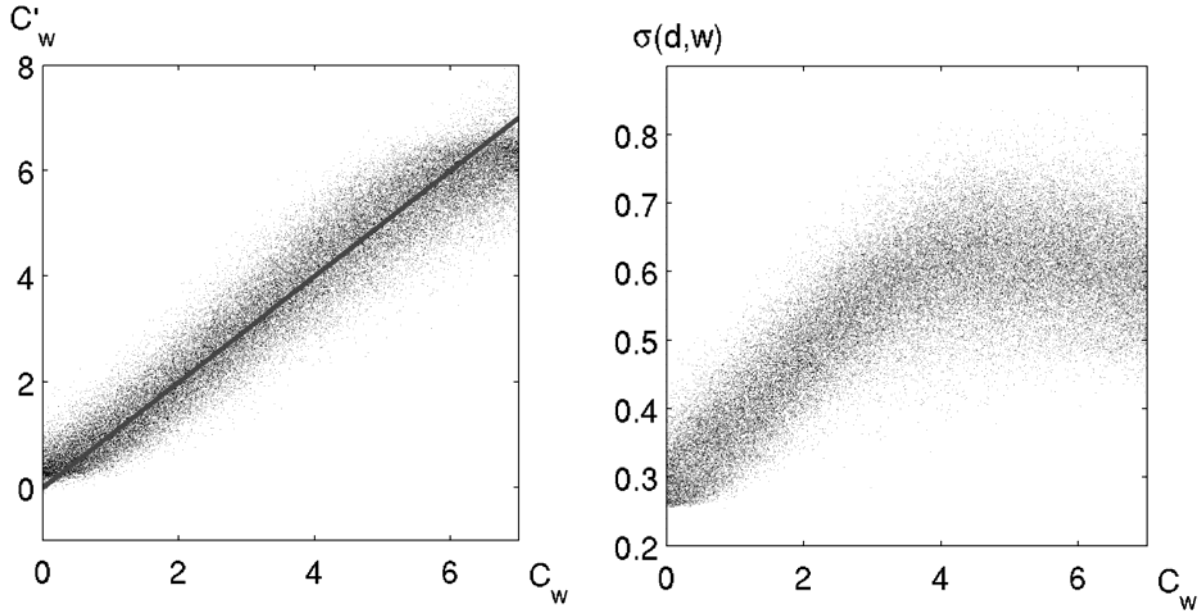
$$J(w) = \frac{1}{n}\sum_{i=1}^{n} -\ln p_{MDN}(x_i \mid d_i, w) \tag{9}$$

Unlike MLP, MDN with even one Gaussian component in mixture can approximate both conditional mean and variance of $p(x \mid d)$ [Kravchenko, 2009].

Numerical experiment with PROSPECT model.

Here we will demonstrate use of MDN to solve inverse problem of leaf moisture estimation. To formulate forward problem we will use PROSPECT leaf radiative transfer model. In this case x vector consists of 4 parameters: $x = (N, C_{ab}, C_w, C_m)^T$, while observation vector d consists of seven leaf reflectances in MODIS-like spectral

bands. To pose inverse problem we will assume uniform a priory density $p(x)$ and independent Gaussian noise model for $\boldsymbol{\varepsilon}$ (5% standard deviation). To estimate plant moisture we will use MDN with 7 neurons in input layer, 5 neurons in hidden layer and one-dimensional mixture containing one Gaussian component. This network is used to estimate mean and variance of conditional density $p(C_w \mid d)$. Increasing number of mixture's components or number of neurons in hidden layer does not improve the quality of solution in this problem.



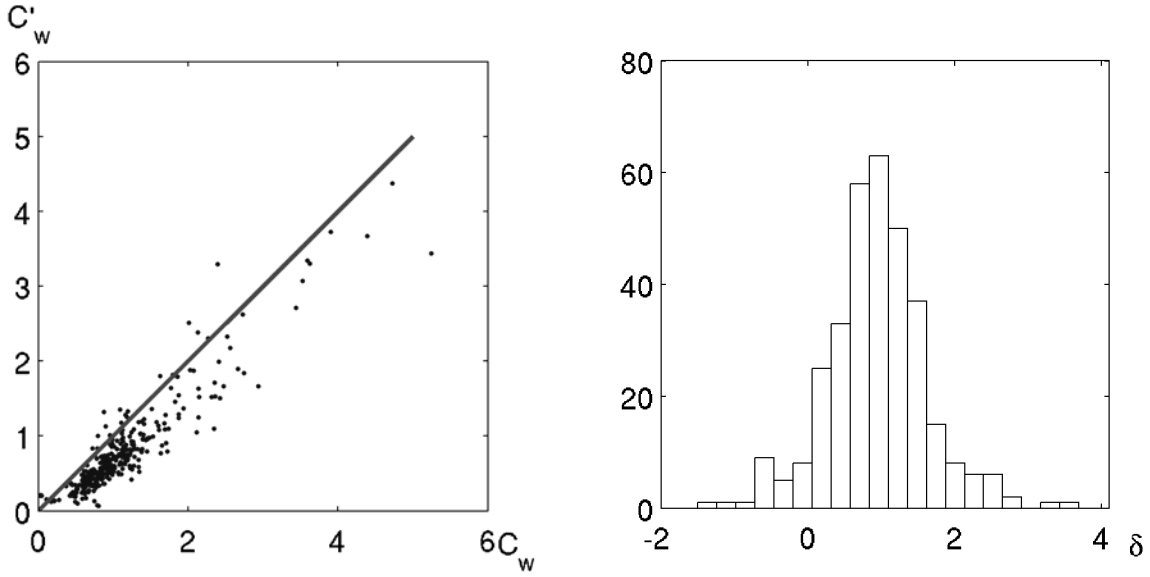
(a) (b)

Figure 5. a — scatter plot of estimated leaf moisture $C_w^{'}$ and true C_w ; b — dependency of estimate of standard deviation of leaf moisture $\sigma(d,w)$ w.r.t. real leaf moisture C_w

Scatter plot of conditional mean of leaf moisture $C_w^{'} = m_1(d_i,w)$ estimated by MDN given observation d_i and true value C_w is shown on fig. 5a (identical dependency is shown by strait line), while dependency of estimate of standard deviation of leaf moisture $\sigma(d_i,w)$ given observation d_i with respect to true value C_w is shown in Fig. 5b. Standard deviation is increased with increase of moisture C_w and stabilized for large C_w (4-7 cg/cm^2). This is in accordance with the fact that sensitivity of SWIR reflectance is decreased for large leaf moisture values.

Validation results. To validate our algorithm we used LOPEX leaf optical properties database (Leaf Optical Properties EXperiment). This database contains over 1250 plant reflectance spectra. For validation purpose 330 fresh leaf spectra of 66 plant species at different moisture level were used. Spectra were aggregated using MODIS band relative spectral response functions. Fig. 6a shows the scatter plot of estimated leaf moisture ($C_w^{'}$) and observed (C_w), while fig. 6b shows the histogram of moisture estimation error normalized by estimate of standard deviation

$\delta = \left(C_w^{'} - C_w\right) \Big/ \sigma(d_i, w)$. Most of the departures (90%) are located in $[-2;2]$ interval (in $\pm 2\sigma$ interval) that confirms adequacy of standard deviation estimates using MDN.



(a) (b)

Figure 6. a — scatter plot of estimated leaf moisture $C_w^{'}$ and true C_w; b — histogram of normalized errors δ

Levels of Integration: Main Problems and Possible Solutions

Modern tendencies of globalization and development of the "system of systems" GEOSS lead to the need of integration of heterogeneous satellite-based monitoring systems. Integration can be done at different levels: (i) data exchange level, (ii) task management level. Data exchange is supposed to provide infrastructure for sharing data and products. This infrastructure enables data integration where different entities provide various kinds of data to support joint solution of complex problems (Fig. 7). Task management level envisages running applications at distributed computational resources provided by different entities (Fig. 8). Since many of the existing satellite monitoring system rely on Grid technologies appropriate approaches and technologies should be evaluated and developed to enable Grid system integration (so called InterGrid).

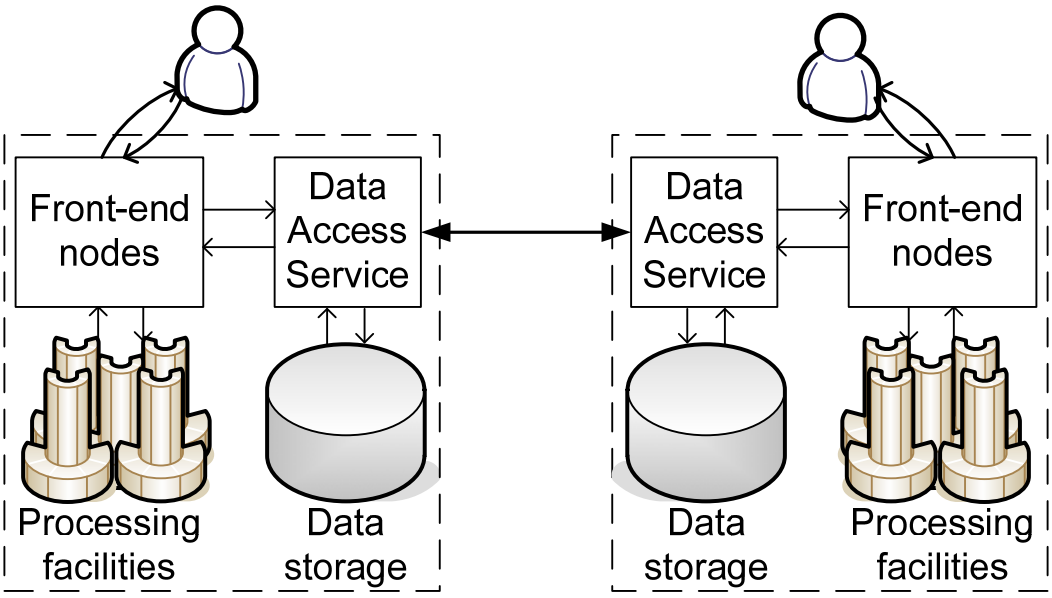


Figure 7. Data integration level

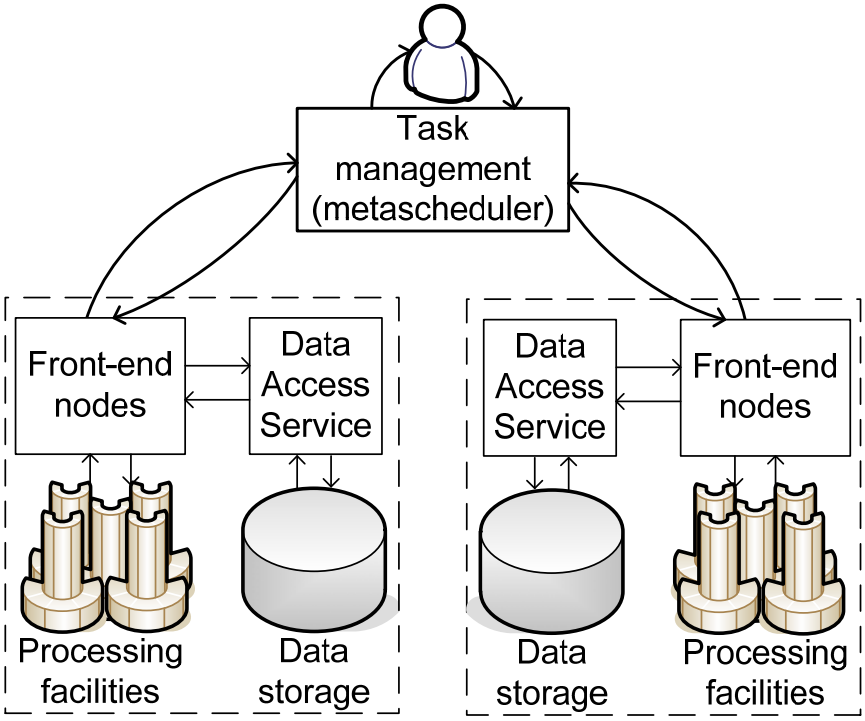


Figure 8. Task management level

This section highlights main challenges and possible solutions for satellite monitoring systems integration at both levels, and provides the case-studies for both cases.

Integration at data exchange level could be done by using common standards for EO data exchange, common user interfaces, and common data and metadata catalogues. Considering the task management level, the following problems additionally should be tackled: the use of joint computational infrastructure; development of jobs submission and scheduling algorithms; load monitoring enabling; security policy enforcement.

Data exchange level. At present the most appropriate standards for data integration is Open Geospatial Community (OGC) standards. Data visualization issues can be solved by using the following set of standards: WMS (Web Map Service), SLD (Style Layer Descriptors) and WMC (Web Map Context). OGC's WFS (Web Feature Service) and WCS (Web Coverage Service) standards provide uniform ways for data delivery. In order to provide interoperability at the level of catalogues CSW (Catalogue for Web) standard can be applied.

Since data are stored at geographically distributed sites there can be issues regarding optimization of visualization schemes. In general, there are two possible ways for distributed data visualization: centralized visualization scheme and distributed visualization scheme. Advantages and faults of each scheme were described in [Shelestov et al., 2008].

Task management level. In this subsection we present main issues and possible solutions for Grid-system integration. Main prerequisite of such kind of integration is certificates trust. It could be done, for example, through EGEE infrastructure that nowadays brings together the resources of more than 70 countries. Another problems concerned with different Grid systems integration are as follows: enabling data transfers and high-level access to geospatial data; development of common catalogues; enabling jobs submission and monitoring; enabling information exchange.

Data transfer. GridFTP is an appropriate and reliable solution for data transfer. The only limitation is the requirement of transparent LAN (local area network) infrastructure.

Access to geospatial data. High-level access to geospatial data can be organised in two possible ways: using pure WSRF services or using OGSA-DAI container. Each of this approach has its own advantages and weaknesses. Basic functionality for WSRF-based services can be easily implemented (with proper tools), packed and deployed. But advanced functionality such as security delegation, third-party transfers, indexing should be implemented by hands. WSRF-based services can also pose some difficulties if we need to integrate them with other data-oriented software.

OGSA-DAI framework provides uniform interfaces to heterogeneous data. This framework makes possible to create high-level interfaces to data abstracting hiding details of data formats and representation schemas. Most of problems in OGSA-DAI are handled automatically, e.g. delegation, reliable transfer, data flow between different sources and sinks. OGSA-DAI containers are easily extendable and embeddable. But comparing to WSRF basic functionality implementation of OGSA-DAI extensions is more difficult. Moreover, OGSA-DAI require preliminary deployment of additional software components.

Task management. There are two possible approaches for task management. One of them is to use Grid portal (Fig. 9) supporting different middleware platforms, such as GT4, gLite, etc. Grid portal is an integrated platform to

end-users that enables access to Grid services and resources via standard Web browser. Grid portal solution is easy to deploy and maintain, but it doesn't provide application interface and scheduling capabilities.

Another approach is to develop high-level Grid scheduler (Fig. 10) that will support different middleware by providing some standard interfaces. Such metascheduler interacts with low-level schedulers (used in different Grid systems) enabling in such way system interoperability. Metascheduler approach is much more difficult to maintain comparing to portals; however, it provides API with advanced scheduling and load-balancing capabilities. At present, the most comprehensive implementation for the metascheduler is a GridWay system. The GridWay metascheduler is compatibility with both Globus and gLite middlewares. Starting from Globus Toolkit v4.0.5 GridWay become standard part of its distribution. GridWay system provides comprehensive documentation for both users and developers that is an important point for implementing new features.

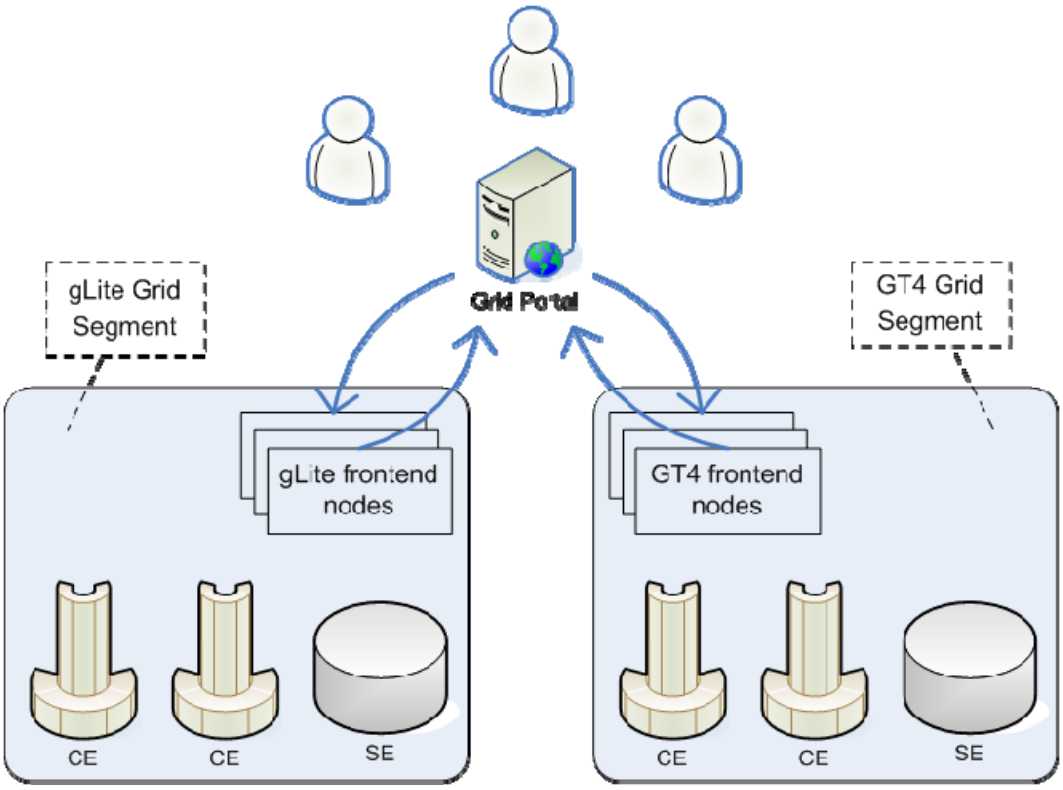


Figure 9. Portal approach to Grid system integration

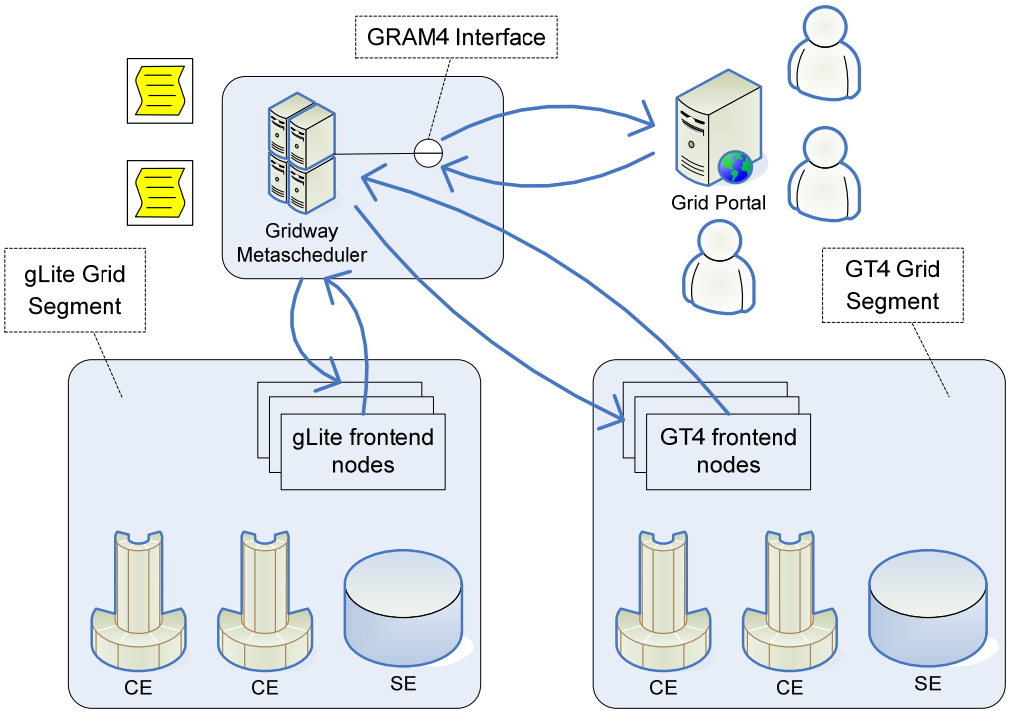


Figure 10. Metascheduler approach

In the next section we show the examples of application of described approaches to integration of satellite monitoring systems and development of InterGrid environment.

