ITHEA

International Journal INFORMATION THEORIES & APPLICATIONS

2007 Volume 14 Number 2

International Journal
INFORMATION THEORIES & APPLICATIONS
Volume 14 / 2007, Number 2

IJ ITA is official publisher of the scientific papers of the members of the Association of Developers and Users of Intellectualized Systems (ADUIS).

IJ ITA welcomes scientific papers connected with any information theory or its application.
IJ ITA rules for preparing the manuscripts are compulsory.
The **rules for the papers** for IJ ITA as well as the **subscription fees** are given on *www.foibg.com/ijita*.
The camera-ready copy of the paper should be received by e-mail: *info@foibg.com* .
Responsibility for papers published in IJ ITA belongs to authors.

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

International Journal "INFORMATION THEORIES & APPLICATIONS" Vol.14, Number 2, 2007

Printed in Bulgaria

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: **Institute of Information Theories and Applications FOI ITHEA**
Sofia, 1000, P.O.B. 775, Bulgaria. www.foibg.com, e-mail: info@foibg.com

ISSN 1310-0513 (printed) **ISSN 1313-0463 (online)** **ISSN 1313-0498 (CD/DVD)**

LOW-POWER TRACKING IMAGE SENSOR BASED ON BIOLOGICAL MODELS OF ATTENTION

Alexander Fish, Liby Sudakov-Boreysha, Orly Yadid-Pecht

Abstract: This paper presents implementation of a low-power tracking CMOS image sensor based on biological models of attention. The presented imager allows tracking of up to N salient targets in the field of view. Employing "smart" image sensor architecture, where all image processing is implemented on the sensor focal plane, the proposed imager allows reduction of the amount of data transmitted from the sensor array to external processing units and thus provides real time operation. The imager operation and architecture are based on the models taken from biological systems, where data sensed by many millions of receptors should be transmitted and processed in real time. The imager architecture is optimized to achieve low-power dissipation both in acquisition and tracking modes of operation. The tracking concept is presented, the system architecture is shown and the circuits description is discussed.

Keywords: Low-power image sensors, image processing, tracking imager, models of attention, CMOS sensors

ACM Classification Keywords: B.7.0 Integrated circuits: General, I.4.8 Image processing and computer vision: scene analysis: tracking

1. Introduction

Real time visual tracking of salient targets in the field of view (FOV) is a very important operation in machine vision, star tracking and navigation applications. To accomplish real time operation a large amount of information is to be processed in parallel. This parallel processing is a very complicated task that demands huge computation resources. The same problem exists in biological vision systems. Compared to the state-of-the-art artificial imaging systems, having about twenty millions sensors, the human eye has more than one hundred million receptors (rods and cones). Thus, the question is how biological vision systems succeed to transmit and to process such a large amount of information in real time? The answer is that to cope with potential overload, the brain is equipped with a variety of attentional mechanisms [1]. These mechanisms have two important functions: (a) attention can be used to select relevant information and/or to ignore the irrelevant or interfering information; (b) attention can modulate or enhance the selected information according to the state and goals of the perceiver. Most models of attention mechanisms are based on the fact that a serial selection of regions of interest and their subsequent processing can greatly facilitate the computation complexity. Numerous research efforts in physiology were triggered during the last five decades to understand the attention mechanism [2]-[10]. Generally, works related to physiological analysis of the human attention system can be divided into two main groups: those that present a spatial (spotlight) model for visual attention [2]-[4] and those following object-based attention [5]-[10]. The main difference between these models is that the object-based theory is based on the assumption that attention is referenced to a target or perceptual groups in the visual field, while the spotlight theory indicates that attention selects a place at which to enhance the efficiency of information processing.

The design of efficient real time tracking systems mostly depends on deep understanding of the model of visual attention. Thus, a discipline, named neuromorphic VLSI that imitates the processing architectures found in biological systems as closely as possible was introduced [11]. Both spotlight and object-based models have been recently implemented in analog neuromorphic VLSI design [12]-[23]. Most of them are based on the theory of selective shifts of attention which arises from a saliency map, as was first introduced by Koch and Ullman [12]. Object-based selective attention systems VLSI implementations in 1-D and lately 2-D were presented by Morris et al [13]-[16]. An additional work on an analog VLSI based attentional search/tracking was presented by Horiuchi and Niebur in 1999 [17].

Many works on neuromorphic VLSI implementations of selective attention systems have been presented by Indiveri [19]-[21] and others [22]-[23]. In 1998 Brajovic and Kanade presented a computational sensor for visual tracking with attention. These works often use winner-take-all (WTA) [24] networks that are responsible for selection and tracking inputs with the strongest amplitude. This sequential search method is equivalent to the spotlight attention found in biological systems.

Most previously presented neuromorphic imagers utilize image processing implemented on the focal plane level and employ photodiode or phototransistor current-mode pixels. Typically, each pixel consists of a photo detector and local circuitry, performing spatio-temporal computations on the analog signal. These computations are fully parallel and distributed, since the information is processed according to the locally sensed signals and data from pixel neighbors. This concept allows reduction in the computational cost of the next processing stages placed in the interface. Unfortunately, when image quality and high spatial resolution are important, image processing should be performed in the periphery. This way a high fill factor (FF) can be achieved even in small pixels.

This paper presents implementation of a low-power tracking CMOS image sensor based on a spotlight model of attention. The presented imager allows tracking of up to N salient targets in the field of view. Employing image processing at the sensor focal plane, the proposed sensor allows parallel computations and is distributed, but on the other hand most of the image processing is performed in the array periphery, allowing image quality and high spatial resolution. The imager architecture is optimized to achieve low-power dissipation both in acquisition and tracking modes of operation. This paper is a continuation of the work presented in [25], where we proposed to employ a spotlight model of attention for the bottleneck problem reduction in high resolution "smart" CMOS image sensors and of the work presented in [26], where the basic concept for an efficient VLSI tracking sensor was presented.

Section 2 briefly describes spotlight and object-based models of attention and presents system architecture of the proposed sensor. Low-power considerations, as well imager circuits description are shown in Section 3. Section 4 discusses advantages and limitations of the proposed system. Conclusions and future work are presented in Section 5.

2. Tracking Sensor Architecture

The proposed tracking sensor operation is based on the imitation of the spotlight model of visual attention. Because this paper presents concepts taken from different research disciplines, first, a brief description of existing models of attention is presented for the readers that are not familiar with this field. Then, the proposed sensors architecture is shown.

2.1 Existing Attention Models

Much research was done in attention during the last decades and numerous models have been proposed over the years. However, there is still much confusion as to the nature and role of attention. All presented models of attention can be divided to two main groups: spatial (spotlight) or early attention and object-based, or late attention. While the object-based theory suggests that the visual world is parsed into objects or perceptual groups, the spatial (spotlight) model purports that attention is directed to unparsed regions of space. Experimental research provides some degree of support to both models of attention. While both models are useful in understanding the processing of visual information, the spotlight model suffers from more drawbacks than the object-based model. However, the spotlight model is simpler and can be more useful for tracking imager implementations, as will be shown below.

2.1.1 The Spatial (Spotlight) Model

The model of spotlight visual attention mainly grew out of the application of information theory developed by Shannon. In electronic systems, similar to physiological, the amount of the incoming information is limited by the system resources. There are two main models of spotlight attention. The simplest model can be looked upon as a spatial filter, where what falls outside the attentional spotlight is assumed not to be processed. In the second model, the spotlight serves to concentrate attentional resources to a particular region in space, thus enhancing processing at that location and almost eliminating processing of the unattended regions. The main difference between these models is that in the first one the spotlight only passively blocks the irrelevant information, while in the second model it actively directs the "processing efforts" to the chosen region.

Figure 1(a) and Figure 1(b) visually clarify the difference between the spatial filtering and spotlight attention.

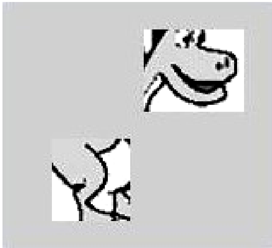


Figure 1 (a). An example of spatial filtering

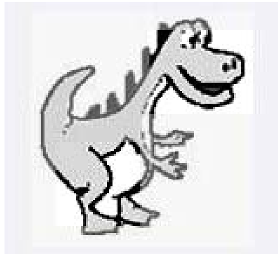


Figure 1 (b). An example of spotlight model of attention

A conventional view of the spotlight model assumes that only a single region of interest is processed at a certain time point and supposes smooth movement to other regions of interest. Later versions of the spotlight model assume that the attentional spotlight can be divided between several regions in space. In addition, the latter support the theory that the spotlight moves discretely from one region to the other.

2.1.2 Object-based Model

As reviewed above, the spotlight metaphor is useful for understanding how attention is deployed across space. However, this metaphor has serious limitations. A detailed analysis of the spotlight model drawbacks can be found in [1]. An object-based attention model suits more practical experiments in physiology and is based on the assumption that attention is referred to discrete objects in the visual field. However being more practical, in contrast to the spotlight model, where one would predict that two nearby or overlapping objects are attended as a single object, in the object-based model this divided attention between objects results in less efficient processing than attending to a single object. It should be noted that spotlight and object-based attention theories are not contradictory but rather complementary. Nevertheless, in many cases the object-based theory explains many phenomena better than the spotlight model does.

The object-based model is more complicated for implementation, since it requires objects' recognition, while the spotlight model only requires identifying the regions of interest, where the attentional resources will be concentrated for further processing.

2.2 System Architecture

The proposed sensor has two modes of operation: target acquisition and target tracking. In the acquisition mode *N* most salient targets of interest in the FOV are found. Then, *N* windows of interest with programmable size around the targets are defined. These windows define the active regions, where the subsequent processing will occur, similar to the flexible spotlight size in the biological systems. In the tracking mode, the system sequentially attends only to the previously chosen regions, while completely inhibiting the dataflow from the other regions.

The proposed concept permits choosing the attended regions in the desired order, independent on the targets saliency. In addition it allows shifting the attention from one active region to the other, independent of the distance between the targets. The proposed sensor aims to output the coordinates of all tracking targets in real time. Similar to biological systems, which are limited in their computational resources, the engineering applications are constrained with low-power dissipation. Thus, maximum efforts have been done to reduce power consumption in the proposed sensor. This power reduction is based on the general idea of "no movement – no action", meaning that minimum power should be dissipated if no change in the targets position occurred.

Figure 2 shows the architecture of the proposed tracking sensor. The sensor includes (a) A Pixel array with a two dimensional resistive network , (b) Y-Addressing and X-Addressing circuitry consisting of N digital registers for target windows definition and for image readout control, (c) an analog front end (AFE) for image readout, (d) A current Looser-take-all (LTA) circuit for target detection during the acquisition mode, (e) two analog 1-D center of mass (COM) computation circuits for X and Y COM coordinates update (each consists of 1-D analog current mode Winner-take-all (WTA) circuit), (f) Analog memory for temporal storage of loser values of all rows during the acquisition mode and digital memory for targets coordinates storage, (g) acquisition mode control and target mode control blocks.

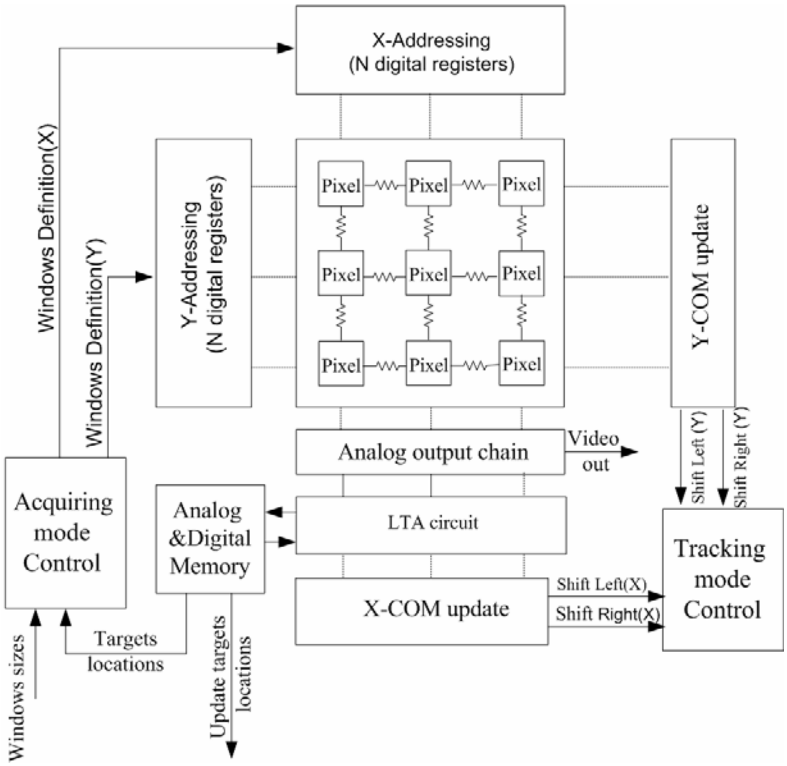


Figure 2. Architecture of the proposed tracking sensor.

In the acquisition mode the sensor finds the *N* most salient targets in the FOV and calculates their centroid coordinates. This is achieved in the following way: all neighboring pixels are connected by resistors (implemented by transistors), creating the resistive network. The pixel voltage becomes smaller if it is more exposed, and the local minimum of the voltage distribution can be regarded as the centroid of the target, corresponding to the exposed area. This minimum is found using an analog looser-take-all circuit (LTA). At the first stage of the acquisition mode, all pixels of the whole image are activated. The global minimum, corresponding to the brightest target is located using a one dimensional LTA circuit. To achieve this purpose, the whole image is scanned row by row (using one of the digital shift registers in the Y-addressing circuitry), finding the local minimum in each row. Then, the row local minima are input to the same LTA circuit again and the global minimum is computed. A more detailed description of this concept can be found in [27], where the two dimensional WTA computation was performed using two 1-D WTA circuits. Once the first brightest target is found, the system defines a small size programmable window, with the center located at the target centroid coordinates. The size of this window is predefined by the user before the acquisition mode starts and depends on the target size. While finding the second bright target in the FOV, all pixels of the first window, consisting of the brightest target found during the first search, are deactivated. This way, the bright pixels of the first target do not influence the result of the second search. The remains *N-1* targets are found in the same way. As a result, at the end of the acquisition mode all centroid coordinates of the *N* most salient targets in the FOV are stored in the memory and *N* small windows around these coordinates are defined. The window definition is performed using two digital shift registers. Thus, 2N shift registers are required to define N different windows. The acquisition mode control block is responsible for defining and positioning these active windows. Note, that the acquisition mode is very inefficient in terms of power dissipation because the whole sensor array is activated and the LTA operation and windows definition are power inefficient operations. On the other hand, the acquisition is a very rare operation and its time can be neglected in comparison with the tracking period.

Once the sensor has acquired *N* salient targets, the tracking mode is initiated. The predefined windows serve as a spotlight in biological systems, such that only the regions inside the windows are processed. Opposite to biological systems, these "spotlights" attend only to the regions predefined in the acquisition mode. Thus, even if new more salient objects appear during the tracking, the attention to the chosen regions is not influenced.

Because the sensor is in the tracking mode most of the time, it is very important to achieve very low-power dissipation in this mode. In the proposed system this is achieved in the following ways:

1. Only pixels of active windows and the circuitry responsible for proper centroid detection and pixels readout are active. The remaining circuits (including most pixels of the array) are disconnected from the power supply.
2. All shift registers in Y-addressing and X-addressing circuitries are optimized for low frequencies operation by leakage currents reduction.
3. During the tracking mode the sensor doesn't calculate new centroid coordinates. A simple analog circuit (COM update block in Figure 2) checks if the new centroid location differs from the centroid location of the previous frame. In the case that no difference was found, the circuit does not need to perform any action, significantly reducing system power dissipation. This principle suits the general idea of "no movement – no action". If the target changes its position, the "shift left" or "shift right" (both for x and y) signals are produced by the COM update blocks. These signals are input to the tracking mode control block and the appropriate shift register performs movement to the right direction, correcting the location of the window.
4. Each active window definition is performed using two shift registers. This windows definition method allows switching from one target of interest to another without any need in accessing the memory and loading the new target coordinates. This way the switching time between different objects does not depend on the distance between the targets and sensor power dissipation is reduced.

3. Circuits Description

In this Section we present some of the most important circuits, utilized by the sensor. This includes current mode LTA circuit, current mode WTA circuit, X-COM and Y-COM update circuits and ultra low-power shift registers. The pixel is implemented as a standard global-shutter active pixel sensor (APS) [28] with current mode readout instead of a conventional source follower amplifier inside the pixel. This current mode readout allows parallel read out and summarization of currents from all pixels in the active window at the same time. These summarized currents then are used for further processing by the X-COM and Y-COM update circuits, as described below. In addition, a conventional video data readout is available.

3.1 Current Mode Loser-Take-All (LTA) circuit

As previously mentioned, LTA circuit is responsible for targets detection and their COM computation during the acquisition mode. Since the COM computation is done in a serial manner and should be performed accurately, high speed and high precision LTA circuit is required. In addition, current mode LTA is required (pixels outputs are currents). Most of the previously presented LTA solutions are not suitable for the proposed sensor design. As a result, we use a LTA circuit that we have recently developed to achieve high speed and high precision [29]. Figure 3 shows cells 1 and k (out of the N interacting cells) of the LTA circuit. Cell k receives a unidirectional input current, I_{ink}, and produces an output voltage V_{outk}. This output has a low digital value if the input current I_k is identified as loser, and high, otherwise. The circuit applies two feedbacks to enhance the speed and precision: the excitatory feedback ΔI_k and inhibitory feedback ΔI_{avgk}. The basic operation of the LTA is based on input currents average computation and comparison of the input current of each cell to that average. The local excitatory feedback works to increase the compared average value in each cell, allowing that cell to be a loser. Oppositely, the inhibitory feedback works globally by reduction of the input currents average value and thus allowing inhibition of non-losing cells. A more detailed description of the circuit operation is provided below. The LTA circuit operates as follows: the drains of M_1 transistors of all N cells of the array are connected to the drains of M_4 transistors by a single common wire with voltage Vcom. The circuit starts the competition by applying Rst='1' for a short period of time. This way the excitatory feedback ΔIin and the inhibitory feedback $\Delta Iavg_k$ are cancelled. Assuming that all N cells in the array are identical and Rst='1' is applied, $\Delta Iavg_k$ = 0A and the current I'avg, through $M1_k$, is equal to the average of all input currents of the array, neglecting small deviations in the referenced input currents. I'avg is copied to $M3_k$ by the NMOS current mirror ($M1_k$ and $M3_k$) and is compared with the input current Iin_k copied by the PMOS current mirror ($M2_k$ and $M5_k$). If Iin_k=I'avg then $Vout_k$=VDD/2, assuming the same drivability factor K of $M3_k$ and $M5_k$ transistors.

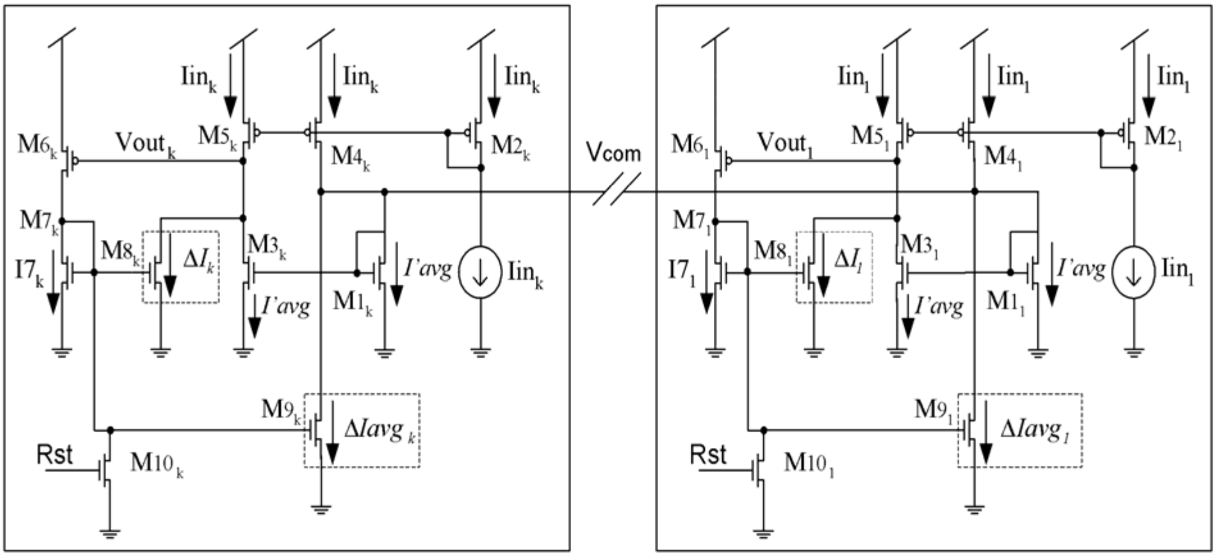


Figure 3. Cells 1 and k (out of N) of the LTA circuit.

An increase in input current Iin_k relatively to I'avg causes an increase in $Vout_k$ due to the Early effect. This way, during the reset phase, input currents of all cells are compared to the average of all input currents of the array, producing a unique output $Vout_k$ for every cell. The cell having the smallest input current value produces the smallest $Vout_k$ voltage. With the completion of the reset phase, i.e. Rst='0', the excitatory feedback ΔI_k and the inhibitory feedback $\Delta Iavg_k$ are produced. The $Vout_k$ node inputs to the gate of $M6_k$ PMOS transistor, thus the cell with the smaller Vx_k (smaller input current) produces a higher current $I7_k$ through $M6_k$ and $M7_k$. This current is copied by the NMOS current mirror ($M7_k$ and $M8_k$), creating the excitatory feedback ΔI_k. On the other hand, $I7_k$ is copied by the NMOS current mirror ($M7_k$ and $M9_k$), resulting in inhibitory feedback $\Delta Iavg_k$. ΔIk is added to the I'avg flowing through $M3_k$ and $\Delta Iavg_k$ is subtracted from the average of all input current by connection $M9_k$ transistor to the COM node, decreasing the I'avg value. This way, every cell produces a new Voutk voltage value, according to the comparison between the input current Iin_k and a sum of a current produced by the excitatory feedback ΔI_k and a new value of current I'avg, that is now given by:

$$I'avg = \frac{\sum_{k=1}^{N} Iin_k - \sum_{k=1}^{N} \Delta Iavg_k}{N} = Iavg - \frac{\sum_{k=1}^{N} \Delta Iavg_k}{N} \quad (1)$$

where Iavg is the average of all input currents of the array and N is the number of array cells. For the cell, having the smallest input current, the difference between Iin_k and a sum of ΔI_k and I'avg grows, thus decreasing $Vout_k$ value.

The computation phase is finished after one cell only is identified as a loser, producing $Vout_k$='0'. All other cells are identified as winners with $Vout_k$='1'. In this steady-state the excitatory and inhibitory feedbacks of the all winner cells and I'avg are approximately equal to zero, while $\Delta Iavg_k$ of the loser cell is approximately equal to the sum of all input currents. This way the circuit states stable preventing the selection of other potential losers unless the next reset is applied and a new computation starts. A more detailed description on the circuit operation can be found in [29].

To examine the presented LTA circuit it was designed, simulated and fabricated in 0.35um, 3.3V, n-well, 4-metal, CMOS, TSMC technology process supported by MOSIS. Table 1 summarizes the main characteristics of the circuit. As can be seen, the circuit achieves both high precision and high speed.

Parameter	Typical value	Worst case value (if exists)
Range of input currents	4 – 25 [µA]	------
Voltage supply	1.8V	------
Power Dissipation	58uW per cell	75 µW per cell
Delay	5nsec	95nsec
Precision	0.1 µA	0.5 µA
Occupied area (per cell)	26µm*22µm	------

Table 1: The main characteristics of the designed circuit.

3.2 X-COM and Y-COM update circuits

As mentioned, during the tracking mode the sensor doesn't calculate the new centroid coordinates. Instead, very simple X-COM and Y-COM update circuits (see Figure 4) check if the new X or Y centroid locations (respectively) differ from the centroid locations of the previous frame. In the case that no difference was found, the sensor does not perform any action, significantly reducing system power dissipation.

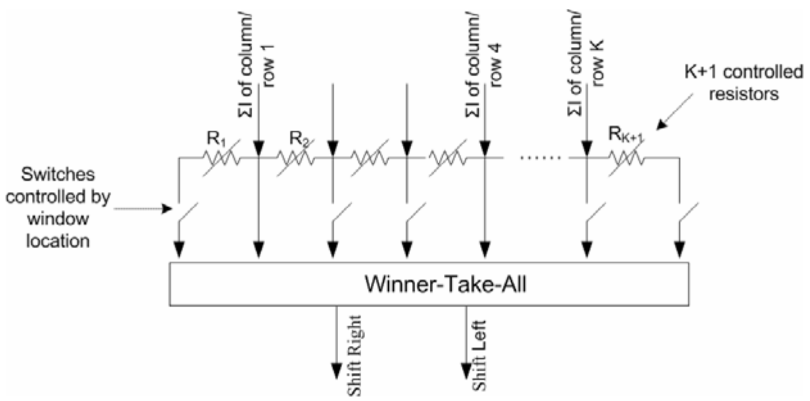


Figure 4. The X-COM and Y-COM update circuit implementation

The X-COM and Y-COM circuits have the same implementation, consisting of (K+1) controlled resistors (implemented by transistors), (K+2) digital switches, controlled by window location and (K+2) size analog current mode WTA circuit for the array, having K columns/row, respectively. Each COM update circuit receives K unidirectional input currents from the sensor array. The input current I_{ink} to the X-COM update circuit is the sum of all output pixel currents of the column K, while from the input current I_{ink} to the Y-COM update circuit is the sum of all output pixel currents of the row K. During the tracking mode (the COM update circuits are activated only in this mode), only pixels inside the windows of interest are activated. Therefore, in this mode, the input current to the COM circuit I_{ink} represents the sum of all row/column K active window output currents, for the case where the row/column K falls into the window of interest. In case, where the row/column K falls out the window of interest, the value of I_{ink} is zero. All input currents input to the resistive network, consisting of (K+1) controlled resistors (can either have a normal resistance R or very high resistance R_{high}) and then routed to the WTA circuit by switches. Both switches and the resistors are controlled by the windows of interest locations. For the active window, spread between column *i* and column (*i*+c), the resistors 1 to (*i*-1) and (*i*+2) to (K+1) have very high resistance R_{high}, while the resistors *i* to (*i*+1) are set to have a normal resistance R. This way resistive network, consisting of R value resistors is created around the active window of interest. Only two switches (*i*-1) and (*i*+1) are set to be on, while the others are set to be off. As a result, the WTA circuit receive only two non-zero input currents, representing the X/Y COM coordinates of the target located inside the window of interest. In case these currents are equal (the WTA circuit does not succeed to find the winner), the COM coordinates are exactly in the center of the window and therefore window location update is not required. On the other hand, if (i-1) current is larger than (i+1), it means that the target has moved right relatively to its location in the previous frame and Shift Right output is activated. In case if (i-1) current is smaller than (i+1), the Shift Left output is activated.

3.3 Current Mode Winner-take-all (WTA) circuit

As mentioned, the X-COM and Y-COM update circuits utilize current mode WTA circuit. The implementation of the WTA circuit (see Figure 5) is very similar to the implementation of the LTA circuit, presented in sub-section 3.1 and detailed description of its operation can be found in [30]. Similarly to the LTA, the WTA circuit employs inhibitory and local excitatory feedbacks based on input currents average computation, enhancing precision and speed performance of the circuit. Local excitatory feedback provides a hysteretic mechanism that prevents the selection of other potential winners unless they are stronger than the selected one by a set hysteretic current. The WTA circuit can be useful for integration with circuits operating in the strong inversion region and supplying input currents of 3µA-50µA, as well as for subthreshold applications with inputs of 0nA-50nA. It achieves very high speed (32nsec for high currents of 3µA-50µA (measured) and 34nsec for subthreshold currents – (simulated)) in case when a very small difference between two input currents is applied (30nA for high currents and 1.8nA for subthreshold applications). These circuit performances are the direct result from very strong feedbacks applied in the circuit. The circuit ability to cope with wide range of input currents is very important for the COM update circuit implementation since the inputs to this circuit can have very wide range.

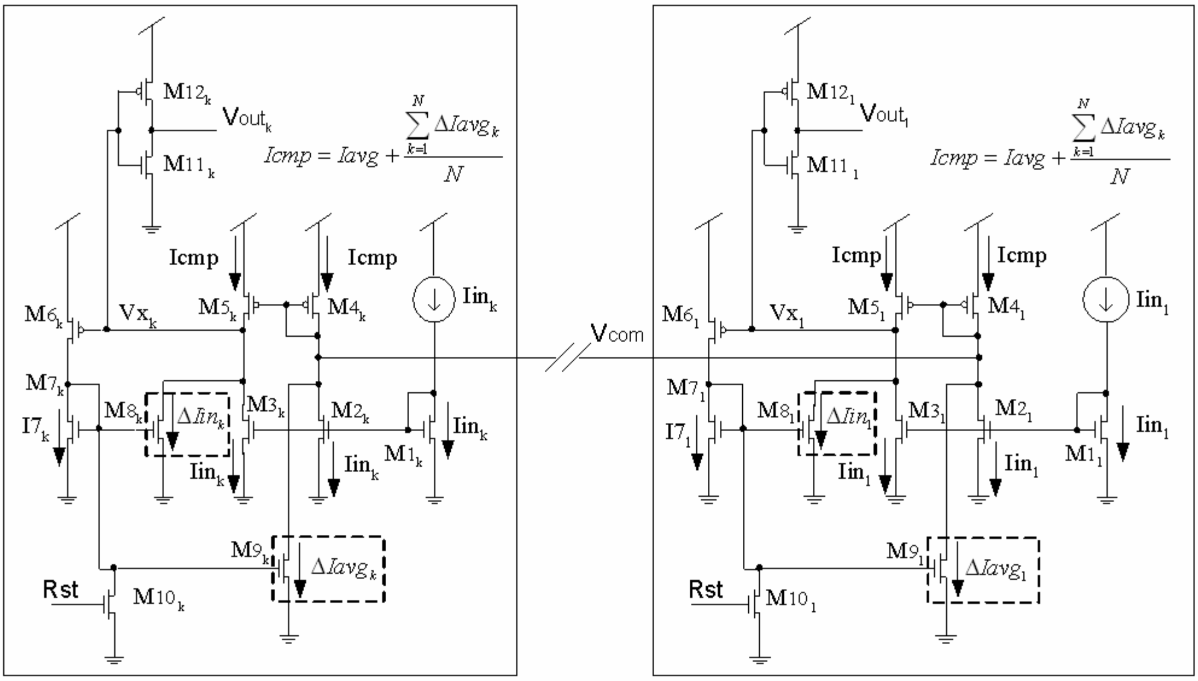


Figure 5. Cells 1 and k (out of n) of the WTA circuit

3.4 Y-addressing and X-Addressing circuitry implementation

Generally, the Y-Addressing and X-Addressing circuitry can be implemented using digital decoders. However, using shift registers for read out control reduces power dissipation and the number of global buses. In addition, windows of interest can be easily defined using shift-register. As previously mentioned, 2N shift registers are required to define N different windows. The architecture of shift register, used in the discussed tracking sensor (see Figure 6) , is based on the conventional simple shift register structure and utilizes low-power D-Flip-Flops (DFFs), described in details in [31].

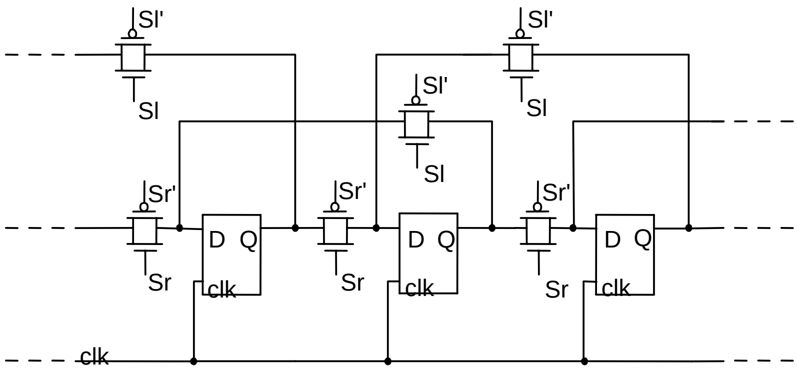


Figure 6. Architecture of the conventional shift-right and shift-left register

This register allows shifting of the vector of bits right or left – very important function in windows definition. The power dissipation of the shift register can be reduced by examining the nature of the inputs to the register. When the register is used for signal readout control, its input vector consists of a single '1' and of (N-1) digital zeros ('0'), assuming an N size register. Thus, in steady state, only one (out of N) DFF has '1' in its output. For the case, when the register is used for window definition, its input vector consists of K high digital bits and (N-K) low digital bits, assuming an N size register defines K*K size window. Usually, K<<N, resulting in the same solution for both cases. Figure 7 shows the DFF, optimized for these kinds of input vectors. This master-slave FF is constructed by cascading two different transmission gate latches: the first one is the dynamic latch and the second one is pseudo-static latch. Having only 15 transistors, this circuit is optimized for leakage current reduction for the case of '0' at the FF output. For this case all possible leakage currents in the circuit (signed by arrows in Figure 7) are reduced due to connection of two series connected "off" transistors. Note, that for the case of the shift register, having an input vector with almost all "high" bits, the presented FF can be optimized in a similar way, taking in account the high digital value in the FF output.

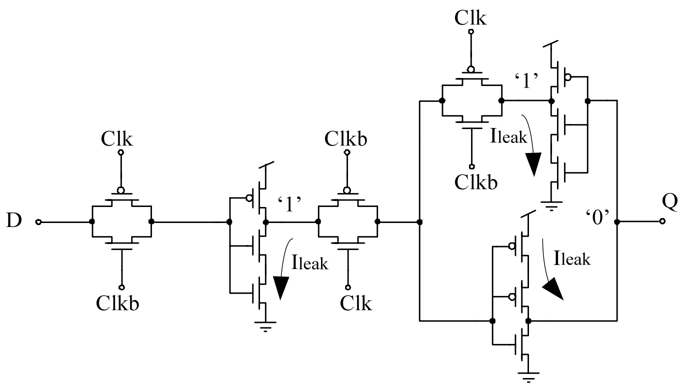


Figure 7. DFF, optimized for an input vector consisting of a large number of zeros

To check the suitability of the mentioned DFF circuit for the proposed tracking system, it has been designed and implemented in 0.18μm technology to compare the FF design with a set of representative flip-flops, commonly used for high performance design. In addition, the low leakage shift register is compared to the register, based on conventional FFs. Simulation results show up-to 10% power reduction in case of 10KHZ operation and up-to 60% reduction in case of 5 pixels size window definition at 30Hz frequency. More detailed description of the FF and SR performances can be found in [31].