Implementation: Lessons Learned

Integration of satellite monitoring systems. The first case-study refers to the integration of satellite monitoring systems of NSAU (Ukraine) and IKI RAN (Russia). The overall architecture for integration of data provided by two organizations is depicted in Fig. 11. The proposed approach is applied for the solution of problems for agriculture resources monitoring and crop yield prediction. Within integration NSAU provides WMS interfaces to NWP modelling data (using WRF model) [Kussul et al., 2008b], in-situ observations from meteorological ground stations in Ukraine, and land parameters (such as temperature, vegetation indices, soil moisture) derived from satellite observations from MODIS instrument onboard Terra satellite. IKI RAN provides WMS interfaces to operational land and disaster monitoring system. Both NSAU and IKI RAN provides user Web-interfaces to monitoring systems that support OGC WMS standards.

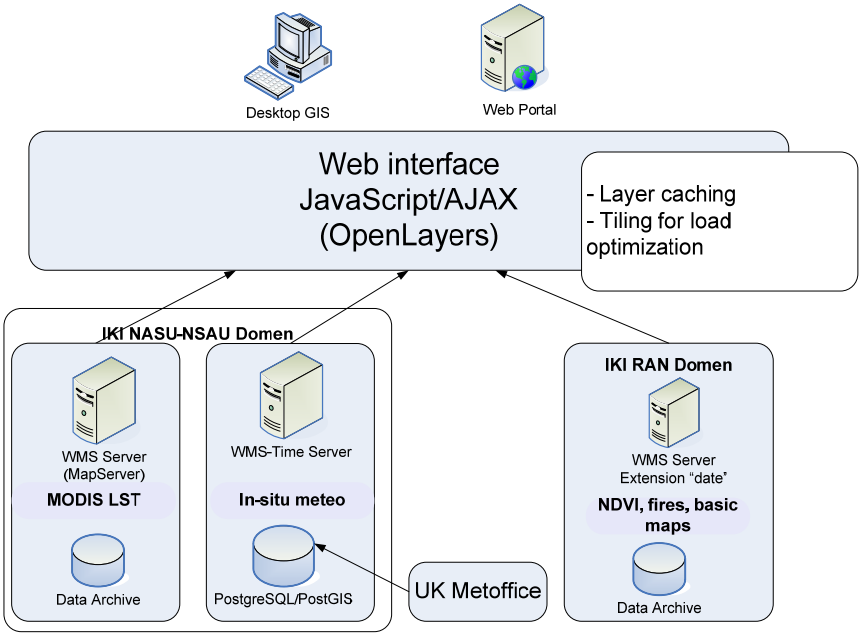


Figure 11. Architecture of satellite monitoring system integration

In order to provide user interface that will enable visualization of data from multiple sources we use open-source OpenLayers framework (http://www.openlayers.org). OpenLayers is "thick client" software based on JavaScript/AJAX and fully operational on client side. Main OpenLayers features also include: support for several WMS servers, support for different OGC standards (WMS, WFS), cache and tiling support to optimize visualization, support for of both raster and vector data. The provided data and products are accessible via Internet http://land.ikd.kiev.ua. The example of OpenLayers visualization of data from multiple sources is depicted in Fig. 12.

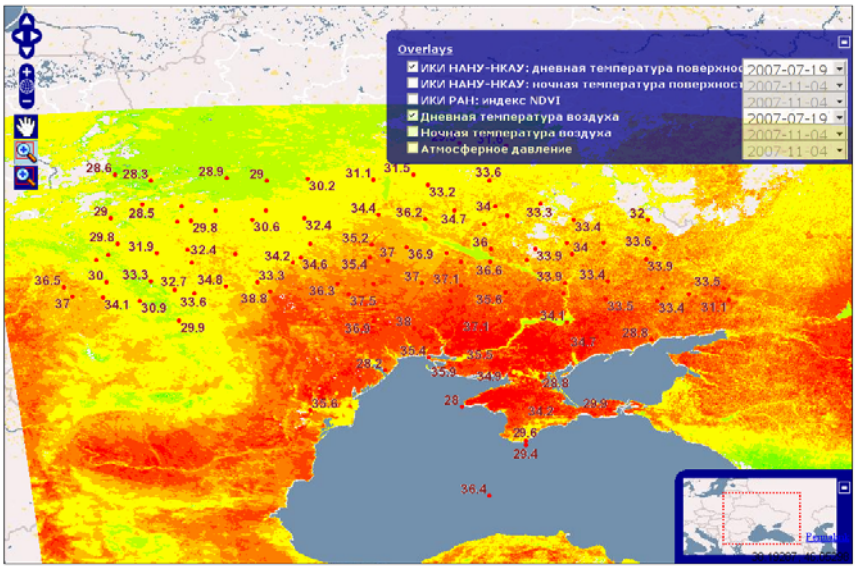


Figure 12. OpenLayers interface to multiple data

InterGrid testbed development. The second case-study refers to the development of InterGrid for environmental and natural disaster monitoring. InterGrid integrates Ukrainian Academician Grid (with Satellite data processing Grid segment) and CEODE Grid (Chinese Academy of Sciences) and is considered as a testbed for Wide Area Grid (WAG) implementation—a project initiated within CEOS Working Group on Information Systems and Services (WGISS).

The important application that is being solved within InterGrid environment is flood monitoring and prediction. This task requires adaptation and tuning of existing hydrological and hydraulic models for corresponding territories and the use of heterogeneous data stored at multiple sites. Flood monitoring and prediction requires the use of the following data sets: NWP modelling data (provided by Satellite data processing Grid segment), SAR imagery from Envisat/ASAR and ERS-2/SAR satellites (provided by ESA), products derived from optical and microwave satellite data such as soil moisture, precipitation, flood extent etc., in-situ observations from meteorological ground stations and digital elevation model (DEM). The process of model adaptation can be viewed as a complex workflow and requires the solution of optimization problems (so called parametric study). Satellite data processing and products generation tasks also represent complex workflow and require intensive computations. All these factors lead to the need of using computational and informational resources of different organizations and their resources into joint InterGrid infrastructure. The architecture of proposed InterGrid is depicted in Fig. 13.

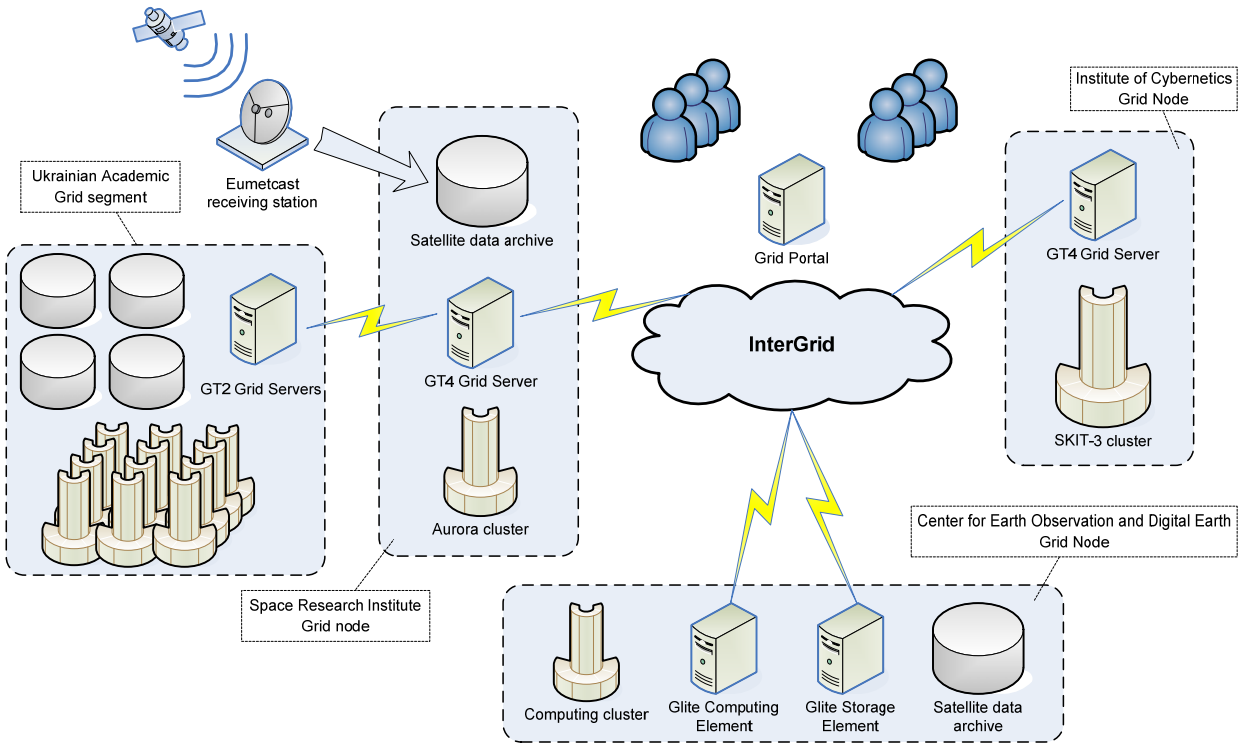


Figure 13. InterGrid architecture

GridFTP was chosen to provide data transfer between Grid systems. In order to enable interoperability between different middleware (for example, Satellite data processing Grid segment is using GT4; CEODE Grid is using gLite 3.x; Ukrainian Academician Grid is based on NorduGrid) we developed Grid portal that is based on GridSphere portal framework (http:// www.gridsphere.org). The developed Grid portal allows users to transfer data between different nodes and submit jobs on computational resources of the InterGrid environment. The portal also provides facilities to monitor statistics of the resources such as CPU load, memory usage, etc. The further works on providing interoperability between different middleware are directed to the development of metascheduler using GridWay system. In the nearest future we are intended to provide integration with ESA's EO Grid-on-Demand infrastructure.

Conclusions

This paper presented different approaches to multi-source data integration for the solution of complex applied problems in the Earth Science domain. In particular, we considered two problems, flood mapping and vegetation state estimation that requires the use of heterogeneous data acquired from multiple sources: remote-sensing from space, modelling and in-situ observations. We used satellite SAR imagery and DEM to extract flood extent from satellite data. To segment and classify the imagery we used self-organising Kohonen maps that provide such useful features as effective software tool for the visualization of high-dimensional data, automatically discover of statistically salient features of pattern vectors in data set, and possibility to find clusters in training data pattern space which can be used to classify new patterns. We tested our approach for various SAR instruments and for a number of flood events covering various geographical regions. The achieved classification rate was from 85.40% to 98.52% depending on the SAR instrument used.

Another application was vegetation state estimation from satellite and modelling data. We use physical modelling of satellite signal using canopy radiative transfer models. Under this approach the estimation problem is considered as inverse to the problem of simulation of satellite signal. To solve inverse problems we apply neural networks, namely Mixture Density Networks (MDNs) that allow the modelling of conditional density as a mixture of Gaussian densities. Another useful property of MDNs is that they can approximate both conditional mean and variance in the output density. We run different numerical experiments using PROSPECT model and LOPEX leaf optical properties database. Most of the departures (90%) were located in $\pm 2\sigma$ interval that confirms adequacy of standard deviation estimates using MDNs.

Since both these applications are data- and computation-intensive, we use Grid computing technologies. In such a case computational and informational resources are geographically distributed and may belong to different organisations. For this purpose, we also investigated benefits of different approaches to the integration of satellite-based monitoring systems. We investigated two possible levels of integration, namely data level and task management level. As to data integration level, we found that integration could be provided by using existing standards for geospatial data, in particular OGC standards. We demonstrated applicability and usability of this approach to the integration of existing satellite monitoring systems of Ukraine and Russia for agriculture applications. The use of standard OGC interfaces makes it possible to standardise and facilitate the development of integrated satellite monitoring systems (based on existing systems) to exploit the synergy and acquire information of new quality. As to integration at task management level, we reviewed two solutions: portal-based and metascheduling approach. We implemented portal solution based on the GridSphere framework to the InterGrid environment that integrates several regional and national Grid systems. In order to provide advanced scheduling and load-balancing capabilities the further works will be directed to the implementation of metascheduler based on GridWay system.

Further investigations will be directed to the integration of distributed monitoring systems with Sensor Web to provide automatic delivery of data from heterogeneous sources and their processing in the Grid environment.

Acknowledgment

This work is supported by the joint project of the Science & Technology Center in Ukraine (STCU) and the National Academy of Sciences of Ukraine (NASU), "Grid Technologies for Multi-Source Data Integration" (No. 4928), and the Ministry of Education and Science of Ukraine, "Development of Integrated Remote Sensing Data Processing System using Grid Technologies" (No. M/72-2008).

Bibliography

[Bacour et al., 2002] C. Bacour, S. Jacquemoud, Y. Tourbier, et al. Design and analysis of numerical experiments to compare four canopy reflectance models. Ibid, V. 79, 72-83, 2002.

[Bacour et al., 2006] C. Bacour, F. Baret, D. Béal, M. Weiss, K. Pavageau. Neural network estimation of LAI, fAPAR, fCover and LAI×Cab, from top of canopy MERIS reflectance data: principles and validation. Ibid, 105, 313–325, 2006.

[Bishop, 1996] Bishop C. Neural Networks for Pattern Recognition. Oxford University Press, 1996.

[Ceccato et al., 2002] P. Ceccato, N. Gobron, S. Flasse, et al. Designing a spectral index to estimate vegetation water content from remote sensing data: part 1 theoretical approach. Ibid, 82, 188-197, 2002.

[Combal et al., 2002] B. Combal, F. Baret, M. Weiss, et al. Retrieval of canopy biophysical variables from bidirectional reflectance using prior information to solve the ill-posed inverse problem. Ibid, 84, 1-15, 2002.

[Cunjian et al., 2001] Y. Cunjian, W. Yiming, W. Siyuan, Z. Zengxiang, H. Shifeng. Extracting the flood extent from satellite SAR image with the support of topographic data. In: Proc of Int Conf on Inf Tech and Inf Networks (ICII 2001), vol. 1, 87-92, 2001.

[Feret et al., 2008] J.B. Feret, C. François, G.P. Asner, et al. PROSPECT-4 and 5: advances in the leaf optical properties model separating photosynthetic pigments. Ibid, 112, 3030-3043, 2008.

[Foster and Kesselman, 2004] I. Foster, C. Kesselman. The Grid: Blueprint for a New Computing Infrastructure. 2nd Edition, Morgan Kaufmann, 2004.

[Fukui, 2007] H. Fukui Sentinel Asia/Digital Asia: Building Information Sharing Platform by Geo web services and contributing to Disaster Management Support in the Asia-Pacific Region, 2007.

[Fusco et al., 2003] L. Fusco, P. Goncalves, J. Linford, M. Fulcoli, A. Terracina, G. Terracina. Putting Earth-Observation on the Grid. ESA Bulletin, 114, 86-91, 2003.

[Gao, 1996] B.C. Gao. NDWI – a normalized difference water index for remote sensing of vegetation liquid water from space. Remote Sensing of Environment, 58, 257–266, 1996.

[Haykin, 1999] S. Haykin. Neural Networks: A Comprehensive Foundation. Upper Saddle River, New Jersey: Prentice Hall, 1999.

[Horritt, 1999] M.S. Horritt. A statistical active contour model for SAR image segmentation. Image and Vision Computing, 17, 213-224, 1999.

[Horritt, 2006] M.S. Horritt. A methodology for the validation of uncertain flood inundation models. J of Hydrology, 326, 153-165, 2006.

[Kogan et al., 2004] F. Kogan, R. Stark, A. Gitelson, E. Adar, L. Jargalsaikhan, C. Dugrajav, S. Tsooj. Derivation of Pasture Biomass in Mongolia from AVHRR-based Vegetation Health Indices. Int. J. Remote Sens, 25(14), 2889-2896, 2004.

[Kohonen, 1995] T. Kohonen. Self-Organizing Maps. Series in Information Sciences, Vol. 30. Springer, Heidelberg, 1995.

[Kopp et al., 2007] P. Kopp, I. Petiteville, A. Shelestov, G. Li. Wide Area Grid (WAG). In: Proc. The 7th Ukrainian Conference on Space Research, National Flight and Control Center, Evpatoria, Ukraine, 209, 2007.