4. Discussion

In this section we present the expected performance of the proposed tracking imager and briefly discuss advantages and limitations of the current architecture. Table 2 summarizes the expected characteristics of the proposed system. The sensor will be fabricated in a standard 0.18μm CMOS technology and will be operated using 1.8V supply voltage. The pixel size is expected to be 7 x 7μm and to achieve fill factor of at least 60%. At this stage the test chip will include a relatively small array of 64 x 64. The reason for the small array is failure probability reduction and limited budget. On the other hand, this array size still allows showing the proof of concept. At the first fabrication phase the system will able to track up to 3 salient targets of interest at 30 frames per second. In future designs we are working to increase the number of the tracked targets and to improve real time operation, allowing tracking at up to 100 frames per second. As mentioned, the proposed imager employs spatial filtering version of the spotlight models of attention, where what falls outside the attentional spotlight is assumed not to be processed. The drawback of this method is that during the tracking mode the sensor filters all information outside windows of interest, including potential targets that appear in the FOV during the tracking. In our future implementations we plan to upgrade the tracking sensor with the spotlight attention model, where the spotlight serves to concentrate attentional resources to a particular region in space, thus enhancing processing at that location and almost eliminating processing of the unattended regions (but still checking theses regions). An additional limitation is that the proposed system does not utilize Correlated Double Sampling (CDS) circuit to reduce Fixed Pattern Noise (FPN). CDS implementation in such kind of tracker is not trivial and thus it will be implemented only in the next version of the system to reduce failure probability at this stage. Finally, the system is expected to achieve very low-power dissipation of less than 2mW. The next generation of this tracking system will achieve power dissipation of less than 1mW.

Parameter	Expected value
Technology	0.18μm standard CMOS technology
Array size	64 x 64
Voltage supply	1.8V
Pixel Size	7 x 7 μm
Fill Factor	> 60%
No. of tracked targets	3
Real time operation	60 frames/second
Sensor read out method	Global Shutter
Utilized Attention model	Spatial Filtering
Bad Pixel Elimination	Yes
FPN Reduction	No
Power Dissipation	< 2mW

Table 2: The expected characteristics of the tracking system.

5. Conclusions

Implementation of low-power tracking CMOS image sensor based on biological models of attention was presented. Imager architecture and principle of operation were discussed, as well designs of the most important circuits, like Winner-Take-All, Looser-Takes-All, COM update and X/Y-Addressing circuits, utilized by the tracking system were shown. A brief description of the spatial and object-based models of attention was also presented. The expected system performance was discussed, showing advantages and drawbacks of the proposed sensor. The presented imager allows tracking of up to N salient targets in the field of view. The imager architecture is optimized to achieve low-power dissipation both in acquisition and tracking modes of operation. Further research includes improvement of the current sensor architecture and its realization in an advanced CMOS technology.

Acknowledgements

We would like to thank the Israeli Ministry of Science and Technology for funding this project.

Bibliography

[1] Chun, M. M., & Wolfe, J. M. Visual Attention. In E. B. Goldstein (Ed.), Blackwell's Handbook of Perception, Vol. Ch 9, pp. 272-310). Oxford, UK: Blackwell, 2001.

[2] R. W. Remington, L. Pierce. Moving attention: Evidence for Time-invariant shifts of visual selection attention. Perception and Psychophysics, vol. 35, pp. 393-399, 1984.

[3] J. Moran, R. Desimone. Selective attention gates visual processing in the Extrastriate Cortex. Science, vol. 229, pp. 784-787, 1985.

[4] G. Sperling, E. Weichselgartner. Episodic theory of the dynamics of spatial attention. Psychological review, vol. 102, no. 3, pp. 503-532, 1995.

[5] A. Treisman. Features and targets: The fourteenth Barlett memorial lecture. Quarterly Journal of Experimental Psychology, 40A, pp. 201-237, 1988.

[6] A. Treisman. Feature binding, attention and target perception. Philosophical Transactions of the Royal Society of London, vol. 353, pp. 1295-1306, 1998.

[7] A. Treisman, G. Gelade. A feature-integration theory of attention. Cognitive Psychology, vol. 12, pp. 97-136, 1980.

[8] A. Treisman, D. Kahneman, J. Burkell. Perceptual targets and the cost of filtering. Perception & Psychophysics, vol. 33, pp. 527-532, 1983.

[9] S. Yantis. Targets, attention, and perceptual experience in "Visual Attention". R. D. Wright (Eds.), pp. 187-214. Oxford, NY: Oxford University Press, 1998.

[10] M. I. Posner. Orienting of attention. Quarterly Journal of Experimental Psychology, vol. 32,pp. 3-25, 1980.

[11] C. Koch, B. Mathur. Neuromorphic vision chips. IEEE Spectrum, vol. 33, pp. 38–46, May 1996

[12] C. Koch and S. Ullman, "Shifts in Selective Visual Attention: Towards the Underlying Neural Circuitry," Human Neurobiology, vol. 4, pp. 219–227, 1985.

[13] T. G. Morris, T. K. Horiuchi, and P. DeWeerth, "Target-Based Selection Within an Analog Visual Attention System", IEEE Transactions on circuits and systems-11, Analog and Digital Signal processing, vol.45, no.12, December 1998.

[14] T. G. Moris and S. P. DeWeerth, "A Smart-Scanning Analog VLSI Visual-Attention System", Analog Integrated Circuits and Signal Processing, vol.21, p.67-78, 1999.

[15] T. G. Morris and S. P. DeWeerth, "Analog VLSI excitatory feedback circuits for attentional shifts and tracking." Analog Integrated Circuits and Signal Processing, 13(1/2), pp. 79–92, May/June 1997.

[16] C. S. Wilson, T. G. Morris, and P. DeWeerth, "A two-dimensional, target-based analog VLSI visual attention system", Twentieth Anniversary Conference on Advanced Research in VLSI, IEEE Computer Society Press: Los Alamitos, CA. Vol. 20. pp. 291-308. March 1999.

[17] T. Horiuchi and E. Niebur, "Conjunction search using a 1-D analog VLSI based attentional search/tracking chip," In Wills, S. (Ed.), Proceedings of the 1999 Conference in Advanced Research in VLSI., pp. 276-90 Georgia Tech, Atlanta, 1999.

[18] V. Brajovic and T. Kanade, "Computational sensor for visual tracking with attention", IEEE journal of Solid-State Circuits, Vol.33, No.8, August 1998.

[19] G. Indiveri, "Neuromorphic analog VLSI sensor for visual tracking: Circuits and application examples," IEEE Trans. on Circuits and Systems II, vol. 46, no. 11, pp. 1337–1347, November 1999.

[20] G. Indiveri, A. M. Whatley, and J. Kramer, "A reconfigurable neuromorphic VLSI multi-chip system applied to visual motion computation," in Proc. 7th Int. Conf. Microelectronics Neural, Fuzzy and Bio-Inspired Systems; Microneuro'99. Los Alamitos, CA: IEEE Computer Society Press, Apr. 1999, pp. 37–44.

[21] G. Indiveri, "A 2D neuromorphic VLSI architecture for modeling selective attention", Proc. IJCNN2000, 2000.

[22] R. Etienne-Cummings, J. Van der Spiegel, and P. Mueller,"A visual smoot pursuit tracking chip," Advances in Neural Information Processing Systems, D. S. Touretzky, M. C. Mozer, and Hasselmo M. E., Eds. 1996, vol. 8, MIT Press.

[23] C. Higgins and C. Koch, "A modular multi-chip neuromorphic architecture for real-time visual processing," J. Analog Integrated Circuits Signal Process, vol. 26, no. 3, pp. 195–211, Sept. 2000.

[24] J. Lazzaro., S. Ryckebusch, M. A. Mahowald, and C.A. Mead, "Winner-tale-all networks of O(n) complexity", ed. D.S.Touretzky, Morgan Kaumann, San Mateo, CA, 1:703-711, 1989.

[25] A. Fish and O. Yadid-Pecht, "Bottleneck Problem Solution using Biological Models of Attention in high resolution tracking sensors", to be appear International Journal on Information Theories and Applications.

[26] A. Fish, A. Spivakovsky, A. Golberg and O. Yadid-Pecht, "VLSI Sensor for multiple targets detection and tracking", IEEE ICECS, December 2004, CD ROM.

[27] A. Fish, D. Turchin, O. Yadid-Pecht, " An APS with 2-Dimensional winner-take-all selection employing adaptive spatial filtering and false alarm reduction", IEEE Trans. on Electron Devices, Special Issue on Image Sensors, Vol. 50, No. 1, pp. 159-165, January, 2003.

[28] O. Yadid-Pecht, R. Etienne-Cummings. CMOS imagers: from phototransduction to image processing. Kluwer Academic Publishers, 2004.

[29] A. Fish, V. Milrud and O. Yadid-Pecht, "High speed and high resolution current Loser-take-all circuit of o(N) complexity", IEEE ICECS, December 2004, CD ROM.

[30] A. Fish, V. Milirud and O. Yadid-Pecht, "High speed and high resolution current winner-take-all circuit in conjunction with adaptive thresholding", IEEE Transactions on Circuits and Systems II, vol. 52, no. 3, pp. 131-135, March 2005.

[31] A. Fish, V. Mosheyev, V. Linkovsky and O. Yadid-Pecht, "Ultra Low-Power DFF based Shift Registers Design for CMOS Image Sensors Applications", IEEE ICECS, Tel-Aviv, Israel, December 2004, CD ROM.

Authors' Information

Alexander Fish – e-mail: afish@ee.bgu.ac.il

Liby Sudakov-Boreysha –e-mail: libys@bgu.ac.il

The VLSI Systems Center, Ben-Gurion University, Beer Sheva, Israel;

Orly Yadid-Pecht – The VLSI Systems Center, Ben-Gurion University, Beer Sheva, Israel, and Dept. of Electrical and Computer Engineering, University of Calgary, Alberta, Canada; e-mail: oyp@ee.bgu.ac.il

IMAGE SENSORS IN SECURITY AND MEDICAL APPLICATIONS

Evgeny Artyomov, Alexander Fish, Orly Yadid-Pecht

Abstract: This paper briefly reviews CMOS image sensor technology and its utilization in security and medical applications. The role and future trends of image sensors in each of the applications are discussed. To provide the reader deeper understanding of the technology aspects the paper concentrates on the selected applications such as surveillance, biometrics, capsule endoscopy and artificial retina. The reasons for concentrating on these applications are due to their importance in our daily life and because they present leading-edge applications for imaging systems research and development. In addition, review of image sensors implementation in these applications allows the reader to investigate image sensor technology from the technical and from other views as well.

Keywords: CMOS image sensors, low-power, security applications, medical applications.

ACM Classification Keywords: B.8.1.i Hardware - Integrated Circuits - Types and Design Styles - VLSI

1. Introduction

Fast development of low-power miniature CMOS image sensors triggers their penetration to various fields of our daily life. Today we are commonly used to meet them in digital still and video cameras, cellular phones, web and security cameras, toys, vehicles, factory inspection systems, medical equipment and many other applications (see Figure 1). The advantages of current state-of-the-art CMOS imagers over conventional CCD sensors are the possibility in integration of all functions required for timing, exposure control, color processing, image enhancement, image compression and analog-to-digital (ADC) conversion on the same chip. In addition, CMOS imagers offer significant advantages in terms of low power, low voltage, flexibility, cost and miniaturization. These features make them very suitable especially for security and medical applications. This paper presents a review of image sensors utilization in part of the security and the medical applications.

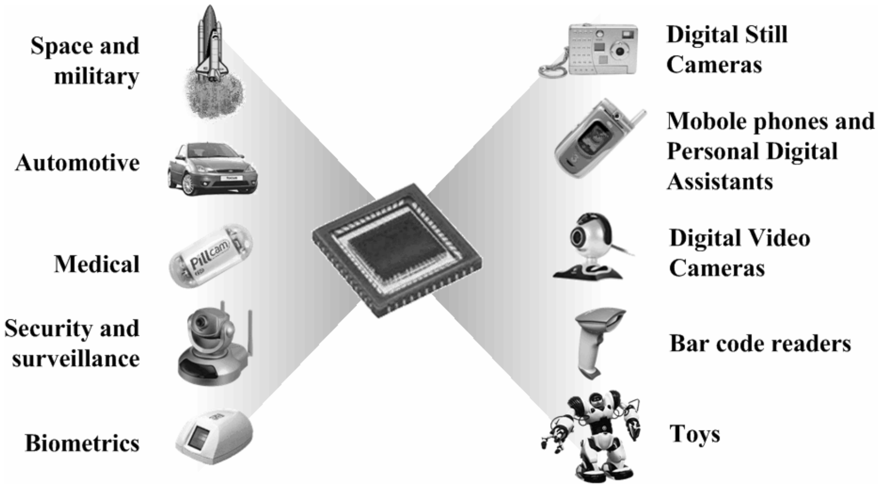


Figure 1. Image sensors applications

During the last few years imaging systems for security applications have been significantly revolutionising. Large, high cost and inefficient cameras mostly used for specific military and government applications have been replaced with compact, low-cost, low-power smart camera systems, becoming available not only for military and government, but for wide spreading in civilian applications. In this paper we will concentrate on two major categories: (a) surveillance systems – usually used for observation, anomaly detection and alarming, employing one or multiple cameras, (b) biometrics systems – used for access control and person identification. Each of the presented categories requires sensors having different specifications: for example, while low-power and compactness are the most important features for some surveillance systems, robustness and high image quality are the most important requirement in biometric systems.

Medical applications also benefit from the fast image sensors technology development. Introduction of miniature, ultra-low power CMOS image sensors have opened new perspectives to minimally-invasive medical devices, like wireless capsules for gastrointestinal tract observation [1]. Here we will review two very important medical applications:

(a) artificial retina – used as an artificial replacement or aid to the damaged human vision system,

(b) wireless capsule endoscopy – used in minimally invasive gastrointestinal tract diagnostics.

The remainder of the paper is organized as follows: Section II briefly presents CMOS image sensor technology with reference to "smart" CMOS image sensor architecture. The role of image sensors in security applications is described in Section III. Section IV reviews medical applications employing state-of-the-art CMOS imagers. Section V concludes the paper.

2. CMOS Image Sensor Technology in a Glance

The continuous advances in CMOS technology for processors and DRAMs have made CMOS sensor arrays a viable alternative to the popular charge-coupled devices (CCD) sensor technology. Standard CMOS mixed-signal technology allows the manufacture of monolithically integrated imaging devices: all the functions for timing, exposure control and ADC can be implemented on one piece of silicon, enabling the production of the so-called "camera-on-a-chip" [2]. Figure 2 is a diagram of a typical digital camera system, showing the difference between the building blocks of commonly used CCD cameras and the CMOS camera-on-a-chip [3]. The traditional imaging pipeline functions—such as color processing, image enhancement and image compression—can also be integrated into the camera. This enables quick processing and exchanging of images. The unique features of CMOS digital cameras allow many new applications, including network teleconferencing, videophones, guidance and navigation, automotive imaging systems, robotic and machine vision and of course, security and bio-medical image systems.

Most digital cameras still use CCDs to implement the image sensor. State-of-the-art CCD imagers are based on a mature technology and present excellent performance and image quality. They are still unsurpassed for high sensitivity and long exposure time, thanks to extremely low noise, high quantum efficiency and very high fill factors. Unfortunately, CCDs need specialized clock drivers that must provide clocking signals with relatively large amplitudes (up to 10 V) and well-defined shapes. Multiple supply and bias voltages at non-standard values (up to 15 V) are often necessary, resulting in very complex systems.

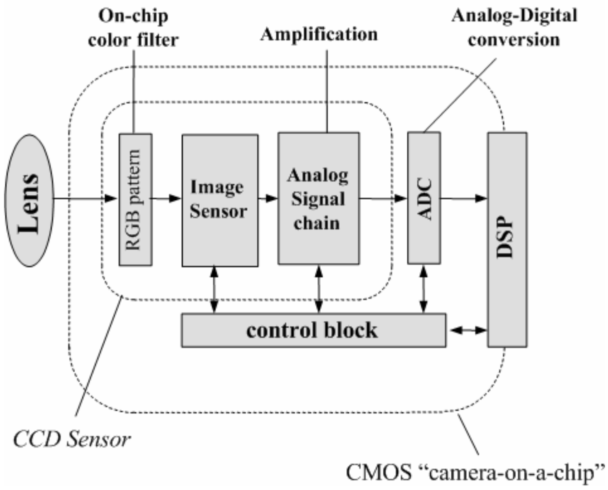


Figure 2. Block diagram of a typical digital camera system.

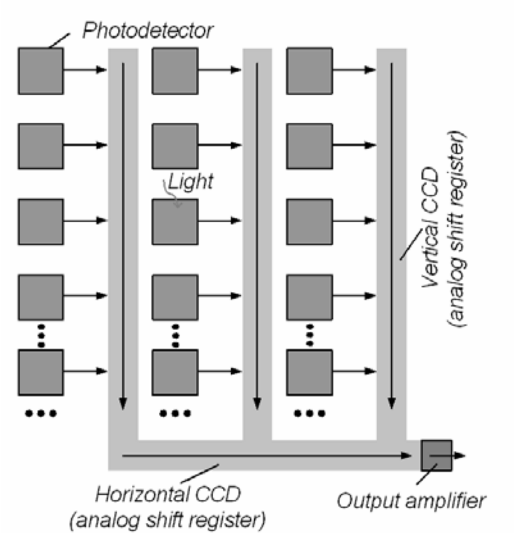


Figure 3. Block diagram of a typical interline transfer CCD image sensor.

Figure 3 is a block diagram of a widely used interline transfer CCD image sensor. In such sensors, incident photons are converted to charge, which is accumulated by the photodetectors during exposure time. In the subsequent readout time, the accumulated charge is sequentially transferred into the vertical and horizontal CCDs and then shifted to the chip-level output amplifier. However, the sequential readout of pixel charge limits the readout speed. Furthermore, CCDs are high-capacitance devices and during readout, all the capacitors are switched at the same time with high voltages; as a result, CCD image sensors usually consume a great deal of power. CCDs also cannot easily be integrated with CMOS circuits due to additional fabrication complexity and increased cost. Because it is very difficult to integrate all camera functions onto a single CCD chip, multiple chips must be used. A regular digital camera based on CCD image sensors is therefore burdened with high power consumption, large size and a relatively complex design; consequently, it is not well suited for portable imaging applications.

Unlike CCD image sensors, CMOS imagers use digital memory style readout, using row decoders and column amplifiers. This readout overcomes many of the problems found with CCD image sensors: readout can be very fast, it can consume very little power, and random access of pixel values is possible so that selective readout of windows of interest is allowed. The power consumption of the overall system can be reduced because many of the supporting external electronic components required by a CCD sensor can be fabricated directly inside a CMOS sensor. Low power consumption helps to reduce the temperature (or the temperature gradient) of both the sensor and the camera head, leading to improved performance.

An additional advantage of CMOS imagers is that analog signal and digital processing can be integrated onto the same substrate, allowing fabrication of so called "smart" image sensors. Many "smart" image sensors have already been demonstrated in the literature. They performed functions of real time object tracking [4]-[11], motion detection [12]-[13], image compression [14]-[15], widening the dynamic range of the sensor [16]-[20] and others. These functions are usually performed by digital or nonlinear analog circuits and can be implemented inside the pixels and in the periphery of the array. Offloading signal processing functions makes more memory and DSP processing time available for higher-level tasks, such as image segmentation or tasks unrelated to imaging.

CMOS pixels can be divided into two main groups, passive pixel sensors (PPS and active pixel sensors (APS). Each individual pixel of a PPS array has only a photosensing element (usually a photodiode) and a switching

MOSFET transistor. The signal is detected either by an output amplifier implemented in each column or by a single output for the entire imaging device. These conventional MOS-array sensors operate like an analog DRAM, offering the advantage of random access to the individual pixels. They suffer from relatively poor noise performance and reduced sensitivity compared to state-of-the-art CCD sensors. APS arrays are relatively novel image sensors that have amplifiers implemented in every pixel; this significantly improves the noise parameter.

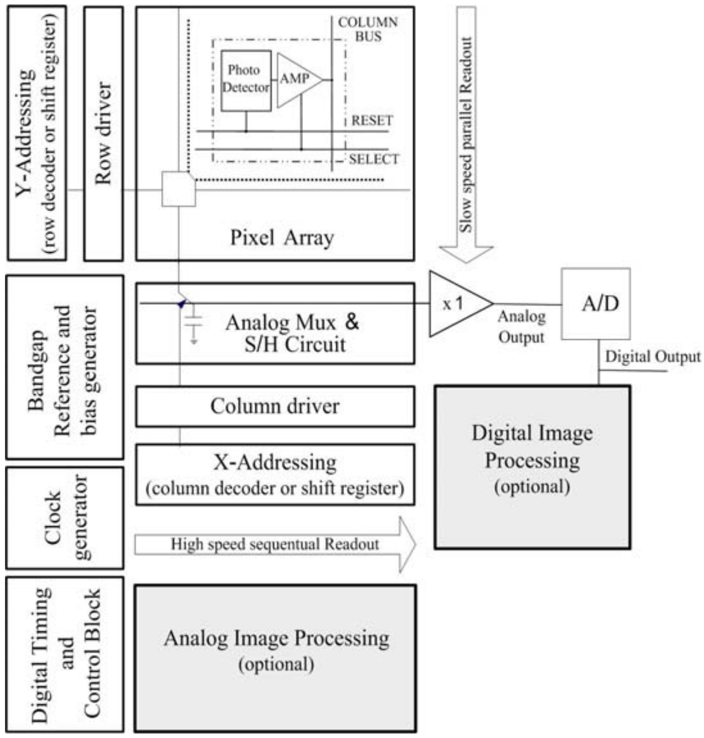


Figure 4. General architecture of the "smart" CMOS APS based image sensor.

Figure 4 shows the general architecture of the "smart" CMOS APS based image sensor. The core of this architecture is a camera-on-a-chip, consisting of a pixel array, a Y-addressing circuitry with a row driver, an X-addressing circuitry with a column driver, an analog front end (AFE), an analog-to-digital converter (ADC), a digital timing and control block, a bandgap reference and a clock generator. Optional analog and digital processing blocks "upgrade" the camera-on-a-chip core to a "smart" imager, and they are used to perform additional functions, that can vary from design to design, depending on the application and system requirements.

The basic imager operation, depends on the chosen photodetector and pixel types, readout mode, Y-addressing and X-addressing circuitries, ADC type and of course, the analog and/or digital image processing. A brief description of the main imager building blocks is presented herein.

A. *APS Pixel Array* – the imager pixel array consists of N by M active pixels, while the most popular is the basic photodiode APS pixel, employing a photodiode and a readout circuit of three transistors. Generally, many types of photodetectors and pixels can be found in the literature. This includes a p-i-n photodiode, photogate and pinned photodiode based pixels, operating either in rolling shutter or in global shutter (snapshot) readout modes. The difference between these modes is that in the rolling shutter approach, the start and end of the light collection for each row is slightly delayed from the previous row, leading to image distortion when there is relative motion between the imager and the scene. On the other hand, the global shutter technique uses a memory element inside each pixel and provides capabilities similar to a mechanical shutter: it allows simultaneous integration of the entire pixel array and then stops the exposure while the image data is read out. A detailed description of both rolling shutter and global shutter pixels can be found in [3].

Note, although most of today's "cameras-on-a-chip utilize" very simple pixels, many "smart" imagers employ more complicated pixels. Some of them perform analog image processing tasks at the pixel level. Very good examples for these imagers are neuromorphic sensors, where each pixel consists of a photo detector and local circuitry, performing spatio-temporal computations on the analog brightness signal. Another example is an imager, where the A/D conversion is performed in the pixel level.

B. Scanning Circuitry – Unlike CCD image sensors, CMOS imagers use digital memory style readout, usually employing Y-Addressing and X-Addressing to control the readout of output signals through the analog amplifiers and allow access to the required pixel. The array of pixels is accessed in a row-wise fashion using the Y-Addressing circuitry. All pixels in the row are read out into column analog readout circuits in parallel and then are sequentially read out using the X-Addressing circuitry (see Figure 4).

C. Analog Front End (AFE) - all pixels in a selected row are processed simultaneously and sampled onto sample-and-hold (S/H) circuits at the bottom of their respective rows. Due to this column parallel process, for an array having M columns, the AFE circuitry usually consists of 2*M S/H circuits, M size analog multiplexer, controlled by the X-Addressing circuitry, and one or M amplifiers to perform correlated double sampler (CDS). The CDS improves the signal-to-noise ratio (SNR) by eliminating the fixed pattern noise (FPN). A programmable- (or variable-) gain amplifier (PGA or VGA) follows the CDS to amplify the signal and better utilize the full dynamic range of the A/D converter (ADC).The number of amplifiers, required to perform the CDS functionality depends on the chosen CDS architecture and is equal to 2*N in case the subtraction is done separately for each column. The choice of an AFE configuration depends on many factors, including: the type of sensor being used, dynamic range, resolution, speed, noise, and power requirements. The considerations regarding making appropriate AFE choices for imaging applications can be found in [21].

D. Analog-to-digital conversion (ADC) – ADC is the inherent part of state-of-the-art "smart" image sensors. There are three general approaches to implementing sensor array ADC:

1. Pixel-level ADC, where every pixel has its own converter [22]-[23]. This approach allows parallel operation of all ADCs in the APS array, so a very low speed ADC is suitable. Using one ADC per pixel has additional advantages, such as higher SNR and simpler design.

2. Column-level ADC, where an array of ADCs is placed at the bottom of the APS array and each ADC is dedicated to one or more columns of the APS array [24]-[25]. All these ADCs are operated in parallel, so a low-to-medium-speed ADC design can be used, depending on the sensor array size. The disadvantages of this approach are the necessity of fitting each ADC within the pixel pitch (i.e., the column width) and the possible problems of mismatch among the converters at different columns.

3. Chip-level ADC, where a single ADC circuit serves the whole APS array [26]-[27]. This method requires a very high-speed ADC, especially if a very large array is used. The architecture shown in Figure 4, utilizes this approach for ADC implementation.

E. Bandgap reference and current generators – these building blocks are used to produce on-chip analog voltage and current references for other building blocks like amplifiers, ADC, digital clock generator and others. It is very important to design high precision and temperature independent references, especially in high resolution state-of-the-art image sensors, where the temperature of the die can vary by many tens of degrees.

F. Digital timing and control block, clock generator - aim to control the whole system operation. Their implementation in the chip level decreases the number of required I/O pads and thus reduces system power dissipation. Synchronized by the generated clock, the digital timing and control block produces the proper sequencing of the row address, column address, ADC timing and the synchronization pulses creation for the pixel data going offchip. In addition, it controls the synchronization between the imager and the analog and digital processing.

G. Analog and Digital Image Processing – although these blocks are optional, they play a very important role in today's "smart" image sensors. Conventional vision systems are put at a disadvantage by the separation between a camera for “seeing” the world, and a computer or DSP for “figuring out” what is seen. In these systems all information from the camera is transferred to the computer for further processing. The amount of processing circuitry and wiring necessary to process this information completely in parallel is prohibitive. In all engineered systems, such computational resources are rarely available and are costly in terms of power, space, and reliability. Opposite to a conventional camera-on-a-chip, which only captures the image and transfer it for the further processing, "smart" image sensors reduce the computational cost of the processing stages interfaced to it by carrying out an extensive amount of computation at the focal plane itself (analog and digital image processing blocks in Figure 4), and transmitting only the result of this computation (see Figure 5).

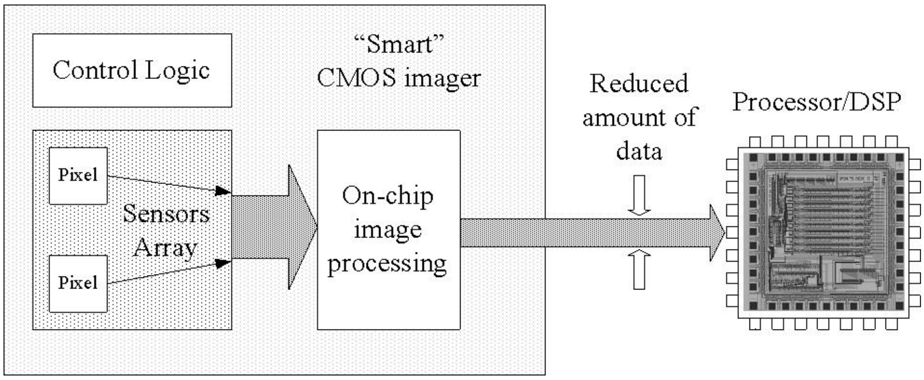


Figure 5. An example of an imaging system, employing a "smart" CMOS image sensor with on-chip processing and processors/DSPs for image processing

Both analog and digital processing can be performed either in the pixel or in the array periphery. There are advantages and disadvantages for both methods. In-pixel digital image processing is very rare because it requires pixel-level ADC implementation and results in very poor fill factor and large pixel size. In-pixel analog image processing is very popular, especially in the field of neuromorphic vision chips. In these chips in-pixel computations are fully parallel and distributed, since the information is processed according to the locally sensed signals and data from pixel neighbors. Usually, neuromorphic visual sensors have very low-power dissipations due to their operation in the subthreshold region, but suffer from low resolution, small fill-factor and very low image quality. Other applications employing in-pixel analog processing are tracking chips, wide dynamic range sensors, motion and edge detection chips, compression chips and others. The periphery analog processing approach assumes that analog processing is performed in the array periphery without penalty on the imager spatial resolution and it is usually done in a column parallel manner. While this approach has computational limitations compared to in-pixel analog processing, it allows better image quality. Periphery digital processing is the most standard and usually simpler. It is performed following the A/D conversion, utilizes standard existing techniques for digital processing and is usually done on the chip level. The main disadvantage of this approach is its inefficiency by means of area occupied and power dissipation. Note, all mentioned techniques can be mixed and applied together on one chip to achieve better results.

3. Image Sensors in Security Applications

The importance of security applications has significantly increased due to numerous terrorists' attacks worldwide. This area also greatly benefits from the achievements in the image sensors field. Today we can meet the cameras not only in military applications, but also in commercial and civilian applications. They are present in the shops and on the streets, in the vehicles and on the robots. The applications are numerous and can not be covered in this short paper. We have decided to concentrate on two important applications that represent a large fraction of the total security market. These applications are surveillance and biometrics. Both of the applications are extensively utilized in military, commercial and civilian fields.

3.1 Surveillance

Surveillance systems enable a human operator [28] to remotely monitor activity over large areas. Such systems are usually equipped with a number of video cameras, communication devices and computer software or some kind of DSP for real-time video analysis. Such analysis can include scene understanding, attention based alarming, colour analysis, tracking, motion detection, windows of interest extraction etc. With recent progress in CMOS image sensor technology and embedded processing, some of the mentioned functions and many others can be implemented in dedicated hardware, minimizing system cost and power consumption. Of course, such integration affects system configurability, but not all applications require configurable systems: some of them benefit from low cost and low power dedicated hardware solutions.

For example, in [29] we have presented an image sensor that can be used for such applications. Due to a specific scanning approach this sensor can be used efficiently for motion detection, tracking, windowing and digital zoom. Figure 6 shows the standard approach for sensor data scan - raster and the alternative – Morton or Z scan.

(a) raster scan – conventional approach in image sensors,

(b) Morton (Z) scan – newly proposed and implemented

Figure 6. Two approaches for data scan

The Morton (Z) scan poses a very valuable feature, neighbor pixels that are concentrated in blocks appear at the output sequentially, one after other. With this scanning approach the image blocks can be easily extracted and processed with simple on-chip hardware. For example, for constructing video camera with $\times 4$ digital zoom, the blocks of 4×4 pixels need to be extracted and averaged. Similarly, cameras with digital zoom $\times 8$ and $\times 16$ can be easily constructed. Figure 7 shows measurements from our test chip.

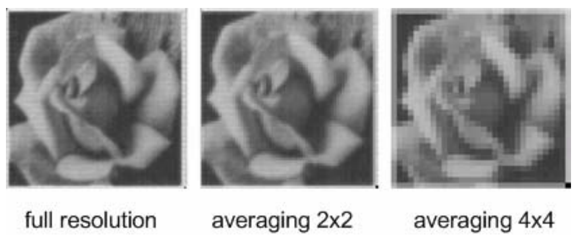


Figure 7. Morton scan chip test results

Another example is a wide dynamic range (WDR) imager. Dynamic range (DR) quantifies the ability of a sensor to image highlights and shadows. If we define the dynamic range of the sensor as 20log(S/N), where S is the maximal signal value and N is the sensor noise, the typical image sensors will have a very limited dynamic range, about 65-75 dB. Wide dynamic range imaging is very important in many surveillance systems. The dynamic range can be increased in two ways: the first one is noise reduction and thus enabling expansion of the dynamic range toward darker scenes; the second method is incident light saturation level expansion, thus improving the dynamic range toward brighter scenes.

Herein we present one of the possible solutions for dynamic range extension in CMOS Active Pixel Sensors (APS) [2]. As in a traditional CMOS APS, this imager is constructed of a two-dimensional pixel array, with random pixel access capability and row-by-row readout rolling shutter method. Each pixel contains an optical sensor to receive light, a reset input and an electrical output representing the illumination received. This imager implements a simple function for saturation detection, and is able to control the light exposure time on a pixel-by-pixel basis, resulting in no saturation. The pixel value can then be determined as a floating-point representation. To do so, the outputs of a selected row are read out through the column-parallel signal chain, and at certain points in time are also compared with an appropriate threshold value, as shown in Figure 8. If a pixel value exceeds the threshold, i.e. the pixel is expected to be saturated at the end of the exposure time; the reset is given at that time to that pixel. The binary information concerning the reset (i.e., if it is applied or not) is saved in a digital storage for later calculation of the scaling factor. Thus, we can represent the pixel output in the following floating-point format: $M \cdot 2^{EXP}$, where the mantissa (M) represents the digitized pixel value and the exponent (EXP) represents the scaling factor. This way a customized, linear, large increase in the dynamic range is achieved.

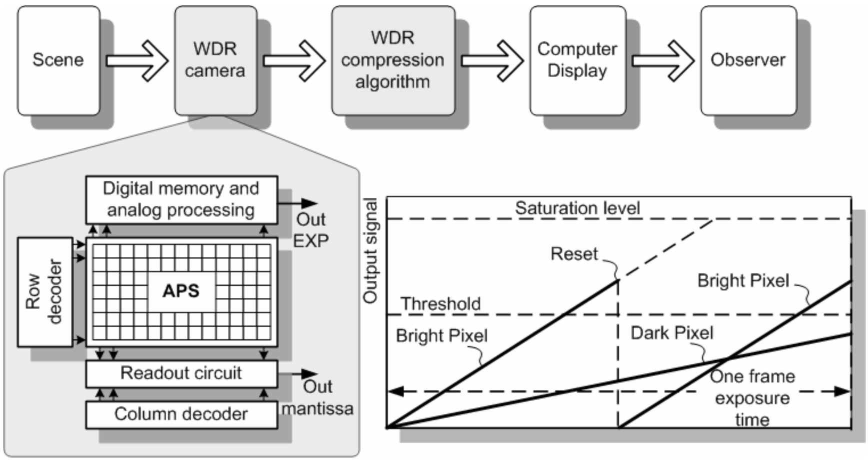


Figure 8. Imaging pipeline, image sensor architecture and work principle.

Figure 9(a) and Figure 9(b) show a comparison between an image captured by a traditional CMOS imager and by the autoexposure system described here. In Figure 9(a), a scene is imaged with a strong light hitting the object; hence, some of the pixels are saturated. At the bottom of Figure 9(b), the capability of the autoexposure sensor for imaging the details of the saturated area in real time may be observed. Since the display device is limited to eight bits, only the most relevant eight-bit part (i.e., the mantissa) of the thirteen-bit range of each pixel is displayed here. The exponent value, which is different for different areas, is not displayed.

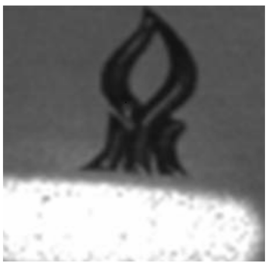


Figure 9. (a) Scene observed with a traditional CMOS APS sensor,

Figure 9. (b) Scene observed with our in-pixel autoexposure CMOS APS sensor.

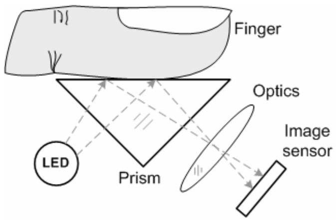
3.2 Biometric personal identification

Biometric personal identification is strongly related to security and it refers to “identifying an individual based on his or her distinguishing physiological and/or behavioral characteristics (biometric identifiers)” [30]. Figure 10 shows the most frequently used biometric characteristics.

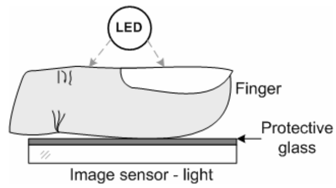
Figure 10. Biometric characteristics

Almost all biometric characteristics, shown in Figure 10, require some kind of sensing. Usually, conventional image sensors with external hardware or software image processing are used. The difficulty for on-chip integration is caused by the complexity of the required image processing algorithms. However, there are some developments that successfully achieve the required goals by parallel processing utilization.

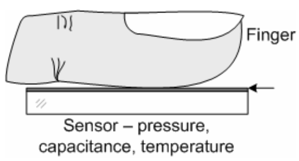
To give some more detailed examples in the field, we concentrate on fingerprint sensors. Generally these sensors can be classified by the physical phenomena used for sensing: optical, capacitance, pressure and temperature. The first two classes are the most popular and both mainly employ CMOS technology.



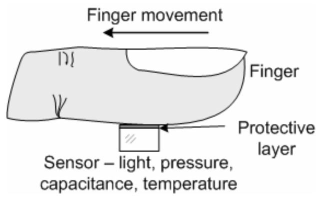
(a) optical - reflection based sensor



(b) optical – transmission based sensor



(c) non-optical – based on pressure, capacitance or temperature sensor



(d) “sweep” sensor

Figure 11. Fingerprint sensors

In Figure 11 various technologies for fingerprint sensing are shown [31]. The most popular approach (see Figure 11 (a)) is based on optical sensing and light reflection from the finger surface. Also, this type provides high robustness to finger condition (dry or wet), but the system itself is tend to be bulky and costly. Alternative solutions that can provide compact and lower cost solutions, are based mostly on solid state sensors where the finger is directly placed on the sensor. However, in these solutions the sensor size needs to be at least equal to the size of the finger part used for sensing. Two sensors of this type are shown in Figure 11 (b) and (c). The first one is based on light transmitted through the finger and then sensed by the image sensor, while the second one is the non-optical sensor that can be implemented either as pressure, capacitance or temperature sensor. The fingerprint sensor, known as a “sweep” sensor and shown in Figure 11 (d), can be implemented using either the optical or other previously mentioned techniques. A “sweep” sensor employs only a few rows of pixels, thus in order to get a complete fingerprint stamp the finger needs to be moved over the sensing part. Such technology greatly reduces the cost of the sensor due to reduced sensor area and solves the problem of fingerprint stamp that needs to be left on the surface in the first two methods.

In all presented methods, the output signal is usually an image and the sensors are composed of pixels that sense either temperature, pressure, photons or change in capacitance. The overall architectures of these sensors are similar to the architecture described in section II and they integrate various image and signal processing algorithms, implemented the same die. Various research papers have been published in this area and numerous companies are working on such integration. For example, in [32] the authors implement image enhancement and robust sensing for various finger conditions. Capacitive sensing CMOS technology is used and data is processed in a column parallel way. The same technology is used also in [34], but the fingerprint identifier is also integrated and the data is processed massively in parallel for all pixels.

Despite the fact that fingerprint technology is quite mature, there is much work to be done to reduce power consumption, to improve technology and image processing algorithms and to achieve better system miniaturization.

4. Image Sensors in Medical Applications

Almost all medical and near medical areas benefit from image sensors utilization. These sensors are used for patients' observation and drug production, inside the dentists offices and during surgeries. In most cases the sensor itself represents only a small fraction (in size and cost) of the larger system, but its functionality plays a major role in the whole system. Figure 12 shows examples of medical applications where CMOS image sensors are used. In this section of the paper we mostly concentrate on applications that push current image sensor technology to the edge of the possibilities. These applications are wireless capsule endoscopy and retinal implants. Both of these applications will play an important role in millions of patients' lives in the near future.

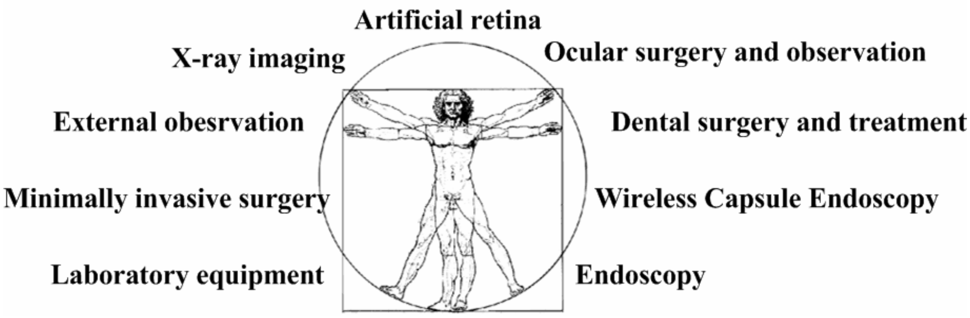


Figure 12. Image sensors applications in medicine

4.1 Wireless Capsule Endoscopy

Conventional medical instrumentation for gastrointestinal tract observation and surgery uses an endoscope that is externally penetrated. These systems are well developed and provide a good solution for inter-body observation and surgery. However, the small intestine (bowel) was almost not reachable using this conventional equipment, leaving it for observation only through surgery or through an inconvenient and sometimes painful push endoscopy procedures. Few years ago the sphere was revolutionized by the invention of the wireless image sensor capsule, which after swallowing, constantly transmits a video signal during its travel inside the body [1]. The capsule movement is insured by the natural peristalsis. According to Gavriel Iddan [1], the founder of Given Imaging™ [35] that commercializes this technology, "The design of the video capsule was made possible by progress in the performance of three technologies: complementary metal oxide silicon (CMOS) image sensors, application-specific integrated circuit (ASIC) devices, and white-light emitting diode (LED) illumination".

The general architecture of the capsule is shown in the Figure 13. It consists of LEDs, optics, camera, digital system processing, transmitter or transceiver and a power source. The dashed blocks represent additional future requirements for such capsules.

All capsule electronic components are required to be low power consumers to enable constant video transmission for a prolonged time (for about 6-8 hours) and/or high capacity batteries. An alternative solution to in-capsule batteries [36] is to use an external wireless power source that supplies energy to the capsule through electromagnetic coils. Such a solution enables to relax power requirements for the capsule electronics. This solution also provides an advantage in freeing space inside the capsule for other useful functions such as biopsy or medication. Also, the capsule position can be controlled externally through a strong magnetic field. But the required strong magnetic field can limit the capsule usage in spite of position control advantages [33].

Currently the Given Imaging™ capsule developers have reached very encouraging results enabling two capsules: one intended for the Esophagus part (the upper part) of the gastrointestinal tract and the second for small intestine observation. The first kind of the capsule is equipped with two CMOS image sensors and can transmit the video signal for about 20 minutes with 14 frames per second for each camera. The second one consists of only one CMOS image sensor and can transmit two frames per second for about eight hours. The company is developing now a new capsule generation that can transmit four frames per second.

Despite these encouraging results, a lot of work should be done to allow further miniaturization, image processing and compression algorithms integration, power reduction by various means (system integration, technology scaling etc.), frame-rate increase, quality improvement and usage of alternative power sources with larger capacity. The ultimate goal that needs to be achieved is full video frame-rate transmission for about 7-8 hours. To achieve these goals, a number of additional research groups work worldwide on wireless capsules development: eStool by Calgary university in Canada [37], MiRO by Intelligent Microsystems Center in Korea [38], EndoPill by Olympus [39].

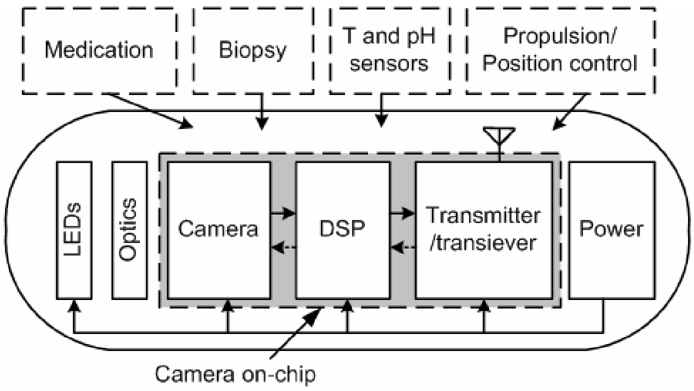
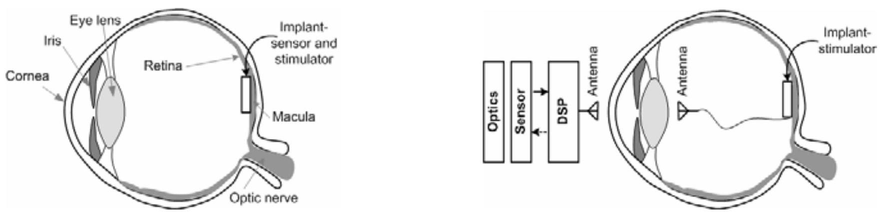


Figure 13. The swalable capsule architecture.
In the dashed boxes additional functionality that will be required in the future is shown

4.2 Artificial Retina

Artificial vision is another example of CMOS image sensors implementation in medical applications. Today millions of people are suffering from full or partial blindness that was caused by various retinal deceases. In the early eighties it was shown that electrical stimulation of the retinal nerves can simulate visual sensation even in the patients with fully degraded receptors. Recently, researchers in a number of research institutes have developed miniature devices that can be implanted into the eye and stimulate the remaining retinal neural cells, returning partial vision ability for the blind patients. Such implants are called artificial retinas. Usually they are implanted in the macula area that normally is densely populated by the receptors and enables high-resolution vision. This break-through was enabled by the progress in electronics, surgical instrumentation, and biocompatible materials. Currently there are two major approaches for artificial retina development (see Figure 14).



(a) (b)

Figure 14. Artificial retinas: (a) implantable sensor (b) external sensor

The first and the most promising one is the integration of sensing and stimulation elements in the same device and the second is separation of sensing and stimulation. In the first approach, an artificial retina device is an autonomous circuitry that does not require external control and the optics that is used for sensing is the natural optics of the eye composed of the cornea and lens. In the second, all the sensing and processing is performed outside of the eye and only stimulating elements are implanted during surgery. The data transfer from the sensing part to the stimulation part is performed through an RF link or through a tiny cable. In both approaches the implant can be subretinal or epiretinal.

Actually there is a number of groups working in the field [40]-[44] but we will concentrate on two that have shown very promising results and are now performing clinical trials and commercialization through companies named Optobionics™ [40] and Second Sight [41]. Both groups already have a number of patients with such implants.

The device developed by Optobionics™ group does not require any power source, integrates about 5000 sensing (microphotodiodes) and stimulation (electrodes) elements, features two millimetres in diameter and is implanted under retina. The basic artificial silicon retina unit is shown in Figure 15 [45]. It is composed of a stimulating electrode and three PIN photodiodes connected in series to increase the output voltage.

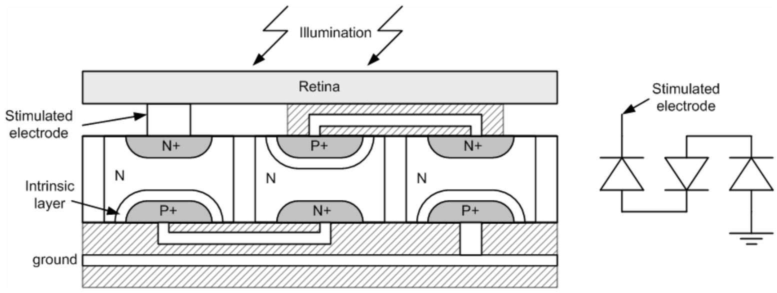


Figure 15. Artificial silicon retina – basic unit

The second group decided to follow the second approach and separate sensing from stimulation. The camera with the processor is situated on the patient glasses and the signal is transmitted through a cable to the eye that has an implanted stimulator. Currently, the implant resolution is not so impressive compared to the first group and features only 16 electrodes, but the developers plan to increase the resolution in the future models to 60 and 1000 electrodes [46].

5. Conclusions

In this paper we have presented a brief review of CMOS image sensors utilization in security and medical applications. In these applications image sensors play a major role and usually define the edge of the imaging technology. Despite the CMOS image sensor technology already exists for more than a decade, it is continuously developing and penetrating into new fields that were unreachable by its predecessor, CCD technology. Although many successes have been achieved during the last decade, a lot of work still needs to be done in this area. It requires extensive collaboration between various fields such as: electrical engineering, materials, computer science, medicine, psychology, chemistry etc. As to the electrical engineering and sensing fields, the work should be concentrated in the directions of power consumption reduction, functionality improvement and system integration. However, like in every multidisciplinary, electrical engineers, developing the electronic devices for medical purposes, are required to understand all above mentioned fields to successfully implement such devices.

References

[1] G. Iddan, G. Meron, A. Glukhovsky, P. Swain, “Wireless capsule endoscopy”, *Nature*, Vol. 405, p. 417, May 2000

[2] E. Fossum, "Low power camera-on-a-chip using CMOS Active Pixel Sensor technology", in *IEEE Symposium on Low Power Electronics*, pp. 74–77, 1995.

[3] O. Yadid-Pecht and R. Etienne-Cummings, "CMOS imagers: from phototransduction to image processing", *Kluwer Academic Publishers*, 2004.

[4] A. Fish, D. Turchin, O. Yadid-Pecht, " An APS with 2-Dimensional winner-take-all selection employing adaptive spatial filtering and false alarm reduction", *IEEE Trans. on Electron Devices*, Special Issue on Image Sensors, January, 2003.

[5] V. Brajovic and T. Kanade, “Computational sensor for visual tracking with attention”, *IEEE journal of solid-state circuits*, Vol.33, No.8, August 1998.

[6] T. Horiuchi and E. Niebur, “Conjunction search using a 1-D, analog VLSI-based attentional search/tracking chip,” *Conference for Advanced Research in VLSI, D. Scott Wills and Stephen P. DeWeerth*, Eds., pp. 276–290. IEEE Computer Society, 1999.

[7] G. Indiveri, “Neuromorphic analog VLSI sensor for visual tracking: Circuits and application examples.” *IEEE Trans. On Circuits and Systems*, II 46(11), pp. 1337–1347, November 1999.

[8] C. S. Wilson, T. G. Morris, and P. DeWeerth, “A two-dimensional, object-based analog VLSI visual attention system", *Twentieth Anniversary Conference on Advanced Research in VLSI, IEEE Computer Society Press*: Los Alamitos, CA. Vol. 20. pp. 291-308. March 1999.

[9] M.Clapp and R..Etienne-Cummings, “A dual pixel-type imager for imaging and motion centroid localozation”, *Proc. ISCAS'01*, Sydney, Australia, May 2001

[10] N. Mei Yu, T. Shibata and T. Ohmi, “A Real-Time Center-of-Mass Tracker Circuit Implemented by Neuron MOS Technology”, *IEEE transactions on circuits and systems*—II, vol. 45, no. 4, April 1998.

[11] R.C. Meitzler, K. Strohbehn and A.G. Andreou, “A silicon retina for 2-D position and motion computation”, *Proc. ISCAS'95*, New York, USA, 1995.

[12] A. Simoni, G. Torelli, F. Maloberti, A. Sartori, S. E. Plevridis and A. N. Birbas, “A Single-Chip Optical Sensor with Analog Memory for Motion Detection”, *IEEE Journal of Solid-State Circuits*, Vol. 30, No. 7, July 1995.

[13] M. Clapp and R. Etienne-Cummings, “Dual Pixel Array for Imaging, Motion Detection and Centroid Tracking,” *IEEE Sensors Journa*l, Vol, 2, No. 6, pp. 529-548, December 2002.

[14] S. Kawahito, M. Yoshida, M. Sasaki, K. Umehara, D. Miyazaki, Y. Tadokoro, K. Murata, S. Doushou, and A. Matsuzawa, “A CMOS Image Sensor with Analog Two-Dimensional DCT-Based Compression Circuits for One-Chip Cameras”, *IEEE Journal of Solid-State Circuits*, Vol. 32, No. 12, 1997.

[15] K. Aizawa, H. Ohno, Y. Egi, T.Hamamoto, M. Hatory, H. Maruyama and J. Yamazaki “On sensor Image Compression,” *IEEE Transaction On Circuits And Systems For Video Technology*, vol 7, no. 3, pp. 543-548, June 1997.

[16] O. Yadid-Pecht, A. Belenky " In-Pixel Autoexposure CMOS APS " *IEEE Journal of Solid-State Circuits*, Vol. 38, No. 8, pp. 1425-1428, August 2003.

[17] A. Fish, A. Belenky and O. Yadid-Pecht, “Wide Dynamic Range Snapshot APS for Ultra Low-Power Applications, *IEEE Transactions on Circuits and Systems II*, vol. 52, no. 11, pp. 729-733, November, 2005.

[18] O. Yadid-Pecht, C. Clark, B. Pain, C. Staller, E. Fossum , Wide dynamic range APS star tracker, in *Proc. SPIE/IS&T Sym. on Electronic Imaging: Science and Technology*, San Jose, California, SPIE Vol. 2654, Jan 29-Feb3, 1996, pp. 82-92.

[19] O. Yadid-Pecht, E. Fossum, "Wide Intrascene Dynamic Range CMOS APS Using Dual Sampling" , *IEEE Trans. Elec. Dev., special issue on solid state image sensors*, Vol. 44, No. 10, pp. 1721-1724, October 1997.

[20] O. Yadid-Pecht, "Wide dynamic range sensors", *Optical Engineering*, Vol. 38, No. 10, pp.1650-1660, October 1999.

[21] K. Buckley, "Selecting an Analog Front-End for Imaging Applications", *Analog Dialogue*, vol. 34-6, pp. 1-5, 2000.

[22] D.X.D. Yang, A. El Gamal, B. Fowler and H. Tian, “A 640x512 CMOS Image Sensor with Ultra Wide Dynamic Range Floating Point Pixel Level ADC,” IEEE ISSCC, WA 17.5, 1999.

[23] B. Pain, S. Mendis, R. Scober, R. Nixon, and E. Fossum, "Low-power low-noise analog circuits for on-focal-plane signal processing of infrared sensors," IEEE Workshop on Charge Coupled Devices and Advanced Image Sensors, June, 1995.

[24] A. Dickinson, S. Mendis, D. Inglis, K. Azadet, and E. Fossum, "CMOS Digital Camera with Parallel Analog-to Digital Conversion Architecture," IEEE Workshop on Charge Coupled Devices and Advanced Image Sensors, April, 1995.

[25] A. Krymski and N. Tu, "A 9-V/Lux-s 5000-Frames/s 512x512 CMOS Sensor," IEEE Trans. Electron Devices, vol. 50 pp. 136-143, Jan. 2003.

[26] S. Smith, J.Hurwitz, M. Torrie, D. Baxter, A. Holmes, M.Panaghiston, R. Henderson, A. Murray, S. Anderson, and P. Denyer, "A single-chip 306x244-pixel CMOS NTSC video camera," ISSCC Digest of Technical Papers, pp. 170-171, February 1998.

[27] M. Loinaz, K. Singh, A. Blanksby, D. Inglis, K. Azadet, and B. Acland," A 200mW 3.3V CMOS color camera IC producing 352x288 24b Video at 30frames/s," ISSCC Digest of Technical Papers, pp 186-169, February 1998.

[28] G. L. Foresti, C. Micheloni, L. Snidaro, P. Remagnino, and T. Ellis, “Active video-based surveillance system”, *IEEE Signal Processing Magazine*, pp. 25-37, March 2005.

[29] E. Artyomov, Y. Rivenson, G. Levi, Orly Yadid-Pecht, “Morton (Z) Scan Based Real-Time Variable Resolution CMOS Image Sensor”, *IEEE Transactions on Circuits and Systems for Video Technology*, Volume 15, Issue 7, July 2005, pp. 947 – 952.

[30] A. Jain, L. Hong, S. Pankanti, “Biometric identification”, *Communications of the ACM*, Vol. 43, No. 2, 2000.

[31] K. Uchida, “Fingerprint identification”, *NEC Journal of Advanced Technology*, Vol. 2, No. 1, pp. 19-27, 2005.

[32] S.J. Kim, K.H. Lee, S.W. Han, and E. Yoon, "A 200×160 Pixel CMOS Fingerprint Recognition SoC with Adaptable Column-Parallel Processors,", *IEEE International Solid-State Circuits Conference (ISSCC)*, pp. 250-251, February 2005.

[33] “Robotics road map”, EURON Technology Roadmaps, April 23, 2004, www.org.id.tue.nl/IFIP-SG16/robotics-roadmap-2004.pdf

[34] S. Shigematsu, H. Morimura, Y. Tanabe, T. Adachi, and K. Machida, "A Single-Chip Fingerprint Sensor and Identifier", *IEEE JOURNAL OF SOLID-STATE CIRCUITS*, Vol. 34, No. 12, pp. 1852-1859, December 1999

[35] www.givenimaging.com

[36] www.rfnorika.com

[37] K.N.C. Hin, O. Yadid-Pecht, M. Mintchev, "e-Stool: Self-Stabilizng Capsule for Colonic Imaging", *Neuro-stimulation Conf*, France, July 2005.

[38] microsystem.re.kr/main_eng/menu04/sub_menu01.asp

[39] "Olympus Launches High-resolution Capsule Endoscope in Europe", *Olympus press release*, October 13, 2005, www.olympus.co.jp/en/news/2005b/nr051013capsle.cfm

[40] www.optobionics.com

[41] www.2-sight.com

[42] www.eye-chip.com

[43] www.iip-tec.com/

[44] www.bostonretinalimplant.org/

[45] U.S. Patent US-007003354-B2, given to Optobionics™

[46] S. Derra, "Bring sight to the blind", *R&D magazine*, 2005, www.rdmag.com

Authors' Information

Evgeny Artyomov – e-mail: artemov@bgu.ac.il

Alexander Fish – e-mail: afish@ee.bgu.ac.il

Orly Yadid-Pecht – e-mail: oyp@ee.bgu.ac.il

The VLSI Systems Center, Ben-Gurion University, Beer Sheva, Israel

MANOMETRY-BASED COUGH IDENTIFICATION ALGORITHM

Jennifer A. Hogan, Martin P. Mintchev

Abstract: Gastroesophageal reflux disease (GERD) is a common cause of chronic cough. For the diagnosis and treatment of GERD, it is desirable to quantify the temporal correlation between cough and reflux events. Cough episodes can be identified on esophageal manometric recordings as short-duration, rapid pressure rises. The present study aims at facilitating the detection of coughs by proposing an algorithm for the classification of cough events using manometric recordings. The algorithm detects cough episodes based on digital filtering, slope and amplitude analysis, and duration of the event. The algorithm has been tested on in vivo data acquired using a single-channel intra-esophageal manometric probe that comprises a miniature white-light interferometric fiber optic pressure sensor. Experimental results demonstrate the feasibility of using the proposed algorithm for identifying cough episodes based on real-time recordings using a single channel pressure catheter. The presented work can be integrated with commercial reflux pH/impedance probes to facilitate simultaneous 24-hour ambulatory monitoring of cough and reflux events, with the ultimate goal of quantifying the temporal correlation between the two types of events.

Keywords: Biomedical signal processing, cough detection, gastroesophageal reflux disease.

ACM Classification Keywords: I.5.4 Pattern Recognition: Applications – Signal processing; J.3 Life and Medical Sciences

1. Introduction

The latest comprehensive statistics provided by the National Institutes of Health suggests that gastroesophageal reflux (GER) and related symptoms affect approximately 20% of the US population, resulting in over 700,000 annual hospitalizations [1]. GER, which occurs in healthy individuals without causing discomfort, is characterized by the movement of gastric content into the esophagus. Reflux episodes are considered a disorder, known as gastroesophageal reflux disease (GERD), when the episodes become frequent, prolonged, and/or are ineffectively cleared back into the stomach. An individual suffering from GERD shows problematic symptoms, such as heartburn, discomfort, and chest pain, which result due to prolonged exposure of the distal esophageal mucosa to the refluxate. Excessive exposure can lead to further consequences, such as esophagitis and esophageal ulceration.

GERD is commonly cited as a significant cause of chronic cough [2-5]. In multiple studies, GERD has been documented to be a cause of chronic persistent cough in 38 to 82% of patients [2]. Irwin et al. report that GER can cause cough either by aspiration of gastric content, or by stimulation of the distal esophagus due to repetitive or prolonged GER events [3]. It has also been documented that cough may precipitate GER by stimulating transient lower esophageal sphincter (LES) relaxation [2]. To determine if a patient suffering from chronic cough should be treated for GERD, it is important to quantify whether there is a causal relationship between cough and reflux. In addition to facilitating the diagnosis, more accurate identification of the temporal correlation between cough and reflux can lead to a more appropriate treatment [6].

Numerous studies have focused on determining this temporal correlation [5-8]. However, they rely on manual identification of cough, either through the interpretation of user diaries and user-triggered events [5, 7, 8], or through the manual analysis of manometric recordings, identifying cough as simultaneous, short duration, rapid pressure rises (time to peak less than 1 second) across multiple manometric recording sites [6]. This results in a time-consuming process, prone to user error.

In this paper, we present an algorithm for accurately identifying coughs based on a single-channel manometric recording using a custom pressure probe optimized for cough detection [9]. The algorithm can be implemented in real-time, allowing a simultaneous indication of cough events with the recording of reflux episodes. The two-stage algorithm, which consists of a filtering stage followed by a decision stage, has some common elements with the real-time QRS detection algorithm proposed by Pan and Tompkins [10, 11]. In particular, the cough detection algorithm has a similar filtering stage to its QRS counterpart, with both algorithms sharing the goal of reducing high-frequency noise components and removing low-frequency baseline drifts and pressure changes, thus amplifying the pressure response of interest. The filtering stage of the proposed algorithm differs mainly in the bandpass frequency, which is specific to coughs, and in the utilization of integration in a time window corresponding to the typical duration of a cough. The decision stage of the cough detection algorithm has also been adapted specifically to isolate the pressure response characteristics of a cough.

2. Methodology

2.1 Algorithm Description

2.1.1 Overview

Initially, the presented cough detection algorithm employs a combination of bandpass filtering, differentiation, and moving window integration to magnify the short, rapid pressure rises characteristic of cough events and to reduce the baseline drift and the low-frequency pressure variations characteristic of esophageal peristalsis. The second stage of the algorithm consists of decision logic which performs the recognition process, determining the occurrence of a cough based on a combination of dual-threshold and width detection. To demonstrate feasibility of the algorithm on acquired manometry tracings, development was initially performed in Matlab (The MathWorks, Natick, MA).

In the following sections, each stage of the algorithm is described in more detail.

2.1.2 Filtering Stage

The analysis of manometry tracings acquired using a single-channel fiber optic pressure catheter showed that the frequency components of the pressure rise associated with a cough typically range from 2.5 Hz to 5 Hz. An initial bandpass filter stage is used to attenuate signal components and noise artifact outside this frequency range. The bandpass filter is achieved using cascaded low-pass and high-pass filters. Both filters are zero-phase, least-squares approximations, 17th order FIR filters. The low-pass filter, which has a 3-dB frequency of 5 Hz, reduces high-frequency noise. The high-pass filter, which has a 3-dB frequency of 2.5 Hz, reduces lower-frequency peristaltic contractions and baseline drifts. Following the bandpass filtering stage, the signal is differentiated. The differentiation stage focuses on identifying cough by its characteristically steep slope. The resulting signal is then rectified, making all data points positive facilitating the decision process. Finally, the signal is integrated to provide an average of the output, thus incorporating the width of the cough episode into the output signal.

2.1.3 Decision Stage

The output of the filtering stage results in the amplification of cough events and the reduction of high-frequency noise and low-frequency pressure rises due to peristalsis. Because the filtering stage dramatically improves the cough signal-to-noise ratio, the decision stage of the cough detection process is simplified. Cough detection on the filtered data relies on first determining the maximum peak within a given sampling interval. This can be determined recursively in real-time. The threshold is then adjusted automatically based on the current maximum peak. The threshold has been empirically set to 0.33 of the maximum peak of the current interval. By periodically adjusting the threshold based on the maximum peak of the current sampling interval, the algorithm adapts to changing characteristics in the signal. After determining the threshold, the algorithm searches for a dual crossing of the threshold, with the first crossing registering a positive slope, the second crossing registering a negative slope, and the time interval between the crossings being within the cough duration parameter. The cough duration has been set to 400 ms, based on measuring cough durations on manometry tracings.

2.2 Evaluating the Algorithm

Evaluation of the discussed algorithm was performed using esophageal manometry recordings acquired with a single-channel pressure catheter optimized for cough detection. The catheter [9], comprises a miniature white-light interferometric fiber optic pressure sensor encapsulated in such a manner as to optimize sensitivity to pressure at the catheter tip, such as that experienced during a cough event, while reducing sensitivity to circumferential force, such as that experienced during esophageal peristalsis. The esophageal probe interfaces to an optical signal conditioner which converts the optical signal to an analog output. The analog output is sampled at 50 Hz using a PCMCIA data acquisition board (National Instruments, Austin, TX) and a custom-designed real-time software application on a laptop computer. A block diagram of the system is shown in Figure 1.

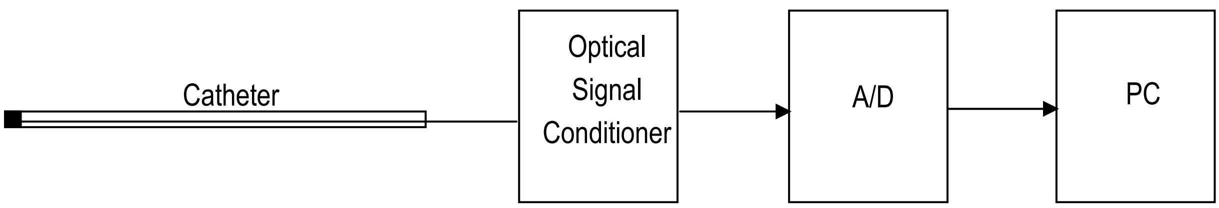


Figure 1 - Block diagram of experimental setup.

In vivo esophageal recordings were acquired from a healthy volunteer. Three separate trials were performed to ensure repeatability. In each trial, the volunteer performed a series of respiratory and gastrointestinal events, such as swallowing, belching, heavy breathing, coughing, throat clearing, and laughing. The volunteer also initiated movement (i.e. head rotation, bending, and standing) to attempt to introduce unwanted artifacts. All events were marked on the recording at the exact time of their occurrence.

3. Results

Preliminary testing of the algorithm on in vivo data samples demonstrated a consistent detection of cough events. Figure 2 shows the original signal and the output of each filtering step for one in vivo sampling interval recorded from the healthy volunteer.

In Figure 2a, the raw signal is shown, with gastrointestinal and respiratory events that occurred during the sampling interval indicated. The output of the low pass filter can be observed in Figure 2b, which reduces the high-frequency noise ripples. Figure 2c, which shows the output of the high pass filter, demonstrates the amplification of the cough events, while attenuating the slower pressure variations resulting from talking and swallowing. The differentiation process amplifies slope information, shown in Figure 2d, which is then rectified resulting in a signal comprising positive data points, shown in Figure 2e. The final output of the moving window integrator is shown in Figure 2f, where it can be observed that the cough signals have been amplified, while other pressure variations have been reduced.

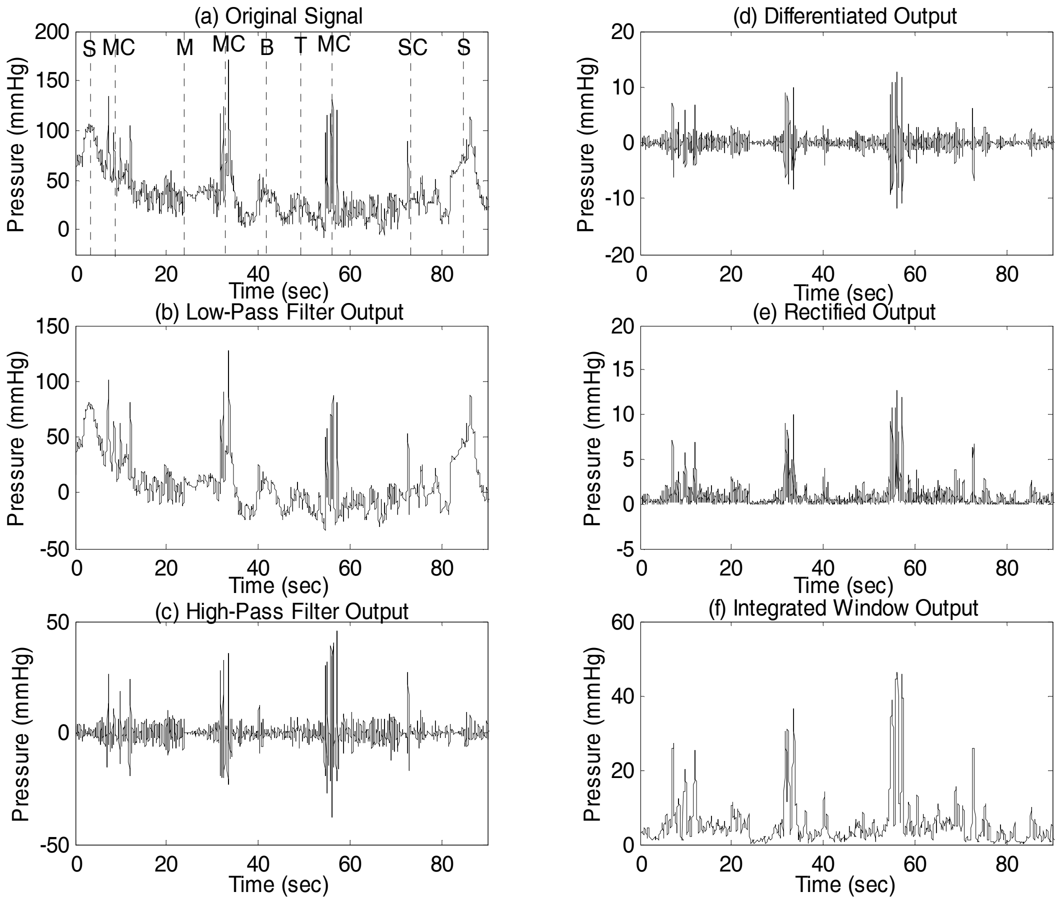


Figure 2 - Sample data of (a) original signal, (b) low-pass filter output, (c) high-pass filter output, (d) differentiated output, (e) rectified output, and (f) integrated window output.
S indicates a peristaltic contraction resulting from a swallow,
MC indicates the occurrence of multiple coughs, or a cough burst,
M indicates patient movement,
B indicates a belch,
T indicates patient talking,
SC indicates the occurrence of a single cough.

The final result of the cough detection algorithm is depicted in Figure 3, where the cough identification results (Figure 3c) are compared with the raw signal (Figure 3a) and the output of the filtering stage (Figure 3b), which supplies the input signal to the decision process. Despite the presence of noise artifacts and two high pressure variations resulting from peristalsis, the algorithm has successfully identified all cough events in the given sample by targeting a dual crossing of the established threshold within the set duration. This can be observed in Figure 3c.