[Kravchenko, 2009] A. Kravchenko. Neural networks method to solve inverse problems for canopy radiative transfer models. Cybernetics and System Analysis, N 3, 159-172, 2009. (in Russian)

[Kussul et al., 2008a] N. Kussul, A. Shelestov, S. Skakun. Grid System for Flood Extent Extraction from Satellite Images. Earth Science Informatics, 1(3-4), 105-117, 2008.

[Kussul et al., 2008b] N. Kussul, A. Shelestov, S. Skakun, O. Kravchenko. Data Assimilation Technique For Flood Monitoring and Prediction. International Journal on Information Theory and Applications, 15(1), 76-84, 2008.

[Liang, 2004] S. Liang. Quantitative Remote Sensing of Land Surfaces. Wiley, 2004.

[Liang, 2008] S. Liang (ed.). Advances in Land Remote Sensing. Springer, 2008.

[Martinez and Le Toan, 2007] J.M. Martinez, T. Le Toan. Mapping of flood dynamics and spatial distribution of vegetation in the Amazon floodplain using multitemporal SAR data. Remote Sensing of Environment, 108, 209-223, 2007.

[Niedermeier et al., 2000] A. Niedermeier, E. Romaneeßen, S. Lenher Detection of coastline in SAR images using wavelet methods. IEEE Transactions Geoscience and Remote Sensing, 38(5), 2270-2281, 2000.

[Qingyuan et al., 2005] Z. Qingyuan, X. Xiangming, B. Braswell, et al. Estimating light absorption by chlorophyll, leaf and canopy in a deciduous broadleaf forest using MODIS data and a radiative transfer model. Remote Sensing of Environment, 99, 357-371, 2005.

[Rees, 2001] W.G. Rees. Physical Principles of Remote Sensing, Cambridge University Press, 2001.

[Scheuren et al., 2008] J.-M. Scheuren, O. le Polain de Waroux, R. Below, D. Guha-Sapir, S. Ponserre. Annaul Disaster Statistical Review – The Number and Trends 2007. Center for Research of the Epidemiology of Disasters (CRED). Jacoffsaet Printers, Melin, Belgium, 2008.

[Shelestov et al., 2006] A.Yu. Shelestov, N.N. Kussul, S.V. Skakun. Grid Technologies in Monitoring Systems Based on Satellite Data. J. of Automation and Information Science, 38(3), 69-80, 2006.

[Shelestov et al., 2008] A. Shelestov, O. Kravchenko, M. Ilin. Distributed visualization systems in remote sensing data processing GRID, International Journal "Information Technologies and Knowledge", 2(1), 76-82, 2008.

[Tarantola, 2005] Tarantola A. Inverse problem theory problem and methods for model parameter estimation. SIAM, 2005.

[Verhoef et al., 2007] W. Verhoef, Q. Xiao, L. Jia, et al. Unified optical-thermal four-stream radiative transfer theory for homogeneous vegetation canopies. IEEE Transactions on Geoscience and Remote Sensing, V. 45, 1808–1822, 2007.

[Wagner et al., 2007] W. Wagner, C. Pathe, D. Sabel, A. Bartsch, C. Kuenzer, K. Scipal. Experimental 1 km soil moisture products from ENVISAT ASAR for Southern Africa. ENVISAT & ERS Symposium, Montreux, Switzerland, 2007.

Authors' Information

Kussul Nataliia – Deputy Director, Space Research Institute NASU-NSAU, Glushkov Prospekt 40, build. 4/1, Kyiv 03680, Ukraine;
e-mail: inform@ikd.kiev.ua

Major Fields of Scientific Research: Grid computing, design of distributed software systems, parallel computations, intelligent data processing methods, Sensor Web, neural networks, satellite data processing, risk management and space weather.

Shelestov Andrii – Senior Scientist, Space Research Institute NASU-NSAU, Glushkov Prospekt 40, build. 4/1, Kyiv 03680, Ukraine;
e-mail: inform@ikd.kiev.ua

Major Fields of Scientific Research: Grid computing, intelligent methods of estimation and modeling, satellite data processing, Sensor Web, neural networks, multi-agent simulation and control, software engineering, distributed information systems design.

Skakun Sergii – Senior Scientist, Space Research Institute NASU-NSAU, Glushkov Prospekt 40, build. 4/1, Kyiv 03680, Ukraine;
e-mail: serhiy.skakun@ikd.kiev.ua

Major Fields of Scientific Research: Remote sensing data processing (optical and radar), image processing, Grid computing, Sensor Web, neural networks, analysis of computer system's users behavior.

Kravchenko Oleksii – PhD Student, Space Research Institute NASU-NSAU, Glushkov Prospekt 40, build. 4/1, Kyiv 03680, Ukraine;
e-mail: inform@ikd.kiev.ua

Major Fields of Scientific Research: inverse modelling, pattern recognition, remote-sensing data processing, parallel algorithms, GRID computing.

STRUCTURAL MODEL OF HALFTONE IMAGE AND IMAGE SEGMENTATION EXPERIMENTS

Vitaly Vishnevsky, Vladimir Kalmykov, Tatyana Vlasova

Abstract: The structural model of gray-scale picture is offered. The structural model supposes the selection of objects in the image, thus their description is invariant in relation to affine transformations. Object form is exhaustive its description of. Object form is defined by its contour and the optical density function, which is certain within its contour bounds. The determination of halftone picture contour is proposed, as a sequence, consisting of straight line segments and curve arcs, and the straight line segments and the curve arcs are the critical lines of surface which corresponds to the halftone picture. An example, how to use the structural model of halftone picture for the medical preparations image processing got on the method of Kyrlyan, is considered.

Key words: structural analysis of halftone picture, contour, row model, segmentation

ACM Classification Keywords: I.5.1 Models, I.3.5 Computational Geometry and Object Modeling.

Introduction

Visual information, particularly halftone pictures processing is one of the most difficult artificial intelligence problems and, at the same time, more and more urgent for the practical use in the most different branches of science and technologies.

The raster mode of halftone pictures is presently used in artificial intelligence means.

To process the raster half-tone pictures, the heavy computational resources are needed.

The raster half-tone pictures processing, for example, objects identification, which have different values of affine transformations - a scale, position, rotation, require the heavy computational resources or in general view are impossible.

In modern visual halftone information processing such concepts as *object contours* are not utillized practically (except for the object contours of halftone image preliminary turned into binary one).

The ability to segment a picture in the eyeshot, to stand out objects, which differ against a background by optical density, color, texture, other, is one of basic and the most natural feature of human visual perception. The form which is defined by a contour - border between an object and background is the basic attribute of any object. A contour, in same queue, is accepted by a human as a sequence of line segments and curve arcs. The form of half-tone and color objects is determined, in addition, by the function of optical density taking into account a color, texture within a contour each of objects. These human visual perception features are reproduced in the offered structural model for halftone image.

A structural model enables to present arbitrary images as regular description which consists of background and object descriptions.

Task of adduction to the structural model of arbitrary images, set in a raster mode, distorted hindrances in general case yet not decided. However in separate, numerous enough cases, bringing images over to the structural model allows substantially to promote speed and quality of visual information processing that, in same queue, provides the high-quality functioning of utilizing these facilities of information technologies. The objects of halftone picture, transformed in a structural mode, invariant in relation to affine transformations , by the best appearance befit as basic data for processing by the methods, based on growing pyramidal networks [Gladun,1] and theory of recognition and memory [Rabinovich,2]

Structural Model of Halftone Image

The structural model of half-tone image corresponds to the known representations about visual perception mechanisms. Objects, located on an image background, which have been determined by the two-dimensional function of optical density, are the basic structural display elements. Objects, in same queue, have been determined by the contours which bound objects and by the two-dimensional optical density function, within the bounds of object. Contours are the closed sequences which are formed by the segments of lines and by the curve arcs.

Under an image it is understand a part of plane, bounded with some geometrical figure, usually by a rectangle, which the value of optical density is defined for every point of. In other words, a rectangle with the sizes of X,Y on a plane is the function domain of $\rho = f(x,y)$, $(0 \le x \le X ; 0 \le y \le Y)$. It is possible to put some surface of $z = f(x,y)$ in correspondence to this function.

Let us present some information over from an area analytical geometry in space [Korn,Korn,3]. The set of points P(x,y,z), the co-ordinates of which satisfy the system of equations

$$x=x(u,v),\ y=y(u,v),\ z=z(u,v) \qquad (1)$$

at the suitable values of actual parameters of u,v, is named *a continuous surface*, if right parts of equations are the continuous functions of parameters. It is possible to define a surface as the equation

$$j(x,y,z)= 0 \text{ or } z = f(x,y).$$

A surface can have more than one cavity. *A continuous surface*, consisting of one cavity and not having self-intersections (multiple points) is named as the simple surface. It is having in mind, that simple surfaces are bilateral (one-sided surfaces, such as a sheet of Mebiusa, are eliminated).

The point of surface (1) is named *a regular point*, if at some parametric surface representation the function (1) has in a sufficient closeness to the examined point continuous partial first derivatives and, at least, one of determinants

$$\begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial y}{\partial u} \\ \frac{\partial x}{\partial v} & \frac{\partial y}{\partial v} \end{vmatrix}, \quad \begin{vmatrix} \frac{\partial y}{\partial u} & \frac{\partial z}{\partial u} \\ \frac{\partial y}{\partial v} & \frac{\partial z}{\partial v} \end{vmatrix}, \quad \begin{vmatrix} \frac{\partial z}{\partial u} & \frac{\partial x}{\partial u} \\ \frac{\partial z}{\partial v} & \frac{\partial x}{\partial v} \end{vmatrix}$$

is different from a zero. The simple surface piece, bounded with the closed curve, is named *regular*, if all of its internal points are regular. Bilateral simple (closed or unclosed) surface, composed of the finite number regular pieces with common regular arcs and points is named *a regular surface*.

Thus, regular unclosed surface in space of Oxyz, which consists of simple pieces of surface, can be put in correspondence to every halftone picture. It takes a place the next contingency for the surface, which corresponds a halftone picture. One and only one value of function r(x,y) corresponds every value of co-ordinates (x,y), that a perpendicular to the plane of image in any point x,y crosses an imaginary surface one and only one time.

The every piece external contour (border) of regular surface is the closed sequence of regular curve arcs and line segments. The contour points are not the regular points of simple surfaces pieces. The contour points are boundary points of simple surfaces pieces. The contour points form the special lines of surfaces which are boundaries, dividing the different pieces of simple surfaces. Unlike binary images the points of which can have two values of optical density only - black or white, the areas of half-tone pictures can have different change laws of optical density. Accordingly, an amount of different from each other nearby pieces of simple surfaces, as a rule, is more than two. So, contours of gray-scale picture can't be, in general case, the simply connected sequences, consisting of regular curve arcs and line segments, and consist of branches, connecting graph junctions, which form graph. The branches are the special lines of image. The contour points, except graph junctions, are the regular points of branches. Graph junctions are the special points of contour and all of image. Branches and graph junctions together with the law of change of optical closeness of every piece simple a surface fully determine a regular surface and proper by it area of image. In many practical cases the contours of gray-scale picture are the simply connected sequences of regular curve arcs and line segments, when simple, not contact with each other objects are located on background, that simplifies the task of structural analysis substantially.

It is always possible to select areas in a gray-scale picture, having permanent, or changing on a certain law value of optical density. So, a gray-scale picture can be represented as some area of regular surface, consisting of regular pieces of simple surfaces, thus every object of image corresponds to one or a few pieces of simple surfaces.

Every simple surface piece is fully determined some function of two arguments and contour. Within the boundary of every simple surface piece the law of change an optical density can be represented as a n degree polynomial function. A contour is the closed sequence of branches – line segments and curve arcs, and every branch can be represented as a n degree polynomial function. The representation of simple surface piece as a set of polynomial functions is invariant in relation to affine transformations. Consequently, the offered half-tone image representation is invariant in relation to affine transformations.

Digital Line Model of Arbitrary Gray-scale Picture

Let the grate of N×M, having discreteness value of *t,* is overlapped the image plane Oxy. That is the vertical lines of the grate {0,N}, distance *t* between them, are parallel to Ox axis. The horizontal lines of the grate {0,M}, distance *t* between them also, are parallel to Oy axis. The grate cells correspond image pixels, having integer co-ordinates *n,m* (n→{0,N}, *m*→{0,M}), thus $x = n \cdot t$; $y = m \cdot t$. Discrete, digital presentation of image substantially differs from the mathematical model of halftone picture. Discrete image elements are pixels, having finite sizes, unlike infinitely small points, used in a mathematical model. The lines of discrete image are usually formed with pixels, having finite sizes. At the same time contour lines of mathematical model do not have a thickness.

To eliminate the contradictions arising up when using the halftone picture structural model for discrete image, let represent image, as two-dimensional cell complex [Kovalevskiy, 4].

An *D*-dimensional cell complex is a structure consisting of abstract elements called cells. Each cell is assigned an integer value from 0 to *D* called its dimension. There is a bounding relation imposed onto the cells: a cell of a lower dimension may bind some cells of a higher dimension. An example of a two-dimensional complex is shown in Fig. 1. In this case pixels are two-dimensional elements. For every pixel a value of optical density r is the basic determining pixel attribute.

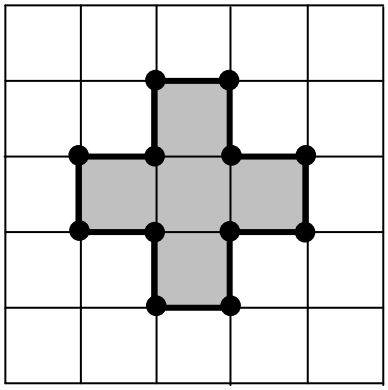


Fig.1

The pixels are represented in Fig. 2 as the interiors of the squares, the cracks as the sides of the squares and the points, i.e. the 0-cells, are the end points of the cracks and simultaneously the corners of the pixels.

The *boundary* (*crack boundary*), of object contour is the set of all boundary cracks and all end points of these cracks. A boundary contains no pixels and is therefore a „thin" set, whose area is zero. Contour of object (Fig. 2) in this case is the connected closed sequence of *contour cracks*, boundary between object pixels and background.

As shown in [2], representation of discrete image as a cell complex gives a lot of advantages; in particular, the contour of object becomes a thin curve with a zero area. The images of objects in discrete space can be presented as areas with some set function of optical density, bounded by the contour lines.

Let $\rho_m(n)$; (n = 0,N) – is the optical density function of horizontal pixel row number m (m = 0,M), and $\rho_n(m)$; (n = 0,N) – is the optical density function of vertical pixel line number n (n = 0,N). The regular and critical points in regular surfaces can be detected by means of structural analysis of $\rho_m(n)$ and $\rho_n(m)$ functions for all of vertical and horizontal picture rows [Kalmykov 5]. It ensues from determination of regular surface, that a point of surface is regular, if it is the regular point of horizontal and vertical lines of crossing. It ensues from determination of regular surface, that a point of surface is regular, if it is the regular point of horizontal and vertical crossing rows. If the point of surface is a critical point at least one of rows - horizontal and/or vertical crossing rows, such point is the critical, boundary point of regular surface that is the area of halftone picture.

The halftone picture areas boundary points (i.e. his regular surface) form the contour lines. The contour lines, in same queue, contain regular and critical points. The next operations must be executed to realize the developed method of halftone picture structural analysis.

1. The critical points of regular surfaces (i.e. areas of image) detecting.

2. The image critical lines (contours) construction, that are bounded objects, using the critical points of regular surfaces.

3. The contour structural elements – line segments and curve arcs detecting.

On the fig. 2 an example of halftone image structural analysis, using row model, is presented, namely contours of gray-scale picture selection. The next operations are showed over an image. The optical density function graphs for each vertical and horizontal rows of image are built, the examples of which are represented on the fig. 2 c,d. For every graph the sequence of elements which he consists of is determined, - digital line segments and digital curve arcs. The boundary points of the graph elements are its critical points of this row and all of halftone picture. The critical points of image are selected on the fig. 2b as a white color. The critical points belong to the halftone picture contour lines. The halftone picture contours are built, using the critical points. On the fig. 2e the contours in a raster mode, formed of the separate special points, are presented, and corresponded contours in a vector mode - fig. 2f.

Image Processing Experiments Using Row Model of Gray-scale Processing Picture

The digital image processing on the basis structural row model will be considered using as examples of the images of medical preparations, got on the Kirlian method. The images of medical preparations contain objects the form of which is very changeable, but, at the same time, there is diagnostic information exactly in a form, specialists to determine confidently enough at a visual estimation.

As a rule, such images are distorted by noise. The medical preparation images are using decision making in the medical diagnostic systems. Images which are got in the process of functioning of such systems, far not always there can be enough high quality that considerably reduces possibility of their rapid and complete perception by experts at the reasonable time.

But only the large quantity of such images to be analyzed, there can be the got new knowledge's the health state of population about. The processing and decision-making time, similarly as well as the amount of experts in the medical

diagnostic systems, as a rule, are limited. It is impossible to process such volumes of visual information without automation, to be high quality processing.

Images, got on the Kirlian method (farther Kirlian image), are pictures, executed on the special film, size of A4, on which luminescences are registered from each of ten fingers (fig.3).

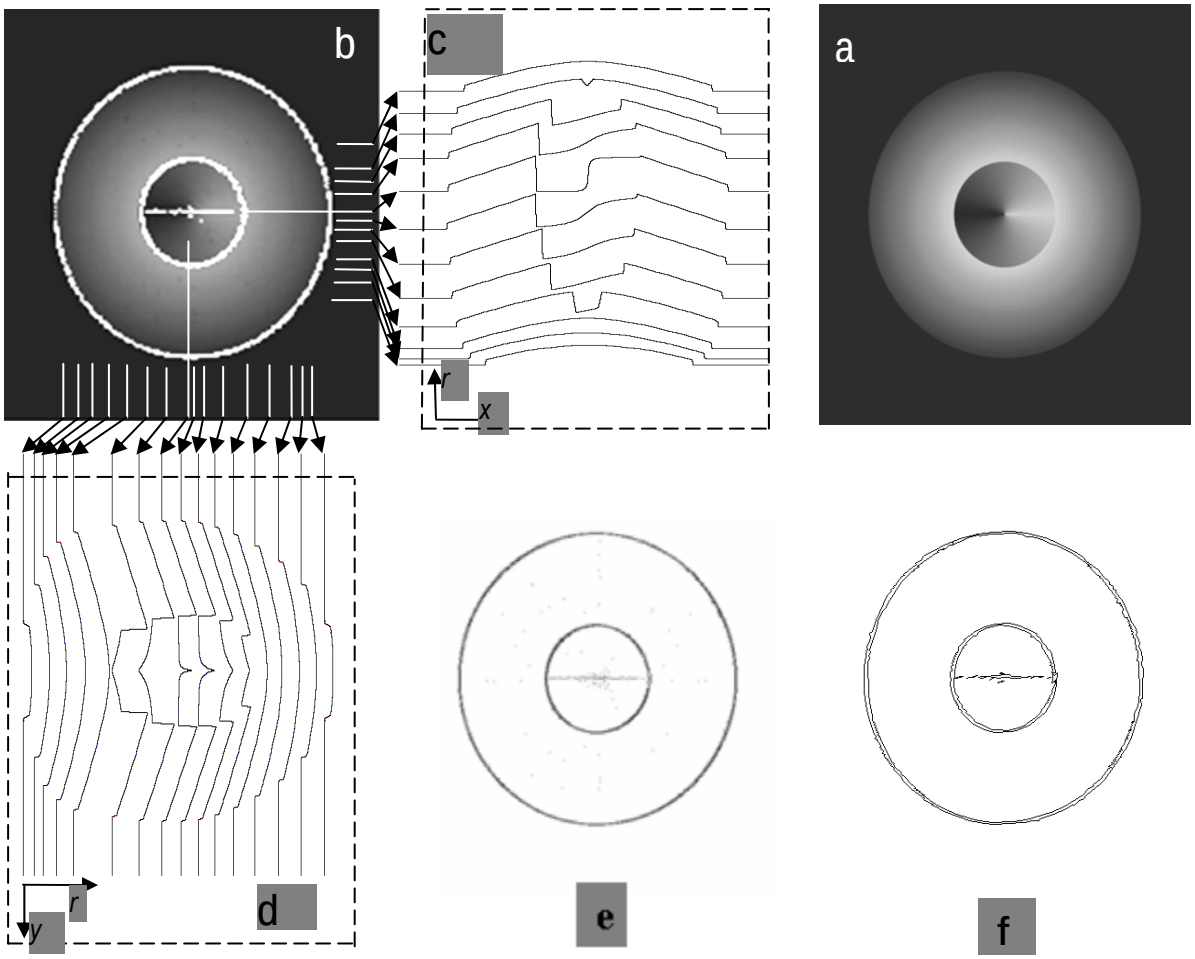


Fig.2. Contours detection on the halftone picture is executed by the programs, realizing the row model of image, and contours processing: a - model halftone picture; b - the same image with the detected critical points; c – the curves of optical density - *r* - horizontal lines; d - curves of optical density - *r* - vertical lines; e - contours images, formed of separate critical points; f - connected contours, consisting of line segments.

The images are of very bad quality: background variations, many hindrances which on intensity and size are comparable with objects, objects form and intensity instability and so on. Although on processing content these images would be considered binary, however, even such images binarization task cannot be considered easy, not to mention about the objects selection and identification tasks. There is the software for processing of Kirlian images,

got on the special devices separately for every finger [Korotkov, 5]. However the use of such devices complicates diagnostics, as the organism state can substantially change himself for the recording time each of ten fingers luminescence in consecutive order.

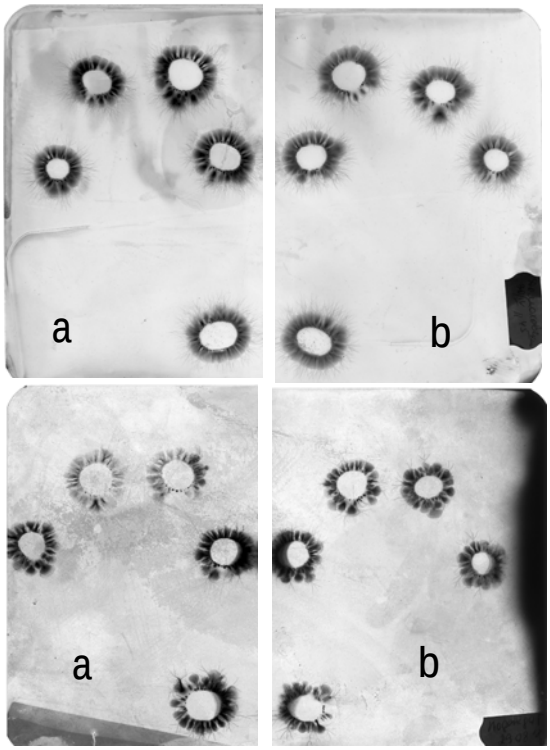


Fig. 3. Examples of Kirlian images: fingers emissions: a)- left, b) - right hands

M_o, m_f is mean values of objects brightness and background respectively; o_{min}, o_{max}, f_{min}, f_{max} are minimum and maximal brightness values of objects and background respectively.

To utilize the available bundled software for the Kirlian images, which all hand fingers (fig. 3) are simultaneously registered on, it is preliminary necessary to segment the images and define the turn corner of every finger luminescence in relation to a vertical line.

The developed software utilizes the halftone picture row model and it is intended for automatic segmentation of Kirlian images. The next operations are executed while Kirlian images processing.

The optical density histogram of the Kirlian image to be processed must be built (fig.4). Minimum objects brightness value of o_{min}, as minimum image brightness value, maximal background brightness value of f_{max}, as a maximal image brightness value is determined on a histogram. Objects mean brightness value of M_o as first maximum, while

brightness values increase since a zero, and background mean brightness *of* M_f, as the first maximum while brightness values decrease since maximal (255) are determined.

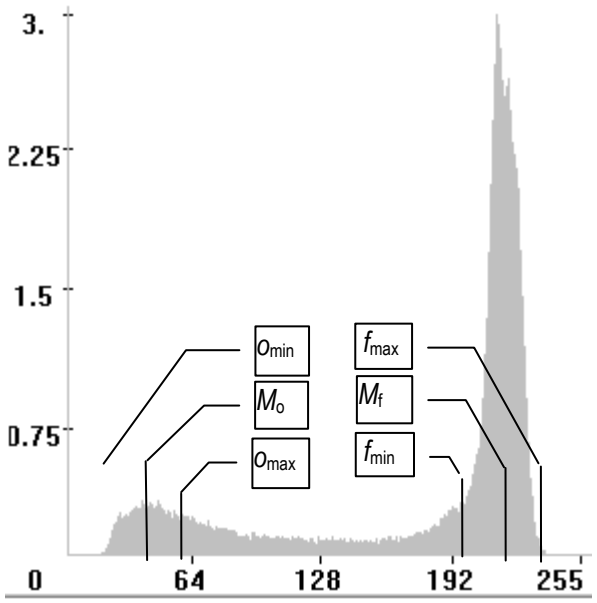


Fig. 4. Histogram of the image brightness (optical density). The values of brightness are built along abscise axis. The values, proportional the pixel amounts for this value of brightness along ordinate axes.

Background brightness minimum value as $f_{min}= M_f -(f_{max}- M_f)$ and objects brightness maximal value as $o_{max}= M_o +(M_f - o_{min})$ are calculated taking into account the next supposition. The random distribution of brightness values are symmetric as for background pixels, so for objects pixels.

– The next function for the image $v_b(m,n) = \begin{vmatrix} f_{\min}, \text{for } r(m,n) \geq f_{\min}; \\ o_{\max}, \text{for } r(m,n) \leq o_{\max}. \end{vmatrix}$ is calculated. This function is not the binarization function, as pixels, having intermediate brightness values *of* $omax<r(m,n)<fmin$ are not taken into account in the calculation process. Such pixels do not matter for the objects detection on an image, at least, for the decision of tasks on Kirlian images processing.

This transformation is the nonlinear change of quantum levels amount from 256 to 3 and allows to a great extent to eliminate influencing of noise in the image. The examples of function *of* $vb(m,n)$ are represented on fig.5 as the polyline. Numbers 1,4,5,8,9,12,13,16 mark the critical image points, which belong to the background. Numbers 2,3,6,7,10,11,14,15 mark the critical image points which belong to the objects.

The internal and, if necessary, external object contours are built, using the selected critical points (fig.5). If contours are selected, the objects are detected successfully.

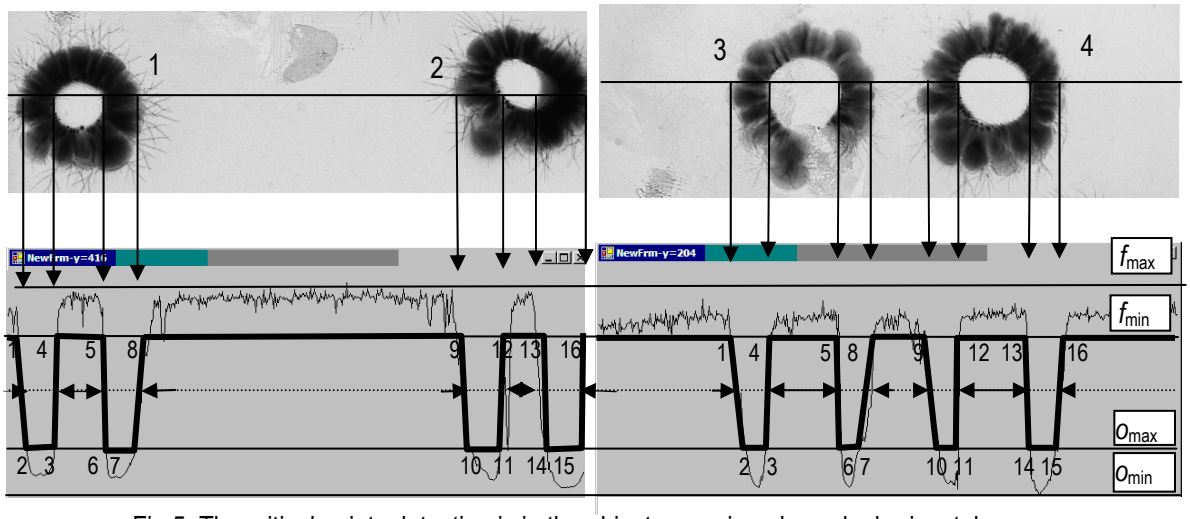


Fig.5. The critical points detection is in the objects crossing places by horizontal rows.

Objects positions (fingers emission) in relation to the palm center are detected. Internal contours are approximated as ellipses, the turn corner of every finger is detected, the finger emission images are turned to finger vertical position accordance (fig.6) and the resulting image files are formed.

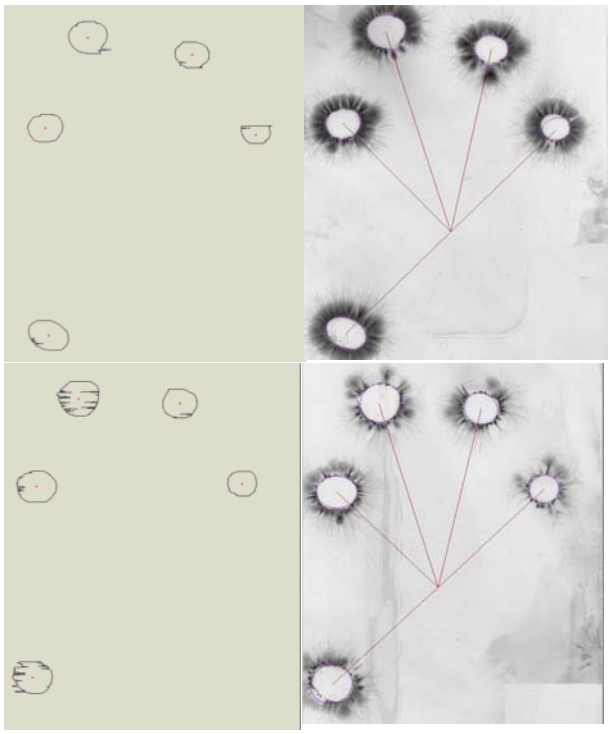


Fig.6. Internal object contours of (on the left).
Approximation of contours as ellipses (on the right).

Conclusion

1. The offered structural model allows halftone picture objects representations to be invariant to affine transformations.

2. The experiments confirmed possibility to process the half-tone images, which are different from each other such affine transformations as scale, rotation, location, using structural model.

Bibliography

[1] Gladun V.P. Planning of decisions. Kiev: Naukova dumka 1987. – 167p.

[2] Rabinovich Z.L. About the natural mechanisms of thought and intellectual computers // Cybernetic and system analysis. - 2003. - №5 P.82-88

[3] Korn G.A., Korn T. M. Mathematical handbook for scientists and engineers. - Moscow: Nauka, 1974.

[4] Kovalevsky V.A. Applications of Digital Straight Segments to Economical Image Encoding, In Proceedings of the 7-th International Workshop, DGCI'97, Montpellier, France, December 3-5, 1997, Springer 1997, P. 51-62.

[5] Korotkov K.G. The bases of GRV bioelectrography. - SPb., 2001. - 134 p.

Authors' Information

Vitaly Vishnevsky - head of division, senior researcher, candidate of engineering sciences, Institute of problems of mathematical machines and systems, prosp. akad. Glushkova 42, 03680, Kiev 187, Ukraine; e-mail: vit@immsp.kiev.ua

Major Fields of Scientific Research: Information technologies, Decision support systems

Vladimir Kalmykov - senior researcher, candidate of engineering sciences, Institute of problems of mathematical machines and systems, prosp. akad. Glushkova 42, 03680, Kiev 187, Ukraine; e-mail: kvg@immsp.kiev.ua

Major Fields of Scientific Research: Image processing, Visual information systems

Tatyana Vlasova - researcher, Institute of mathematical machines and systems, prosp. akad. Glushkova 42, 03680, Kiev 187, Ukraine;

e-mail: chery@immsp.kiev.ua

Major Fields of Scientific Research: Software technologies, Information technologies

AN ONTOLOGY-BASED APPROACH TO THE INCOMPLETE SIMULATION MODEL ANALYSIS AND ITS AUTOMATIC COMPLETION

Alexander Mikov, Elena Zamyatina, Evgenii Kubrak

Abstract: *The subsystem of simulation model completion is discussed. This subsystem is one of the components of computer-aided design and simulation system Triad.Net. Triad.Net is dedicated to computer systems design. And it is well-known, that very often investigators do not know about the behavior in details of each component of computer system being under design in the early stages of this process. The paper considers an ontology-based approach for the incomplete simulation model analysis and its automatic completion. A behavior procedure for the undefined element is searched for in special database and included in simulation model. The paper considers the method of model completion, namely, introduces the concept of a "semantic type" and some conditions that should be fulfilled for an appropriate behavior procedure to be chosen. The base ontology of simulation model representation is discussed and the choice of language OWL is explained. The presented example shows the process of simulation model automatic completion, illustrates the use of semantic type and additional conditions. Besides, the paper describes the programming tools which provide an ontology-based automatic completion, presents the semiautomatic completion of partly described simulation model in simulation system Triad and the architecture of the program subsystem being under consideration.*

Keywords: Simulation, ontology, simulation model uncertainty, automatic completion, OWL.

ACM Classification Keywords: I.6 SIMULATION AND MODELING I.6.2 Simulation Languages:J.6 COMPUTER-AIDED ENGINEERING I.2 ARTIFICIAL INTELLIGENCE I.2.5 Programming Languages and Software - Expert system tools and techniques

Introduction

The researchers using simulation as a method of investigations of complex systems often confronted with a problem of analysis of partly described models. Usually the behavior (rules of operation) of some model components is unknown. For example, it is not known, how much time the database search will take and will it be successful (information system is an object of investigations). Another example: a designer of computer networks does not know the exact behavior of router or another device in the early stages of computer networks design. A designer needs only a rough algorithm of data transfer. He doesn't know yet this algorithm in details.

It is clear, that under such conditions simulation process will not provide accurate account of complex system processes. However, despite this fact, a researcher wants to carry out a simulation experiment, and to obtain some results, which should be considered approximate.

The problem is that the simulation system cannot perform a simulation experiment if it lacks even one procedure describing behavior of any element of designed complex system. Therefore substitution of lacking behavior procedures with some "appropriate" ones taken from the standard library is needed to bring the model up to full strength.

This paper describes a process that is known as an automatic completion of a simulation model and considers an ontology based approach towards the problem solving used in simulation system Triad.Net [Mikov, 1995, 2003].

It is necessary to mention, that knowledge-based approach and automation are two factors increasing simulation modeling effectiveness and flexibility.

Related works

Some papers, dedicated to motivations for using ontologies in simulation modeling and role of ontologies in simulation process, appeared last time, for example [Silver, 2006], [Fishwick, 2004, 2005], [Miller, 2005].

So, [Benjamin, 2006] describes an ontology-based solution framework for simulation and modeling and analysis and outlines the benefits of this solution approach.

First of all let us introduce ontologies: «an ontology is an inventory of the kinds of entities that exist in a domain, their salient properties, and the salient relationships that can hold between them», every domain-typically, in this context, some piece of the actual world such as a manufacturing system, an university, a business – has its own ontology, which we refer to simply as a domain ontology [Benjamin,1995]. Ontology represents knowledge in a structured manner, within the structure of a semantic network, consisting of a diagram composed of nodes and arcs. Nodes are dedicated for representation of concepts, but arcs – the relationships among concepts within a particular knowledge domain. So we can see an illustration of simple ontology which relates various Petri nets models together (fig.1.) [Fishwick, 2005].

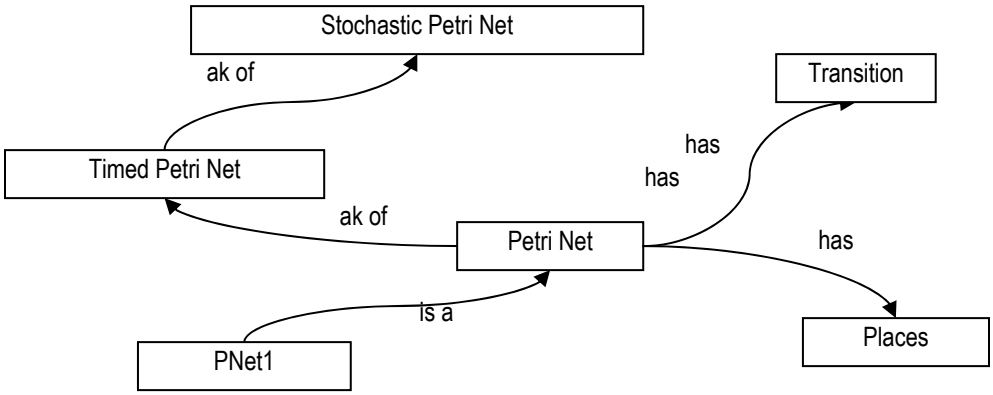


Fig.1.An Ontology for Petri Nets

One can extract such a knowledge from an ontology, represented by fig.1.: (1) A Petri net is a kind of (ak of) Timed Petri Net, which in turn is a kind of Stochastic Petri Net (Timed Petri Nets are types of Stochastic Petri Nets); (2) Each Petri Net is composed of Places and Transitions; and (3) One particular Petri Net is labeled Pnet1 (instance Pnet1 defines one specific Petri Net and may be associated with any business process).