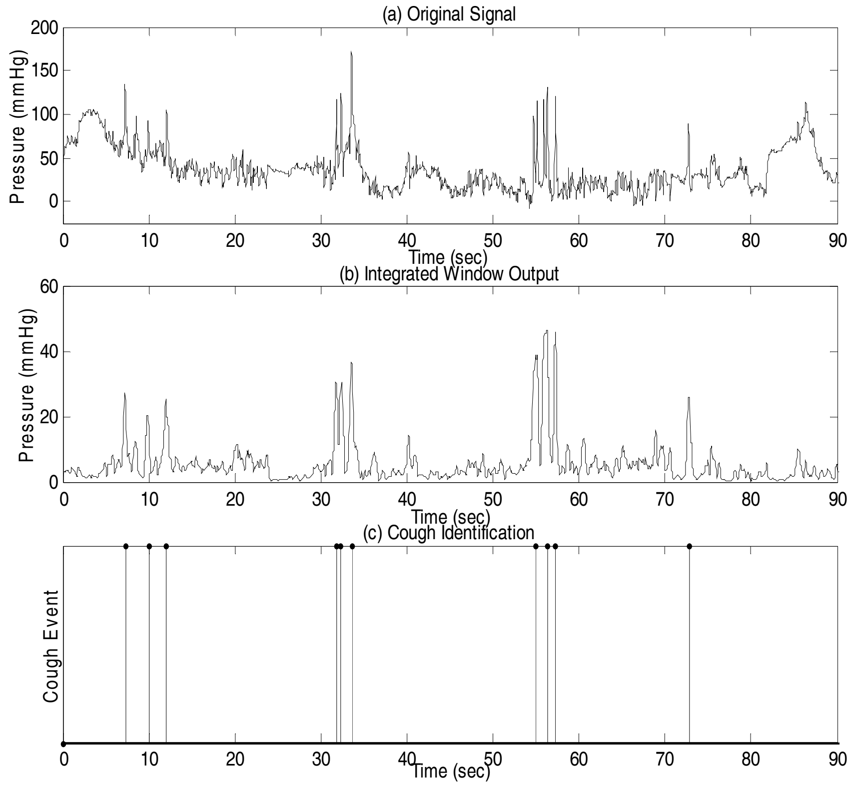


Figure 3 - Sample data showing cough identification,
including (a) original signal, (b) integrated window output, and (c) final algorithm results.

4. Discussion

In recent years, numerous studies have attempted to define the temporal correlation between cough events and reflux events in a pursuit to fully understand the relationship between chronic cough and GERD [5-8]. In these investigations, a tool to facilitate the identification of cough episodes without manual analysis of sound recordings, diary annotations, or manometry tracings is still lacking. In the present study, we developed an algorithm for conveniently identifying coughs based on a single manometric channel recording.

Using an algorithm consisting of a filter stage to amplify frequency components associated with the characteristics of cough while attenuating other pressure artifacts, followed by decision logic to identify the cough events from the filtered signal, cough episodes have been correctly identified in samples acquired via manometry. The successful demonstration of the feasibility of the algorithm paves the way for more robust testing using a larger database of manometric readings. Using the algorithm structure presented, specific parameters, including the filter cut-off frequencies, the duration of cough events, and the adaptive threshold setting, may be adjusted to more accurately satisfy the characteristics of cough based on a larger sampling pool of test recordings. Also, a further qualifier can be added to the real-time implementation, in which the definition of a cough can be modified to include bursts of two, three, or more cough events.

5. Conclusion

An innovative algorithm for identifying cough events based on manometric recordings was developed. Initial testing of the algorithm was performed on in vivo data samples acquired using a single-channel pressure catheter. Preliminary studies indicate that the algorithm is suitable for cough detection. By integrating the technique with a single-channel pressure and pH/impedance catheter for 24-hour ambulatory monitoring of cough and reflux, a suitable diagnostic tool for GERD could be achieved.

Acknowledgement

This study was supported in part by the Natural Sciences and Engineering Research Council of Canada, the Alberta Ingenuity Fund, and the Informatics Centre of Research Excellence.

Bibliography

[1] United States National Institute of Health, “Digestive disease statistics”, [Online]. Available: http://digestive.niddk.nih.gov/statistics/statistics.htm. Accessed on January 26, 2006.

[2] K. F. Chung, J. G. Widdicombe, and H. A. Boushey, *Cough: Mechanisms and Therapy*, Oxford, Great Britain: Blackwell Publishing, pp. 97-106, 2003.

[3] R. S. Irwin, C. T. French, F. J. Curley, J. K. Zawacki, and F. M. Benett, “Chronic cough due to gastroesophageal reflux: clinical, diagnostic, and pathogenetic aspects,” *Chest*, vol. 105, pp. 1511-1517, 1993.

[4] Y. W. Novitsky, J. K. Zawacki, R. S. Irwin, C. T. French, V. M. Hussey, and M. P. Callery, “Chronic cough due to gastroesophageal disease: efficacy of antireflux surgery,” *Surgical Endoscopy*, vol. 16, no. 4, pp. 567-571, 2002.

[5] A. W. Wunderlich and J. A. Murray, “Temporal correlation between chronic cough and gastroesophageal reflux disease”, *Digestive Diseases and Sciences*, vol. 48, no. 6, pp. 1050-1056, 2003.

[6] D. Sifrim, L. Dupont, K. Blondeau, X. Zhang, J. Tack, and J. Janssens, “Weakly acidic reflux in patients with chronic unexplained cough during 24 hour pressure, pH, and impedance monitoring,” *Gut*, vol. 54, pp. 449-454, 2005.

[7] S. M. Harding, M. R. Guzzo, and J. E. Richter, “24-h esophageal pH testing in asthmatics,” *Chest*, vol. 115, no. 3, pp. 654-659, 1999.

[8] B. Avidan, A. Sonnenberg, T. G. Schnell, and S. J. Sontag, “Temporal associations between coughing or wheezing and acid reflux in asthmatics,” *Gut*, vol. 49, pp. 767-772, 2001.

[9] J. A. Hogan and M. P. Mintchev, “Method and apparatus for intra-esophageal cough detection,” Technical Report submitted to University Technologies International, Calgary, Alberta, March, 2006.

[10] J. Pan and W. J. Tompkins, “A real-time QRS detection algorithm,” *IEEE Transactions in Biomedical Engineering*, vol. 32, no. 3, pp. 230-236, 1985.

[11] P. S. Hamilton and W. J. Tompkins, “Quantitative investigation of QRS detection rules using the MIT/BIH arrhythmiac database,” *IEEE Transactions in Biomedical Engineering*, vol. 33, pp. 1157-1165, 1986.

Authors' Information

Jennifer A. Hogan – M.Sc. graduate student, Department of Electrical and Computer Engineering; University of Calgary, Calgary, Alberta, Canada, T2N 1N4

Martin P. Mintchev – Prof., Dr., Department of Electrical and Computer Engineering; University of Calgary; Calgary, Alberta, Canada, T2N 1N4; Department of Surgery, University of Alberta; Edmonton, Alberta T6G 2B7

Phone: (403) 220-5309; Fax (403) 282-6855; e-mail: mmintche@ucalgary.ca

MULTIMODAL MAN-MACHINE INTERFACE AND VIRTUAL REALITY FOR ASSISTIVE MEDICAL SYSTEMS

Adil Timofeev, Alexander Nechaev, Igor Gulenko, Vasily Andreev, Svetlana Chernakova, Mikhail Litvinov

Abstract: The results of research the intelligence multimodal man-machine interface and virtual reality means for assistive medical systems including computers and mechatronic systems (robots) are discussed. The gesture translation for disability peoples, the learning-by-showing technology and virtual operating room with 3D visualization are presented in this report and were announced at International exhibition "Intelligent and Adaptive Robots–2005".

Keywords: multimodal man-machine interface, virtual reality, assistive medical systems.

ACM Classification Keywords: I.2. Artificial Intelligence

1. Introduction

The modern medicine is actually required the development of new information technologies and the multimodal man-machine interface (MMI) for control of medical robots and mechatronic systems, for automation of surgery with virtual reality means for creating telemedical diagnostic systems, etc.

The NATO-grant № PST.CLG 975579 “The man-machine interface for assistive systems in neurosurgery”, executed in 1999-2000 by partners from the St.-Petersburg Institute for Informatics and Automation of Russian Academy of Science (SPIIRAS), the State University of Aerospace Instrumentation, University of Karlsruhe (Germany) and Harvard Medical School (USA) has been directed on development of the man-machine interface, adaptive robots and multiagent technologies intended for neurosurgery [1]. The results of researches the medical MMI, robotics and mechatronic systems, including the results submitted in this report, have been presented at the international exhibition and a symposium “Intelligent and Adaptive Robots - 2005” [2].

2. Video Capture of Motion for Translating from Sign Language on Natural Language and Back

The development of anthropomorphous robots and their animation are very important for solution of various problems. Innovative results in this area, based on models and means of virtual reality, may be applied at such human activities, as medicine, sports, learning, computer and cognitive graphics. For example, in medicine the motion analysis of person (patient) can be essentially for automatic diagnostics and treatment of orthopedic diseases.

The researches of human’s or anthropomorphous robot’s motion are very important in the computer graphics for animation of virtual actors with methods and means of video capture was offered in [3]. Currently these methods being adapted for solution the problem of translation from a natural language to a sign language and back and for development of MMI for disable peoples.

The communication for people with restrictions on hearing or speech is taking place in completely different way, than it is for people without such restrictions. Peoples with such restrictions cannot watch TV, listen to radio without additional means and communicate by phone, as it done usually by people. So it is necessary to develop the approach to realization of specialized interfaces, which could be used in any telecommunication devices. The interfaces being discussed include the input-output informative means and data links.

As for data links, it is widely used a highly developed and reliable Information Telecommunication (IT) technologies. The advanced input-output informative systems for disability peoples are the novel MMI technologies.

As for gesture exchange with MMI, it is offered the output of gesture translation to be carried out by means of animated “Avatar”, the simplified 3D computer model of human. Avatar can reproduce gestures by two ways [3, 4]:

- Generating gestures from corresponding text expressions in natural language using script sequences from database;
- Reproducing animation of gestures commands transmitted directly under the data link.

The gesture input means also can work by two following ways:

- Text generating in natural language according to analysis of gestures reproduced by operator;
- Transforming gestures of operator to gesture command sequences of animation for avatar.

3. Man-machine Interface for Aassistive Medical Systems Based on Video Capture of Motions

The important task in the development of robotic systems is a design of the man-machine interface. The new approach in creation of such interface is based on technologies of video capture of robot’s motion and a virtual reality means [5, 6].

The general structure of man-machine interface and robot control system designed by using means of video capture and virtual reality technology includes two subsystems, see Fig. 1.

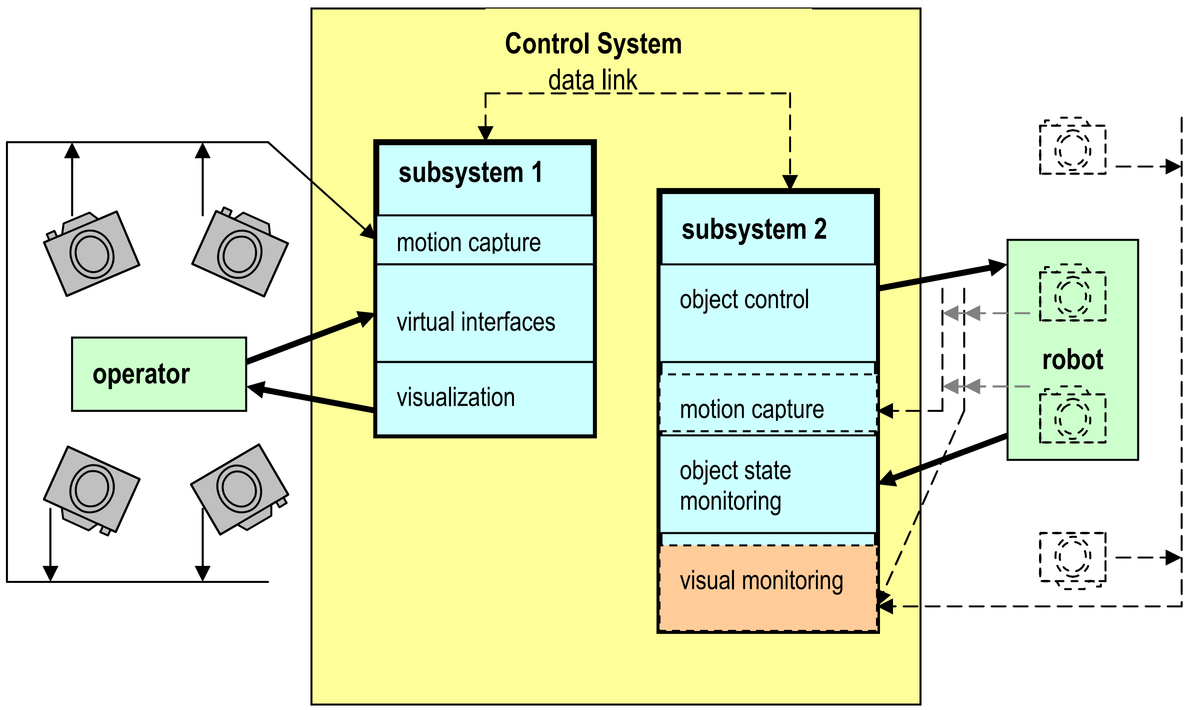


Fig. 1

Subsystem 1 (subsystem of interaction with operator) includes the following components:

- System of video capture of operator’s motions,
- Virtual control devices,
- Visualization procedures.

Subsystem 2 (subsystem of interaction with robot) has a various structures depending on application. However it is necessary to include the following components in this subsystem:

- Means of processing and distribution of robot control signals,
- Means of supervision over a condition of robots.

It is also proposed to develop the following components increasing the control efficiency and reliability of MMI systems [3, 4]:

- Active multi-channel system of video capture of motion;
- Means of visual observation.

4. The Learning Technology Using Multimodal Man-machine Interface

The intelligence learning technology with showing the natural movements (gestures) of operator [7, 8] is developed to creating natural and accessible means for people's communication with technical and information systems (fig. 2).

The offered learning technology for medical mechatronic systems (robots) is based on the intelligence MMI [9], which provide the following features for medical applications:

- Simple and correct understanding of gesture and speech commands of doctor or patient [10],
- The intuitive teaching the medical mechatronic system by means of natural operator's hand movements (without traditional movement's programming) [8],
- The visualization of real and virtual 3D medical images (X-ray, endoscopic, thermography, etc),
- The doctor's automatic workplace (AWP) with 3D viewing of X-ray images in real time mode [13];
- The novel stereo-projective systems for observation X-ray, endoscopic and other medical images;
- The human interaction with computer's models and real medical equipment by means tactile and force-torque sensors during diagnostic or planning of medical operation [6, 12],
- The information and telecommunication support of medical equipment control systems using the Virtual Operating Room (VOR).

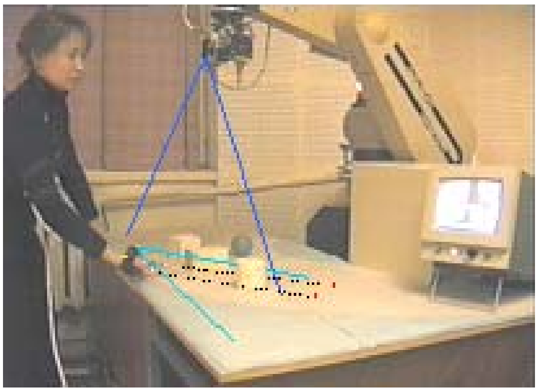
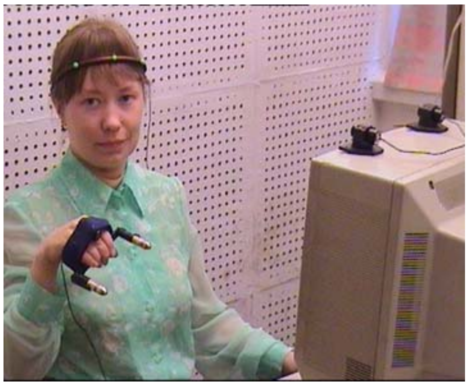



Fig.2

The development of the Multimodal Man-machine Interface (MMI) for telemedicine and medical robotics assumes the creation of highly realistic effect of doctor's presence in remote environment of patient.

For this purpose virtual models and computer-synthesized 3D-images of virtual objects composed with images of real environment (Augmented Reality technologies, AR) are used. The basic problems of AR technology realization is a problem of exact registration of a computer-synthesized image of geometrical model with real image and a problem of 3D image visualization in a real time mode.

It is suggested to use the following means to 3D visualization of medical images (fig.3):

- Stereo-glasses for observation on the computer monitor;
- Stereo-displays glasses for 3D viewing color images without computer monitor;
- Means of augmenting real medical images by virtual images (AR).

Tactile-force interaction with virtual medical objects is necessary when it is not only required to observe the environment, but it is also necessary to perform any actions in it. These manipulations in remote environment will be more successful, if it will be possible to create realistic and adequate perception of objects in environment surrounding the patient and medical robot, to give an ability of feeling a virtual object with mass, shape, elastic and friction features as a real object.

Fig. 3

These technologies have partially been developed within the Partner Project 1992p with EOARD (European Office of Aerospace Research and Development, London, United Kingdom) and based on a long-term experience of medical system development [14].

The Head Tracking System (HTS) and Hand Tracking System (HTS+) prototypes have been developed for accurate measurement of human-operator's head and hands movements. The basic aim of this development is to create the simple (cost-effective), reliable and steady means for human interaction with telecontrol objects.

For medicine the development of so-called "haptic" interface, intended for tactile and force-torque displaying on hand of doctor (surgeon) is especially actual. The developed handle with force-torque feedback sensors reflects the real tactile and force interaction of mechatronic system (robots) instrument with real objects.

The prototype of handle with force-torque sensor for controlling and learning of medical mechatronic systems is shown on Fig. 4. As an example of medical application of novel force-torque sensor is a prototype of robot-masseur with exact control of force contact to patient's muscles and a high level safety of robot-masseur movements near the patient or personnel.

Fig. 4

The offered technologies of 3D visualization, virtual reality means and learning-by-showing technology are especially useful for navigation, programming of movements and control of medical mechatronic systems [4, 11]. Thus information technologies with creating a highly realistic effect of presence the doctor in remote patient's environment are useful for telemedicine application too.

Using of medical robots in surgery or telemedicine may be dangerous to patients. Therefore it is necessary to create virtual models of robot, patient and operating room for testing, planning and safety providing of medical operation with usage of mechatronic systems (medical robots).

5. Virtual Model of Operating Room

Virtual operating room (VOR) should be similar to real one as much as possible, because the doctor's work in usual conditions would be more effective and productive. Therefore the VOR must be "filled" with familiar medical tools, equipment and assistive robots in their computer representation.

In this case there is an opportunity of carrying out of trial educational operations not on the real patient body, but on his virtual model. Carrying out of operation on the virtual "electronic patient" can be considered as a planning stage of real surgical operation. The second important problem being solved using VOR is an increasing professionalism of medical personnel. The dynamic virtual model of neurosurgical VOR is presented on fig. 5.

The information on real and virtual operation and clinical data of patient are united in a uniform picture of medical operation. It is possible to return to past operation for comparative analysis and finding-out the efficiency of different methods of medical operation. Due to this not only doctors, but also students and post-graduate students can to training with VOR due to improve their skill [3-6, 11].

During the virtual operation the surgeon can to observe the trajectory of medical tool, for example, under brain tomogram of patient. In a case when the virtual medical tool to reach zone of high risk for virtual patient, the doctor can request the VOR program to find other variant of virtual tool targeting to necessary point of brain on a more safety trajectory or to take a conclusion on impossibility of carrying out of operation.

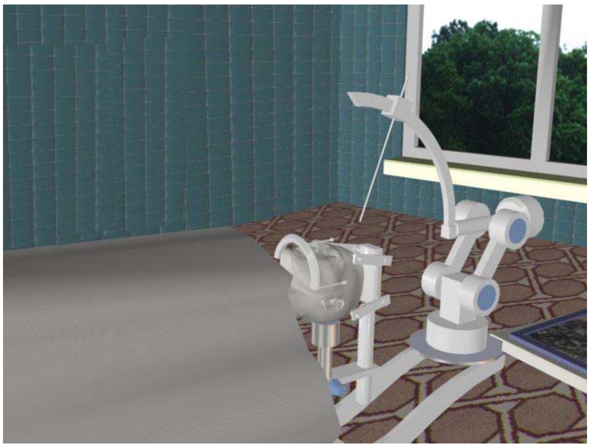


Fig. 5

6. Conclusion

The offered information technologies, the multimodal man-machine interface and a virtual reality technique find more and more wide applications not only in medicine, but also in other areas (service, home assistance, telecommunications, space, etc.). Thus the intelligence man-machine interface, dynamic models of a virtual reality and multi-agent technologies will play especially important role [5, 6].

At the International exhibition «Intelligent and Adaptive Robots – 2005» the authors of this report have been awarded with 4 Gold Medals and Diplomas «for the development and introduction of innovative technology of man-machine interaction for telemedicine and medical robotics».

The presented work is executed at partial support of the grant RFBR № 05-01-08044_ofi and the grant № 17 "Development of Man-Machine Interface and Control Methods for Neurosurgical Robots on the Base of Virtual Reality Models" of Saint-Petersburg Scientific Center of Russian Academy of Sciences.

Bibliography

[1] C. Burghart, O. Schorr, S. Yigit, N. Hata, K. Chinzei, A. Timofeev, R. Kikinis, H. Wörn, U. Rembold A Multi-Agent-System Architecture for Man-Machine Interaction in Computer Aided Surgery. – Proceedings of the 16th IAR Annual Meeting (Strasburg, November 22-23, 2001), pp. 117-123.

[2] Intelligent and Adaptive Robots – 2005. Official Exhibition Catalogue – M.: Open Company "ZETP", 2005, p. 45 (In Russian).

[3] Gulenko I.E., Shugina V.S. Application of technologies of video capture of movement in medicine and sports. – Works of conference "New information technologies" (Sudak, Crimea, May, 22-29, 2005), pp. 318-320 (In Russian).

[4] Timofeev A.V., Gulenko I.E. and Litvinov M.V. Analysis, Processing and Transfer of Dynamic Images in Virtual Reality Models. – Pattern Recognition and Image Analysis, 2004, vol. 16, No.1, pp.97-99.

[5] Timofeev A.V., Andreev V.A., Gulenko I.E., Derin O.A., Litvinov M.V. Design and Implementation of Multi-Agent Man-Machine Interface on the Base of Virtual Reality Models. – Proceedings of 9-th International Conference “Speech and Computer” (20-22 September, 2004), Saint-Petersburg, Russia, pp. 670-676.

[6] Timofeev A.V. Intellectualization for Man-Machine Interface and Network Control in Multi-Agent Infotelecommunication Systems of New Generation. – Proceedings of 9-th International Conference “Speech and Computer” (20-22 September, 2004), Saint-Petersburg, Russia, pp. 694-700.

[7] Kulakov F.M., Nechaev A.I., Chernakova S.E. Modelling of Environment for the Teaching by Showing Process. In Proc. of SPIIRAS, Russia, St. Petersburg, 2002, Issue № 2, pp. 105–113.

[8] Chernakova S.E., Kulakov F.M., Nechaev A.I. Learning of the robot by method of showing with use «sensing» gloves. – Works of the First international conference on mechatronics and a robotics (Saint Petersburg, May, 29 – June, 2, 2000), pp. 155-164 (In Russian).

A. Karpov, A. Ronzhin, A. Nechaev, S. Chernakova. Assistive multimodal system based on speech recognition and head tracking. In Proceedings of 9-th International Conference SPECOM’2004, St. Petersburg, 2004, pp. 521-530.

[9] A. Karpov, A. Ronzhin, A. Nechaev, S. Chernakova. Multimodal system for hands-free PC control. In Proc. of 13-th European Signal Processing Conference EUSIPCO–2005, Antalya, Turkey, 2005.

[10] Litvinov M.V., Timofeev A.V., Popov A.B. Virtual model of the adaptive robot for neurosurgery. – Proceedings of the International conference “Adaptive robots and the general logic theory of systems - AR&GSLT, (Saint Petersburg, on May, 8-11 2004), pp. 119–122 (In Russian).

[11] Timofeev A.V., Shibzuhov Z.M., Sheojev A.M. Designing and learning multi-agent diagnostic systems. – Proceedings of the First international conference on mechatronics and a robotics “M&R.

[12] Nechaev A.I., Nazaruk V.P., Chernakova S.E. Method of registration and visualization of 3D X-ray images in real time for technical nondestructive checking and medical diagnostics. – "Information technologies", 11, 2005, Moscow, pp. 11-21, in Russian.

[13] Chernakova S.E., Kulakov F.M., Nechaev A.I. Advanced man-machine interface for telerobotic based on man-operator’s motions and gaze direction tracking – Proceedings of Sixth ISTC Scientific Advisory Committee Seminar “Science and Computing”, Moscow. 15–17 September 2003, Volume 1, part II, pp. 303-308.

Authors’ Information

Timofeev Adil Vasilievich – Dr. Sc., Professor, Honoured Scientist of Russian Federation, Saint-Petersburg Institute for Informatics and Automation of Russian Academy of Sciences, 199178, Russia, Saint-Petersburg, 14-th Line, 39, phone: +7-812-328-0421; fax: +7-812-328-4450, e-mail: tav@iias.spb.su

Nechaev Alexander Ivanovich – Scientific Researcher, Saint-Petersburg Institute for Informatics and Automation of Russian Academy of Sciences, 199178, Russia, Saint-Petersburg, 14-th Line, 39, phone: +7-812-328-0421; fax: +7-812-328-4450, e-mail: nechaev@iias.spb.su

Gulenko Igor Evgenievich – Post Graduate Student, Saint-Petersburg Institute for Informatics and Automation of Russian Academy of Sciences, 199178, Russia, Saint-Petersburg, 14-th Line, 39, phone: +7-812-328-0421; fax: +7-812-328-4450, e-mail: gig@yandex.ru

Andreev Vasily Alexandrovich – Post Graduate Student, Saint-Petersburg Institute for Informatics and Automation of Russian Academy of Sciences, 199178, Russia, Saint-Petersburg, 14-th Line, 39, phone: +7-812-328-0421; fax: +7-812-328-4450, e-mail: vasiliy-a@yandex.ru

Chernakova Svetlana Eduardovna – Minor Scientist, Saint-Petersburg Institute for Informatics and Automation of Russian Academy of Sciences, 199178, Russia, Saint-Petersburg, 14-th Line, 39, phone: +7-812-328-0421; fax: +7-812-328-4450, e-mail: chernakova@iias.spb.su

Litvinov Mikhail Vladimirovich – Post Graduate Student, Baltic State Technical University “Voenmech”, 190005,Saint-Petersburg, Russia, 1-st Krasnoarmeyskaya, 1, e-mail: sid-4d@inbox.ru

APPLICATION OF NETWORK TECHNOLOGIES FOR DEVELOPMENT OF MEDICAL DATA-ADVISORY CLINIC “MED-HEALTH”

Anatoly Bykh, Elena Visotska, Tatjana Zhemchuzhkina, Andrey Porvan, Alexander Zhook

Abstract: In this article the medical data-advisory web-resource developed by authors is considered. This resource allows carrying out information interchange between consumers of medical services and the medical establishments which give these services, and firms-manufacturers of medical equipment and medicaments. Main sections of this web-site, their purposes and capabilities are considered in this article.

Keywords: web-resource, information interchange, medical advice.

ACM Classification Keywords: J.3 Life and medical sciences: Medical information systems

Introduction

Since electronic global networks appeared it is possible to watch the tendency of the Internet-technologies applications in different fields of human activity. However, application of such technologies in medicine remained aloof until recently. It is bound with following: not any information can be transferred to a network, and the transmitted information not always corresponds to the general requirements to the data transfer.

One of the guidelines in modern Internet-technology applications is development of medical data-advisory resources which allow getting a different medical information, as well as remote advice of the expert in certain field of medicine. We can refer to these progressive guidelines basic medical sites, different municipal information portals, specialized sites of the medical services, sites of firms-manufacturers of medical equipment or their trade representatives, etc.

Advanced technologies are of invaluable help for the medical and healthcare industry. There are really many possibilities and ways to use the information technologies in the healthcare, research and development areas. Computer systems and networks help scientists to perform experiments with greater accuracy, collect and analyze huge amounts of medical information, develop and create new medical tools and drugs.

Urgency of the problem

Among known web-resources in Ukraine we can mark sites of private clinics “Dr. Alex” [1], “Nebosvod” [2], and specialized sites “Mednet”, “Doctor-home” [3]. These sites give the information about departments of these medical establishments, medical personnel, the lists of services with prices and medical equipment which is available there. However, these sites contain information concerning only certain medical establishment, and guest of these sites often can’t get complete information about necessary question in one place, and this takes expense of time and means. Besides, many institutions save on information about other medical establishments and their services with aim to reduce costs for site making. That’s why we can’t say about working efficiency of these sites.

In that way, making of the medical data-advisory resource which takes into account all noted above lacks, gives interest and is actual problem.

Working up of the medical data-advisory resource

If to consider medical web-resources (sites) as integrated data-advisory system it is necessary to distinguish key components of this system [4]. It is offered to consider this system as integration of representatives of purposeful groups. Regarding these groups the medical institution has or can have communicative aims. Traditionally distinguish external and internal web-resource environments (fig. 1).

In case of medical data-advisory web-resource external environment contains:

- clients-consumers of services, proposed by medical institutions (population, physicians, pharmacologists, medical establishments of different types);
- clients-providers, giving the information for allocation in web-resource (medical establishments, drugstores, firms-manufacturers of medical equipment and pharmaceutical preparations, etc.).

The internal environment of medical data-advisory web-resource is essentially information resource, containing information of the following contents:

- the index of medical establishments; parts selected in structure of these establishments; existing medical cabinets; medical private offices;
- the personnel of medical establishment (personal data of treating and advising doctors, attendants, etc.);
- the medical equipment (diagnostic and therapeutic equipment, which is available there, capabilities of this equipment, the list of the procedures spent with this equipment; the advertising information of firms-manufacturers of the medical equipment);
- services given by medical establishments (consultations of doctors, diagnostic and therapeutic procedures, etc.).

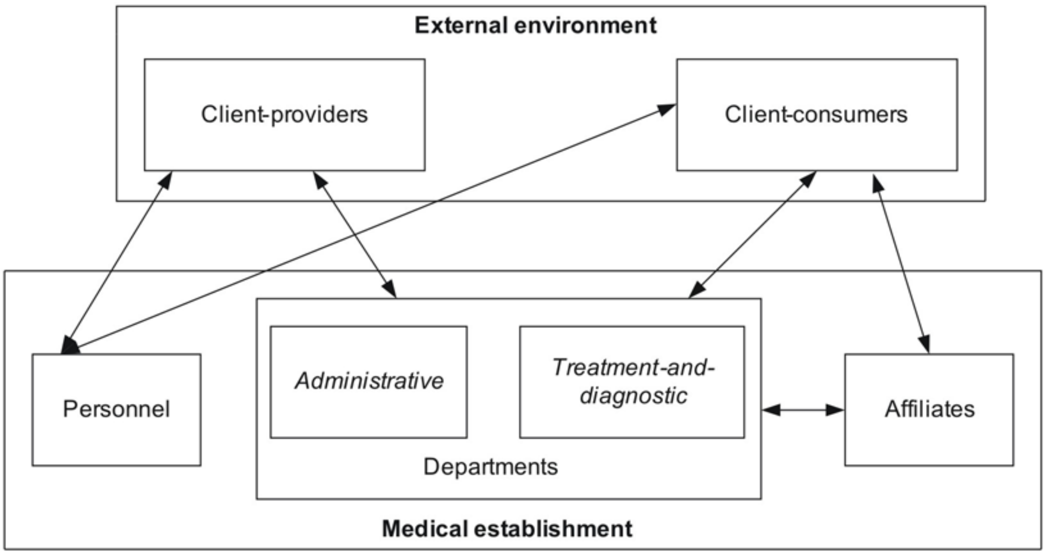


Figure 1 - External and internal web-resource environments

It is possible to allocate three base functions of an offered Web-resource:

1. Information function - realization of information and advertising activity of the Web-resource (giving of the information about the medical institutions, given services, consultations, the diagnostic and therapeutic equipment, medical preparations, etc.),
2. Communicative function - function of information interchange between clients and a resource (this resource is the information intermediary between consumers of medical services and establishments, which these services give);
3. Service function - giving of advisory services both in on-line and off-line mode, giving of the information about capabilities of diagnostics and treatment in different medical institutions, function of electronic payments and an electronic drugstore.

Web-site "Med-Health" suggested in this work based on the structure of a medical data-advisory Web-resource considered above. This web-resource has joined a plenty of ideas which in individual variants have been realized repeatedly, but together and in such volume is realized for the first time. The general block diagram of this web-resource is given on fig. 2. The block diagram contains all elements of a medical infrastructure and logically coordinates them in uniform cooperating structure.

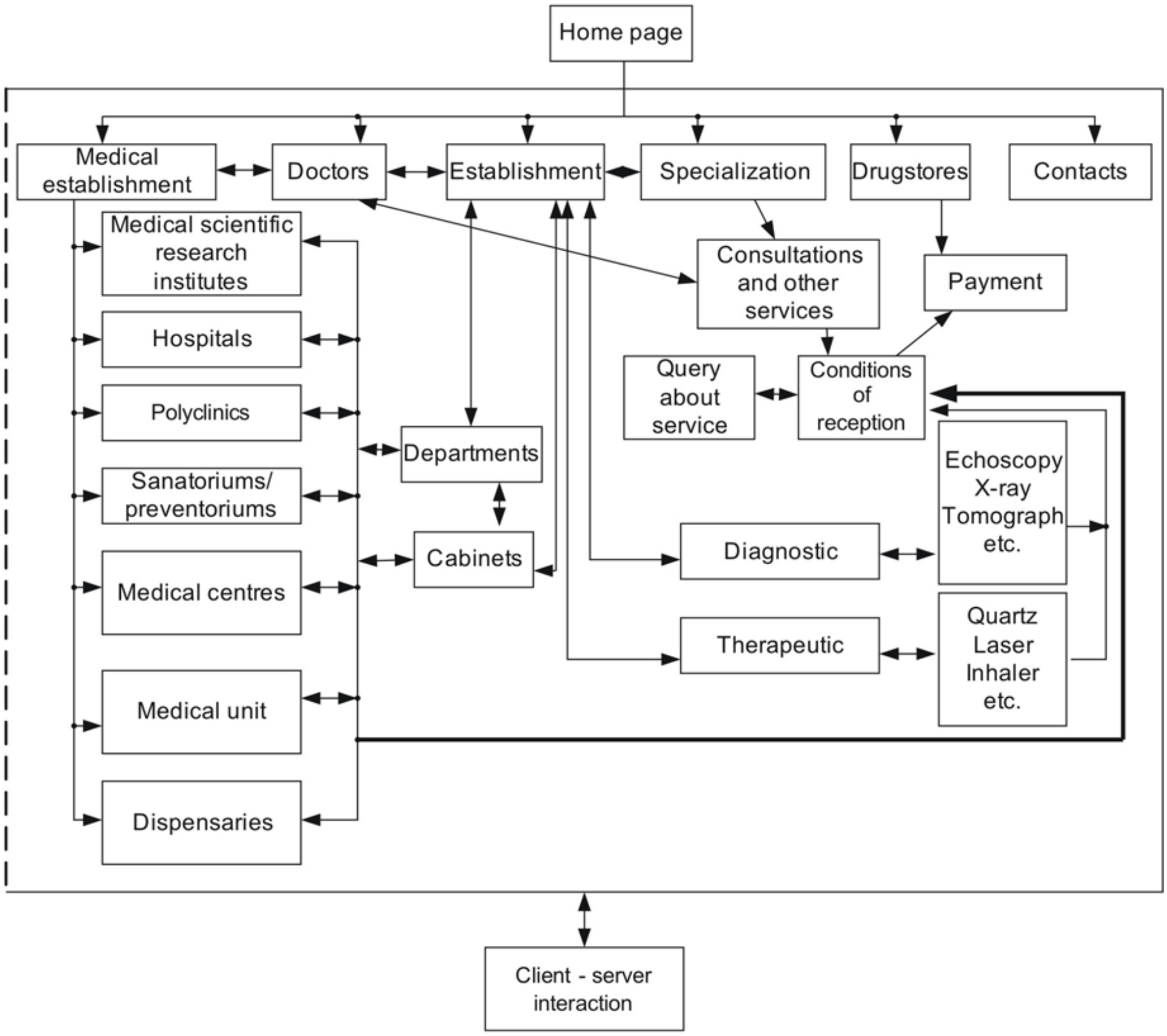


Figure 2 - The block diagram of the medical data-advisory web-resource “”

The site consists of six basic sections (fig. 2):

- medical establishments,
- doctors,
- medical equipment,
- medical specialization,
- drugstores,
- contacts.