Let us consider the benefits of ontologies. So we can conclude that ontologies may be used to share a common understanding of the structure of information, enable reuse of domain knowledge, make domain assumptions explicit, separate domain knowledge from operational knowledge, and analyze domain. Several of modeling researches may use the same ontology providing links between the concepts and so it increases the potential for interoperability, integration and reuse of simulation artifacts. Ontologies are useful across the simulation modeling and analysis lifecycle, particularly in the problem analysis and conceptual model design phases, they are essential in facilitating simulation model interoperability, composition, and information exchange at the semantic level.

First of all we will show how ontologies may be used in simulation process consisting of several parts: purpose of simulation model defining, acquiring and analysis of system description, determining and classification modeling objectives, determining object roles, boundaries and level of detail, data collection and data analysis, detailed model designing.

So while modeler defines purpose of simulation model design the role of ontologies may be determine as "terminology harmonization to enable shared and clear understanding".

The stage of conceptual model design includes such activities as system description validation, model boundaries and level of modeling abstraction identification, model objects and their roles identification, model structure and logic determination. Ontology knowledge is used here to determine the unambiguously differentiated abstraction levels, to map system objects to model objects and to identify an appropriate model objects role. Moreover ontological analysis helps reason with system constraints to facilitate determination of model logic.

The stage of data analysis suppose the fulfilling such an activities as data reduction, statistical analysis performing, data and text mining. Ontologies play a major role in detailed data analysis.

Next stage of simulation process is detailed model design. It have to provide a refinement of simulate model objects, their validation, moreover, a refinement of model structure and model logic. The primary role of ontologies in detailed model design is in the detailed analysis of information about objects and constraints. This involves mapping the simulation model constraints to specifications of real world constraints that are found within the domain system descriptions.

With increased use of distributed intelligence approaches to simulation modeling (distributed simulation, federated simulation, agent based simulation, etc.), ontologies play a critical role in simulation integration and simulation composability (for example, HLA [Gustavson, 2004], [Rathnam, 2004]). "Composability is the capability to select and assemble simulation components in various combinations into simulation systems to satisfy specific user requirements" [Petty, 2003]. The components to be composed may be drawn from repository. One can distinguish two kinds of composabilities: syntactic and semantic. Syntactic composability deals with the compatibility of implementation details such as parameter passing mechanisms, external data accesses, and timing mechanisms. Semantic composability, on the other hand, deals with the validity and usefulness of composed simulation models [Petty, 2003].

Nowadays simulation models may be composed using components which were developed by different teams in different domains. It is necessary to provide valid interaction between components. One can name components being integrated as federate and simulation model which was developed with these components (federates) is often called a federation. There are multiple challenges involved in component-based simulation. The component models may or

may not have been developed with the federation in mind. The principal technical challenge that must be addressed "in dealing with this comprehensive and critical void include modeling and simulation composability, semantic interoperability and information sharing, and model composition at multiple levels of abstraction" [Benjamin, 2006].

An ontology-driven approach to component-based simulation includes the following steps: (a) component ontology building up; (b) a repository creation; (c) repository access to the M&S community providing. These steps suppose next activities:

(a) Define an ontology for all components that explicitly captures the definition of goals of the component, detailed descriptions of net inputs required for execution, detailed descriptions of total outputs generated, system requirements, constraints on successful execution, and preconditions and post conditions, etc.

(b) Create a virtual repository of components and component ontologies.

(c) An M&S expert attempting to compose a federation must be able to easily identify, based on ontology descriptions, the components required for the composition. The repository must allow for the downloading of component copies for such uses.

Almost all papers consider the process of creating ontologies and discuss the special languages such as OWL (a language for OWL ontologies building) [Dean, 2002], [Lacy, 2004]. The benefits of ontology management methods and tools, the role of ontology-based analysis and the architecture of OSFM are described in [Benjamin,2006] . OSFM is an Ontology-driven Simulation Modeling Framework (OSMF) solution that provides a "visual programming environment" to rapid compose, build, and maintain distributed, federated simulation. The role of ontologies in trajectory simulation is discussed. The ontology is regarded as the domain model component of the reuse infrastructure, and is being developed to be a reusable knowledge library on trajectory simulations [Durak, 2006].

The authors of paper [Liang, 2003] represent port ontology. It is suite program tool for simulation model composition with components (or subsystems) from repository. Port is an interface; this interface describes the boundaries of component (subsystem). System is a configuration of subsystems that are connected to each other through well-defined interfaces. The configuration interface of a component object consists of ports. Ports define the intended interaction between a component and its environment; interaction consists of the exchange of energy, matter or signals (information). For instance, the configuration interface of the motor may has ports for the stator, the shaft of the rotor, and the electrical connectors. It is through its ports that a component (sub-system) interacts with other components (sub-systems), as is indicated in the graph by the connection between ports. The fact that these interactions have been abstracted into ports does not imply that only components with standardized connectors can be defined in this fashion. When interfaces are not completely standardized (e.g. a weld between two structural elements), the interaction can still be abstracted into one of a relatively small set of general interactions types (e.g. a rigid mechanical connection).

Port ontology usage allows automating the process of model composition with appropriate components and subsystems because of the fact that ontologies save knowledge about ports connection.

Later we shall consider the representation of model in simulation system Triad, main features of this system, after that we will regard the example of partly described model and show the application of ontology approach to solve the problem of simulation model automatic completion.

Simulation system Triad.Net architecture

Let us consider simulation modeling system Triad.Net, its appointment, its components and functions of each component. So distributed simulation system Triad.Net is a modern version of previous simulation modeling system Triad [Mikov, 1995] dedicated to computer aided design and simulation of computer systems. Triad.Net is designed as distributed simulation system, so various objects of simulation model may be distributed on the different computer nodes of a computer system. One more specific characteristic of Triad.Net – remote access, so several investigators may fulfill a certain project from different computers situating in different geographical points.

Distributed simulation system Triad.Net consists of some subsystems: compiler (TriadCompile), core of simulation system (TriadCore), graphical and text editors, subsystem of testing and debugging (TriadDebugger), subsystem of distributed simulation (synchronization of simulation model objects which are situated on different calculating nodes of computer system, conservative and optimistic algorithms realization), subsystem for equal workload of calculating nodes (TriadBalance), subsystem of remote and local access (TriadEditor), subsystem of automatic and semiautomatic simulation model completeness (TriadBuilder). Components of simulation system Triad.Net are represented on fig. 2. Triad.Net is suitable for computer aided design of computer system and may be applied in various domains.

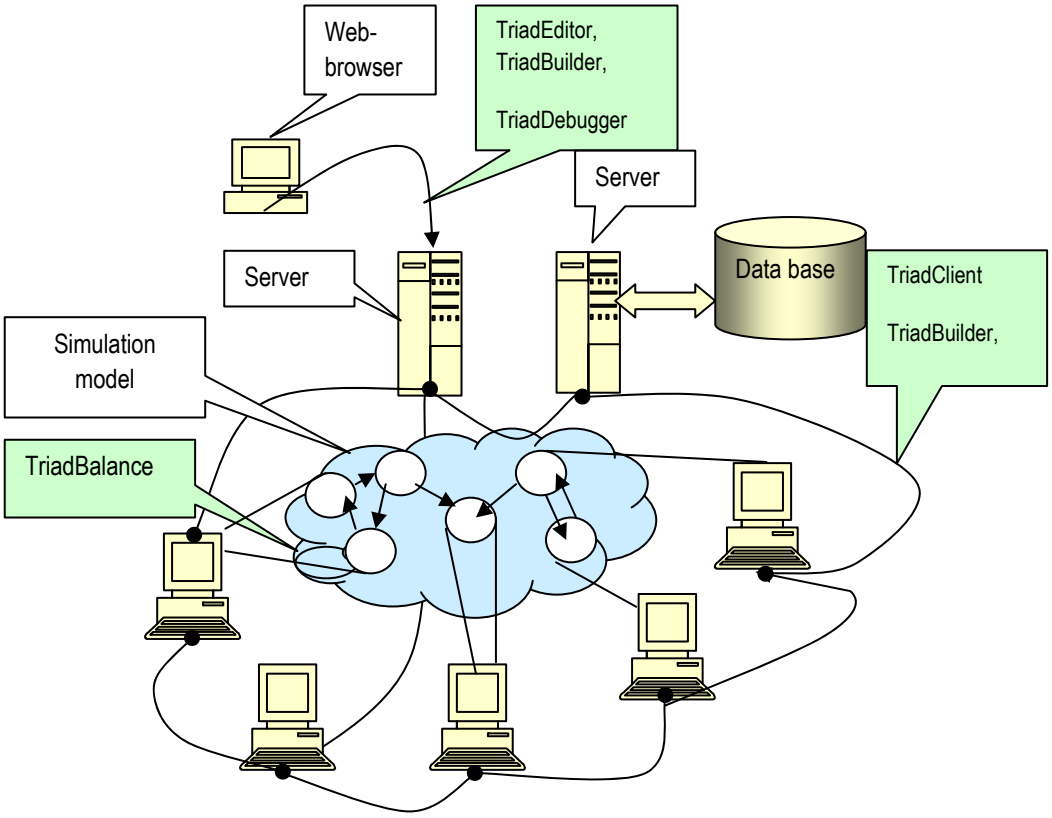


Fig.2. Simulation modeling system Triad.Net architecture

Now we are going to consider in details one of above components – the subsystem of automatic and semiautomatic simulation model completeness (TriadBuilder). Initially we address to the specific characteristics of simulation model in Triad.Net.

Model description

A simulation model $\mu = \{STR, ROUT, MES\}$ consists of three layers, where *STR* is a layer of structures, *ROUT* – a layer of routines and *MES* – a layer of messages.

The layer of structures is dedicated to describe the physical units and their interconnections, but the layer of routines presents its behavior. Each physical unit can send a signal (or message) to another unit. So each object has input and output poles. Input poles are used for sending messages. Output poles serve for receiving messages. A message of simple structure can be described in the layer of routines. A message of a complex structure can be described in the layer of messages only. Many objects being simulated have a hierarchical structure. Thus their description has a hierarchical structure too. One level of the system structure is presented by graph $P = \{U, V, W\}$. P-graph is named as graph with poles. *V* is a set of nodes, presenting the physical units of an object to be designed, *W* – a set of connections between them, *U* – a set of external poles of a graph. Internal poles of nodes are used for information exchange within the same structure level; in contrast the set of external poles serves to send signals (or more complex information) to the objects situated on higher or underlying levels of description.

Poles are very important part of a model. They represent "interfaces" of model components: objects, routines and graphs: communication links are being established through the poles of structure nodes; applying routine instance to a structure node consists of relating set of routine's poles to the set of node's poles; also, to complete operation of opening of structure node with a graph we need to relate outer poles of a graph to the poles of node (outer pole of a graph is a set of poles of it's nodes used to communicate with the rest of the model). Thus, when a node is opened with a graph or routine is set for a node, this node or routine is "inserted" into model and interacts with the rest of the model through the "interface" described by poles of the node object.

A set of routines is named as routine layer *ROUT*. Special algorithms – elementary routines – define a behavior of a physical unit and are associated with particular node of graph $P = \{U, V, W\}$. Each routine is specified with the set of events (*E*-set), the linearly partly ordered set of time moments (*T*-set) and the set of states {*Q*-set}. Each state is specified with the values of local variables. Local variables are defined in each routine.

The state is changed only if any event occurs. One event schedules another event(s). Routine (as an object) has input and output poles too. An input pole serves to receive messages, output – to send them. A special statement *out* (*out* <message> *through* <pole name>) is used to send a message. An input event e_{in} has to be emphasized among the other events. All input messages are processed by the input event, and output messages – by the ordinary events.

System Triad.Net [Mikov, 2003] is advanced discrete-event simulation system Triad [Mikov, 1995], but it is the distributed/parallel one. Conservative and optimistic algorithms were designed in Triad.Net. Besides, Triad.Net is characterized by the following [Mikov, 1995]:

- Triad language includes the special type of variables – type "model". There are several operations with the variable type "model". The operations are defined for the model in general and as well as for each layer. For

example, one may add or delete a node, add or delete an edge (arc), poles, create a union or an intersection of graphs. Besides, one or another routine (routine layer) using some rules can be assigned to the node (structure layer). The behavior of the object associated with this node would be changed. Besides, there's no need to retranslate the model. Thus a simulation model can be described by linguistic structures or built as a result of a model transformation algorithm.

- A simulation model is hierarchical, so each model (node) in a structure layer can be associated with some substructure.
- The model analysis subsystem has to provide a user with the possibility to formulate not regulated request. So the investigator may avoid the information superfluity or its insufficiency. The investigator can change the set of collected data within the simulation run, but model remains invariant. The model analysis subsystem has to posses smart software tools to analyze the simulation run results and to recommend the policy for the following simulation runs.

Partly described model

An ordinary simulation system is able to perform a simulation run for a completely described model only. At the initial stage of designing process an investigator may describe a model only partly omitting description of behavior of a model element $\mu_r^* = \{STR, ROUT^*, MES\}$). Simulation model may be described without any indication on the information flows effecting the model ($\mu_s^* = \{STR^*, ROUT^*, MES\}$) or without the rules of signal transformation in the layer of messages ($\mu_m^* = \{STR, ROUT^*, MES\}$). However for the simulation run and the following analysis of the model all these elements have to be described may be approximately.

For example, in a completely described model each terminal node $v_i \in V$ has an elementary routine $r_i \in ROUT$. An elementary routine is represented by a procedure. This procedure has to be called if one of poles of node v_i receives a message. But some of the terminal nodes v_i of partly described model do not have any routines. Therefore the task of an automatic completion of a simulation model consists either in "calculation" of appropriate elementary routines for these nodes, i.e. in defining $r_i = f(v_i)$, either in "calculation" of a structure graph $s_i = h(v_i)$ to open it with (in order to receive more detailed description of object being designed). It was mentioned above that the routine specifies behavioral function assigned to the node, but the structure graph specifies additional structure level of the model description. And at the same time, all structures s_i must be completely described as the sub-models.

These actions have to be fulfilled by the subsystem TriadBuilder. Fig.3 represents processing of simulation model in Triad.Net.

One may differentiate automatic and semiautomatic completeness of partly described simulation system. Semiautomatic completeness supposes to use a special linguistic construction and appropriated program component. It is "*conditions of simulation*". Another strategy: to put simulation model into the «environment of simulation».

Semiautomatic supposes that investigator have to include some statement in the part '*initial*' of *'conditions of simulation'*. We can denominate these statements: (a) an imposition of a routine on a node; (b) an imposition of a message layer on a model; (c) node description with substructure; (d) statements changing the structure of simulation model (adding and removing of some nodes, arcs, inputs and outputs and so on). One may use statement *simulate* to initiate simulation run. But at first simulator fulfill completion of partly defined simulation model

(processing of statements in part *'initial'* of linguistic construction *'conditions of simulation'*). So, the semiautomatic completeness allows changing the behavior of the simulation model during the same simulation experiment (node description with substructure, an imposition of a routine on a node). Moreover semiautomatic completeness allows changing the structure of messages (an imposition of a message layer on a model). The information flow impact on model and algorithms of signal transformation may be changed if investigator will use another 'conditions of simulation'. Let us consider an example of simulate statement*: simulate M on condition of simulation New_Condition*(*M.N1.a*, *M.N2.b*), where *M.N1.a*, *M.N2.b* – are actual arguments. These actual arguments are the simulation model variables being under monitoring by information procedures of Triad.Net.

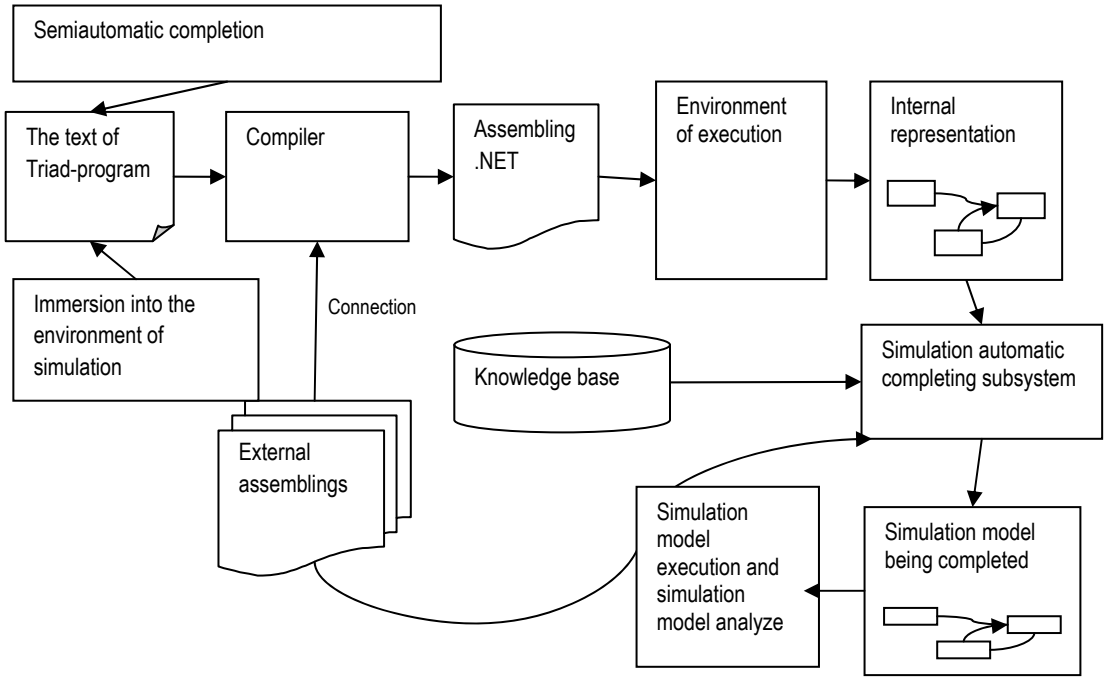


Fig 3. Simulation system Triad.Net and model completion subsystem

Let's consider an example of computer network fragment simulation. This fragment consists of workstations and routers (fig.4.). Each workstation has one neighbor (router). Workstations attempt to transfers data to another one. Data have to pass some routers, but the behavior of routers is unknown. So it is necessary to detect all nodes of simulation model, find out nodes (v_i) without routines, search out the appropriate one in data base of routines and fulfill the completion of model. Let's discuss the method of model completion suggested by authors.

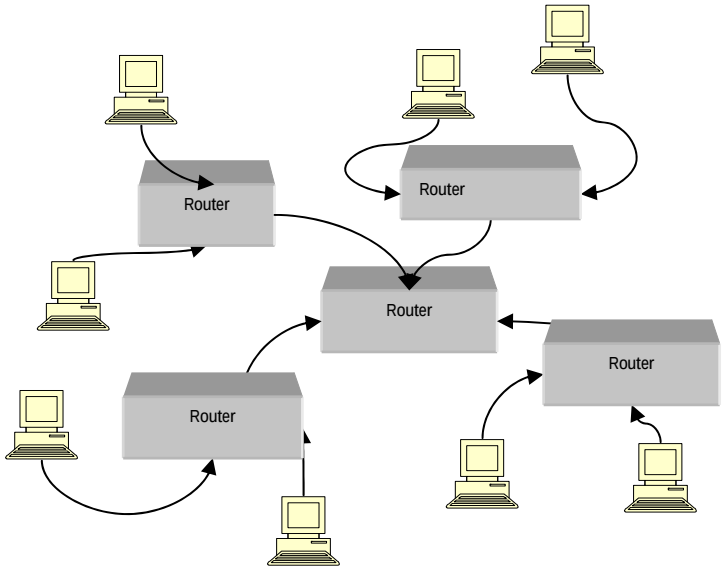


Fig. 4. The fragment of computer network consists of workstations and routers

The method of model completion

The method of a model *M* completion at a node *v* is based on following.

Node *v* is represented by sets *In*(*v*) and *Out*(*v*) of input and output poles respectively, its name and semantic type. Its poles are described by types of messages passing through them. Node *v* is located in some "environment" of the model *M*, it is adjacent to certain nodes v_k of the model. Nodes v_k for their part are also described by their poles, names, semantic types and environments.

All this information imposes restrictions on possible routine for *v*. It must not be chosen randomly. Despite some remaining freedom of choice, it should be picked out from some limited set.

An elementary routines library is created for a given simulation domain. The library should be consistent with domain ontology.

The "semantic type" concept is used to provide means for selection of an appropriate objects "opening". A semantic type of the node defines a possible object class from certain domain. The given node can represent objects of this class in a model. For example, when computing systems are simulated, one can define such semantic types as "processor", "register", "memory unit". Or, when queuing systems are simulated, such semantic types as "queue", "server" and "generator" would be appropriate. Using information, obtained from semantic type of a node, one can pick out an appropriate specification for it.

It is necessary to use special statement <object name> => <semantic type name> in order to declare semantic type of an object.

Semantic type could be declared by statement type <semantic type name>.

The fragment of Triad program can be given below. This fragment illustrates the statements mentioned above (to declare semantic type and to denote semantic type).

```
Type Router,Host;
integer i;
  M:=dStar(Rout[5]<Pol[4]>);
  M:=M+node Hst[8]<Pol>;
  M.Rout[0]=>Router;
  for i:=1 by 1 to 4 do
        M.Rout[i]=>Router;
        M:=M+edge(Rout[i].Pol[1]—Hst[2*i-2]);
        M:=M+edge(Rout[i].Pol[2]—Hst[2*i-1]);
  endf;
  for i:=0 by 1 to 7 do
        M.Hst[i]=>Host;
  endf;
M:=M+edge(Rout[1].Pol[2]--Rout[4].Pol[0])+edge(Rout[0].Pol[2]--Rout[3].Pol[0]);
```

Fig. 5. The fragment of Triad program (Program 1) with the semantic types.

An internal form of the simulation model can be gained as a result of a program run (fig.3.): it is a graph, each node of the graph represents a workstation or a router. Each workstation has a semantic type “host”, each router – semantic type “router”.

Corresponding routine instance search

So called specification, configuration and decomposition conditions are used to test the routine instance for each undefined node in order to complete simulation model. First of all, consider a specification condition. Closely related to specification condition is such a concept as “semantic type”. This concept was discussed earlier.

Let us assume that v – terminal node, r –routine instance from the knowledge base. Let us introduce the function *equtype*(v, r), defining specification condition performance. The result of this function is true, if the semantic type *Type*(v), assigned to the node corresponds to a semantic type *Type*(r), associated to routine instance found in the knowledge base.

Semantic type *T*1 corresponds to semantic type *T*2, if *T*1 is a superclass *T*2 (i.e. $T2 \subset T1$).

By this means the condition of specification is true if found routine instance corresponds to the semantic type of the node or to more special type.

Some specific type Object is introduced to provide the processing of the objects with unknown semantic type. Semantic type Object is a parent to all other semantic types. Any node is considered to be belonging to this parent type Object. If several particular semantic types are denoted the node is marked as belonging to each of them. Therefore routine instance without specified semantic type (more precisely, semantic type Object is specified) can be applied only to the node without definite semantic type. However, only routine instances belonging to the intersection of several semantic types can be applied to the nodes with these several types. For example, if we want to describe the node with a several functions (working as a router and host at the same time) we should declare both semantic types for it. In this case routine instances belonging only to one of these types would not be sufficient. We should find routine instance belonging to the intersection of these types, i.e. implementing both router and host functions.

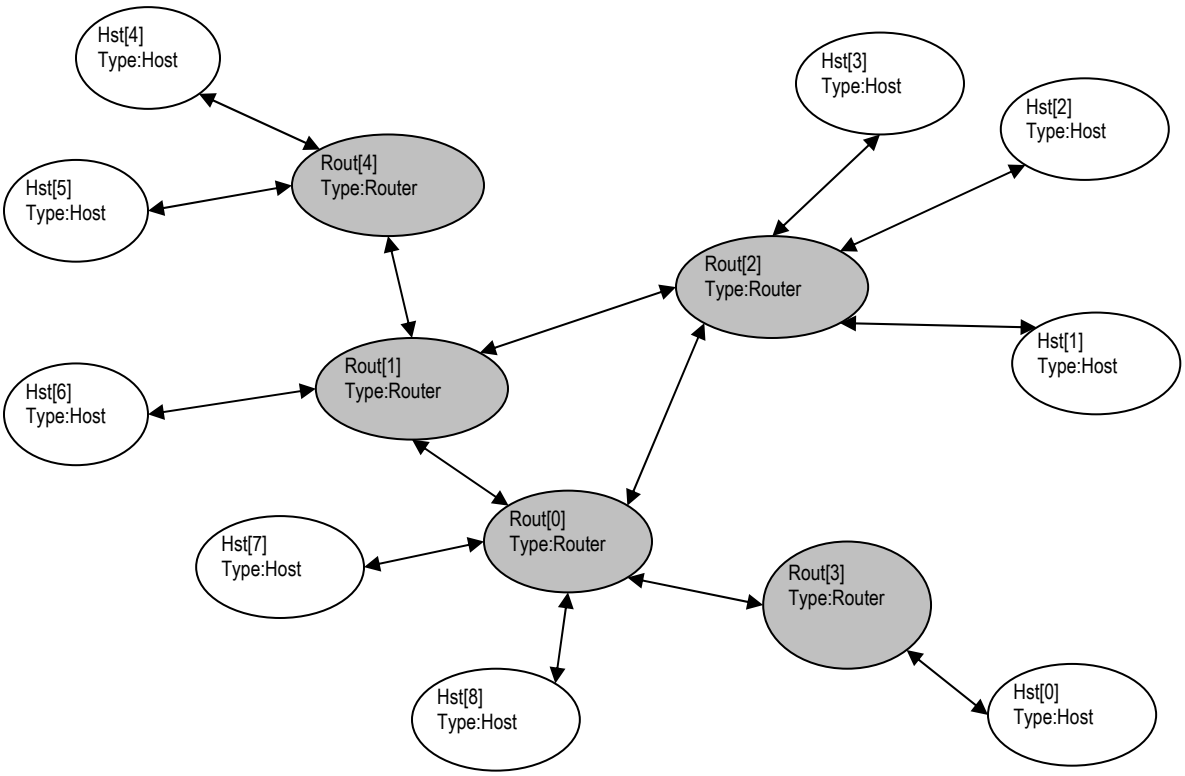


Fig.6. The internal form of computer network and semantic types

Configuration condition supposes the following: the amount of input and output poles of a node has to be equal to the amount of input and output poles of an appropriate routine instance. When we apply routine r to the node v the following relations are defined:

$$L_i: In(v) \rightarrow In(r) \quad (1)$$

$$L_o : Out(v) \rightarrow Out(r) \quad (2)$$

It is significant that these mappings are not functional ones. They define a set of related pairs (p_1, p_2), where $p_1 \in v$, $p_2 \in Pol(r)$, thus each pole can belong to any number of pairs or not to be used at all. Let $D(L_i/o)$ be cardinal numbers of input and output poles of a node, related to mappings *Li* and *Lo* respectively. Depending on these values the undefined node has to have some definite amount of input and output poles, to be more precise, the following mappings are significant:

$$|In(v)| \geq |D(L_i)| \quad (3)$$

$$|Out(v)| \geq |D(L_o)| \quad (4)$$

Nonrigourous equation allows using the nodes with an excessive amount of input and output poles when the routine instance is applied. The presence of the necessary minimum of poles is tested only, so some of inputs and outputs can be left "hanging", and all the messages, sent through them will not be processed.

Nodes *Rout*[1 to 4] are declared as nodes with 4 poles in our example (*M* := *dStar*(*Rout*[5] < *Pol*[4]>);). However, each of them is connected only with 3 neighbors, so one of its poles will not be used.

Decomposition condition defines rules of node connections in a model graph, and is derived from node adjacency relations.

To define decomposition conditions we can introduce the concept of surrounding graph of some node *v*. Let $G = \{V;W;U\}$ be graph and node ($v \in V$) belonging to graph *G*. The relation *S* determines the adjacency of nodes in graph *G*, i.e.:

$$\forall v_1, v_2 \in V : (v_1; v_2) \in S \leftrightarrow \exists p_1 \in v_1, \exists p_2 \in v_2 : ((p_1; p_2) \in W \vee (p_2; p_1) \in W) .$$

Function Sub(w,v) defines a set of poles of the node w, connected with poles v:

$$Sub\ (w, v) = \{p \in w \mid \exists p_0 \in v : (p; p_0) \in W \vee (p_0; p) \in W\} \quad (5)$$

Then graph $GG(v) = \{V';W';\varnothing\}$ - is a surrounding graph for v if the following conditions are observed:

$$V' = \{v\} \cup \{w' \mid w' = Sub(w,v);\ (w;v) \in S\} \quad (6)$$

$$\forall w : (w;v) \in S \rightarrow Sub(w,v) \in V' \quad (7)$$

$$\forall p_1, p_2 : [(p_1;p_2) \in W] \wedge [p_1 \in v \vee p_2 \in v] \rightarrow (p_1;p_2) \in W' \quad (8)$$

$$W' \subset W \quad (9)$$

Eq (6) limits the set of nodes of the surrounding graph of the node *v*: it includes the node *v* itself and subsets of nodes adjacent with it. It is significant to take into account only those poles of adjacent nodes which are in accordance with Eq (5) directly connected with poles of node *v*.

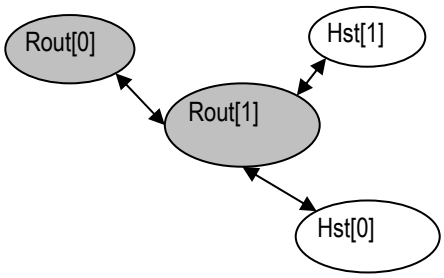


Fig.7. The surrounding graph of a node Rout[1]

Eq (7) specifies that this sort of subset exists for each adjacent node.

Eq (8) informs that each edge adjacent to the node v would be included to the surrounding graph. Eq (6) and Eq (7) state that all adequate poles would be included in the surrounding graph too.

Eq (9) asserts that there are no any excessive edges in the surrounding graph but only edges corresponding to Eq (8) are included.

So decomposition condition checks up following:

- which semantic types are set to the nodes connected to given node *v*,
- an isomorphism of two graphs: the surrounding graph of a node *GG*(*v*), and the pattern graph *GG*'(*r*) taken from the domain ontology.

The pattern graph is stored in a knowledge base and associated with a routine instance.

The fulfillment of decomposition condition is determined by function *iso*(*GG*'(*r*), *GG*(*v*)), which searches the environment graph of node *v* for a subgraph isomorphic to *GG*'(*r*), and also checks some additional restrictions for environment graph.

Graph *GG*'(*r*) is a pattern graph taken from the knowledge base and it should be relevant to the actual surrounding graph. If *v*' is a central node of this pattern graph then it shouldn't have any "hanging" poles, i.e. each pole of node *v*' corresponds at least one pole of routine.

Thus the task of model completion subsystem implies the following: to find the proper routine instance from the knowledge base for each undefined node in a partly described model. Therewith the conditions of specification, configuration and decomposition have to be followed. The information required to test these conditions should be stored in knowledge base. The ontology approach is used to represent this information.

Base ontology

The base ontology describes a model representation in Triad.Net: classes such as Model, Object, Routine, Polus and so on are specified in it.

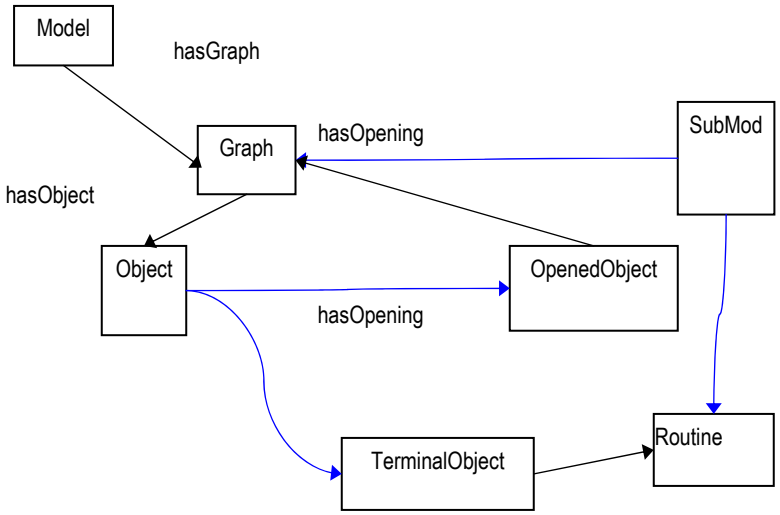


Fig.8. The fragment of ontology

Special class SubMod is presented describing everything the model object can be specified with. Its subclasses are the Routine and Graph classes representing the set of routines and structure graphs of one hierarchical level. Their common superclass allows unifying operations of opening object with a graph and applying the routine to an object.

The Web Ontology Language (OWL) is used to store the base ontology and all the domain ontologies. It was chosen because it allows composing information from different sources and is widespread and well known.

A part of the base ontology including most common concepts of ontology is depicted on Fig.5. “Has_subclass” relations are shown as blue arrows.

The poles hierarchy and its connection to the rest of ontology are missing on this picture.

Each semantic type defines a structure graph or a routine suitable for opening.

In the view of ontology knowledge representation semantic type is some class of nodes possessing certain properties. For example “request generator” is a node having at least one output pole connected to a node with semantic type of “queue”. Possible subclasses of “request generator” are generators of specific requests. To define them one can apply restrictions on message types corresponding to generator’s output only.

Thus, for any specific domain we should create ontology expanding the basic one which should contain information on how one should model the concepts from this domain.

Since we are mostly using the *Has_subcalss* relation it is convenient to have a hierarchy representation of an ontology. With the use of this approach, child nodes will represent concepts specifying concepts of parent nodes.

One could bind the most common concepts of a domain, for example, “device” when modeling the computing systems domain to root nodes. The nodes residing on lower levels of hierarchy would represent classes defining more precise concepts of a domain: devices would divide to “processor”, “memory unit” etc. “Processors” would be divided depending on their architecture and so on.

Simulator subsystem for model completion

The model completion subsystem includes the following components:

- Model analyzer. It searches through model and looks for terminal nodes without routines to mark them as needed to process.
- Model converter. It converts model from its internal form to the ontology representation, it is used to save surrounding graph of a routine instance.
- Inference module. It analyses the model ontology and searches it for an appropriate routine instances for each node marked by model analyzer.
- Model builder. Applies found routine instances to the nodes.
- Knowledge base, containing information about semantic types and routine instances with their pattern graphs.

Let us consider the algorithm of model completion subsystem in our example (fig. 3).

Model completion subsystem starts when the internal form of simulation model is built according to a Triad code.

First, model analyzer searches the model for incomplete nodes, and marks them. Assume that only the router nodes don’t have routines applied. Thus, the model analyzer will mark all *Rout* nodes.

The inference module starts looking for an appropriate routine instance for each of marked nodes. We’ll take *Rout*[1] as an example. Assume, that there are several routine instances for a Router semantic type, describing routers with 2, 3… 10 neighbors.

According to specification condition, inference module picks out these 9 routine instances discarding the ones with other semantic types.

Then, according to configuration condition it discards the instances of a router routine with 5 or more neighbors (*Rout*[1] is declared having 4 poles).

Lastly, it starts checking decomposition condition for remaining routine instances. Instance with 4 neighbors will be discarded, because the surrounding graph of *Rout*[1] has only 4 nodes, and the pattern graph for the instance has 5 nodes. All others will suffice for the condition. In order to avoid ambiguity and to choose the most appropriate instances, they are sorted before checking the decomposition condition, according to several heuristics (by the number of nodes for example).

After the appropriate instance has been found, it is applied to the node.

Conclusions

The paper considers a problem of completeness of partly defined simulation model. Partly defined simulation model and it's analyze and investigation may be actual at the beginning of simulation process. An investigator may not know all details of simulation model being under design. For example, an investigator cannot describe in detail the behavior of some components of computer network. But he wants to receive some results while simulations run. So it is necessary to complete simulation model description. Authors suggest automatic and semiautomatic completeness. Automatic and semiautomatic completeness of partly defined simulation model are the function of special component (TriadBuilder) of simulation system Triad.Net.