The section “Medical establishments” comprises the information about all existent medical institutions (medical scientific research institutes; hospitals; polyclinics; sanatoriums / preventoriums; the medical centers; medical units; dispensaries. This section contains the information about a type of the establishment, its personnel; the contact information; the other help information concerning the medical establishment.

The section “Doctors” contains the personal information about doctors, a place of job, their medical specializations, the experience of job, the achievements, used methods of treatment, the contact information. The opportunity to find out about consultations and other kinds of the services offered by the doctor, including services online is given in this section.

The section “Medical specialization” includes the basic directions of diagnostics and the therapies selected in activity of medical establishments and doctors (for example, cardiology, surgery, ophthalmology, etc.).

The section "Medical equipment" contains the information about the equipment contained in concrete medical establishments, with the description of its capabilities and services given with this equipment; the information of firms-manufacturers about the medical equipment produced by them.

The section "Drugstore" includes all assortments of the goods of medical purpose offered by drugstores, as the information about an address of the drugstore. The opportunity of orders of the goods through the Internet with home delivery is given.

The section "Contacts" contains the information for communication with administration of the site at occurrence of questions or offers.

The connections shown on the block diagram of a resource (fig. 2) are entered for universality of a resource. The site is based on a unique universal database which allows to create inquiries with any logic interrelations, and not only what were projected originally, but also anyone additional which have arisen after designing a database. This provides flexibility and universality of database. The logic structure of a database is not defined by physically existing connections. The information is stored in a plenty of the interconnected tables with the maximal fragmentation for maintenance of flexibility of a database. The relevant database provides the multiuser access to this resource with differentiation of the rights of access [5-7].

The developed database consists of set of tables. Sets of the structured data are stored in these tables. For tables the indexes facilitating search of the necessary lines are created. The declarative restrictions providing reference integrity and, hence, a coordination of the interconnected data in various tables are added to tables. Storage in a database of the procedures intended for performance of some actions with a database for preservation of representations, providing the specialized access to the data of the table is stipulated also.

Network relations between a database and a Web-resource are under construction on the basis of architecture a client - server. The distributed structure of servers is more preferred as the fail of one of servers does not result in the termination of functioning of all Web-resource. Other reason for the decision of the organizational structure consisting of several servers is geographical separation of medical institutions. As consequence, the main server it is not attached rigidly to server stations of the basic level and can be established practically on any of existing operational systems on removed enough distance.

Filling of a database with the information is carried out as follows. Representatives of medical institutions of various type fill in specially developed questionnaires: the questionnaire of medical establishment, the questionnaire of department, the questionnaire of the doctor. The information from these questionnaires is placed in the certain sections of a database on which the site is based.

Access to the information of medical character which can be chosen by different criteria, building inquiries in a database, allows to receive instant answers to the put questions. The web-resource is multilanguage, that means, that a segment of users of the web-resource are users from all world. A resource is dynamical, that allows to update, change and correct its contents at any time.

Segmentation of users of a suggested web-resource:

– first of all it is people which activity is directly connected to medicine - doctors, medical science officers, manufacturers of the medical equipment and medical products;

– people who need the information on where is possible to pass diagnostic examination, treatment or to find out more information on this or that disease, methods of treatment, medicines, a way of their application and where this or that preparation can be got, how to register for the surgery, what is necessary for this purpose, or to carry out appointment or consultation on-line.

The resource contains huge volume of the information of medical character. These are medical articles, abstracts, researches of doctors, the description of various medical products, diseases and ways of their treatment. The information on a plenty of medical establishments of a various types is collected.

The web-resource allows, except for information search, to carry out medical teleconferences, to carry out consultations, register for the surgery, register for the examination.

The big attention in the resource is given to the friendly interface that the visitor who has got for the first time on the web-resource, could without problems understand navigation and easily find that information which is necessary for him.

For convenience of dialogue and safety of the confidential information a lot of measures and means is used [8,9]. First of all - differentiation of the rights of access of users. Enciphering of the data of the information which transfer or is received by users.

For a realization of monitoring, storage of the information on a state of health, results of examination, diagnoses and the other medical information the separate database which is based on relevant database Cache which has been initially conceived as base for storage, enciphering and processing of the medical information is used.

Access to the closed sections of this resource can be received with the help of a key which can be received only after accreditation [10]. The key has a binding to the person, to the IP-address of a computer, from which it comes, and to other parameters as much as possible to protect the web-site and the confidential information from the non-authorized access.

On a portal there is a separate line of consultations which it is possible to receive on-line. Constantly there is an on duty expert who can competently answer an interesting and exciting question.

The created Web-resource is designed as for those who cares of the health and interests in novelties in the field of medicine and for experts of a medical structure and other visitors. The user of a web-resource who is the consumer of the services given by medical institutions with the help of the offered web-site can receive if necessary the information about what medical institutions accessible to this user renders service required to him (service of diagnostic, therapeutic or advisory type), what conditions of rendering of this service, what type of the equipment is used there. Due to complete volume of the necessary information, the consumer chooses service optimally suitable for him on quality and availability. Medical institutions receive an opportunity to illuminate all spectrums of the services rendered by these establishments for a wide range of consumers of these services. Besides as on the same site firms-manufacturers of the medical equipment place advertising of their production, medical institutions get access to the information on the newest medical technologies, that allows to improve quality of medical services.

Conclusion

The offered medical information resource provides two-way contact between consumers of medical services and the establishments giving these services, and also manufacturers of the equipment and medical preparations. Complete resource of this kind is interest for various groups of users, first of all for those who wish to choose necessary and optimally accessible medical service among great number of the services given by different medical institutions. Besides this, resource enables medical institutions to give the information about them. Firms-manufacturers of the medical equipment and of the pharmacological preparations also receive an opportunity to acquaint the broad audience of consumers and experts with produced production that enables medical institutions to have the information on the newest medical technologies. The resource provides general infrastructure and basic components for research and development of medical and healthcare information systems with remote access. One of the main advantages of the web-resource “Med-Health” over miscellaneous other resources of such kind, is that the “Med-Health” is interesting both doctors and patients, and for manufacturers and researchers.

Bibliography

1. http:// www.doctor-alex.ua/dep_6.php
2. http:// www.nebosvod.com
3. http:// n-nov.mednet.com/resources.htm
4. http://www.weboptima.ru/med.htm
5. Джеймс Ли, Брент. Использование Linux, Apache, MySQL и PHP для разработки Web-приложений, 2004;
6. Поль Дюбуа. Применение MySQL и Perl в Web-приложениях, 2002;
7. Марк Шпеник, Оррин Следж. Руководство администратора баз данных Microsoft SQL Server 2000, 2001.
8. Скотт Бармен. Разработка правил информационной безопасности, 2002;
9. Вильям Столлингс. Основы защиты сетей. Приложения и стандарты, 2002;
10. Ричард Э. Смит. Аутентификация: от паролей до открытых ключей, 2002;

Authors' Information

A. I. Bykh – Doctor of Physics and Mathematics, professor, Head of Biomedical Electronics sub-faculty of Kharkov National University of Radio Electronics

E. V. Visotska – PhD, associate professor of Biomedical Electronics sub-faculty of Kharkov National University of Radio Electronics

T. V. Zhemchuzhkina – PhD, senior lecturer of Biomedical Electronics sub-faculty of Kharkov National University of Radio Electronics

A. P. Porvan – engineer of Biomedical Electronics sub-faculty of Kharkov National University of Radio Electronics

A. V. Zhook – student of Biomedical Electronics sub-faculty of Kharkov National University of Radio Electronics

Kharkov National University of Radio Electronics, Ukraine, 61166, Lenin Avenue, 14, Biomedical Electronics sub-faculty, e-mail: diagnost@kture.kharkov.ua

OPEN SOURCE INFORMATION TECHNOLOGIES APPROACH FOR MODDELING OF ANKLE-FOOT ORTHOSIS

Slavyana Milusheva, Stefan Karastanev, Yuli Toshev

Abstract: Computer modeling is a perspective method for optimal design of prosthesis and orthoses. The study is oriented to develop modular ankle foot orthosis (MAFO) to assist the very frequently observed gait abnormalities relating the human ankle-foot complex using CAD modeling. The main goal is to assist the ankle-foot flexors and extensors during the gait cycle (stance and swing) using torsion spring.

Utilizing 3D modeling and animating open source software (Blender 3D), it is possible to generate artificially different kind of normal and abnormal gaits and investigate and adjust the assistive modular spring driven ankle foot orthosis.

Keywords: biomechanics; 3D computer modeling, ankle-foot orthosis

ACM Classification Keywords: I.3.7 Three-Dimensional Graphics and Realism, I.6.5 Model Development

Introduction

Open source software refers to computer software available with its source code and under an open source license. Such a license permits anyone to study, change, and improve the software, and to distribute the unmodified or modified software. It is the most prominent example of open source development. This software gives an outstanding flexibility in terms of extensibility and modularity.

The study is based of 3D modeling technology provided by one of the most advanced open source software – **Blender.** Blender is a free 3D modeler program. It is used for modeling and rendering three-dimensional graphics and animations. Blender is available for several operating systems, including FreeBSD, IRIX, GNU/Linux, Microsoft Windows, Mac OS X, Solaris, SkyOS, and MorphOS. In addition, Blender's recent burst of new features in the last few versions has actually brought it close in feature set comparison to high-end 3D software such as 3D Studio Max and Maya. Among these features and user interface ideas are, for example, complex fluid and cloth effects, a comprehensive and well-thought out hotkey program, which rivals that of most higher end applications, and a wide range of easily accessible and creatable extensions using Python scripting. Regardless of lack of natively implemented CAD functionality there are a lot of possibilities for development of Python based helper scripts for precise engineer modeling.

Ankle-foot orthosis (AFO) is commonly used to help subjects with weakness of ankle dorsiflexor muscles due to peripheral or central nervous system disorders. Both these disorders are due to the weakness of the tibialis anterior muscle which results in lack of dorsiflexion assist moment. The deformity and muscle weakness of one joint in the lower extremity influences the stability of the adjacent joints, thereby requiring compensatory adaptation.

During level plane ambulation the ankle should be close to a neutral position (right angle) each time the foot strikes the floor. Insufficient dorsiflexion may be the result of hyperactive plantarflexors that produce very high plantarflexion moment at the ankle, or weakness of the dorsiflexion muscles. This affects the ability of the ankle to dorsiflex. As result the patient make a forefoot contact instead of normal “heel-strike”. If there is a weak push-off, the stride length reduces, and the gait velocity decreases. Similarly, during the gait swing phase, the ankle is dorsiflexed to allow the foot to clear the ground while the extremity is advanced. Hyperactive or weak dorsiflexors may result in insufficient dorsiflexion, which must be compensated by alterations in the gait patterns so that the toes do not drag. This insufficient dorsiflexion during the gait swing phase is termed as “foot-drop”. In addition to the toes dragging, the foot may become abnormally supinated, which may result in an ankle sprain or fracture, when the weight is applied to the limb. Foot-drop is commonly observed in subjects after a stroke or personal nerve injury.

There are several possible treatments for foot-drop - medicinal, orthotic, or surgical. It is to note that the most common is the orthotic treatment. Orthotic devices are intended to support the ankle, to correct deformities, and to prevent further occurrences. A key goal of orthotic treatment is to assist the patient achieving normal gait patterns.

Different orthoses are used to enhance the ankle-foot position and mobility. The most common types are hingeless and hinge orthoses. Using springs, the hinge orthoses could assist ankle flexion/extension during gait, i.e. they are pseudo-active orthotic devices. The standard ankle foot orthoses (AFO) is a rigid polypropylene structure which prevents any ankle motion.

Methods

The study is oriented to develop modular ankle foot orthosis (MAFO) with two units (shank brace and foot brace) connected with lateral and medial adjustable hinged joints.

Gait analysis

Gait analysis is useful in objective documentation of walking ability as well as identifying the underlying causes for walking abnormalities in patients with cerebral palsy, stroke, head injury and other neuromuscular problems. The results of gait analysis are useful in determining the best course of treatment in these patients.

Normal gait

The gait cycle begins when one foot contacts the ground and ends when that foot contacts the ground again. Thus, each cycle begins at initial contact with a stance phase and proceeds through a swing phase until the cycle ends with the limb's next initial contact. Stance phase accounts for approximately 60 percent, and swing phase for approximately 40 percent, of a single gait cycle.

Each gait cycle includes two periods when both feet are on the ground. The first period of double limb support begins at initial contact, and lasts for the first 10 to 12 percent of the cycle. The second period of double limb support occurs in the final 10 to 12 percent of stance phase. As the stance limb prepares to leave the ground, the opposite limb contacts the ground and accepts the body's weight. The two periods of double limb support account for 20 to 24 percent of the gait cycle's total duration.

Stance phase of gait is divided into four periods: loading response, midstance, terminal stance, and preswing. Swing phase is divided into three periods: initial swing, midswing, and terminal swing. The beginning and and ending of each period are defined by specific events.

Each subphase is accompanied by a change in position, ground reaction force, and/or internal muscular forces. Gait cycle analysis in this sense is essentially a sagittal plane function.

The ankle is plantarflexed 10 degrees at heelstrike, with further plantorflexion dampened by the ankle dorsiflexors, aiding with shock absorption. At midstance, ground reaction tends to dorsiflex the ankle which is held rigid by the plantarflexors, controlling forward thrust of the tibia. Ground reaction continues to push the ankle toward dorsiflexion in terminal stance, resisted by the plantarflexors. The ankle passively dorsiflexes as it is unloaded in preswing.

Ankle joint motion (sagittal plane):

- between initial heel contact and foot flat ankle undergoes ~3-4° plantar flexion (first 6% of stride);
- after foot flat ankle dorsiflexes until a little beyond 40% of stride (as hip moves over ankle), reaching maximum of 8-10°;
- ankle plantar flexes for remainder of stance phase until after push-off (reaches maximum plantar flexion of 16-19° just after toe-off);
- after push-off ankle rapidly dorsiflexes during early swing for toe clearance;
- ankle dorsiflexion slows or stops during mid-swing but may continue to dorsiflex slightly in late swing until just prior to heel contact when plantarflexion begins.

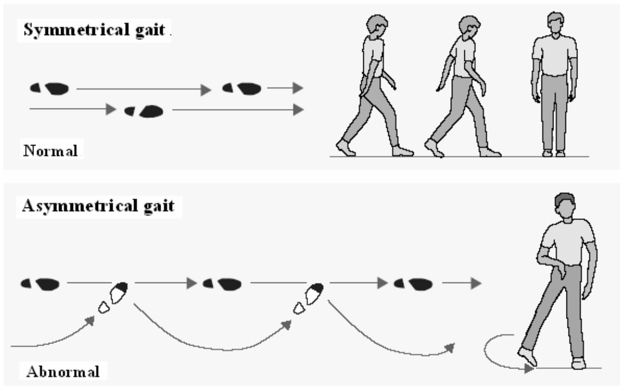


Fig.1.Gait analysis

Abnormal gait

Pathological gait describes altered gait patterns that have been affected by deformity (usually in the form of contractures), muscle weakness, impaired motor control. Any alteration affecting one or more motion or timing pattern can create a pathological gait pattern.

Deviations to normal gait patterns can be observed during both swing and stance phases and requires systematic evaluation for assessment of functional compensations and/or neuromuscular-skeletal factors. Functional compensations are voluntary posturings that attempt to substitute for specific motor weaknesses and joint instabilities. It is important to identify functional compensations from imposed mechanisms for appropriate orthotic design and therapeutic considerations.

Gait analysis can be used to evaluate more objectively the dynamic basis for an observed gait deviation in the patient requiring a lower limb orthosis. It also can be a valuable tool in objectively assessing the impact of different orthotic interventions.

Ankle-foot orthosis

A standard polypropylene AFO is a rigid polypropylene structure which prevents any ankle motion.

Different orthoses are used to enhance the ankle-foot position and mobility. The most common types are hingeless and hinge orthoses. Using springs, the hinge orthoses could assist ankle flexion/extension during gait, i.e. they are pseudo-active orthotic devices.

Which are to assist the quite popular gait abnormalities inherent to a spring-controlled human ankle-foot complex.

Previous studies have shown the DACS (dorsiflexion assist controlled by spring) AFO to have the following desirable characteristics: 1) the magnitude of the dorsiflexion assist moment and the initial ankle angle of the AFO can be changed easily, and 2) no plantarflexion assist moment is generated.

Torsion springs

Torsion springs (Fig.2) can store and release angular energy or statically hold a mechanism in place by deflecting the legs about the body centerline axis. They offer resistance to twist or rotationally applied force.

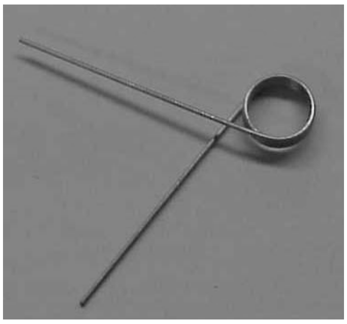


Fig.2. Torsion springs

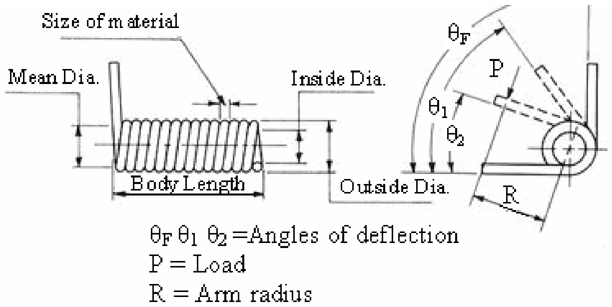


Fig. 3.Linear characteristic of the spring

Figure 3. shows the linear characteristic of the torsion spring

$M_{spring} = -k\theta$ linear spring

where M_{spring} is the spring torque, k is the spring constant and θ is the angular displacement from its rest angle

Dynamic equilibrium at joints

For normal leg, the dynamic equilibrium at each joint can be expressed as:

$M_i = M_g + M_s + M_d + M_a$

where M_i, M_g, M_s, M_d and M_a represent the torque due to moment of inertia of the rotating segment, gravity, joint stiffness, joint viscosity and muscle activation respectively. As there will be no muscle activity during the period considered, M_a becomes zero.

Results

Using the advanced open source 3D modeler (Blender3D) with outstanding user script capabilities (by Python script language extensions), different 3D computer solid models of MAFO were developed (Fig.4, Fig.5, Fig.6 and Fig.7).

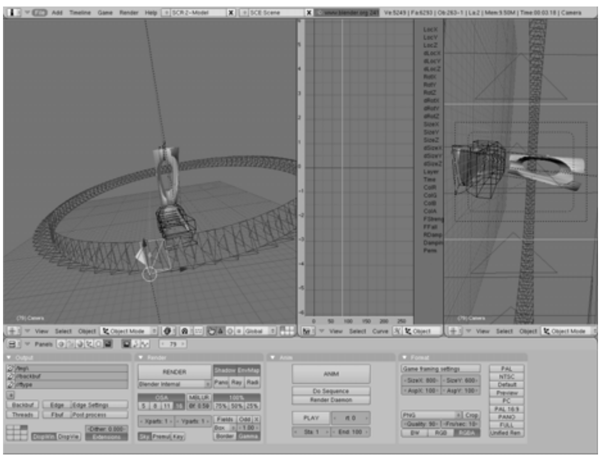


Fig.4. Dump screen of Blender3D during modeling process

Fig.5. 3D computer solid model of MAFO normal state

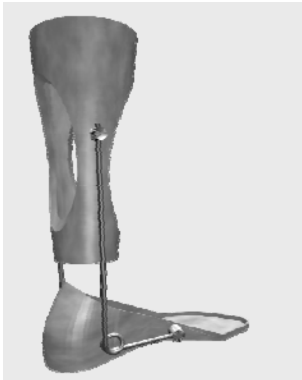


Fig.6. 3D computer solid model of MAFO in state of tension includes two parts: lower part (foot) and upper part (calf) with two torsion spring (lateral and medial).

Fig.7. 3D model of the human ankle-foot segment with a model of orthoses.

The main idea was to design two personalized AFO parts - lower (foot) and upper (calf), using 3D human model (artificially generated by specialized Blender3d script).

On the base of the obtained 3D surfaces two personalized AFO parts were designed and different variants of elastic elements (torsion spring) connecting the two parts of the orthosis.

A modular ankle-foot orthosis (MAFO) with one degree-of-freedom (dorsiflexion-plantarflexion motion) has been developed. The flexion/extension is controlled by springs. Dorsiflexion correction is achieved via the compression force of springs within the assistive device.

The big advantage of the 3D model is the possibility for further dynamics and kinematics development in field of more precise simulation of the real orthosis behavior. This could be achieved by combination of build-in physics engine of Blender3D, which covers both rigid and soft body simulation and user developed scripts.

Discussion

The magnitude of the MAFO dorsiflexion assist moment and the initial ankle angle can be modified by variation of the spring parameters (spring constant, spring rest angle). Regardless the simplicity of MAFO, the results in improvement of plantarflexion during swing phase are similar to the results obtain using commercial AFOs. The proposed modular ankle-foot orthosis is currently under additional mechanical durability tests. Continuous plantarflexion was applied to MAFO to check the durability of each part. At present, more than two million repetitions of plantar flexion have been applied and no serious problems have arisen. The results of the durability test will be use to improve the design of MAFO. Our MAFO can be used by the patients daily, and is also useful for gait training, since various characteristics can be easily modified. Moderately large dorsiflexion assist moment and small dorsiflexion initial ankle angle facilitates the increase of knee extension muscle forces, thus preventing forward foot slap during the initial stance phase.

References

Milusheva S., Tochev D., Karastanev S. (2005) Ankle-foot orthosis with spring elements. Proceedings of the 10th Congress on Theoretical and Applied Mechanics, Varna, pp. 145-149

http://www.fleshandbones.com/readingroom/pdf/1162.pdf

Acknowlegements

The study was supported by grant TN 1407/04 - Bulgarian National Research Fund.

Authors' Information

Slavyana Milusheva - Institute of Mechanics and Biomechanics, BAS, Acad.G.Bonthev St., bl.4, Sofia-1113, Bulgaria; e-mail: slavyana@imbm.bas.bg

Stefan Karastanev - Institute of Mechanics and Biomechanics, BAS, Acad.G.Bonthev St., bl.4, Sofia-1113, Bulgaria; e-mail: stefan@info.imbm.bas.bg

Yuli Toshev - Institute of Mechanics and Biomechanics, BAS, Acad.G.Bonthev St., bl.4, Sofia-1113, Bulgaria; e-mail: ytoshev@imbm.bas.bg

APPLICATION OF THE ARTIFICIAL INTELLIGENCE ALGORITHMS FOR SYSTEM ANALYSIS OF MULTI DIMENSION PHYSIOLOGICAL DATA FOR DEVELOPING POLYPARAMETRIC INFORMATION SYSTEM OF PUBLIC HEALTH DIAGNOSTICS

Nina Dmitrieva, Oleg Glazachev

Abstract: The polyparametric intelligence information system for diagnostics human functional state in medicine and public health is developed. The essence of the system consists in polyparametric describing of human functional state with the unified set of physiological parameters and using the polyparametric cognitive model developed as the tool for a system analysis of multitude data and diagnostics of a human functional state. The model is developed on the basis of general principles geometry and symmetry by algorithms of artificial intelligence systems. The architecture of the system is represented. The model allows analyzing traditional signs - absolute values of electrophysiological parameters and new signs generated by the model – relationships of ones. The classification of physiological multidimensional data is made with a transformer of the model. The results are presented to a physician in a form of visual graph – a pattern individual functional state. This graph allows performing clinical syndrome analysis. A level of human functional state is defined in the case of the developed standard ("ideal") functional state. The complete formalization of results makes it possible to accumulate physiological data and to analyze them by mathematics methods.

ACM Classification Keywords: J.3 Life and Medical Sciences: Health. Medical information systems. I.3.5 Computational Geometry and Object Modeling. I.5.1 Models Geometric.

Introduction

One of problems of the contemporary preventive medicine is the development of an informational system of health diagnostics, which could enable to conduct a system analysis of multitude data, while could be comparable with existing clinical functional diagnostics and corresponding to the modern requirements to medical information systems [Hummel et al. 2000]. The experience obtained by us through the use of the visualized patterns and graphic modeling of functional states of an organism under activity of physiological substances [Dmitrieva et al. 1982] created the basis for development of the polyparametric method for evaluation of a human functional state in terms of the pattern recognition theory [Dmitrieva et al. 1989]. Patient data are presented in graphical formats as visual patterns, which permit to interpret these data in clinical-physiological terms. According to the recommendations of the Word Health Organization we have conducted the comparative research of a health state of students by polyparametric and clinical physiological methods [Dmitrieva, Glazachev, 2000]. These results demonstrated advantages and disadvantages of polyparametric method and lead us to development of new model on the basis of an artificial intelligence algorithms to improve one [Pospelov,1992; Zenkin, 1991].

Case-Based Reasoning

The gist of the polyparametric method for diagnostics of a human functional state consists in polygraphic recording (0.5 minutes) and data processing of objective physiological characteristics (electrocardiogram, electromiogram, electrovasogram and others), parameterization of analog signals, polyparametric description of a functional state with the unified set of the time - amplitude parameters, using artificial intelligence algorithms and graphical modeling and methods of pattern recognition for an analysis of multi dimension data on line mode. Necessity and sufficiency of the set of parameters for a description of functional state have been grounded earlier. The novelty of the new variant of the polyparametric method contains in the original polyparametric cognitive model presenting the intelligence image system as the tool for the system analysis of multi dimension data.

The Intelligence Image model. Absolute values of the unified set of physiological parameters mentioned above are represented as vectors in the system of polar coordinate (Fig.1). Each parameter has its own scale determined by modal level (middle circle). The contour limited with external and internal circumferences (maximal and minimal values of the parameters without pathognomonic signs) is the intelligence transformer performing analysis and classification of every parameter and whole shape, for example nosologic diagnosis. The active part of the intelligence transformer provides relationships of parameters as additional new signs generating new knowledge about a subject.

The changes of sympathetic or parasympathetic regulation are reflected in a displacement of a pattern (dotted circumference on fig.1) to the left and to the right correspondingly.

The model of "ideal functional state" is characterized by the invariant relationships of all parameters (fig.2, left top). This model and patterns of individual functional states are constructed on the basis of the general intelligence model.

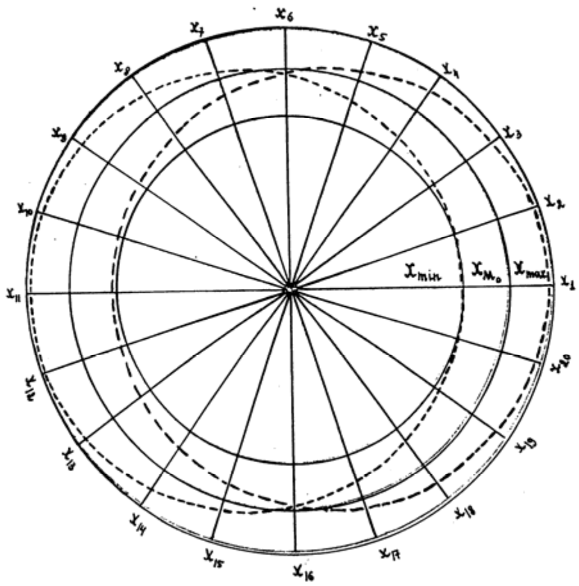


Fig.1. Artificial intelligence image model for analysis of physiological sings.
(Vectors X_1 – X_{20} are physiological parameters; the method of construction is described in the text).

The results of the polyparametric examination are presented to physician in tabular form and as a pattern of functional state. On fig.2 there are 4 protocols of polyparametric examination patients with different functional sate: the first column is the list of physiological parameters and theirs values (in physical dimensions), the second column represents the relationships of parameters giving as the values of deviation (percents) from the invariant, visual graph (the pattern) of individual functional state enables performing clinical analysis of multiple data in interactive mode.

The polyparametric method allows evaluating level of a functional state in on-line mode during 3-5 minutes.The main characteristic of satisfactory state class is relationships of parameters closed to invariant: deviation is less than 5-7% whereas the absolute values of parameters can be in wide range between maximal and minimal rings. Thus the pattern of satisfactory functional state has a round form or similar to it. The deeper the adaptation syndrome the greater the misbalance of parameters is become. The pattern state assumes an irregular form of a different kind (Fig.2). It means that parameter relationships of the vital physiological functions are supplementary diagnostics signs of changes of human functional states. This is the new knowledge about information connections of physiological functions.

The state patterns are classified on decision rules in the PC programs and graded into four classes according to the main stages of adaptation processes development The results of the polyparametric examination according to the classes of functional state were controlled by criterion χ^2 (by program “S-Plus 2000 professional”) and discriminative analysis (table 1).

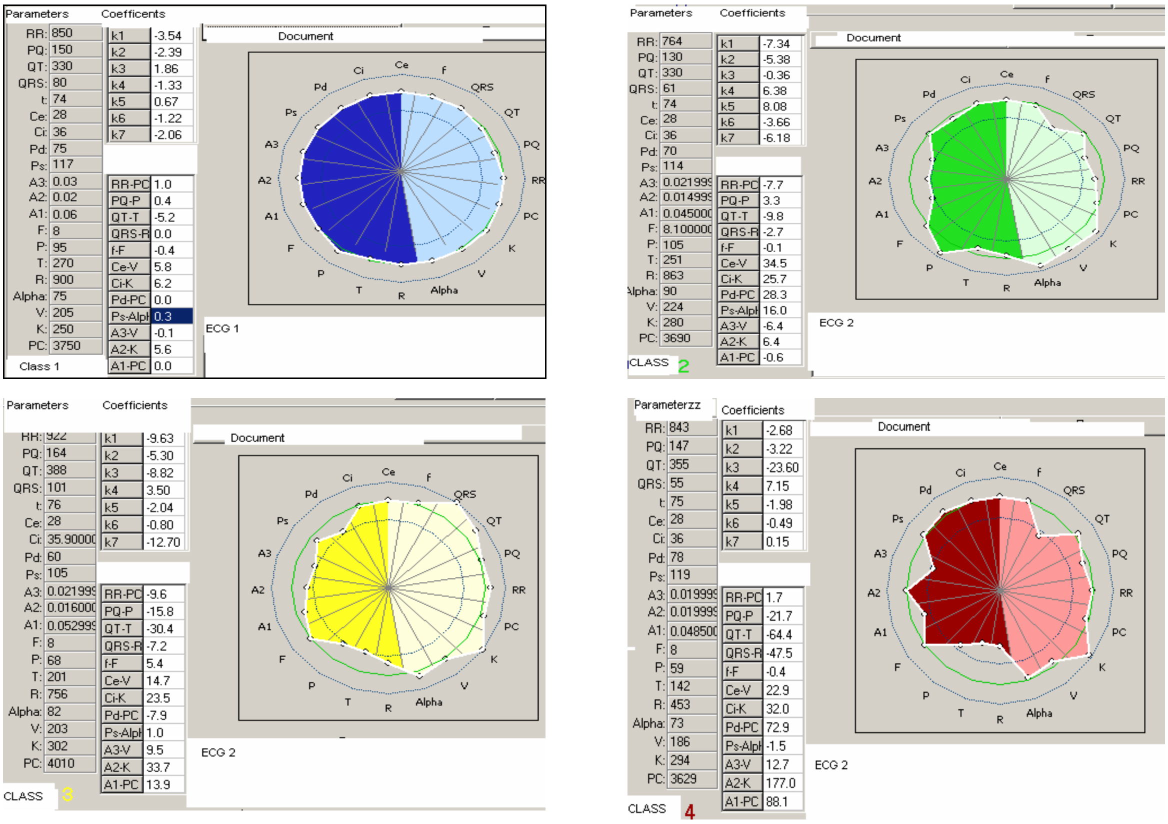


Fig.2. Patterns of different functional states. (1– satisfactory functional state, 2 – strained functional state, 3 – overstrained functional state, 4 – stress. Description of structure is in text)

Table1. Differences between functional classes by criterion χ^2 (by program “S-Plus 2000 professional”)

Classes	4-1	4-2	4-3	3-1	3-2	2-1
Criteria	113.6	82.5	33.95	44.86	26.35	97.2
Mean	0.001	0.001	0.0034	0.001	0.034	0.009

Thus, with χ^2 criterions the main stages of adaptation process have objective differences between classes of functional states.

The discriminative analysis of the polyparametric data has confirmed their subdivision into 4 prescribed classes of functional states with satisfactory differences (under 9%).

As using parameters allow characterizing the autonomic regulation, the patterns of functional states can be considered as syndrome of autonomic status [Veyn, 1998]. For a definition of autonomic regulation every time parameter in the pattern marked with light color and amplitude parameters are darkened. This gives possibility for a physician to definite autonomic status ease and quickly [Dmitrieva, 1999].

Special investigations were found out significant individual differences of the patterns. The results of statistical analysis (mean values, mode, standard deviation and coefficient of variation) of polyparametric data demonstrated highly variable of some parameters for different people. It means the number of combinations of the changed signs can be high. The research of individual variations revealed that they can be satisfactorily systematized into major classes of states in respect of the standard model.

In support of that the polyparametric data were analyzed with the cluster method by the strategy of Word. On the Fig.3 the results of cluster analysis of polyparametric documents (on left) and statistic refined data (on right) are represented.

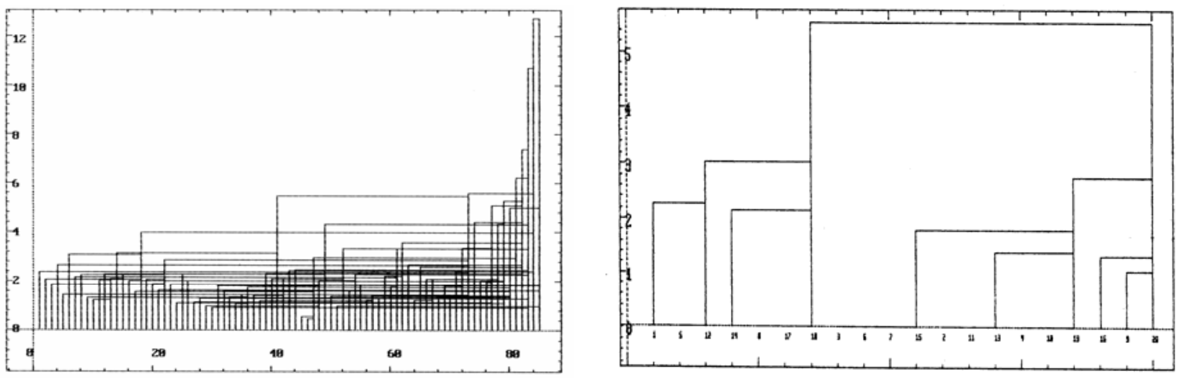


Fig3. Cluster analysis of documents of polyparametric examinations of students (axis X –individual documents, axis Y –cluster ranges).

There were singled out three main and four added clusters. The interpretation of clusters was performed by the visual analysis of individual documents. It was found out that same people examined in different time formed same clusters. Thus there were demonstrated that some syndromes are stable because of their patterns are fluctuated close to errors of a measure of parameters (± 7-8 %). Thus the reproducibility of derived parameters and their relationships has been confirmed. It was shown that a high formalization of the obtained results of the polyparametric examination makes to do systematization and mathematical analysis of multidimensional physiological data.

The patterns of typical adaptation syndromes were selected for using in the information support of physician decisions in diagnostics of a person adaptation syndrome [Seley H. 1976]. These patterns were fixed in phase 2 (Fig.4) to use them as the conceptual models of different adaptation syndromes for comparative analysis of newcomer patterns. The gist of this procedure is “data mining”. The scheme of DATA MINING of polyparametric technology is represented on fig.4.

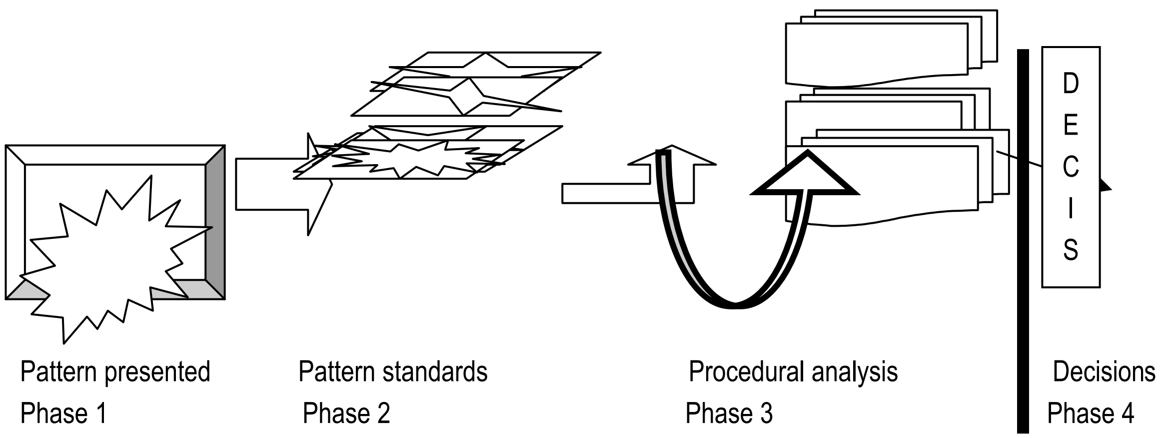


Fig.4. The scheme of polyparametric technology of diagnostics of multitude human functional states.

The phase 3 of the technology is intended for a predictable decision by using the known rules of procedure “if S then A” [Newell, Simon, 1972]. It is easy to see changes of any pattern and to interpret a dynamic of a functional state using these rules working with the model.

But the obtained results do not allow making a definite conclusion about exact role of concrete signs in the forming of pattern.

The polyparametric method and technology are open for further development and improvement based on the bank-accumulated data of polyparametric examinations and evaluations of a functional state under therapeutics and correction mode.

Application of the polyparametric method in condition of the comparative analysis of the results of examination of students with the clinical methods was conducted. The findings demonstrated that 44% of students during their term are in a state of overstrain and 40% in adaptation failure according to the classification of stages of adaptation development processes. Thus the comparative analysis showed the good correlation and correspondence.

Conclusion

The information intelligence image polyparametric system represents the instrument to analyze multidimensional data, performing knowledge engineering and data mining. It allows very quickly define level of a human functional state relatively to the developed standard of functional state. The image of functional state is very easy for interpretation. Complete formalization of the polyparametric results makes it possible to accumulate data and analyze them by math methods. This tool allows determining and evaluating relationships of electrophysiological parameters, which became a new diagnostics signs. The conceptual model of standard (“ideal”) functional state was formed. This model allows to circumvent the indefinite notions of “norm” and “ordinary man” and work out the technique for measurement of various deviations from the ”ideal functional state” as changes of the functional state.

Today the first experimental model of computer polyparametric system is installed in the Moscow State University, where the health of students has been examined for some years. The polyparametric method is open for further development of the preclinical diagnostics of functional disorders on the “training-with-a teacher” basis.

Bibliography

[Hummel et al.2000] Hummel M, van Rossum W., Verkerke G., Rakhorst G. Assessing medical technologies in development. //Intern. J. of Technology Assessment in Health Care. 2000; 16:4.

[Dmitrieva et al. 1982] Dmitrieva N.V., Nizhnii S.V., Ivanova I.V. Chemical (drug) stress – qualitative assessment of toxic effect of physiologically active substances // Izvestiya AN SSSR. Ser. Biol. – 1982. – N 3. – P. 398

[Dmitrieva et al. 1989] Dmitrieva N.V., Voronov E.B., Yakovlev U.V. at al. The Polyparametric method of evaluation of human functional state with image recognition method. // Phisiologiay cheloveka. 1989. 4. P. 103-112.

[Dmitrieva, Glazachev, 2000] Dmitrieva N. V., Glazachev O.S. Individual Health and Polyparametric Diagnosis of Organism’s functional state. M.: Gorizont, 2000. – 214 p.

[Pospelov, 1992] Pospelov D.A. Cognitive graphic – a window in new world // News Art, Intel.. 1992 .3.4.

[Zenkin, 1991] Zenkin A.A. The cognitive computer’s graphic. “Nauka”, Novosybirsk. 1991. -186p.

[Dmitrieva, 1999] Dmitrieva N.V. Syndrome analysis of polyparametric images of functional states of organisms // News of Artificial Intellect. – 1999. – N 1. – P. 120–129.

[Seley, 1976] H. Seley, Stress in health and disease. Boston-London, 1976, -1256 p.

[Newell, Simon, 1972] Newell A., Simon H. Human problems solving. Englewood Cliffs: Prentice Hall, 1972.

Author’s Information

Nina V. Dmitrieva – M.D. Prof. Head of the System Analysis Laboratory Institute of Normal Physiology Russian Academy of Medical Sciences, Mochovay str.11.Moscow, 125009 Russia; e-mail: nvdmitrieva@mtu-net.ru

Oleg S. Glazachev - M.D. Prof. of Moscow Medical Academy, , Mochovay str.11.Moscow, 125009 Russia.

DISTINCTIVE FEATURES OF MOBILE MESSAGES PROCESSING

Ken Braithwaite, Mark Lishman, Vladimir Lovitskii, David Traynor

Abstract: World's mobile market pushes past 2 billion lines in 2005. Success in these competitive markets requires operational excellence with product and service innovation to improve the mobile performance. Mobile users very often prefer to send a mobile instant message or text messages rather than talking on a mobile. Well developed "written speech analysis" does not work not only with "verbal speech" but also with "mobile text messages". The main purpose of our paper is, firstly, to highlight the problems of mobile text messages processing and, secondly, to show the possible ways of solving these problems.

Keywords: mobile text messages, text message analysis, natural language processing

ACM Classification Keywords: I.2 Artificial intelligence: I.2.7 Natural Language Processing – Text Analysis

Introduction

1. The reasons why is very difficult to use the classical linguistic approach for verbal speech analysis have been considered in [1]. In this paper the problems of Mobile Short Message (MSM) analysis will be discussed. MSM represents plain text message of 160 characters or less and provided by mobile SMS (short message service). The year 2005 saw an explosion in the volume of MSM being sent to mobile phones. Mobile's users choose to send MSM rather than talking on a mobile call because [2]:
   - They don't have time to chat phone (74%).
   - To not disturb other patrons on public transportation or at a sporting event or restaurant (53%).
   - To get work done and send quick notes when on the road travelling for business (32%).
   - Less disturbing than phone calls (72.5%).
   - One can reach the other party around the clock (30.4%).

   However, mobile operators need to understand that subscribers give greater priority to the convenience of using the service over the technology and capabilities it offers. Therefore, more effort must be placed on creating user-friendly client interfaces that integrate effectively with the handset features.
2. A wide variety of information services can be provided by SMS, including weather reports, traffic information, inventory management, itinerary confirmation, sales order processing, asset tracking, automatic vehicle location, entertainment information (e.g., cinema, theatre, concerts), financial information (e.g., stock quotes, exchange rates, banking, brokerage services), and directory assistance. SMS can support both *push* (i.e. mobile-terminated (MT)) SM and pull (i.e. mobile-originated (MO)) SM to allow not only delivery under specific conditions but also delivery on demand, as a response to a request.
3. The important distinctive feature of MSM is that the majority of them are bilingual (i.e., using both English words and mobile slang from Tegic's T9 dictionary [3]).
4. We will consider MSM in indissoluble link with Inbound Number (INo) represented by a short code (it is typically a 5 digit number which is accessible by subscribers of any mobile operator) or long code (a usual mobile number– works across all operators).
5. Information services as described above are provided by "Content Providers" who must rent an INo. This can be dedicated to provide a single service or shared to provide multiple services. In they case of multiple services, they are distinguished by the use of a key word that user must provide as the first word of the MSM.
6. The standard 12-key keypad found on many mobile phones today (see Figure 1). On this Figure "Imitator of Mobile" is represented. Alphabetic letters are mapped to keys '2' through '9'. However, this arrangement poses problems for text entry. As three or four letters share the same key, some form of disambiguation is required to determine which letter is intended by the user. There are currently two main methods that are

usually used on mobile phones for text entry. They are the multi-tap method and the predictive text entry method. In the multi-tap method, a user taps the key that contains the letter repeatedly until the desired letter appears. The number of taps required depends on the position of the letter on the key. In predictive text input method (e.g., Tegic's T9 [3]), the user presses the key that corresponds to each letter of a word once. The system uses a dictionary of words to determine which of the possible words the key sequence matches. When MSM is received on a particular INo, then for a dedicated INo the MSM is forwarded to the client renting it. If the INo is shared, the MSM needs to be examined to identify the client and the individual service.

7. First we will describe the types of MSM and the problems encountered examining the MSM. The MSM might be represented by:

- **Letter** or **digit**. For example, a number of promotions are quizzes/competitions and sometimes are also interactive, i.e., multiple messages/responses. If the original message to the customer is a question, such as *"How many legs has my dog got?"* then the customer could reply *1*, *2*, *3*, or *4*. Some promotions are multi-choice answers e.g., *'a'*, *'b'*, or *'c'*.
- **Single word** or **number** (e.g. credit card number).
- **Sequence of words** or **numbers.**
- **Combination of words** and **numbers** in MSM.

The main purpose of this paper is to investigate the **bad pairs INo ↔ MSM** and find ways to restore them.

Figure 1. Standard 12-keys keypad

Let's call pair INo ↔ MSM *bad* if:

- INo does not exist;
- Type of MSM was not recognised or keyword of MSM was not recognised. Very often the first whitespace-delimited word represents keyword (KW) and allows the identification of the client;
- The pair INo ↔ MSM does not exist because (¬INo & MSM) ∨ (INo & ¬MSM),

  where ¬INo and ¬MSM stand for *wrong* INo and *wrong* MSM respectively. Let's call INo and MSM *wrong* if they separately exist but link between INo and KW of MSM does not.The reason of wrong MSM is understandable. For example, a user can tap the 2-key once to get 'a', twice to get 'b' and thrice to get 'c'. If he taped wrongly then instead of desired word *bell* he typed *cell*, or using 6-key instead of *come* was *cone*.
- A special type of MSM (so called *stop MSM)* requires synonyms for recognition e.g., *cancel*, *remove*, etc.
- Finally, we would like to underline the most difficult and dangerous problem when INo ↔ MSM exists but

  $$((INo^T \neq INo^D) \,\&\, (KW^T = KW^D)) \vee ((INo^T = INo^D) \,\&\, (KW^T \neq KW^D)) \vee ((INo^T \neq INo^D) \,\&\, (KW^T \neq KW^D)),$$

  where letters *D* and *T* mean what user *desired* to type and what was actually *typed*.

This problem takes place because of ambiguity of both INo and KW i.e., one INo might link to several KW and many different INo might use the same KW, and vice versa.

Let's investigate these problems and discuss the results of KW, INo and bad MSM analysis. Our investigation was grounded in real data analysis. As a result of this discussion an algorithm to deduce the correct KW from a bad MSM will be described. Also, the result of using of this algorithm will be shown.

Keywords Analysis

The result of KW analysis and KW ambiguity is shown on Figure 2, namely:

- Total (valid + invalid) KW distribution among letters and mobile's keys (2-9). A KW is *invalid* if it currently is not used on the INo but at the same time the same KW might be valid for another INo. For example, KW *red* is valid for INo 81025 and 80039, and invalid for 89095;
- Displaying the list of KW for selected letter or Inbound No by clicking the corresponding letter or digit;
- For any KW (by clicking when the list of KW is displayed, or just simply typing in KW) the corresponding list of INo is displayed.;
- List of the next (= expected) symbols is displayed for the entered symbol (letter or digit);
- List of ambiguity for both valid and invalid KW is displayed.
- INo ambiguity is shown on Figure 3.

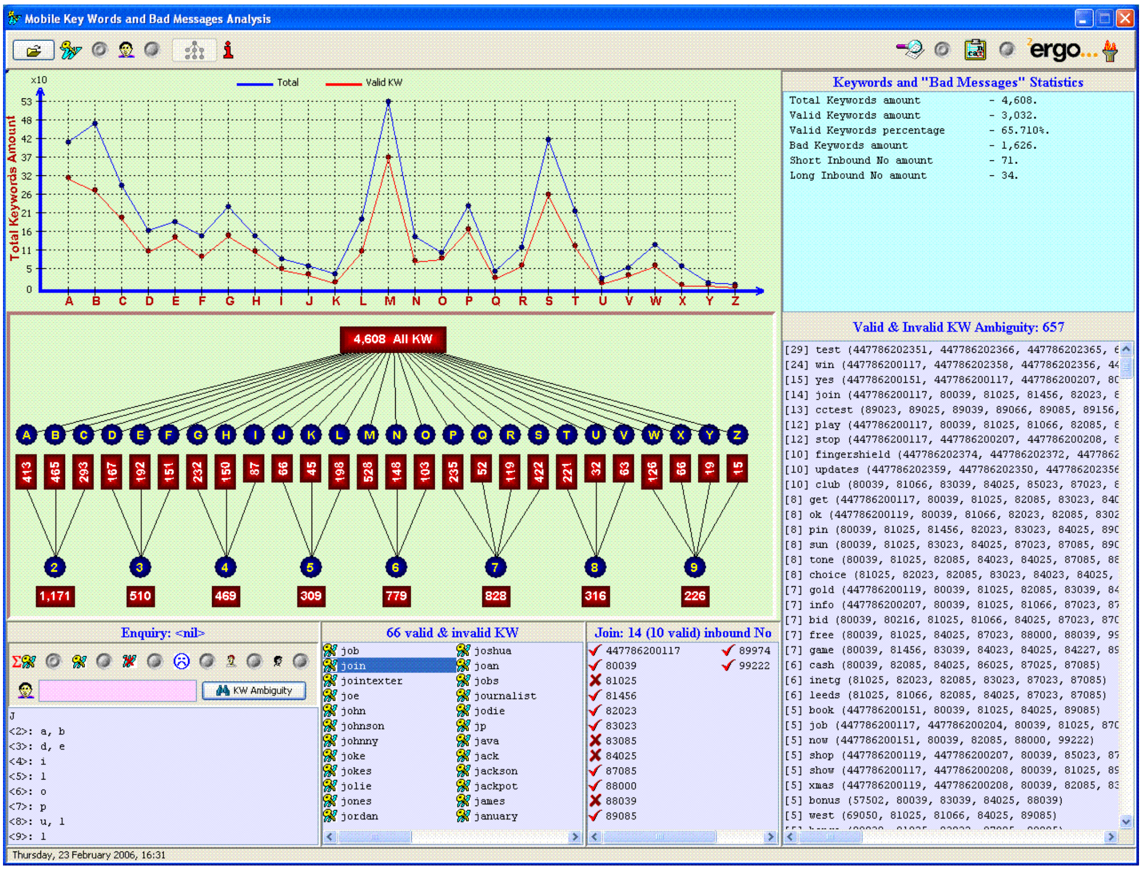


Figure 2. Keywords analysis and KW ambiguity

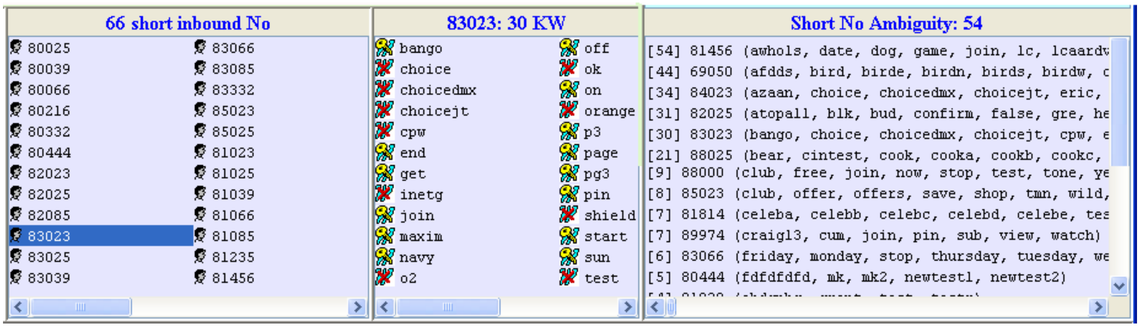


Figure 3. INo ambiguity

To provide such analysis the Knowledge Base (KB) has been created and used for KW, INo and bad MSM analysis, and KW and INo restoration. The main features of KB have been discussed in details in [4]. Here we would like to notice that in our case, under the **KB organisation** we would understand the regularity of data (INo and KW) distribution in memory assuring the storage of various links between them. At any time KB deals only with relatively *small fragments* of the external world. So, the corresponding structures are needed to integrate these fragments separated in time into the integral picture. The structures obtained as a result of integration should contain more information than it had been used for its creation. The organisation of KB should make allowance for such features as:

- associability;
- ability to reflect similar features for different objects and different features for similar objects (where objects are represented by KW and INo);
- heterarchical organisation of information [5]. The idea of heterarchical approach means that a full association of INo and KW represent very complicated net of nodes and unidirectional links between them. The predetermined hierarchy of "super-" and "subclasses" is absent; every node (INo or KW) is a "patriarch" in its own hierarchy if some process of search initiates with it.

Bad Messages Analysis

The main purpose of **B**ad **M**essages (BdM) is to classify BdM and allocate types of BdMs which might be restored. Several hundred thousand BdMs have been detected and result of this is as follows:

- Wrong KW among valid and invalid KW - 42.12%;
- Wrong KW among valid KW - 20.11%;
- Wrong KW among invalid KW - 22.01%;
- Wrong INo - 39.53%;
- "Stop" MSM - 8.78%;
- Empty MSM - 6.47%;
- Wrong alphabet (e.g. Russian) - 2.65%;
- Mobile slang (from T9 dictionary) - 0.37%;
- Rude MSM - 0.08%.

Remark: *Wrong INo* means literally **wrong** INo, e.g. 22120000, or **unknown** INo. So despite that 39.53% of *wrong INo* it would not be effective to spend more effort in trying to decrease this percentage. In the next session of paper some ideas of KW and right INo restoration will be discussed.

Algorithm of KW and/or INo Restoration

1. **INo recognition**. There are four possible type of INo: (i) *valid*; (ii) *invalid*, (iii) *unknown* when either length of INo is different from short or long INo, or INo does not exist in KB. Remark: Checking existing INo in KB would be sufficient to find out if the INo is known or not. But this operation requires more time than simply checking the length of the INo, and (iv) *wrong* INo. Initial analysis of INo does not allow the identification of this type of INo. It would only be possible to do this when KW of the MSM is recognised.
2. **Initial MSM validation**. MSM will be classified as valid if only contains symbols from the Latin alphabet and/or digits are used. Hereafter, only valid MSM will be considered.
3. **Separators elimination** from MSM.
4. **Fillers elimination** from MSM. For example, in MSM: *"I'd like to stop sending messages" I'd like to* is a filler and will be deleted.
5. **Slang elimination** from MSM using T9 dictionary.

6 **Stop** MSM recognition. Remark: In the current version of algorithm MSM *"s_t_o_p"* will not be recognised as a *stop* MSM.

7 **Extracting set of KW** from KB related to **INo**, i.e. $\{KW_{INo}\}$, where $\{KW_{INo}\} \subset \{KW_{KB}\}$. $\{KW_{KB}\}$ represents all existing **KW** in **KB.**

8 **Extracting KW** from MSM, i.e. KW_M. Remark: In the current version of the algorithm only the **first** word of MSM is considered as a KW_M.

9 **Extracting set of INo** from KB related to KW_M, i.e. $\{INo_{KWm}\}$, **where** $\{INo_{KWm}\} \subset \{INo_{KB}\}$.

10 Pair INo ↔ MSM is accepted if $((INo \in \{INo_{KWm}\} \wedge KW_M \in \{KW_{INo}\}) \Rightarrow$ **IS-Correct(MSM))** $\mapsto$ **return**(KW_M), where predicate *IS-Correct(MSM)* is **true** when *"MSM is correct"* and **false** (i.e. ¬IS-Correct(MSM)) - otherwise. Symbol $\Rightarrow$ stands for word *then* and symbol $\mapsto$ means *lead to*. Returned KW_M is used for further analysis.

11 Pair INo ↔ MSM represents BdM, if
$(INo \in \{INo_{KWm}\} \wedge KW_M \notin \{KW_{INo}\}) \oplus (INo \notin \{INo_{KWm}\} \wedge KW_M \in \{KW_{INo}\}) \oplus (INo \notin \{INo_{KWm}\} \wedge KW_M \notin \{KW_{INo}\})$,
where symbol ⊕ means **exclusive or.**

12 After recognition of BdM reason, the attempt to restore BdM is undertaken. To explain this step let us assume that the reason of BdM is:
$INo \in \{INo_{KWm}\} \wedge KW_M \notin \{KW_{INo}\}$.
From this it follows that:
$INo \in \{INo_{KWm}\} \wedge KW_M \notin \{KW_{INo}\} \wedge (KW_M \in \{KW_{KB}\} \oplus KW_M \notin \{KW_{KB}\})$.
If $KW_M \in \{KW_{KB}\}$ then attempts to correct INo should be undertaken. The next step will describe the more complicated case of KW_M correction when $KW_M \notin \{KW_{KB}\}$.

13 **KW_M correction.** There are two different approaches to restore KW_M:

(1) The first approach provides searching $KW_i \in \{KW_{KB}\}$ under several conditions:

- the difference in length of words KW_i and KW_M must be less or equal **1;**
- just two different symbols might be in KW_i and KW_M. This rule covers four possible types of misspelling (the word *attempt* is used to demonstrate the first three types): (i) *attepmt*; (ii) *atempt*; (iii) *attembt*, and (iv) *ozlo*. The last type should be considered more attentively. There are two different reasons for this type of misspelling:

  I. Problem of **symbol recognition**. Very often it is simply impossible for the user to distinguish the letter *'l'* from the digit *'1'*, especially when, for example, the previous symbols are letters but for correct KW digit *'1'* need to be typed in, e.g. *oz10*.

  II. **Easier typing**. For the user it is easier to press the button **0** once than to press the button **6** three times to enter the letter *'o'* in word *bonus*, because for any reader it is still easy to understand word the *b0nus*. Another example, when instead of the letter *'l'* (pressing the button **5** three times), or *'i'* (pressing the button **4** three times) entered digit *1* e.g. *tab1e*.

- **Similarity of words** KW_i and KW_M must be more or equal to some **Threshold of Similarity (TofS), i.e.** $SmIrt(KW_i,KW_M) \geq TofS$. The calculation of $SmIrt(KW_i,KW_M)$ as a percentage is quite simple**:**

  $$SmIrt(KW_i,KW_M) = (ACS_{LR}(KW_i,KW_M) + ACS_{RL}(KW_i,KW_M)) * 2 / (Length(KW_i) + Length(KW_M))*100,$$

  where $ACS_{LR}(KW_i,KW_M)$ and ACS_{RL} stand for **A**mount of **C**ompared **S**ymbols from **L**eft to **R**ight and **R**ight to **L**eft respectively. For example, for considered words: *attepmt*; *atempt*; and *attemppt* the values of $SmIrt(KW_i,KW_M)$ are as follows**:**

  SmIrt(*attempt*,*attepmt*) = (4+1)*2/14*100=71.43%,

  SmIrt(*attempt*,*atepmt*)=(2+4)*2/13*100=92.31%, and

  SmIrt(*attempt*,*attemppt*)= (6+1)*2/15*100=93.33%.

Remark: In the result of comparison of words *attempt* and *attemppt* from *left to right* two sequences remain to be compared from *right to left*: *t* and *pt*. That is why $ACS_{RL}(KW_i,KW_M) = 1$. The compact description of first approach to restore KW_M might be presented in the following manner:

$$\exists\, KW_i\, ((KW_i \in \{KW_{KB}\}) \wedge (Smlrt(KW_i,KW_M) \geq TofS)) \mapsto return(KW_i),$$

where quantifier $\exists$ means *exist*.

To find out an appropriate value for **TofS** thousands of BdMs have been tested for three different values of **TofS – 50.0%, 75.0%,** and **100%**. The decreasing of restored KWs are:

6,370 $\rightarrow$ (-1,137) $\rightarrow$ 5,233 $\rightarrow$ (-709) $\rightarrow$ 4,524.

That is caused by **type 1** of misspelling (wrong sequence of two letters), because **Smlrt(KW_i,KW_M)** is very sensitive to a word's length, e.g. **Smlrt(*node*,*ndoe*)=50.0%, Smlrt(*table*,*tabel*)=60.0%,** and **Smlrt(*axmpridel*,*amxpridel*)=77.78%.** In the current version of the algorithm TofS **= 75.0%** because type 1 misspelling occurs very seldom in short words (i.e. with a length less than 6 characters).

(2) If the previous approach was not success then algorithm is trying to find such $KW_i \in \{KW_{KB}\}$ that is

(i) an initial part of KW_M, i.e. $KW_i \triangleright KW_M$,

(ii) $\forall KW_i\, (KW_i \in \{KW_{KB}\} \wedge KW_i \triangleright KW_M)$ **Select(max(Length(KW_i)),** where quantifier $\forall$ means *from all* and **Select(max(Length(KW_i))** stands for *"select KW_i with maximum length", and*

(iii) $(Length(KW_M) - Length(KW_i)) \leq (Length(KW_M)/2)$, e.g. *airtext* $\triangleright$ ***airtextww3514.***

14 **INo correction.** Result of KW_M correction is shown on Figure 4. To describe the INo correction let us suppose that pair "*81025* $\leftrightarrow$ *cash*" has been entered. This pair has been recognised as BdM because

$INo \notin \{INo_{KWm}\} \wedge KW_M \notin \{KW_{INo}\} \wedge KW_M \in \{KW_{KB}\} \wedge INo \in \{INo_{KB}\}$. **$\{INo_{cash}\}$ = {84025, 86025, 87025, 82085, 87085, 87023}.** It would be not acceptable to advise the user: *"Please try to dial 84025, 86025, 87025, 82085, 87085, or 87023"*. Instead a heuristic approach is used and might be describe as follows:

- For each button define a set of ***"direct neighbour"*** buttons (**DrctN**) and a set of ***"diagonal neighbour"*** buttons (**DgnlN**). Given terms easy to explain by example: DrctN(5) = {2, 4, 6, 8} and DgnlN(5) = {1, 3, 7, 8}.
- Find out the wrongly pressed button. For the considered example, Smlrt(*81025*,*84025*)=80%. The same result that we have for INo *86025* and *87025*. Thus it is very likely that the wrongly pressed button was **1**.
- Now the right button should be selected. DrctN(1) = {2, 4} and DgnlN(1) = {5} associated with button 1. First of all the right button is searching among DrctN(1). It is easy to see that only button 4 could be the right button and that is why INo 84025 is displayed (see Figure 5).

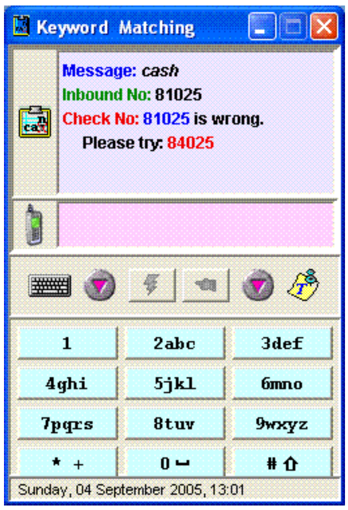


Figure 4. KW_M correction **Figure 5. INo correction**

The result of testing both KW and INo correction is represented on Figure 6.

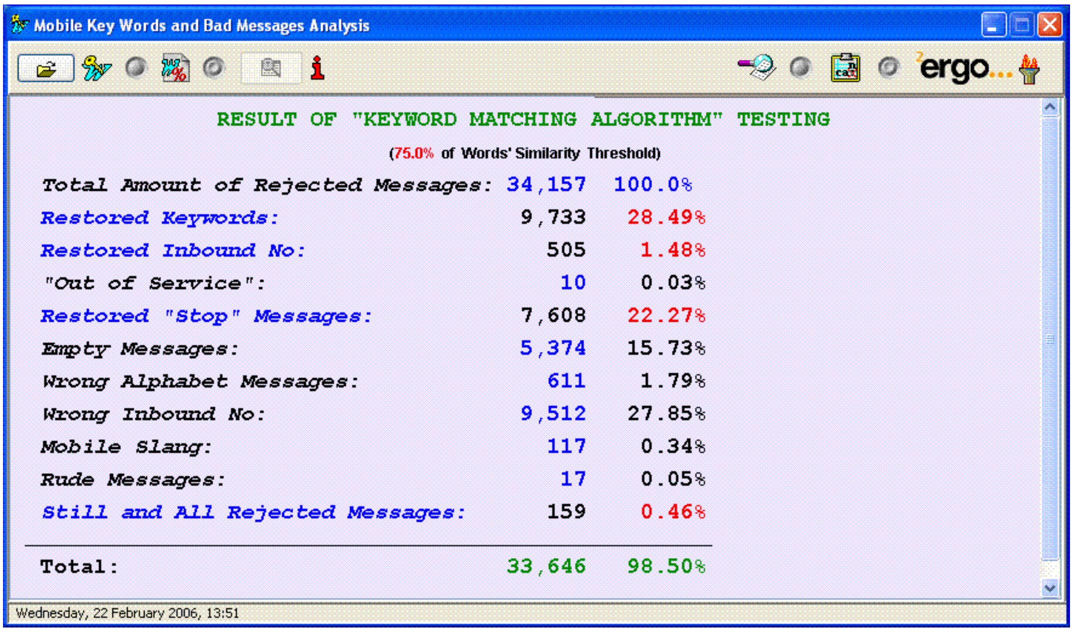


Figure 6. Result of Algorithm Testing

Remark: In Figure 6 the amount of **distinct** *"Still and All Rejected Messages"* is displayed and that is why the initial amount of BdM = 34,157 is more than the total amount of tested and corrected messages (33,646). The described algorithm improved BdM recognition by 52.25%.

Conclusion

The recent development in natural language processing has made it clear that formerly independent technologies can be harnessed together to an increasing degree in order to form sophisticated and powerful information delivery vehicles. Written speech, verbal speech and MSM analysis provide complementary functionalities, which can be combined to meet the modern technologies requirements.

Bibliography

[1] D.Burns, R.Fallon, P.Lewis, V.Lovitskii, S.Owen, "Verbal Dialogue Versus Written Dialogue*, *Proc. of the XI-th International Joint Conference on Knowledge-Dialogue-Solution: KDS-2005*, Varna (Bulgaria), 336-244, 2005.

[2] Opinion Research Corporation, www.orc.co.uk.

[3] Tegic Communication, www.tegic.com.

[4] V.A.Lovitskii and K.Wittamore, "DANIL: Databases Access using a Natural Interface Language", *Proc. of the International Joint Conference on Knowledge-Dialogue-Solution: KDS-97*, Yalta (Ukraine), 282-288, 1997

[5] M.R.Quillian, "Word concepts: A theory and simulation of some basic semantic capabilities", *C.I.P. working paper 79*, Cornegie Inst. of Technol., Pittsburgh, 1965.

Authors' information

Ken Braithwaite – e-mail: ken.braithwaite@2ergo.com

Mark Lishman – e-mail: mark.lishman@2ergo.com

Vladimir Lovitskii – e-mail: vladimir@2ergo.com

David Traynor – e-mail: david.traynor@2ergo.com

2 Ergo Ltd, St. Mary's Chambers, Haslingden Road, Rawtenstall, Lancashire, BB4 6QX, UK

INSTANTANEOUS DATABASE ACCESS

Guy Francis, Mark Lishman, Vladimir Lovitskii, Michael Thrasher, David Traynor

Abstract: The biggest threat to any business is a lack of timely and accurate information. Without all the facts, businesses are pressured to make critical decisions and assess risks and opportunities based largely on guesswork, sometimes resulting in financial losses and missed opportunities. The meteoric rise of Databases (DB) appears to confirm the adage that "information is power", but the stark reality is that information is useless if one has no way to find what one needs to know. It is more accurate perhaps to state that, "the ability to find information is power". In this paper we show how Instantaneous Database Access System (IDAS) can make a crucial difference by pulling data together and allowing users to summarise information quickly from all areas of a business organisation.

Keywords: data mining, natural language, parsing, SQL-query and production rules

ACM Classification Keywords: I.2.1 Application and Expert Systems – Natural language interface

Introduction

1. The rapid advance of computer technology, particularly, the explosive growth of databases (DB), has resulted in the availability of ever increasing amounts of information. Both the number of DB and their contents are growing fast. The total amount of information in the world is estimated to be doubling every 20 months, and much of this is being stored in computer DB. Within businesses what tends to happen is for management to not have access to this information in any user-friendly summary format. This means businesses have built up a reservoir of information but have restricted means for tapping that information for decision making purposes.
2. Data lies at the heart of every modern enterprise. The way it is used, how data is managed, and its quality and accuracy all impact on the success or failure of organisations in every industrial sector. Organisations need information quickly and accurately. But to access and verify records held within a vast DB used to be time consuming, complicated and expensive.
3. More and more people are required to make critical decisions because of increased competitive pressures but they need help to find the relevant data that could guide them in the decision-making process. This situation is termed the *"fact gap"* when many user/managers make decisions in a virtual vacuum using outdated information, borrowed perspectives or pure guesswork.
4. Data rich organisations which have large or varied data sources, often face problems of inconsistency and inaccuracy because they have historically suffered from an unmanageable array of data sources, collected by different people at different times from a variety of channels on a daily basis. Large, disparate DB mean that organisations frequently suffer from a poor standard of data quality and accuracy. Hence, individual records containing, for example, potentially valuable customer information are not harnessed for their true potential and organisations then miss crucial details through lack of knowledge.
5. SQL is the standard query language for accessing data held within a relational DB. With its powerful syntax, SQL represents a leap forward in DB access for all levels of management and computing professionals.
6. Many businesses, from small companies to major multi-nationals, have staff that would benefit from simple access to the organisation's data but are denied it by the complexities of query syntax, such as SQL, and the data structures involved. Training staff can be prohibitively expensive and conventional systems demand higher degrees of computer literacy than may be available. To solve this problem several intelligent tools have been created [1,2].