Semiautomatic completeness supposes special linguistic construction ("conditions of simulation") and appropriate program component using. One can change the initial part of this construction, more precisely, to change some statements in this part. These statements may change the behavior of some simulation model components or it's structure.

The automatic completeness allows searching program components in special data bases following some criteria and thanks to knowledge derived from ontologies.

Ontologies are the convenient path to domain describing. They are efficient for information systems design, data bases and complex programming systems development, computer network design and so on. It is a powerful tool for the system designers. Ontology can be useful to system researcher too. Investigation of a complex system by simulation methods together with ontology allows him/her to get results in difficult cases of incomplete, fuzzy models.

Bibliography

[Mikov,1995] Simulation and Design of Hardware and Software with Triad// Proc.2nd Intl.Conf. on Electronic Hardware Description Languages, Las Vegas, USA, 1995. pp. 15-20.

[Mikov,1995] Mikov A.I.. Formal Method for Design of Dynamic Objects and Its Implementation in CAD Systems // Gero J.S. and F.Sudweeks F.(eds), Advances in Formal Design Methods for CAD, Preprints of the IFIP WG 5.2 Workshop on Formal Design Methods for Computer-Aided Design, Mexico, 1995, pp. 105 -127.

[Mikov,2003] Mikov A.I., Zamyatina E.B., Fatykhov A.H.. A System for Performing Operations on Distributed Simulation Models of Telecommunication Nets // Proc. I Conf. "Methods and Means of Information Processing, Moscow State University (Russia), 2003. pp. 437-443 (in Russian).

[Fishwick,2004] Fishwick P.A. Ontologies For Modeling And Simulation: Issues And Approaches /Paul A. Fishwick, John A. Miller // Proceedings of the 2004 Winter Simulation Conference. pp. 259-264

[Dean,2002] Dean M., Connolly D., van Harmelen F., et al. 2002. Web Ontology Language (OWL) Reference Version 1.0. W3C. //www.w3.org/TR/2002/owl-ref/

[Silver,2006] Silver G.A, Lacy. L.W., J.A. Miller. Ontology Based Representations Of Simulation Models Following The Process Interaction World View. Proceedings of the 2006 Winter Simulation Conference L. F. Perrone, F. P. Wieland, J. Liu, B. G. Lawson, D. M. Nicol, and R. M. Fujimoto, eds., pp.1168-1176.

[Durak,2006] Durak U., Oguztuzun H., Ider S. K..An Ontology For Trajectory Simulation. Proceedings of the 2006 Winter Simulation Conference/ L. F. Perrone, F. P. Wieland, J. Liu, B. G. Lawson, D. M. Nicol, and R. M. Fujimoto, eds. pp.1161-1167

[Benjamin,1995] Benjamin P., Menzel C, Mayer R. J. Towards a method for acquiring CIM ontologies. // International Journal of Computer Integrated Manufacturing, 8 (3) 1995, pp. 225-234.

[Miller,2005] Miller J.A., Baramidze G. Simulation and the Semantic Web // Proceedings of the 2005 Winter Simulation Conference / M. E. Kuhl, N. M. Steiger, F. B. Armstrong, and J. A. Joines, eds.. – pp. 2371-2377

[Benjamin,2006]Benjamin P., Patki M., Mayer R. J. Using Ontologies For Simulation Modeling // Proceedings of the 2006 Winter Simulation Conference/ L. F. Perrone, F. P. Wieland, J. Liu, B. G. Lawson, D. M. Nicol, and R. M. Fujimoto, eds. –pp.1161-1167

[Benjamin, 2005] Benjamin P., Akella K.V., Malek K., Fernandes R. An Ontology-Driven Framework for Process-Oriented Applications // Proceedings of the 2005 Winter Simulation Conference / M. E. Kuhl, N. M. Steiger, F. B. Armstrong, and J. A. Joines, eds.,– pp 2355-2363

[Rathnam,2004]Rathnam T., Paredis C.J.J. Developung Federation Object Models Using Ontologies // Proceedings of the 2004 Winter Simulation Conference / R .G. Ingalls, M. D. Rossetti, J. S. Smith, and B. A. Peters, eds.,– pp 1054-1062

[Gustavson,2004] Gustavson P., Chase T. Using XML and BOMs to Rapidly Compose Simulations and Simulation Environments // Proceedings of the 2004 Winter Simulation Conference / R .G. Ingalls, M. D. Rossetti, J. S. Smith, and B. A. Peters, eds.,– pp. 1467-1475

[Lacy,2004] Lacy L.. Potential Modeling And Simulation Applications Of The Web Ontology Language – Owl / Lee Lacy, William Gerber // Proceedings of the 2004 Winter Simulation Conference. pp. 265-270

[Liang,2003] Liang V-C. A Port Ontology For Automated Model Composition / Vei-Chung Liang, Christiaan J.J. Paredis // Proceedings of the 2003 Winter Simulation Conference, pp. 613-622

Authors' Information

Alexander Mikov – Cuban State University, Head of Department of Computer Technologies, Aksaiskaya str., 40/1-28; e-mail: alexander_mikov@mail.ru

Major Fields of Scientific Research: General theoretical information research, Information Systems, Simulation, Distributed and Parallel programming, Software technologies

Elena Zamyatina – Perm State University, Department of Software of Computer Systems, 614017, Turgeneva str., 33–40, Perm, Russia; e-mail: e_zamyatina@mail.ru

Major Fields of Scientific Research: Simulation, Distributed and Parallel programming, Multiagent Systems

Evgenii Kubrak –Perm State University, Postgraduate of Department of Software of Computer Systems, Leonova str,.20-5, Perm, Russian Federation; e-mail:q_brick@mail.ru

Major Fields of Scientific Research: Simulation, Software technologies, Information systems, Multiagent systems

DUAL PROBLEM OF FUZZY PORTFOLIO OPTIMIZATION

Yuriy Zaychenko, Maliheh Esfandiaryfard

Abstract: The dual problem of fuzzy portfolio optimization is considered and investigated. A mathematical model of this problem was constructed, explored and the sufficient conditions for its convexity were obtained. The sufficient optimality conditions of its solution are presented.

Keywords: fuzzy portfolio, optimization, dual problem, optimality conditions

ACM Classification Keywords: H.4.2. Information system Applications: Types of Systems Decision Support

Introduction

Last years the problem of portfolio optimization attracts great attention due to the development of financial markets in Ukraine. The determining of optimal portfolio enables the investors and financial funds to distribute available financial assets, to obtain high dividends and cut the risk of erroneous solutions. A novel approach to portfolio optimization which is an alternative to classical portfolio model of Markovitz considers this problem under uncertainty and based on application of fuzzy sets theory.

The fuzzy portfolio problem was considered and investigated in [Недосекин, 2003, Зайченко, 2007]. In these works the problem statement was the following: to optimize the total portfolio profitableness under constraints on possible risk. The algorithm of its solution was suggested and explored in [Зайченко, 2007]. In [Зайченко, 2008] the forecasting of stock prices was suggested to utilize in fuzzy portfolio optimization which increased the efficiency of the solution.

The goal of the present work is the consideration of the dual portfolio problem, investigation of the constructed mathematical model and determining the optimality conditions of its solution.

The Dual Fuzzy Portfolio Problem

The initial portfolio optimization problem which is naturally to be called as direct has the following form [Недосекин, 2003, Зайченко, 2007] :

Optimize the expected profitableness of a portfolio

$$\tilde{r} = \sum_{i=1}^{N} x_i \tilde{r}_i \to \max \qquad (1)$$

under constraints on risk

$$\beta(x) \le \beta_{задан}\ 0 < \beta < 1, \tag{2}$$

$$\sum_{i=1}^{N} x_i = 1, \tag{3}$$

$$x_i \ge 0. \tag{4}$$

Let's consider the case when the criterion value r^* meets the conditions

$$\sum_{i=1}^{N} x_i r_{i1} \le r^* \le \sum_{i=1}^{N} x_i \tilde{r}_i = \tilde{r}. \tag{5}$$

Then

$$\beta(x) = \frac{1}{\sum_{i=1}^{N} x_i r_{i2} - \sum_{i=1}^{N} x_i r_{i1}} \left[\left(r^* - \sum_{i=1}^{N} x_i r_{i1} \right) + \left(\sum_{i=1}^{N} x_i \tilde{r}_i - r^* \right) \ln \left(\frac{\sum_{i=1}^{N} x_i \tilde{r}_i - r^*}{\sum_{i=1}^{N} x_i \tilde{r}_i - \sum_{i=1}^{N} x_i r_{i1}} \right) \right]. \tag{6}$$

Now consider the dual portfolio optimization problem dual related to the problem (1)-(4):

to minimize

$$\beta(x) \tag{7}$$

under conditions

$$\tilde{r} = \sum_{i=1}^{N} x_i \tilde{r}_i \ge r_{çŕä} = r^*, \tag{8}$$

$$\sum_{i=1}^{N} x_i = 1, x_i \ge 0. \tag{9}$$

Let's prove that the risk function $\beta(x)$ is convex where $\beta(x) = \left(A(x) + B(x) \ln \frac{B(x)}{C(x)} \right) - D(x)$.

It's necessary to prove that function $D(x) = \frac{1}{\sum_{i=1}^{N} x_i r_{i2} - \sum_{i=1}^{N} x_i r_{i1}}$ is convex and function

$A(x) + B(x) \ln \frac{B(x)}{C(x)}$ is also convex. And besides, both functions are decreasing and nonnegative by x_i,

where $A(x) = r^* - \sum_{i=1}^{N} x_i r_{i1}$; $B(x) = \sum_{i=1}^{N} x_i \tilde{r}_i - r^*$; $C(x) = \sum_{i=1}^{N} x_i \tilde{r}_i - \sum_{i=1}^{N} x_i r_{i1}$.

Really $A(x)$ is linear and therefore is not strictly convex and both $B(x)$ and $C(x)$ are linear.

Besides, $\tilde{r}_i \geq r_{i1}$, $i = \overline{1, N}$, $\sum_{i=1}^{N} x_i \tilde{r}_i - r^* > 0$ by assumption

Consider the function $D(x)$:

$$D(x) = \frac{1}{\sum_{i=1}^{N} x_i r_{i2} - \sum_{i=1}^{N} x_i r_{i1}},$$

$$\frac{\partial D(x)}{\partial x_i} = -\frac{r_{i2} - r_{i1}}{(\sum_{i=1}^{N} x_i r_{i2} - \sum_{i=1}^{N} x_i r_{i1})^2},$$

$$\frac{\partial^2 D(x)}{\partial x_i^2} = \frac{2(r_{i2} - r_{i1})^2}{(\sum_{i=1}^{N} x_i r_{i2} - \sum_{i=1}^{N} x_i r_{i1})^3} > 0.$$

And as $r_{i2} \geq r_{i1}$ then $\frac{\partial^2 D(x)}{\partial x_i^2} > 0$, for all $i = \overline{1, N}$ and $\frac{\partial^2 D}{\partial x_i x_j} = \frac{2(r_{i2} - r_{i1})(r_{j2} - r_{j1})}{(\sum_{i=1}^{N} x_i r_{i2} - \sum_{i=1}^{N} x_i r_{i1})^3}$.

Besides, all the diagonal minors of the second order

$$\begin{vmatrix} \frac{\partial^2 D}{\partial x_i^2} & \frac{\partial^2 D}{\partial x_i x_j} \\ \frac{\partial^2 D}{\partial x_j x_i} & \frac{\partial^2 D}{\partial x_j^2} \end{vmatrix} = 0 .$$

Therefore function $D(x)$ is convex.

Calculate

$$\frac{\partial}{\partial x_i}\left[B(x)\ln\frac{B(x)}{C(x)}\right] = B'(x)\ln\frac{B(x)}{C(x)} + B(x)\frac{B(x)}{C(x)}\frac{B'(x)C(x)-C'(x)B(x)}{C^2(x)} =$$

$$= B'(x)\ln\frac{B(x)}{C(x)} + B'(x) - \frac{C'(x)B(x)}{C(x)} \qquad (10)$$

But

$$C'(x) = \frac{\partial B(x)}{\partial x_i}(\tilde{r}_i - r_{i1}) > 0 ,$$

$$B'(x) = \frac{\partial B(x)}{\partial x_i} = \tilde{r}_i ,$$

and substituting in (10) obtain

$$\frac{\partial}{\partial x_i}\left[B(x)\ln\frac{B(x)}{C(x)}\right] = \tilde{r}_i \ln\frac{B(x)}{C(x)} + \tilde{r}_i - (\tilde{r}_i - r_{i1})\frac{B(x)}{C(x)}. \qquad (11)$$

Find he second partial derivatives

$$\frac{\partial^2}{\partial x_i^2}\left[B(x)\ln\frac{B(x)}{C(x)}\right] = \tilde{r}_i \frac{C(x)}{B(x)}\frac{B'(x)C(x)-C'(x)B(x)}{C^2(x)} - (\tilde{r}_i - r_{i1})\frac{B'(x)C(x)-C'(x)B(x)}{C^2(x)} =$$

$$= \tilde{r}_i\left(\frac{B'(x)}{B(x)} - \frac{C'(x)}{C(x)}\right) - (\tilde{r}_i - r_{i1})\left[\frac{B'(x)}{C(x)} - \frac{C'(x)B(x)}{C^2(x)}\right] = \qquad (12)$$

$$= \tilde{\tilde{r}}_i \left[\frac{\tilde{\tilde{r}}_i}{\sum_{i=1}^{N} x_i \tilde{\tilde{r}}_i - r^*} - \frac{\tilde{\tilde{r}}_i - r_{i1}}{\sum_{i=1}^{N} x_i \tilde{\tilde{r}}_i - \sum_{i=1}^{N} x_i r_{i1}} \right] -$$

$$- (\tilde{\tilde{r}}_i - r_{i1}) \left[\frac{\tilde{\tilde{r}}_i}{\sum_{i=1}^{N} x_i \tilde{\tilde{r}}_i - \sum_{i=1}^{N} x_i r_{i1}} - \frac{(\sum_{i=1}^{N} x_i \tilde{\tilde{r}}_i - r^*)(\tilde{\tilde{r}}_i - r_{i1})}{(\sum_{i=1}^{N} x_i (\tilde{\tilde{r}}_i - r_{i1}^*))^2} \right] = \quad (13)$$

$$= \frac{\tilde{r}^2}{\sum_{i=1}^{N} x_i \tilde{\tilde{r}}_i - r^*} - \frac{2\tilde{r}(\tilde{\tilde{r}}_i - r_{i1})}{\sum_{i=1}^{N} x_i (\tilde{\tilde{r}}_i - r_{i1})} + \frac{(\tilde{\tilde{r}}_i - r_{i1})^2 (\sum_{i=1}^{N} x_i \tilde{\tilde{r}}_i - r^*)}{(\sum_{i=1}^{N} x_i (\tilde{\tilde{r}}_i - r_{i1}^*))^2} = \quad (14)$$

After transferring to common denominator we obtain

$$= \frac{\tilde{\tilde{r}}_i^2 (\sum_{i=1}^{N} x_i (\tilde{\tilde{r}}_i - r_{i1}^*))^2 - 2\tilde{\tilde{r}}_i (\tilde{\tilde{r}}_i - r_{i1}) (\sum_{i=1}^{N} x_i \tilde{\tilde{r}}_i - r^*) \sum_{i=1}^{N} x_i (\tilde{\tilde{r}}_i - r_{i1})}{(\sum_{i=1}^{N} x_i \tilde{\tilde{r}}_i - r^*)(\sum_{i=1}^{N} x_i (\tilde{\tilde{r}}_i - r_{i1}^*))^2} +$$

$$+ \frac{(\tilde{\tilde{r}}_i - r_{i1})^2 (\sum_{i=1}^{N} x_i \tilde{\tilde{r}}_i - r^*)^2}{(\sum_{i=1}^{N} x_i (\tilde{\tilde{r}}_i - r_{i1}^*))^2 (\sum_{i=1}^{N} x_i \tilde{\tilde{r}}_i - r^*)} =$$

$$= \frac{\left[\tilde{\tilde{r}}_i \left(\sum_{i=1}^{N} x_i (\tilde{\tilde{r}}_i - r_{i1}) - (\tilde{\tilde{r}}_i - r_{i1}) (\sum_{i=1}^{N} x_i \tilde{\tilde{r}}_i - r^*) \right) \right]^2}{(\sum_{i=1}^{N} x_i \tilde{\tilde{r}}_i - r^*)(\sum_{i=1}^{N} x_i (\tilde{\tilde{r}}_i - r_{i1}^*))^2} \geq 0 \quad (15)$$

As $\tilde{r}_i > (\tilde{r}_i - r_{i1})$ and $\sum_{i=1}^{N} x_i(\tilde{r}_i - r_{i1}) > \sum_{i=1}^{N} x_i \tilde{r}_i - r^*$, the expression (15) is strictly greater than 0. Thus all partial derivatives of the second order $\Delta_{ii} = \frac{\partial^2}{\partial x_i^2}\left[B(x) \ln \frac{B(x)}{C(x)} \right] > 0$ and respectively $\frac{\partial^2}{\partial x_i^2}\left[A(x) + B(x) \ln \frac{B(x)}{C(x)} \right] > 0.$

Now it's necessary to show that all the diagonal minors of the following form

$$\begin{bmatrix} \Delta_{ii} & \Delta_{ij} \\ \Delta_{ji} & \Delta_{jj} \end{bmatrix} = \Delta_{ii}\Delta_{jj} - \Delta_{ij}\Delta_{ji} = \Delta_{ii}\Delta_{jj} - \Delta_{ji}^{\ 2} \geq 0. \qquad (16)$$

These are sufficient convexity conditions of function $B(x) \ln \frac{B(x)}{C(x)}$, and therefore of the initial function $A(x) + B(x) \ln \frac{B(x)}{C(x)}$.

Find the mixed partial derivatives $\frac{\partial^2}{\partial x_i \partial x_i}\left[B(x) \ln \frac{B(x)}{C(x)} \right]$:

$$\frac{\partial^2}{\partial x_i \partial x_i}\left[B(x) \ln \frac{B(x)}{C(x)} \right] = \frac{\partial}{\partial x_j}\left(B'(x) \ln \frac{B(x)}{C(x)} + B'(x) - \frac{C'(x)B(x)}{C(x)} \right) =$$

$$= \tilde{r}_i \frac{C(x)}{B(x)} \frac{B'_j(x)C(x) - C'_j(x)B(x)}{C^2(x)} - (\tilde{r}_i - r_{i1}) \frac{B'_j(x)C(x) - C'_j(x)B(x)}{C^2(x)} =$$

$$= \tilde{r}_i \left(\frac{B'_j(x)}{B(x)} - \frac{C'_j(x)}{C(x)} \right) - (\tilde{r}_i - r_{i1}) \left[\frac{B'_j(x)}{B(x)} - \frac{C'_j(x)B(x)}{C^2(x)} \right] \qquad (17)$$

where $B'_j(x) = \frac{\partial}{\partial x_j} B(x) = \tilde{r}_j$; $C'_j(x) = \frac{\partial}{\partial x_j} C(x) = \tilde{r}_j - r_{j1}$ and substituting these assignments in (17) obtain

$$\tilde{r}_i\left(\frac{\tilde{r}_j}{\sum_{i=1}^{N}x_i\tilde{r}_i-r^*}-\frac{\tilde{r}_j-r_{j1}}{\sum_{i=1}^{N}x_i(\tilde{r}_i-r_{i1})}\right)-(\tilde{r}_i-r_{i1})\left[\frac{\tilde{r}_j}{\sum_{i=1}^{N}x_i(\tilde{r}_i-r_{i1})}-\frac{(\tilde{r}_j-r_{j1})\sum_{i=1}^{N}x_i\tilde{r}_i-r^*}{\sum_{i=1}^{N}x_i(\tilde{r}_i-r_{i1})^2}\right]=$$

$$=\frac{\tilde{r}_i\tilde{r}_j(\sum_{i=1}^{N}x_i(\tilde{r}_i-r_{i1})^2-\tilde{r}_i(\tilde{r}_j-r_{j1})(\sum_{i=1}^{N}x_i\tilde{r}_i-r^*)(\sum_{i=1}^{N}x_i(\tilde{r}_i-r_{i1})-\tilde{r}_j(\tilde{r}_i-r_{i1})}{(\sum_{i=1}^{N}x_i(\tilde{r}_i-r^*)(\sum_{i=1}^{N}x_i(\tilde{r}_i-r_{i1})^2)}-$$

$$\frac{\sum_{i=1}^{N}x_i(\tilde{r}_i-r_{i1})(\sum_{i=1}^{N}x_i\tilde{r}_i-r^*)+(\tilde{r}_i-r_{i1})(\tilde{r}_j-r_{j1})\sum_{i=1}^{N}x_i(\tilde{r}_i-r_{i1})^2}{(\sum_{i=1}^{N}x_i\tilde{r}_i-r^*)(\sum_{i=1}^{N}x_i(\tilde{r}_i-r_{i1})^2)}=$$

$$\frac{\tilde{r}_i\tilde{r}_j(\sum_{i=1}^{N}x_i(\tilde{r}_i-r_{i1})^2-(2\tilde{r}_i\tilde{r}_j-\tilde{r}_ir_{j1}-\tilde{r}_jr_{i1})(\sum_{i=1}^{N}x_i(\tilde{r}_i-r_{i1})(\sum_{i=1}^{N}x_i\tilde{r}_i-r^*)}{(\sum_{i=1}^{N}x_i\tilde{r}_i-r^*)(\sum_{i=1}^{N}x_i(\tilde{r}_i-r_{i1})^2)}+$$

$$+\frac{(\tilde{r}_i-r_{i1})(\tilde{r}_j-\tilde{r}_{j1})(\sum_{i=1}^{N}x_i\tilde{r}_i-r^*)^2}{(\sum_{i=1}^{N}x_i\tilde{r}_i-r^*)(\sum_{i=1}^{N}x_i(\tilde{r}_i-r_{i1})^2)}. \qquad (18)$$

For convenience and cut of transformations denote the denominator as

$$(\sum_{i=1}^{N}x_i\tilde{r}_i-r^*)\sum_{i=1}^{N}x_i(\tilde{r}_i-r_{i1})^2=E(x). \qquad (19)$$

Substituting the expressions for Δ_{ii} и Δ_{ij} from (15) and (18) into (16) and obtain

$$\Delta_{ii}\Delta_{jj} - \Delta_{ji}^{\ 2} = \frac{\left[\tilde{r}_i \sum_{i=1}^{N} x_i(\tilde{r}_i - r_{i1}) - (\tilde{r}_i - r_{i1})(\sum_{i=1}^{N} x_i\tilde{r}_i - r^*)\right]^2}{E^2(x)} \times$$

$$\times \left[\frac{\tilde{r}_j(\sum_{i=1}^{N} x_i(\tilde{r}_i - r_{i1}) - (\tilde{r}_j - r_{j1})(\sum_{i=1}^{N} x_i\tilde{r}_i - r^*))}{E^2(x)}\right]^2 -$$

$$- \frac{\{\tilde{r}_i\tilde{r}_j(\sum_{i=1}^{N} x_i(\tilde{r}_i - r_{i1})^2 - (2\tilde{r}_i\tilde{r}_j - \tilde{r}_i r_{j1} - \tilde{r}_j r_{i1})(\sum_{i=1}^{N} x_i(\tilde{r}_i - r_{i1})(\sum_{i=1}^{N} x_i\tilde{r}_i - r^*)}{E^2(x)} +$$

$$+ \frac{(\tilde{r}_i - r_{i1})(\tilde{r}_j - r_{j1})(\sum_{i=1}^{N} x_i\tilde{r}_i - r^*)\}^2}{E^2(x)}. \qquad (20)$$

For further simplicity assign

$$\sum_{i=1}^{N} x_i(\tilde{r}_i - r_{i1}) = \tilde{r} - r_{\min}\ ;\ \sum_{i=1}^{N} x_i\tilde{r}_i - r^* = \tilde{r} - r^*. \qquad (21)$$

Substituting them into (20) obtain

$$\Delta_{ii}\Delta_{jj} - \Delta_{ji}^{\ 2} = \frac{\left[\tilde{r}_i(\tilde{r} - r_{\min}) - (\tilde{r}_i - r_{i1})(\tilde{r} - r^*)\right]^2}{E^2(x)} \times$$

$$\times \left[\frac{\tilde{r}_j(\tilde{r} - r_{\min}) - (\tilde{r}_j - r_{j1})(\tilde{r} - r^*)}{E^2(x)}\right]^2 -$$

$$-\frac{\{\tilde{r}_i\tilde{r}_j(\tilde{r}-r_{\min})^2-(2\tilde{r}_i\tilde{r}_j-\tilde{r}_i r_{j1}-\tilde{r}_j r_{i1})(\tilde{r}-r_{\min})(\tilde{r}-r^*)}{E^2(x)}+ \tag{22}$$

$$+\frac{(\tilde{r}_i-r_{i1})(\tilde{r}_j-r_{j1})(\tilde{r}-r^*)\}^2}{E^2(x)}.$$

Further assign

$$\tilde{r}_i(\tilde{r}-r_{\min})-(\tilde{r}_i-r_{i1})(\tilde{r}-r^*)=F;\ \Delta_{ji}=\frac{H}{E},$$

$$\tilde{r}_j(\tilde{r}-r_{\min})-(\tilde{r}_j-r_{j1})(\tilde{r}-r^*)=G.$$

Substituting into (23), obtain

$$\Delta_{ii}\Delta_{jj}-\Delta_{ji}^2=\frac{F^2G^2-H^2}{E^2}=\frac{(FG-H)(FG+H)}{E^2}>0. \tag{23}$$

Non-negativeness condition (24) is following:

$$FG-H>0.$$

Thereof (hence)

$$FG-H=\left[\tilde{r}_i(\tilde{r}-r_{\min})-(\tilde{r}_i-r_{i1})(\tilde{r}-r^*)\right].\left[\tilde{r}_j(\tilde{r}-r_{\min})-(\tilde{r}_j-r_{j1})(\tilde{r}-r^*)\right]-$$

$$-\tilde{r}_i\tilde{r}_j(\tilde{r}-r_{\min})^2-(2\tilde{r}_i\tilde{r}_j-\tilde{r}_i r_{j1}-\tilde{r}_j r_{i1})(\tilde{r}-r_{\min})(\tilde{r}-r^*)-(\tilde{r}_i-r_{i1})(\tilde{r}_j-r_{j1})(\tilde{r}-r^*)^2=$$

$$=\tilde{r}_i\tilde{r}_j(\tilde{r}-r_{\min})^2-\tilde{r}_j(\tilde{r}_i-r_{i1})(\tilde{r}-r^*)(\tilde{r}-r_{\min})-\tilde{r}_i(\tilde{r}-r_{\min})(\tilde{r}_j-r_{j1})(\tilde{r}-r^*)+$$

$$+(\tilde{r}_i-r_{i1})(\tilde{r}_j-r_{j1})(\tilde{r}-r^*)-\tilde{r}_i\tilde{r}_j(\tilde{r}-r_{\min})^2+(2\tilde{r}_i\tilde{r}_j-\tilde{r}_i r_{j1}-\tilde{r}_j r_{i1})(\tilde{r}-r_{\min})(\tilde{r}-r^*)-$$

$$-(\tilde{r}_i-r_{i1})(\tilde{r}_j-r_{j1})(\tilde{r}-r^*)^2=$$

$$= (\tilde{r} - r_{\min})(\tilde{r} - r^*)\left[-\tilde{r}_j(\tilde{r}_i - r_{i1}) - \tilde{r}_i(\tilde{r}_j - r_{j1}) - 2\tilde{r}_i\tilde{r}_j - \tilde{r}_i r_{j1} - \tilde{r}_j r_{i1}\right] = 0 \quad (24)$$

Thus we have obtained the following conditions

$\Delta_{ii} = \dfrac{\partial^2}{\partial x_t^2}\left[B(x)\ln\dfrac{B(x)}{C(x)}\right] > 0$; for all $i = \overline{1,N}$.

Diagonal minors of the form $\mu_{i1} = \begin{bmatrix} \Delta_{ii} & \Delta_{ij} \\ \Delta_{ji} & \Delta_{jj} \end{bmatrix} \geq 0.$

These are the sufficient conditions of convexity of function $B(x)\ln\dfrac{B(x)}{C(x)}$, and therefore the convexity of function

$A(x) + B(x)\ln\dfrac{B(x)}{C(x)}$.

Now it's only left to show that the product of convex functions $A(x) + B(x)\ln\dfrac{B(x)}{C(x)}$ и $D(x)$ will be convex as well at the interval $x_i \in [0,1]$, $i = \overline{1,N}$ taking into account that $D(x) = \dfrac{1}{\sum_{i=1}^{N} x_i(r_{i2} - r_{i1})}$

where $r_{i2} \geq r_{i1}$, $x_i \in [0,1]$, $\sum_{i=1}^{N} x_i = 1$.

Notice that $A(x) + B(x)\ln\dfrac{B(x)}{C(x)}$ и $D(x)$ as proved previously are positive and monotonically decreasing.

For convenience denote $A(x) + B(x)\ln\dfrac{B(x)}{C(x)} = \varphi(x)$.

Let's prove that $\varphi'(x) < 0.$

$$\frac{\partial\varphi}{\partial x_i} = \frac{\partial}{\partial x_i}\left(A(x) + B(x)\ln\frac{B(x)}{C(x)}\right) = A'(x) + B'(x) + B(x)\frac{C(x)}{B(x)}\frac{B'(x)C(x) - C'(x)B(x)}{C^2(x)} =$$

$$= A'(x) + B'(x)\ln\frac{B(x)}{C(x)} + B'(x) - \frac{B'(x)C(x)}{C(x)}. \quad (25)$$

Substituting the values $A'(x)$ and $B'(x)$ obtain

$$\frac{\partial \varphi}{\partial x_i} = -r_{i1} + \tilde{r}_i \ln\frac{B(x)}{C(x)} + \tilde{r}_i - (\tilde{r}_i - r_{i1})\frac{B(x)}{C(x)} =$$

$$= \tilde{r}_i\left(1 + \ln\frac{B(x)}{C(x)}) - r_{i1} - (\tilde{r}_i - r_{i1})\frac{B(x)}{C(x)}\right). \quad (26)$$

As $\frac{B(x)}{C(x)} < 1,\ -r_{i1} + \tilde{r}_{i1}\frac{B(x)}{C(x)} < 0.$

Therefore, after simplifying (27), we obtain

$$\frac{\partial \varphi}{\partial x_i} = \tilde{r}_i(1 + \ln\frac{B(x)}{C(x)} - \frac{B(x)}{C(x)}), \quad (27)$$

$$1 + \ln\frac{B(x)}{C(x)} - \frac{B(x)}{C(x)} = 1 + \ln\frac{\tilde{r} - r^*}{\tilde{r} - r_{\min}} - \frac{\tilde{r} - r^*}{\tilde{r} - r_{\min}}, \quad (28)$$

According previous assumptions $r^* > r_{\min} = \sum_{i=1}^{N} x_i r_{i1}$ and $\tilde{r} > r^*$. Lets show that (28) is greater than 0.

Denote $\tilde{r} - r^* = a$. Then $\tilde{r} - r_{\min} = \tilde{r} - r^* + (r^* - r_{\min}) = a + y$, $y = r^* - r_{\min} > 0$.

Then

$$1 + \ln\frac{\tilde{r} - r^*}{\tilde{r} - r_{\min}} - \frac{\tilde{r} - r^*}{\tilde{r} - r_{\min}} = 1 + \ln\frac{a}{a + y} - \frac{a}{a + y}. \quad (29)$$

Let's show that

$$\Delta = 1 + \ln\frac{a}{a+y} - \frac{a}{a+y} < 0 \text{ , for all y>0.}$$

Evidently, $\Delta = 1 + \ln\frac{a}{a+y} - \frac{a}{a+y} = 0$ by y=0 and function is motonnically decreasing, as

$$\Delta'(y) = -\frac{1}{a+y} + \frac{a}{(a+y)^2} = -\frac{y}{(a+y)^2} < 0 \text{ for all y>0.}$$

Therefore $\Delta(y) < 0$ under y>0. Previously we have proved that $D(x) = \frac{1}{\sum_{i=1}^{N} x_i (r_{i2} - r_{i1})}$ is convex.

Consider

$$\frac{\partial}{\partial x_i}\left(\varphi(x)D(x)\right) = \varphi'(x)D(x) + D'(x)\varphi(x) < 0. \tag{30}$$

Find second partial derivatives

$$\frac{\partial^2}{\partial x_i^2}\left[\varphi(x)D(x)\right] = \varphi''(x)D(x) + \varphi'(x)D'(x) + D''(x)\varphi(x) + D'(x)\varphi'(x) =$$

$$= \varphi''(x)D(x) + \varphi(x)D''(x) + 2D'(x)\varphi'(x). \tag{31}$$

It was proved earlier that $\varphi'(x) < 0$; $D'(x) < 0$; $D''(x) > 0$; $\varphi''(x) > 0$ therefore

$$\frac{\partial^2}{\partial x_i^2}(\varphi(x)D(x)) > 0.$$

Therefore, risk function $\beta(x) = \varphi(x)D(x)$ is convex. End of proof.

Optimality Conditions for Dual Fuzzy Portfolio Problem

As it was earlier shown the dual portfolio problem (7)-(8) is convex programming problem under the corresponding conditions. Taking into account that constraints (8) are linear compose Lagrangian function

$$L(x,\lambda,\mu)=\beta(x)+\lambda(r^*-\sum_{i=1}^{N}x_i\tilde{r}_i)+\mu(\sum_{i=1}^{N}x_i-1). \tag{32}$$

The optimality conditions by Kuhn-Tucker are such:

$$\frac{\partial L}{\partial x_i}=\frac{\partial\beta(x)}{\partial x_i}-\lambda r_i+\mu\geq 0\,;i=\overline{1,N}, \tag{33}$$

$$\frac{\partial L}{\partial\lambda}=-\sum_{i=1}^{N}x_i\tilde{r}_i+r^*\leq 0,$$

$$\frac{\partial L}{\partial\mu}=\sum_{i=1}^{N}x_i-1=0,$$

and conditions of complementary non-fixedness

$$\frac{\partial L}{\partial x_i}x_i=0\,,\ i=\overline{1,N},$$

$$\frac{\partial L}{\partial\lambda}\lambda=\lambda\left(-\sum_{i=1}^{N}x_i\tilde{r}_i+r^*\right)=0, \tag{34}$$

$$x_i\geq 0\,,\ x\geq 0,$$

where $\lambda\geq 0$ и μ are Lagrange multipliers.

This problem may be solved using standard methods of convex programming, for instance by the method of feasible directions or penalty functions method.

Conclusion

The dual fuzzy portfolio problem is considered. The sufficient conditions for this problem to be convex were obtained and investigated. The optimality conditions for the solution were obtained.

Ack nowledgement

The paper is partially financed by the project ITHEA XXI of the Institute of Information Theories and Applications FOI ITHEA and the Consortium FOI Bulgaria. www.ithea.org, www.foibg.com.

Bibliography

[Недосекин, 2003] Недосекин А.О. Система оптимизации фондового портфеля от Сименс Бизнес Сервисез / Банковские технологии. – 2003. – № 5. – Также на сайте: http: //www. finansy. ru/publ/fin/004. Htm

[Зайченко, Малихех, 2008] Юрий Зайченко Юрий, Малихех Есфандиярфард. Оптимизация инвестиционного портфеля в условиях неопределенности на основе прогнозирования курсов акций // Proceedings of XIV-th International Conference “KDS-2008” (Knowledge, Dialogue, Solution). June, 2008, Varna, Bulgaria.-Sophia. - Pp. 212-228.

[Зайченко, 2007] Зайченко Юрий, Малихех Есфандиярфард. Анализ и сравнение результатов оптимизации инвестиционного портфеля при применении модели Марковитца и нечетко-множественного метода // Proceedings of XIII-th International Conference “KDS-2007” (Knolwledge, Dialogue, Solution), Vol.1 , pp.278-286.

[Зайченко, 2008] Зайченко Ю.П., Есфандиярфард М. Оптимизация инвестиционного портфеля в условиях неопределенности.// Системні дослідження та інвестиційні технології. - 2008. - №2. - С.59-76.

Authors' Information

Zaychenko Yuriy P., Dr. of sci., professor, Institute for Applied system analysis NTUU “KPI”, Kiev

e-mail : baskervil@voliacable.com

Maliheh Esfandiaryfard (Iran), post-graduate student , NTUU”KPI”, Kiev, prospect Pobedy,37

e-mail: fard_sem@yahoo.com

TABLE OF CONTENTS OF VOLUME 3, NUMBER 2