The central question to be addressed by this paper is how to improve access to DB for users. Such users may not understand DB, may not know exactly what is in the DB and how data is stored there. Crucially, it follows that they do not possess the means for extracting data. Hereafter, let us use the general term "**A**pplication **D**omain"

(**AD**) to refer to the joined tables of DB and corresponding knowledge about DB contents and metadata (where metadata is a DB value's field' description and tables connectivity). Such knowledge will be stored in a **K**nowledge **B**ase (**KB**). Within this, there are (at least) the following two important issues to discuss: (1) natural User-AD interface, and (2) natural user's enquiry to SQL-query conversion. We feel that many of the concepts we have developed over the years revolve around the problem of representing complex database schemas using simple natural language terminology. This can be of great benefit to any type of data access tool, or to any data access situation.

Natural User-AD Interface

The main requirement for IDAS is **-** to handle non-standard or poorly formed (but, nevertheless, meaningful) user's enquiries. Let us distinguish four different types of **N**atural **U**sers' **E**nquiries (**NUE**): (1) **N**atural **L**anguage **E**nquiry (**NLE**); (2) NLE **T**emplate (**NLET**); (3) **E**nquiry **D**escriptors (**ED**), and (4) **I**mmediate **E**nquiry (**IE**). Such enquiries permit users to communicate with a DB in a natural way rather than through the medium of formal query languages. Obviously issues in these four NUE are related, and the knowledge needed to deal with them may be distributed throughout a NL **I**nterface (**NLI**) system. We want to underline here that the selection of NUE type is not just a user decision because some DB may not be appropriate targets for NLI. It is important to have a clear understanding of these problems so that the NLI can mediate between the user view, as represented by the NLE, and the underlying database structure. Let us consider these four types of NLE.

Natural Language Enquiry provides end users with the ability to retrieve data from a DB by asking questions using plain English. But there are several problems of using NLE:

- The end users are generally unable to describe completely and unambiguously what it is they are looking for at the start of a search. They need to refine their enquiry by giving feedback on the results of initial search e.g. *"I'm looking for a* ***nice*** *city in France for holiday"* (where *Nice* is a city in France but also an adjective in English). Parsing of such simple NLE is quite complicated and requires powerful KB from IDAS [4]. Except lingual ambiguity a lot of problem cause ***"DB field's values"*** ambiguity. For example, in NLE *"I'm looking for the address of an insurance company in Bolton"* the word *Bolton* is a value of the field *City*, part of the value in the field *Company* (e.g. *"Bolton Insurance Company"*), as well as being part of an address (e.g. *"Bolton Road"*);
- Very often a user's NLE cannot be interpreted because the concepts involved are outside of the AD. Therefore IDAS should have an ability to decide whether the NLE is meaningful or not. In the result of analysis of no meaningful NLE, IDAS should describe to the user what is wrong with the NLE and how the enquiry might be rephrased to get the desired information. Such an approach, however, requires a very complicated KB in order to establish a meaningful communication with the user during the dialogue. Moreover, clarifying dialogue for the user creates a bad impression about IDAS because the user wants to be understood by IDAS immediately, without any additional effort on their part.

It is simply impossible to require the users to know the exact values in DB (e.g. name of constituency in ***Election*** AD) in order to ask correctly what is a very simple question: "*Who won the election in Suffolk Central & Ipswich North in 2001?"*. For example, if the user instead of using the symbol ***'&'*** instead types in ***"and"*** IDAS will not find the constituency in DB. But IDAS is an intelligent system and in the result of NLE analysis IDAS understand that user possibly mentioned two different constituencies *Suffolk Central* and *Ipswich North* but both of them incorrectly because there also exists *Suffolk Coastal*, *Suffolk South*, *Suffolk West* and *Ipswich* constituencies. Clarification dialog generated by IDAS irritates user:

IDAS: Do you mean Suffolk Coastal, Suffolk South, or Suffolk West constituency?

User: No, I mean Suffolk Central.

IDAS: Suffolk Central constituency does not exist but there is Suffolk Central & Ipswich North constituency.

User: It's exactly what I meant.

IDAS: Thank you.

- Very often NLE is ungrammatical.

- Direct observation of user NLE shows that <u>all users are lazy</u> i.e. they want to achieve the desired result by using minimum effort. They do not want to type in the long NLE such as *"Identify the parts supplied by each vendor and the cost and sales value of all these items at present on order"*. This is natural behaviour of human being in accordance with the ***principle of simplicity***, or ***Occam's razor principle*** *(Occam's (or Ockham's) razor is a principle attributed to the 14th century logician and Franciscan friar; William of Occam. Ockham was the village in the English county of Surrey where he was born)*. The principle states that "***Everything should be made as simple as possible, but not simpler***" *(The final word is of unknown origin, although it's often attributed to Einstein, himself a master of the quotable one liner)*. Finding a balance between simplicity and sophistication at the input side has been discussed in [5].

  Thus, firstly, NLE does not necessarily mean the enquiry is in plain English, secondly, IDAS should provide different levels of simplicity for NLE. The first step in this direction is NLET.

Natural Language Enquiry Template combines a list of values to be selected when required and generalization of users' NLEs. Examples of some **F**requently **A**sked **Q**uestions (**FAQ**) in AD *Election* are shown below:

- *What was the result in <constituency>?*
- *How many votes did <party> win in <constituency>?*
- *Which party won the election in <constituency>?*
- *Who won an election in <constituency>?*

Initial set of FAQ has been created by export in AD *Election* but in the result of activities new NLE have been collected by IDAS, analysed, generalized and then added to FAQ.

When the user selects an appropriate NLET with some descriptor in angular brackets IDAS immediately displays the list of corresponding values. As soon as the user finds the demand value by simply starting to type it and press button <Enter> result will be displayed (see Figure 1).

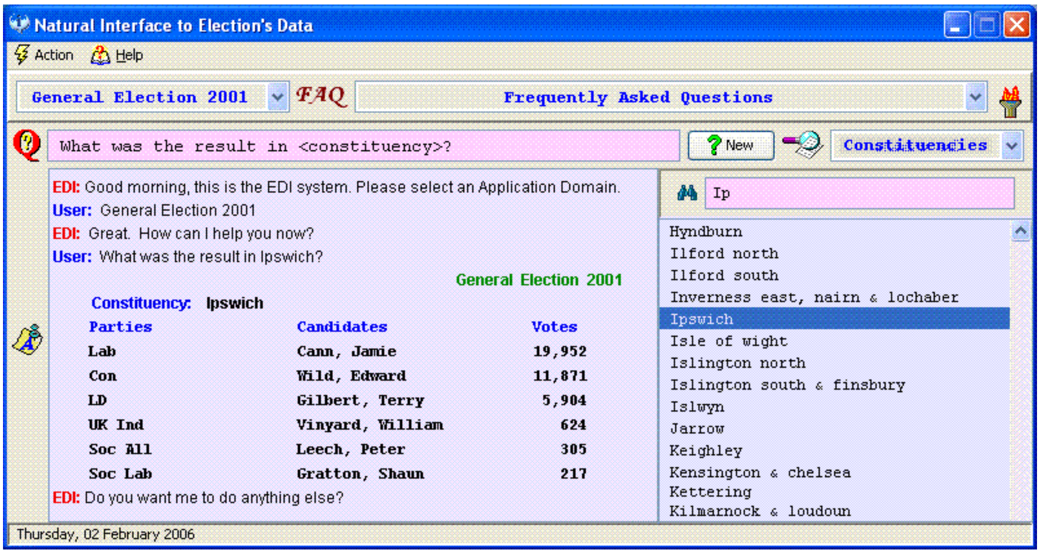


Figure 1. Natural Language Enquiry Template

At first glance, the NLET is an ideal way to communicate with AD but in reality there are some problems, which need to be solved to provide lightness of communication. To highlight such problems is enough to consider quite a simple NLET: "*Who won an election in <constituency>?*". Without knowing *"who is who"* and meaning of *"won election"* IDAS cannot answer this question. To explain it to IDAS the **P**roduction **R**ules (**PR**) need to be involved. Many researchers are investigating how to reduce the difficulty of moving a NLI from one AD to another. The problems in doing this include what information is needed and how the information needs to be represented. From our point of view, **P**reconditioned PR (PPR) is a quite powerful approach to solve this problem. The subset of PPR in format: **<Precondition>** $\mapsto$ **<Antecedent>** $\Rightarrow$ **<Consequent>** is shown below.

1. AD:Election ↦ who ⇒ candidate;
2. AD:Election ↦ [candidate]:<win⊕won> ⇒ [SQL]:<MAX(votes)>;
3. AD:Athletics ↦ [runner]:<win⊕won> ⇒ [SQL]:<MIN(time);
4. AD:Athletics ↦ [shooter]:<win⊕won> ⇒ [SQL]:<MAX(distance);
5. AD:Election & DB:MS Access ↦ votes ⇒ [Field]:<CANDIDATE.VOTE>;
6. AD:Election & DB:MS Access ↦ candidate ⇒ [Field]:<CANDIDATE.[CANDIDATE NAME]>;
7. AD:Election & DB:Oracle ↦ [party]:<win⊕won> ⇒ [SQL]:<MAX(SUM(votes))>;
8. AD:Election & DB:MS Access ↦ [party]:<win⊕won> ⇒ [SQL]:<TOP1, SUM(votes), SUM(votes) DESC>,

where ⊕ - denotes "*exclusive OR*". **Precondition** consist of **class$_1$:value$_1$ {& class$_i$:value$_i$}**. **Antecedent** might be represented by: (i) **single word** (e.g. *who, won, August*, *seven*, *etc.*), (ii) **sequence of words** (e.g. *as soon as*, *create KB*, *How are you doing*, etc.), or (iii) **pair - [context]:<value>**. Context allows one to avoid word ambiguity and thereby distinguish difference between "*Candidate won an election*" and "*Party won an election*". Presentation of **Consequent** is similar to Antecedent structure except (iii). For Consequent pair represents **[descriptor]:<value>**.

For AD *Election* subset (1, 2, 5..8) of PPR is used. PPR 3 and 4 in fact show another meaning of the same word *won* but for a different AD. The last two PPR show the simplest way to cover the difference in SQL for different DB. Result of parsing considered NLET using selected PPR is shown on Figure 2.

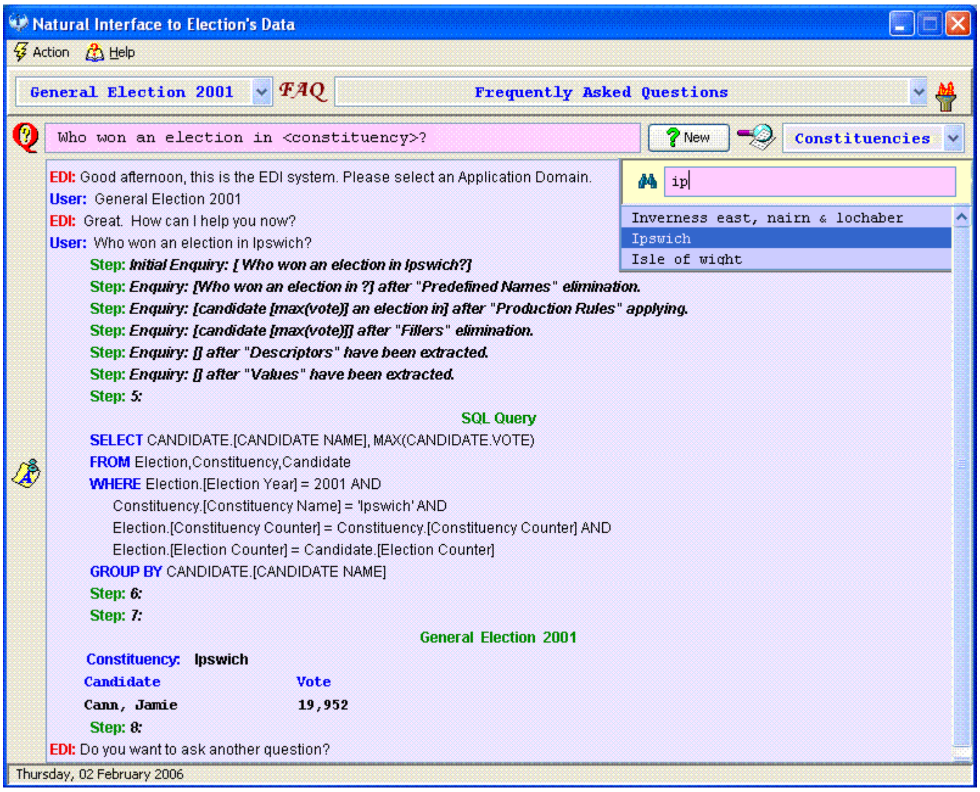


Figure 2. Natural Language Enquiry Parsing

Thus, NLET allows the user to “be lazy” but requires great effort to create the proper set of PPR as part of KB to describe better the more meaningful words. But using NLE and NLET we cannot say that all meaningful words have been described even for quite restricted AD. As a result some users will be disappointed by the IDAS reply. ED is a step in the direction towards simplifying KB and increasing the reliability of IDAS.

Before moving to ED it would be sensible once more to address some NLE and NLET problems. The cognitive process of understanding is itself not understood. First, we must ask: “*What it means to understand a* NLE?”. The usual answer to that question is to model its meaning. But this answer just generates another question: “*What does meaning means?”.* The meaning of a NLE depends not only on the things it describes, explicitly and implicitly, but also on both aspects of its causality: “*What caused it to be said*” and “*What result is intended by saying it*”. In other words, the meaning of a NLE depends not only on the sentence itself, but also on **Who** is asking the question and **How** the question is phrased.

From the linguistics point of view the process of understanding is possible under the following, as a minimum, three conditions [6]:

- IDAS must comprehend and understand separate words but lexical ambiguity sometimes makes such understanding difficult. A classic example of lexical ambiguity is the sentence: *“Time flies like an arrow”*. Each of the first three words could be the main verb of the sentence, and *“time”* could be a noun or an adjective, *“flies”* could be a noun, and *“like”* could be a preposition. Thus, the sentence could have various interpretations other than the accepted proverbial one. It could, for example, be interpreted as a command to an experimenter to perform temporal measurements on flies in the same way they are done on arrows. Or it could be a declaration that a certain species of fly has affection for a certain arrow.
- IDAS must understand the structure of the whole sentence but sometimes that is not a simple matter. If we have an ambiguous phrase such as: *“John saw the woman in the park with a telescope”*, then we usually understand one meaning and ignore the alternative interpretations.
- An empirical study revealed that only **0.53%** of possible sentences considered being grammatical are actually produced [7, p.823]. Note that the capacity to cope with <u>ungrammatical NLE</u> is one of the important requirements of NLE processing.

For artificial system like IDAS the power of natural language to describe the same events in different ways is a great problem. For example, the primitive event: *“Delete a cursor from the screen”* might be described as: *“eliminate a cursor”, get rid of a cursor”, “remove a cursor from the screen”, erase a cursor”, “makes a cursor hidden”, “set the cursor size to 0”, “take away a cursor from the screen”*, etc. Therefore the ED might release IDAS from such problems.

Enquiry Descriptors is especially useful when AD is not simple (e.g. AD *Mobile Messages* on Figure 3). And another important point of using ED is that modern technology has completely changed the way that people use the telephone to exchange dialogue with information held on computers. Well developed *“written speech analysis”* does not work with *“verbal speech”* [3]. For example, the first step of Speech Recogniser to parse NLE *“I’m looking for address of insurance company in Bolton”* will be filler deleting i.e. *“I’m looking for”*. Finally, initial NLE will be represented as a set of descriptors, which represent the NL description of meaningful fields of AD.

The definition of “meaningful fields” depends on AD objectives. For the considered AD *Mobile Messages* is a list of descriptors: {*company, account, network,* etc.}. Between descriptors and meaningful fields exist one-to-one attitude. The procedure for creating ED is very simple (see Figure 3):

- Select desirable descriptors. In the result of selection the corresponding <Table>.<Field> (Descriptor) will be displayed;
- Select field, value for which needs to be assigned, enter value in square brackets and press <Enter>. For descriptor *Date* value [*February*] was defined;
- If some mathematical function need to be involved press corresponding button. To summarize all delivered messages button <*SUM*> has been clicked for selected descriptor *Delivered*;
- Click button SQL to convert ED to SQL-query.

If objectives of using AD changes then set of descriptors need to be extended, which requires effort of KB administrator. But this is the simplest way of extracting data from AD – using **IE**.

Immediate Enquiry is useful for users who are familiar with AD structure and know the meaning of tables and their fields. To create IE, firstly, select table, secondly, select desirable field (see Figure 3). Pair <Table>.<Field> will be displayed. Now user can add a descriptor and do the same procedure as for ED.

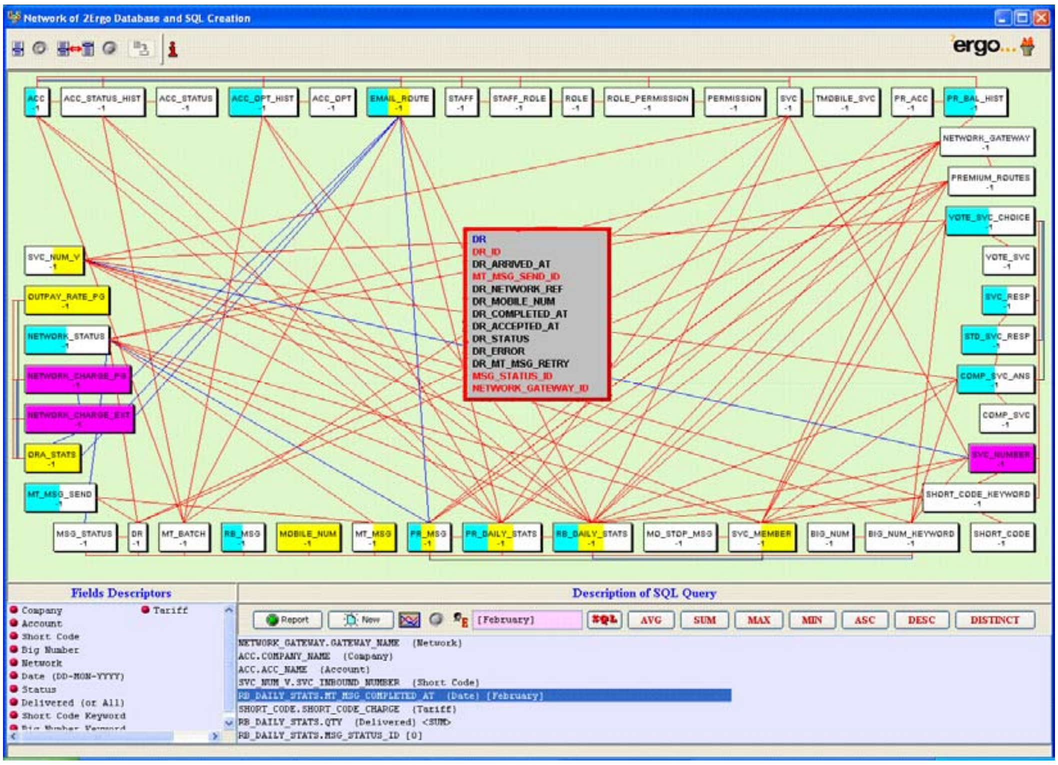


Figure 3. AD "*Mobile Messages"* and example of Enquiry Descriptors

Natural User's Enquiry to SQL Query Conversion

The steps of NLE to SQL query are well defined [1]: **(NLE $\vee$ NLET) $\rightarrow$ ED $\rightarrow$ IE $\rightarrow$ SQL-query.** The final step is quiet complicated because the necessity to access data from many different tables within an AD and join those tables together in a report needs to be implemented. This is extremely important because non-technical users do not know how to join tables to get a more comprehensive view of their data. Quite often a very simple question in English can turn into a very complicated SQL-query e.g. conversion of NLE *"Display all messages amount for all networks in the last month"* gives SQL-query shown on Figure 4.

```
SELECT   NETWORK_GATEWAY.GATEWAY_NAME "Network",SUM(CASE WHEN RB_DAILY_STATS.MSG_STATUS_ID = 0 THEN RB_DAILY_STATS.QTY ELSE 0 END) "Delivered",
         SUM(CASE WHEN RB_DAILY_STATS.MSG_STATUS_ID = 4 THEN RB_DAILY_STATS.QTY ELSE 0 END) "Pending",
         SUM(CASE WHEN RB_DAILY_STATS.MSG_STATUS_ID = 1 THEN RB_DAILY_STATS.QTY ELSE 0 END) "Expired",
         SUM(CASE WHEN RB_DAILY_STATS.MSG_STATUS_ID = 2 THEN RB_DAILY_STATS.QTY ELSE 0 END) "Rejected",
         SUM(CASE WHEN RB_DAILY_STATS.MSG_STATUS_ID = 3 THEN RB_DAILY_STATS.QTY ELSE 0 END) "Undeliverable",SUM(RB_DAILY_STATS.QTY) "Total"
FROM     NETWORK_GATEWAY,RB_DAILY_STATS
WHERE    TRUNC(RB_DAILY_STATS.MT_MSG_COMPLETED_AT) BETWEEN '01-JAN-2006' AND '31-JAN-2006' AND NETWORK_GATEWAY.NETWORK_GATEWAY_ID = RB_DAILY_STATS.NETWORK_GATEWAY_ID
GROUP BY NETWORK_GATEWAY.GATEWAY_NAME
```

Figure 4. Result of NLE to SQL-query conversion

Even the simplest ED like "W*hite thick sliced bread*" cannot be directly converted to SQL-query because AD's data might contain any combination of wrong and correct words and, therefore, four PPR ("*white* $\Rightarrow$ *wht* ", "*thick* $\Rightarrow$ *thk* ", "*sliced* $\Rightarrow$ *slcd* $\oplus$ *sld*", and "*bread* $\Rightarrow$ *brd*") is required [3]. Theoretically, for the considered example there are 16 possible combinations of data, namely: (1) "*White thick sliced bread*", (2) "*White thick sliced brd*", … , (16) "wht thk (slcd ⊕ sld) brd". Result of such conversion is shown in Figure 5.

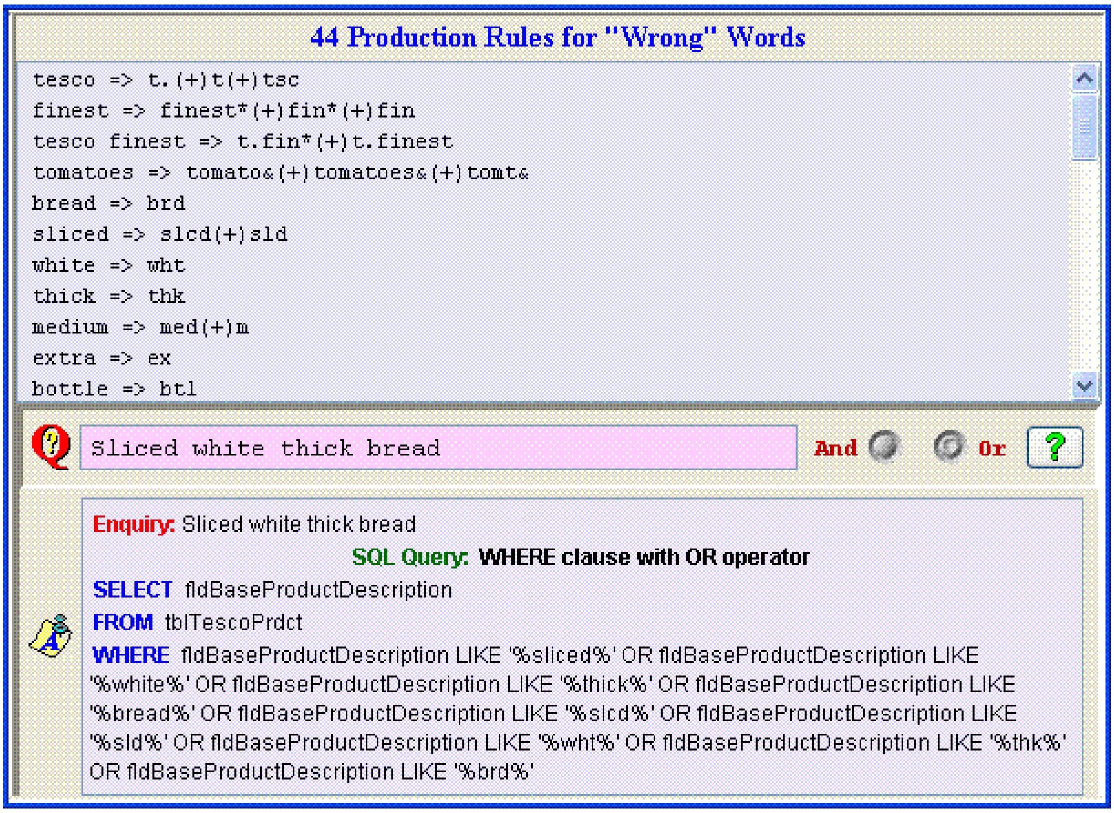


Figure 5. ED to SQL-query conversion using PPR

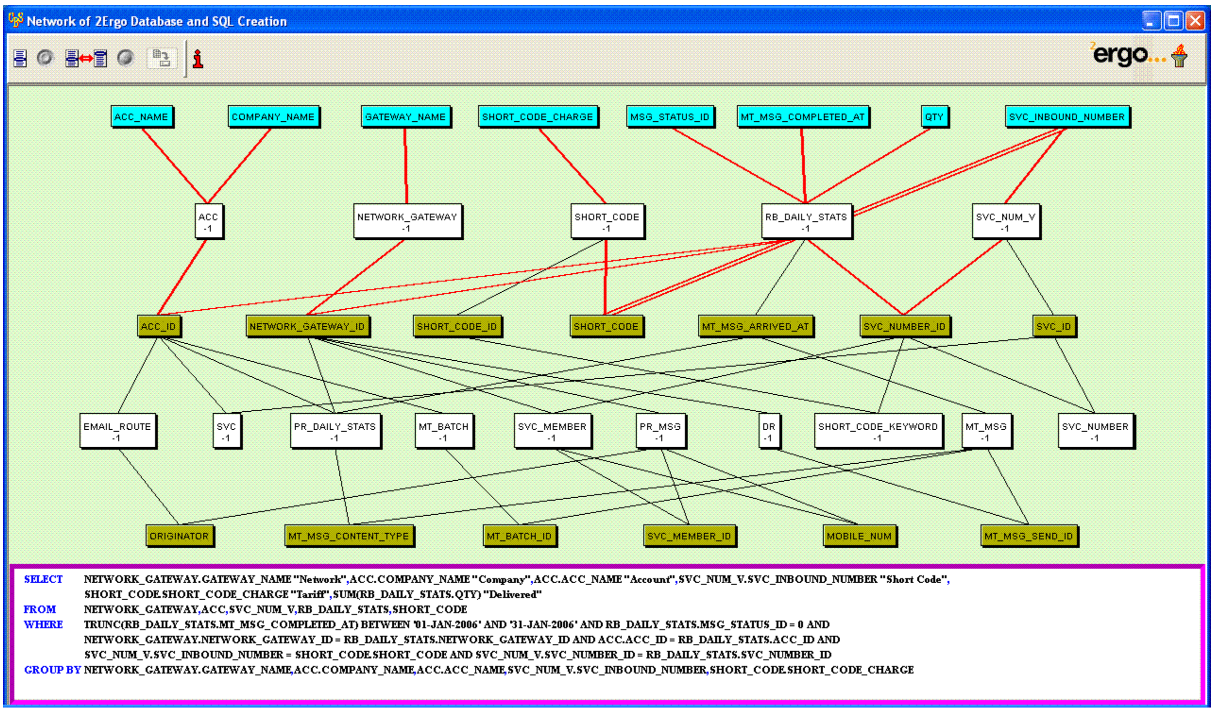


Figure 6. TC Decision Tree and TC Solution

The idea of joining tables in SQL is that individual rows in one table are attached to some corresponding rows in another table. The criteria for joining rows are decided by the highly skill SQL user. IDAS provides automatic Tables Coupling (TC). The main problem of TC is to select the right tables link from a huge number of possible links. Result of conversion ED from Figure 3 to SQL-query is shown as Figure 6. It is easy to see on TC decision tree shown the amount of possible TC Solutions (TCS). It is important to underline that output produced by SQL-query with different TCS might be different. In such situations a critical question arises: *"What is a criteria of selection of theTCS, which provides the right output?"*. IDAS activities are based on the hypothesis that ***"the right output might be produced by SQL-query with the best TCS"***, where the definition of *the best TCS* is obvious. *Let us call TCS the best if for each pair of tables the shortest link was used*. The given definition follows *the principle of simplicity* described earlier. Red lines on Figure 6 indicate TCS. TC decision tree was created using the *breadth-first method*. *Unipath heuristics rule* has been involved for selection of the best TCS. Two different type of fields are used as *foreign keys* to provide TCS:

- *Primary keys* e.g. ACC.ACC_ID = RB_DAILY_STATS.ACC_ID i.e. primary key ACC_ID from table ACC had been placed into table RB_DAILY_STATS as a foreign key.
- *Value fields*. Sometimes for different reasons, the DB has data redundancy i.e. in different tables there are fields with the same data (data duplication). The names of such fields are not necessarily the same. In that case at the stage of KB creation such field names should be described as synonyms. In the considered example, fields SVC_INBOUND_NUMBER from table SVC_NUM_V and SHORT_CODE from table SHORT_CODE are synonyms. Figure 6 has a double red line which shows the links between them.

Conclusion

IDAS effectively allows us to place information directly into the hands of business users - eliminating the need for technical support specialists continually to address *ad hoc* requests from end users. To do it properly all four types of enquiries should be provided. IDAS shields the user from the complexity of the underlying technology and itself acts as an intelligent user assistant.

Bibliography

[1] V.A.Lovitskii and K.Wittamore, "DANIL: Databases Access using a Natural Interface Language", *Proc. of the International Joint Conference on Knowledge-Dialogue-Solution: KDS-97*, Yalta (Ukraine), 282-288, 1997.

[2] G.Coles, T.Coles, V.A.Lovitskii, "Natural Interface Language", *Proc. of the VIII-th International Conference on Knowledge-Dialogue-Solution: KDS-99*, Kacivelli (Ukraine), 104 -109, 1999.

[3] D.Burns, R.Fallon, P.Lewis, V.Lovitskii, S.Owen, "Verbal Dialogue versus Written Dialogue*, *Proc. of the XI-th International Joint Conference on Knowledge-Dialogue-Solution: KDS-2005*, Varna (Bulgaria), 336-244, 2005.

[4] T.Coles, V.A.Lovitskii, "Text Searching and Mining", J. of *Artificial Intelligence,* National Academy of Sciences of Ukraine, *Vol. 3,* 488-496, 2000.

[5] L.Huang, T.Ulrich, M.Hemmje, E.Neuhold, "Adaptively Constructing the Query Interface for Meta Search Engines", Proc. of the *Intelligent User Interface Conf.,* 2001.

[6] Harris R.J. Monaco G.E., "Psychology of Pragmatic Implication: Information Processing between the Lines", Journal *"Exp. Psychol. General"*, 107, 1978, pp.1-22.

[7] Kitano H., "Challenges of Massive Parallelism", *Proc. Of the 13th International Joint Conference on Artificial Intelligence (IJCAI-93)*, Vol. 1, 1993, pp.813-834.

Authors' Information

Guy Francis – e-mail: guy.francis@2ergo.com

Mark Liashman – e-mail: mark.liashman@2ergo.com

Vladimir Lovitskii – e-mail: vladimir@2ergo.com

David Traynor – e-mail: david.traynor@2ergo.com

2 Ergo Ltd, St. Mary's Chambers, Haslingden Road, Rawtenstall, Lancashire, BB4 6QX, UK,

Michael Thrasher – University of Plymouth, Plymouth, Devon, PL4 6DX, UK e-mail: mthrasher@plymouth.ac.uk

WEB-BASED INFORMATION SYSTEMS IN THE STOCK MARKET FINANCIAL INFORMATION DOMAIN

George Blanas, Anthea Blanas

Abstract: The information domain is a recognised sphere for the influence, ownership, and control of information and it's specifications, format, exploitation and explanation (Thompson, 1967). The article presents a description of the financial information domain issues related to the organisation and operation of a stock market. We review the strategic, institutional and standards dimensions of the stock market information domain in relation to the current semantic web knowledge and how and whether this could be used in modern web based stock market information systems to provide the quality of information that their stakeholders want. The analysis is based on the FINE model (Blanas, 2003). The analysis leads to a number of research questions for future research.

Introduction

The development of advanced information systems (IS) serving the stakeholders of the various stock markets around the world is of great significance because they reflect the standards, the regulations, the objectives and the mechanisms for the exploitation and usage of money for the benefit of participating companies and investors. Money management is a capability required by all organisations regardless of the goods they produce or the services they offer. The stock market financial knowledge domain is probably the most commonly accessed and discussed by people of all educational backgrounds, cultures and incomes.

The critical economic role and the complexity of financial information described in the stock exchange financial information systems has led to the development of a number of standards. Unfortunately different groups of players have been using different standards or different interpretations of the same standard (Jayasena, Madnick & Bressan, 2004).

Organisations in the financial industry, like banks, stock exchanges, government agencies, or insurance companies operate their own financial information system or systems. Often they develop different contexts that may result in semantically heterogeneous representations of the same knowledge (Madnick, 1996). On the other hand, when making investment decisions, it is important that the information communicated between participantss is understood by both. Context mediation is one way to translate the meaning on a standard form. In COIN (Siegel & Madnick, 1991), sources and receivers describe their context to the mediator independently, which then uses metadata to detect conflicts in their data representation and to apply relevant conversions.

Interoperability of financial information systems is not the only problem in the international stock exchange system (Jayasena, Madnick & Bressan, 2004). Information quality is a very important aspect. Information quality includes dimensions like accuracy, reliability, bias, age, purpose, value, symmetry and is a very important aspect of information exchange. Systems reflect various changing regulations and laws that may differ from country to country or between regions. Coverage and accuracy of this reflection depends on many dimensions, like information systems building capability of software firms, age of the system software being used and support capability of tools employed for development. Access time and access restrictions to information result in asymmetries that can produce situations that will allow the development of illegal profits. The international networking of big brokers and other players gives them relative advantage to decision making comparing to listed firms or small investors, unless symmetrical information provision has been build into the relevant network of systems. Open standards in design and development of these systems is of critical importance for the increasingly globalised network of stock exchange institutions, but also open standards in the development and operation of the institutions themselves is of equal importance. The issue of time difference in the operation of the various stock exchanges around the globe adds a significant case for asymmetric information provided to the international stock market network stakeholders.

In the following paragraphs we try to evaluate the existing stock market information systems as nodes of an international financial network where stock exchange capabilities have become global for both institutions and investors.

FINE - Framework for Information Network nodEs

Blanas (2003) has developed a general framework on the evaluation of organisations as information systems. The framework is based on the networking paradigm and focuses on the operation, management and evolution of an organisational information network node. Figure 1 shows an expanded node linked to two other nodes presented as circled Ns.

The node is able to enclose a number of capabilities and quality characteristics in various extents and intensities. In figure 1 we can distinguish first and second level feedbacks.

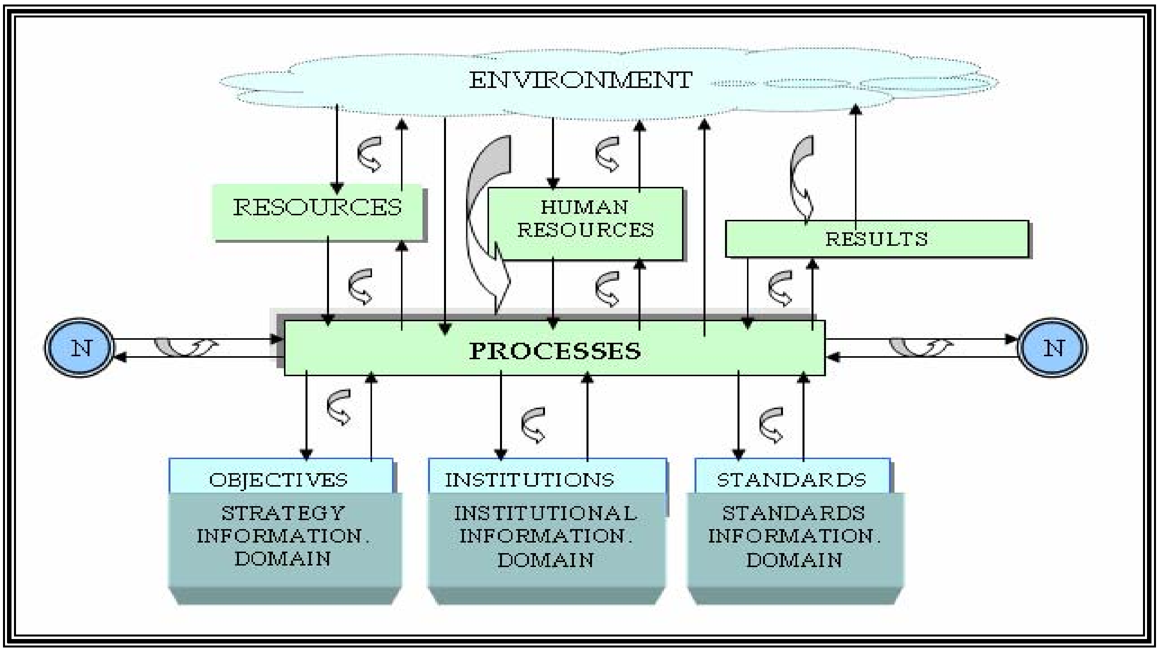


Figure 1 **F.I.N.E. - F**ramework for **I**nformation **N**etwork nod**E**s (Blanas, 2003)

The first level feedbacks are the following:

[1.1] Processes → Human Resources → Processes

Human resource management on the development of leadership and motivation, evaluation, education, and training.

[1.2] Processes → Resources → Processes

Algorithm application, processing, storage requirements, access, evaluation, can be of strategic, managerial, administrative, operational type in respect to the value of information, level of automation, disaggregation and security, processing speed, storage capacity, and cost. Information has immediate relationship to the storage media and the access mechanisms.

[1.3] Processes → [Results →] Environment → Processes

Interface processes with shareholders, clients and citizens, operations, also on processes of service, evaluation, and development of new products and technologies. The level of understanding of the interface information depends on the level of asymmetry between institutional structures for the communication and processing of information.

[1.4] Processes → Objectives → Processes

The processes (conformance, participation, evaluation) refer to the level of recognition of environmental problems or chances, and the adaptation of the node to them, and they contribute to the acceptance, development and application of strategy. The configuration of objectives can become the main second level process for a strategic information domain.

[1.5] Processes → Institutions → Processes

Processes of institutional-managerial type (compliance, participation, evaluation) related to policy making, compliance nad participation to existing institutions and recording of any adaptation difficulties.

[1.6] Processes → Standards → Processes

Organization, operation, administration, assurance, logistics of projects and procedures, based on compliance, use and evaluation of standards. The standardization of information flows reflects the regulations and conventions of management. Many standards are immediate results of political-legislative instititionalization..

[1.7] Processes → Node → Processes

Outsourcing of subprojects and procedures, access to information residing in other nodes (networking) and on the interactions with other nodes' processes.

The second level feedbacks are the following:

[2.1] Results → Environment → Results

Environmental evaluation of node results to the environment.

[2.2] Resources → Environment → Resources

Environmental evaluation of procurement management and environmental issues management.

[2.3] Human Resources → Environment → Human Resources

Environmental evaluation of human resource management criteria and processes.

[2.4] Objectives → Strategic Information Domains → Objectives

Configuration of information domain strategies.

[2.5] Objectives → Institutional Information Domains → Objectives

Configuration of information domain institutions.

[2.6] Objectives → Standards' Information Domains → Objectives

Configuration of information domain standards.

In the proposed framework, we consider that problems are developed in cases of loss of equilibrium or asymmetry in communication across mechanisms. Equilibrium can be lost in cases of lesser capabilities or lost opportunities for learning or adaptation. Asymmetry can be developed from incomplete information or from control of the information flows. The capabilities of the various feedback mechanisms configure the capability maturity level of the node.

The first level feedbacks continuously improve the processes that comply with the current objectives, standards and institutions, using the evaluation processes and collect meta-information for the evaluation system.

The second level feedbacks use the meta-information from the first level and the existing environmental knowledge with probable use of benchmarking and propose the development of new objectives, institutions and standards.

If some feedback falls behind, that is an indication of a deficiency in resources, institutions, capabilities, environmental scanning, or will.

It is profound that the ability of a node to selectively diffuse or protect the information residing in its local memory depends on the corresponding abilities of the related nodes within the network

In the following section, we review the stock market information domain under the FINE network.

The Stock Market Information Domain under the FINE framework

The design of stock market information systems is not covered adequately in the literature. Very little is conveyed regarding the design issues, the system software and the tools applied in their development. One probable reason is to avoid attacks based on known flaws of these systems. Another reason is that there are no international institutions or regulations that demand that such information is published. There are many issues in stock market financial information systems development that require the attention of their stakeholders. We present some of these issues and we examine their possible impact using the description of the second feedback level in the FINE framework. Our analysis is focused on the strategic information domain with certain extensions to the institutional and standards' information domains.

Strategic information domain

The development of the stock market information domain is a very sensitive area for economic and security interests. The tendency for the development of stronger stock market nodes will accelerate with the provision of extra capabilities and support built in their information systems that other nodes will not be able to provide. These capabilities also depend on various political and strategic networks that relate to the companies decision making and their economic environment. The differentiation of the stock market nodes that are strategically positioned in the international stock market network today is already happening. There are a number of questions related to the operation and survival of smaller stock exchange nodes in relation to their information systems' capabilities.

The question of a semantic domain that will be able to describe the advanced stock market information has not been answered yet. Most of the currently represented information is in textual form, that is impossible to be processed in time by most users. Some IS are more advanced than others either in their designs or the richness of the provided information or the functional user interface. There are huge differences between the lower and the higher end of these systems. There is a need for a higher level of semantics at the design and interface level in order to be able to describe and operationalise complex financial information. These leads either in the development of "explanation" or "translation" mechanisms that will always suffer from deficiencies, or the development of ontologies for the semantic web, that is formalised designs of information that can be incorporated in suitable IS for easier searching, processing, and interpretation.

These ontologies must be able to face the needs of both the stock market exchanges and the customers. The outcome of such an effort cannot be anything else but an open design. Until now, none of the existing stock market IS is based on an open design. The only openness may be the use of some XML based tools and standards. There are a number of questions on the type of strategies that should follow by the stock markets around the world:

Should they develop semantic web representations for all their financial information domains?

If yes how is that going to be enforced? If not, who benefits from the existing asymmetry and inequality of information provision?

If enforced, what will be possible results for the various stakeholders? What will be the future of smaller stock exchanges? If not, what will be the dynamics for possible convergence?

Institutional information domain

To answer the questions in the strategic information domain may require institutional or other answers. Any institutional measure cannot succeed unless there is sufficient knowledge capability and power to support it.

Quality labor and sufficient capital are the main inputs for the development of a stock market information system. One could add knowledge, creativity, quality and standards that may not be direct derivatives of labor and capital. The question on where and how to complement the need for more and better input and output is crucial. We need to answer questions on the requirements for state support, institutional measures, and knowledge creation on an international basis.

The development of a common semantic domain for advanced stock market information manipulation requires the cooperation of a large number of stakeholders in both public and private spaces at an international level. There are a major number of institutions involved that will cooperate on such a strategy only if there is a win-win situation for everybody. In case of widely adopted semantic designs, the customers with the lower information processing capability will not be as disadvantaged as they are in the current operation. On the other hand, when this happen, possibly the evidence for fewer stock markets will be profound and the international stock market network might end in fewer larger nodes. There are a number of questions on the type of institutional measures that should develop:

Will these advanced semantic domains be able to be operated by lower capability exchanges?

What institutions will be able to enforce the development and application of new standards?

We are currently going through the strengthening of organised hedge funds that have advanced sourcing and exploitation of information capabilities. When such a change happens will there be a more stable control of the financial markets by their stronger players? What institutions will be able to alleviate the development of oligopoly market structures?

As we can see in the next paragraph, the relevant standards are already in place. There is a need for institutional cooperation in deciding on accepting and using these tools, and on the acceptance of common open designs for the stock market financial domain. The development of IS that will have common designs on a desirable scale of selective features and using different IS development tools can still be left in the discretion and the capability of the independent stock market institutions.

Standards' information domain

The quality of information systems depends very much on the type of development tools being used. It is very important, on what tools should be used, and whether there are available quality designers and programmers to apply them and be able to cooperate effectively.

Programmers can achieve the same user interface and functionality using different tools and algorithms, the important part is the design and the information systems environment. There are fundamental differences in the operating systems, the interfaces, and the structure and the capabilities of the programming tools being used under the different environments.

The application of XML related tools is a unique opportunity for standardisation in financial market information exploitation. XML can be the vehicle for the development of the semantic financial domain. XML extensions and standards for the financial domain like FpML, XBRL, RIXML, ebXML, NewsML, IFX, OFX, MarketsML, ISO 15022, swiftML, MDDL, have been defined (Castells, Foncillas, Lara, Rico & Alonso (2004). Tools for XML usage have been in use like XBRL, RIXML, and NewsML in the last few years. What is missing? A common approach to the user interface domain regardless of the stock market IS that [s]he is logged in.

Conclusions

From the above analysis we gather that development of stock market information systems requires certain strategies to be followed in order to be able to maintain some level of competition on the control of information domains that may be monopolised in certain strategic areas of economy and security. But even if the stock market information systems design and application succeeds, there are great dangers for the control of information domains to remain oligopolised due to unequal distribution of critical masses of stakeholders, and other factors like distribution of wealth and usage of financial tools. It seems that for the moment there are more questions than answers.

Acknowledgements

This research has been carried out under the uaspices of the ARCHIMEDES project that is co-funded by the European Social Fund and National Resources- EPEAEK II

Bibliography

Blanas G. (2003) Total Quality Networking: Total Quality Management, Human and Information Systems Networks, Athens: Patakis Publications

Castells P., Foncillas B., Lara R., Rico M., J. Alonso L. (2004) Semantic Web Technologies for Economic and Financial Information Management. Springer Verlag Lecture Notes in Computer Science, Vol. 3053. Davies, J.; Fensel, D.; Bussler, C.; Studer, R. (Eds.), 2004, XIII, ISBN: 3-540-21999-4, pp. 473-487.

Jayasena S., Madnick S. & Bressan, S. (2004) “Financial Information Mediation: A case study of standards integration for Electronic Bill Presentment and Payment using the COIN mediation technology (TES/VLDB)”, Working Paper CISL# 2004-12, Sloan School of Management, MIT

Madnick S. (1996) "Are we moving toward an information superhighway or a Tower of Babel? The challenge of largescale semantic heterogeneity," Proc. IEEE International Conference on Data Engineering, 1996, pp. 2-8.

Siegel M.& Madnick S.(1991) "A metadata approach to resolving semantic conflicts," Proc. 17th International Conference on Very Large Data Bases, 1991, pp. 133-145.

Thompson, J. (1967) Organisations in Action, New York: McGraw-Hill

Authors’ Information

Dr. **George Blanas** – Associate Professor, School of Economics and Business, T.E.I. of Larissa, Greece, e-mail: blanas@teilar.gr

Anthea Blanas – B.Arts, MSc stud. in Multimedia Informatics, Aristoteles Univ. of Thessaloniki, Greece, e-mail: Anthea_blanas@yahoo.com

RECENT RESULTS ON STABILITY ANALYSIS OF AN OPTIMAL ASSEMBLY LINE BALANCE

Yuri Sotskov

Abstract: Two assembly line balancing problems are addressed. The first problem (called SALBP-1) is to minimize number of linearly ordered stations for processing n partially ordered operations $V = \{1, 2, ..., n\}$ within the fixed cycle time c. The second problem (called SALBP-2) is to minimize cycle time for processing partially ordered operations V on the fixed set of m linearly ordered stations. The processing time t_i of each operation $i \in V$ is known before solving problems SALBP-1 and SALBP-2. However, during the life cycle of the assembly line the values t_i are definitely fixed only for the subset of automated operations $V \backslash \widetilde{V}$. Another subset $\widetilde{V} \subseteq V$ includes manual operations, for which it is impossible to fix exact processing times during the whole life cycle of the assembly line. If $j \in \widetilde{V}$, then operation times t_j can differ for different cycles of the production process. For the optimal line balance **b** *of the assembly line with operation times $t_1, t_2, \ldots, t_n$, we investigate stability of its optimality with respect to possible variations of the processing times t_j of the manual operations $j \in \widetilde{V}$.*

Keywords: Scheduling, robustness and sensitivity analysis, assembly line.

ACM Classification Keywords: F.2.2 Nonnumerical algorithms and problems: Sequencing and scheduling.

Introduction

A single-model paced assembly line, which manufactures homogeneous product in large quantities, is addressed (we use terminology given in monograph [Scholl, 1999]). The assembly line is a sequence of *m* linearly ordered stations, which are linked by a conveyor belt or other material handling equipment. Each station of the assembly line has to perform the same set of operations repeatedly during the life cycle of the assembly line. Set of operations *V,* which has to be processed on the assembly line within one cycle time *c,* is fixed. Each operation $i \in V$ is considered indivisible: An operation has to be completely processed on one station within one cycle time. All the *m* stations start simultaneously the sequences of their operations and buffers between stations are absent. *Simple Assembly Line Balancing Problem* is to find an optimal balance of the assembly line for the fixed cycle time *c*, i.e., to find a feasible assignment of operations *V* into a minimal possible number *m* of stations. In [Scholl, 1999], abbreviation SALBP-1 is used for such a problem.

In this paper, it is assumed that set *V* includes operations of two types. Subset $\widetilde{V} \subseteq V$ includes all the operations for which it is impossible to fix exact processing times for the whole life cycle of the assembly line (manual operations). An operation $i \in V \setminus \widetilde{V}$ is one with operation time t_i being fixed during the life cycle of the assembly line (automated operations). The technological factors define a partial order on the set of operations *V*. Digraph *G* = (*V, A*) with vertices *V* and arcs *A* defines partially ordered set of operations *V* = {1, 2, ..., *n*}, which have to be processed on the assembly line within cycle time *c*. Without loss of generality, it is assumed that $\widetilde{V} = \{1, 2, ..., \widetilde{n}\}$, $V \backslash \widetilde{V} = \{\widetilde{n} + 1, \widetilde{n} + 2, ..., n\}$, and $0 \le \widetilde{n} \le n$. The vectors of the operation times are denoted as follows: $\widetilde{t} = (t_1, t_2, ..., t_{\widetilde{n}})$, $\bar{t} = (t_{\widetilde{n}+1}, t_{\widetilde{n}+2}, ..., t_n)$, $t = (\widetilde{t}, \bar{t}) = (t_1, t_2, ..., t_n)$. The set of *n* operations is presented as follows: $V = \{1, 2, ..., \widetilde{n}, \widetilde{n}+1, ..., n\}$. If $\widetilde{n} = 0$, then $\widetilde{V} = \varnothing$. If $\widetilde{n} = n$, then $\widetilde{V}$ = *V*. The assignment $V = V_1 \cup V_2 \cup \ldots \cup V_m$ of the set of *n* operations to *m* linearly ordered stations *S* = (S_1, S_2,…, S_m) (i.e., partition of set *V* into *m* mutually disjoint non-empty subsets V_k, $k \in \{1, 2, \ldots, m\}$) is feasible operation assignment (called also *line balance*) if the following two conditions hold.

Condition 1: Feasible operation assignment does not violate the precedence constraints given by digraph G = (V, A), i.e., inclusion (i, j) ∈ A implies that operation i is assigned to station S_k: $i \in V_k$, and operation j is assigned to station S_r: $j \in V_r$, such that $1 \le k \le r \le m$.

Condition 2: *Cycle time c is not violated for each station S_k, $k \in \{1, 2, \ldots, m\}$, i.e., sum of the processing times of all the operations assigned to station S_k (called station time), has to be not greater than cycle time c:*

$$\sum_{i \in V_k} t_i \le c \,. \tag{1}$$

For problem SALBP-1, line balance ***b*** is optimal when it uses the minimal number of *m* stations and when both Condition 1 and Condition 2 are satisfied.

Lemma 1: *Constructing an optimal line balance for problem SALBP-1 is binary NP-hard problem even for the case of two stations used in the optimal line balance (S = (S_1, S_2)), empty precedence constraints (A = ∅), and fixed processing times of all the operations V processed on the assembly line ($\widetilde{V}$ = ∅).*

Lemma 1 may be easily proven by polynomial reduction of NP-complete partition problem [Garey, Johnson, 1979] to problem SALBP-1 with two stations and *A* = ∅ (see, e.g., [Scholl, 1999]).

For the sake of simplicity, notation

$$t(V_k) = \sum_{i \in V_k} t_i \tag{2}$$

is used for the original vector *t* = (t_1, t_2, …, t_n) of the operation times. We assume that, if $j \in \widetilde{V}$, then operation time t_j is a given non-negative real number: $t_j \ge 0$. The value of this operation time can vary during life cycle of the assembly line and can even be equal to zero. Zero operation time t'_j will mean that operation $j \in V_k \bigcap \widetilde{V}$ will be processed (e.g., by an additional worker) in such a way that processing operation *j* will do not increase station time for S_k for the new vector $t = (\widetilde{t}\,', \bar{t}) = (t'_1, t'_2, \ldots, t'_{\widetilde{n}}, t_{\widetilde{n}+1}, \ldots, t_n)$ of the operation times: $\sum_{i \in V_k} t_i' = \sum_{i \in V_k \setminus \{j\}} t_i'$. The latter equality is only possible if t'_j = 0.

If $i \in V \setminus \widetilde{V}$, then operation time t_i is given real number fixed during the whole life cycle of the assembly line. We assume that $t_i > 0$ for each operation $i \in V \setminus \widetilde{V}$. As far as the processing time of an automated operation is fixed, one can consider only automated operations, which have strictly positive processing times. Indeed, an operation with fixed zero processing time has no influence on the solution to problem SALBP-1. In contrast to usual stochastic problems (see surveys [Erel, Sarin, 1998; Sarin, Erel, Dar-El, 1999]), we do not assume any probability distribution known in advance for the random processing times of the manual operations. Moreover, it is assumed that optimal line balance ***b*** is already constructed for the given vector ***t*** = ($t_1, t_2, ..., t_n$) of the operation times and the aim is to investigate the stability of optimality of line balance ***b*** with respect to independent variations of the processing times of the manual operations $\widetilde{V} = \{1, 2, \ldots, \widetilde{n}\}$. More precisely, we investigate *stability radius* of an optimal line balance ***b***, which may be interpreted as the maximum of simultaneous independent variations of the manual operation times with definitely keeping optimality of line balance ***b***.

Motivation and Definitions

Problem SALBP-1 arises when a new assembly line must be installed and the internal demands and properties of the assembly line have to be estimated. Cycle time *c* is defined on the basis of customer demands in the finished products. The value of *c* may be calculated as the ratio of available operating time of the assembly line and production volume for the same calendar interval. This problem may also arise when cycle time *c* of acting assembly line has to be changed because of changing customer demands in the finished product.

In the real-world assemble lines, processing times of some operations may be known exactly and fixed for a long time (if an operation has to be done by fully-automated or semi-automated machine). Indeed, modern machines and robots are able to work permanently at a constant speed for a long time. However, in some cases it is not realistic to assume constant processing time for an operation, if it has to be done by a human operator with rather simple tools. In the case of a human work, operation time is subject to physical, psychological, and other factors. Due to the learning of operators, the operation times during the first days (weeks, months) of a life cycle of the assembly line may differ from the processing times of the same operations during the later days (weeks, months). Moreover, some workers can leave the plant and new workers with lower or higher skills have to replace them.

In the case of changeable operation times, it is important to know the credibility of the optimal line balance at hand with respect to possible variations of all or a portion of the operation times. Line balance ***b***, which is optimal for the original vector ***t*** = ($\widetilde{t}, \bar{t}$) of the operation times, may lose its optimality or even feasibility for a new vector $t' = (\widetilde{t}\,', \bar{t})$ of the operation times. For example, due to increasing of operation times, line balance ***b*** may become infeasible for cycle time *c* since inequality (1) may be violated. In such a case, it is necessary to look for another line balance and to use it for a suitable modification of production process on the assembly line. Also, line balance ***b*** may lose its optimality with saving feasibility. It may occur if another operation assignment b_s becomes feasible for the new vector $t' = (\widetilde{t}\,', \bar{t})$ of the operation times and b_s uses less stations than line balance ***b*** uses.

Of course, each re-engineering of the assembly line being in process takes an additional time and other expenditure. So, assembly line modification has to be started, if it is really necessary: When the income from the re-engineering will be larger than the total expenditure caused by this re-engineering. Therefore, an evaluation of expenditures and benefits should be conducted before deciding whether re-engineering of the assembly line is necessary. However, these expenditures and benefits are difficult to evaluate before the end of the re-engineering process. In this paper, we survey some sufficient conditions for keeping the optimality of the line balance being in process.

To test whether line balance ***b*** remains feasible for the new vector $t' = (\widetilde{t}\,', \bar{t})$ of the operation times takes $O(\widetilde{n})$ time (if station times are included in the input data) or $O(n)$ time (otherwise). Indeed, for the new operation times we have to verify inequality (1) for each subset $V_k, k = 1,2,...,m,$ that includes at least one manual operation with changed processing time in vector t'. In the case of feasibility of the line balance ***b*** for the new vector t', in order to test its optimality for t' one has to solve NP-hard problem SALBP-1. Intuitively, it is clear that sufficiently small changes of the operation times $t_1, t_2, ..., t_{\widetilde{n}}$ may keep line balance ***b*** optimal for the new vector $t' = (\widetilde{t}\,', \bar{t})$ of the operation times. Our aim is to estimate or (what is better) to calculate the largest simultaneous and independent variations of the operation times $t_i, i \in \widetilde{V}$, that do not violate the feasibility and optimality of the line balance ***b***. At the stage of the design of the assembly line, there exists a lot of optimal line balances. Using stability analysis, one can select such an optimal line balance, which optimality is more stable with respect to possible variations of the operation times $t_i, i \in \widetilde{V}$.

Let B denote the set of all assignments of operations V to stations S_1, S_2, … , S_m (for possible numbers m of the stations: $1 \le m \le n$), which satisfy Condition 1. Subset of set B of all operation assignments which also satisfy Condition 2 for the given vector $t = (t_1, t_2, \dots, t_n)$ of the operation times is denoted by $B(t) = \{b_0\ b_1, \dots, b_h\}$ where b_k, $k \in \{0, 1, ..., h\}$, means line balance $V = V_1^{\boldsymbol{b}_k} \cup V_2^{\boldsymbol{b}_k} \cup \ldots \cup V_{m_{\boldsymbol{b}_k}}^{\boldsymbol{b}_k}$. Let subset of set $B(t)$ of all the optimal line balances be denoted by $B_{opt}(t)$. Inclusion $\boldsymbol{b} \in B_{opt}(t)$ implies that line balance $\boldsymbol{b}$ with partition $V = V_1^{\boldsymbol{b}} \cup V_2^{\boldsymbol{b}} \cup \ldots \cup V_{m_{\boldsymbol{b}}}^{\boldsymbol{b}}$ satisfies Condition 1, Condition 2, and the following optimality condition for the vector $t = (t_1, t_2, \dots, t_n)$ of the operation times.

Condition 3: $m_b = \min\{ m_{\boldsymbol{b}_k} : b_k \in B(t)\}$.

Since line balance $\boldsymbol{b}$ is contained in the set $B(t)$, we obtain $\boldsymbol{b} = \boldsymbol{b}_r \in B(t)$ for some index $r \in \{0, 1, \dots, h\}$. As a matter of convenience, index r will be omitted for the optimal line balance $\boldsymbol{b}$, which stability will be investigated. In both definitions of set B and set $B(t)$, number m of stations is not fixed: For line balance $\boldsymbol{b}_k$ from the set $B(t)$, inequalities $m_{\boldsymbol{b}} \le m_{\boldsymbol{b}_k} \le n$ must hold and number of stations in an operation assignment from set B has to belong to set {1, 2, … , n}. The question under consideration may be formulated as follows. How much can be modified the components of the vector $\tilde{t}$ simultaneously and independently from each other that the given line balance $\boldsymbol{b}$ remains optimal?

Let $\boldsymbol{R}^n$ ($\boldsymbol{R}_+^n$ respectively) denote the space of n-dimensional real vectors $t = (t_1, t_2, \dots, t_n)$ (the space of n-dimensional non-negative real vectors) with the maximum metric, i.e., distance $d(t,t^*)$ between vector $t \in \boldsymbol{R}^n$ and vector $t^* = (t_1^*, t_2^*, \dots, t_n^*) \in \boldsymbol{R}^n$ is calculated as follows: $d(t,t^*) = \max\{ | t_i - t_i^* | : i \in V\}$, where $| t_i - t_i^* |$ denotes the absolute value of the difference $t_i - t_i^*$. Let line balance $\boldsymbol{b}$ be optimal for the given non-negative real vector $t = (\tilde{t}, \bar{t}) = (t_1, t_2, \dots, t_n) \in \boldsymbol{R}_+^n$ of the operation times, i.e., $\boldsymbol{b} \in B_{opt}(t)$. For problem SALBP-1, the definition of stability radius of an optimal line balance is introduced as follows.

Definition 1: *The open ball $O_\rho(\tilde{t})$ with radius $\rho \in \boldsymbol{R}_+^1$ and center $\tilde{t} \in \boldsymbol{R}_+^{\tilde{n}}$ in the space $\boldsymbol{R}^{\tilde{n}}$ is called a stability ball of the line balance $\boldsymbol{b} \in B_{opt}(t)$, if for each vector $t^* = (\tilde{t}^*, \bar{t})$ of the operation times with $\tilde{t}^* \in O_\rho(\tilde{t}) \cap \boldsymbol{R}_+^{\tilde{n}}$ line balance $\boldsymbol{b}$ remains optimal. The maximal value of the radius ρ of stability ball $O_\rho(\tilde{t})$ of the line balance $\boldsymbol{b}$ is called stability radius denoted by $\rho_b(t)$.*

In Definition 1, vector $\bar{t} = (t_{\tilde{n}+1}, t_{\tilde{n}+2}, \dots, t_n)$ of the processing times of the automated operations and the complete vector $t = (\tilde{t}, \bar{t}) = (t_1, t_2, \dots, t_n)$ of the operation times are fixed, while vector $\tilde{t}^* = (t_1^*, t_2^*, \dots, t_{\tilde{n}}^*)$ may vary within the intersection of the open ball $O_\rho(\tilde{t}) \subset \boldsymbol{R}^n$ with the space $\boldsymbol{R}_+^{\tilde{n}}$ of the non-negative real vectors. Stability radius $\rho_b(t)$ is equal to the minimal upper bound of independent variations ε_i of the processing times t_i of all the manual operations $i \in \tilde{V}$ which definitely keep the optimality of the line balance $\boldsymbol{b}$, i.e., inclusion $\boldsymbol{b} \in B_{opt}(t^*)$ holds with $t_i^* = \max\{0, t_i - \varepsilon_i\}$ or $t_i^* = t_i + \varepsilon_i$.

In the rest of this paper, we survey recent results proven in [Sotskov, Dolgui, 2001; Sotskov, Dolgui, Portmann, 2006; Sotskov *et al*, 2005] on stability analysis of an optimal line balance for problems SALBP-1 and SALBP-2.

Stability Radius of an Optimal Line Balance for Problem SALBP-1

Let $\tilde{V}_k^{\boldsymbol{b}}$ denote the subset of manual operations of set $V_k^{\boldsymbol{b}}$ and let $\bar{V}_k^{\boldsymbol{b}}$ denote the subset of automated operations of set $V_k^{\boldsymbol{b}}$. For each index $k \in \{1, 2, \dots, m_b\}$, equality $V_k^{\boldsymbol{b}} = \tilde{V}_k^{\boldsymbol{b}} \cup \bar{V}_k^{\boldsymbol{b}}$ holds. The following remark is used in stability analysis of an optimal line balance for problem SALBP-1.

Remark 1: *Let us consider the line balance* **b** $\in B_{opt}(t)$ *being in process and the modified vector* $t' = (\tilde{t}\,', \bar{t})$ *of the operation times. If there exists subset* $V_k^{\boldsymbol{b}}$, $k \in \{1, 2, ..., m_{\boldsymbol{b}}\}$, *in the line balance* **b** *such that*

$$\sum_{i \in V_k^{\boldsymbol{b}}} t_i' = 0, \qquad (3)$$

we continue to affirm that line balance **b** *uses* m_b *stations for the modified vector* $t' = (\tilde{t}\,', \bar{t})$ as well.

We can argue Remark 1 as follows. In spite of equality (3) valid for the new vector $t' = (\tilde{t}\,', \bar{t})$, station S_k is still exists in the assembly line with line balance **b**. At very least to delete station S_k causes additional cost and additional time for re-engineering the assembly line. Moreover, after deleting station S_k we obtain another line balance, say, $b^* \in B$: $V = \bigcup_{i \in \{1,2,...,m_b\}, i \neq k} V_i^{\boldsymbol{b}} = V_1^{\boldsymbol{b}^*} \cup V_2^{\boldsymbol{b}^*} \cup ... \cup V_{m_{\boldsymbol{b}^*}}^{\boldsymbol{b}^*}$, where $m_{\boldsymbol{b}^*} = m_{\boldsymbol{b}} - 1$. Due to the validity of inequality $t_i > 0$ for each automated operation $i \in V \backslash \tilde{V}$, equality (3) is only possible if $V_k^{\boldsymbol{b}} = \tilde{V}_k^{\boldsymbol{b}}$.

In [Sotskov, Dolgui, Portmann, 2006], it was proven the following necessary and sufficient condition for the case when optimality of the line balance $b \in B_{opt}(t)$ is unstable.

Theorem 1: *For line balance* $b \in B_{opt}(t)$ *equality* $\rho_b(t) = 0$ *holds if and only if there exists a subset* $V_k^{\boldsymbol{b}}$, $k \in \{1, 2, \ldots, m_{\boldsymbol{b}}\}$, *such that* $\tilde{V}_k^{\boldsymbol{b}} \neq \emptyset$ *and* $t(V_k^{\boldsymbol{b}}) = c$.

To present a formula for exact value of stability radius $\rho_b(t)$ of optimal line balance $b \in B_{opt}(t)$ we need the following notation:

$$\delta^{\boldsymbol{b}} = \min \{\delta_k^{\boldsymbol{b}} : \tilde{V}_k^{\boldsymbol{b}} \neq \emptyset,\ k \in \{1, 2, ..., m_{\boldsymbol{b}}\}\}, \qquad (4)$$

where $\delta_k^{\boldsymbol{b}} = \dfrac{c - t(V_k^{\boldsymbol{b}})}{|\tilde{V}_k^{\boldsymbol{b}}|}$ and value $t(V_k)$ is defined in (2). It is easy to see that testing criterion given in Theorem 1 takes $O(n)$ time. This asymptotic bound is defined due to calculating station times $t(V_p^{\boldsymbol{b}}), p = 1, 2, ..., m_{\boldsymbol{b}}$. If station times are included in the input data, then testing criterion given in Theorem 1 takes $O(m_b)$ time.

The following lower bound of stability radius has been obtained within the proof of Theorem 1 given in [Sotskov, Dolgui, Portmann, 2006].

Corollary 1: *If optimality of line balance* **b** $\in B_{opt}(t)$ *is stable, then* $\rho_{\boldsymbol{b}}(t) \geq \min\{\delta^{\boldsymbol{b}}, \Delta^{\boldsymbol{b}}\}$, *where*

$$\Delta^{\boldsymbol{b}} = \min \{\Delta(V_p^{\boldsymbol{b}^*}) : \boldsymbol{b}^* \in B\} \text{ and } \Delta(V_p^{\boldsymbol{b}^*}) = \frac{t(V_p^{\boldsymbol{b}^*}) - c}{|\tilde{V}_p^{\boldsymbol{b}^*}|}.$$

Let $\tilde{V}_p^{\boldsymbol{b}(d)} = \{i_1, i_2, ..., i_u\}$, where $u = |\tilde{V}_p^{\boldsymbol{b}(d)}|$, and indices v of operations i_v are assigned in such a way that the following inequalities hold: $t_{i_1} \leq t_{i_2} \leq ... \leq t_{i_u}$. It is assumed that $t_{i_0} = 0$. Vector $t'' = (\tilde{t}\,'', \bar{t}) \in \boldsymbol{R}_+^n$ closest to t, for which subset $V_p^{\boldsymbol{b}(d)}$ is feasible (inequality (1) holds for subset $V_p^{\boldsymbol{b}(d)}$ with vector $t'' = (\tilde{t}\,'', \bar{t})$), can be obtained if for each operation $i_q \in \tilde{V}_p^{\boldsymbol{b}(d)}$ we set $t''_{i_q} = \max\{0, t_j - \hat{\Delta}(V_p^{\boldsymbol{b}(d)})\}$ where j and i_q denotes the same manual operation ($j = i_q$) and value $\hat{\Delta}(V_p^{\boldsymbol{b}(d)})$ is calculated as follows:

$$\hat{\Delta}(V_p^{\boldsymbol{b}(d)}) = \max \left\{ \frac{\sum_{i \in V_p^{\boldsymbol{b}(d)}} t_i - c - \sum_{\alpha=0}^{\beta} t_{i_\alpha}}{|\tilde{V}_p^{\boldsymbol{b}(d)}| - \beta} : \beta = 0, 1, ..., |\tilde{V}_p^{\boldsymbol{b}(d)}| - 1 \right\}$$

Here maximum is taken among right-hand fractions calculated for $\beta = 0, 1, ..., \left| \widetilde{V}_p^{\mathbf{b}^{(d)}} \right| - 1$. We define

$$\Delta(\mathbf{b}^{(d)}) = \max\{ \hat{\Delta}(V_p^{\mathbf{b}^{(d)}}) : t(V_p^{\mathbf{b}^{(d)}}) > c \},$$
$$\hat{\Delta}^{\mathbf{b}} = \min\{ \Delta(\mathbf{b}^{(d)}) : \mathbf{b}^{(d)} \in B^{(m_{\mathbf{b}}-1)} \}. \quad (5)$$

In [Sotskov, Dolgui, Portmann, 2006], the following formula for calculating the exact value of stability radius $\rho_b(t)$ has been derived.

Theorem 2: *If optimality of line balance b ∈ $B_{opt}(t)$ is stable, then $\rho_b(t)$ = min{$\delta^{\mathbf{b}}, \hat{\Delta}^{\mathbf{b}}$ } with $\delta^{\mathbf{b}}$ being defined in (4) and $\hat{\Delta}^{\mathbf{b}}$ in (5).*

Let $\lceil a \rceil$ denote the smallest integer greater than or equal to real number a. Theorem 2 implies the following five corollaries.

Corollary 2: *If* $m_b = \left\lceil \frac{t(V)}{c} \right\rceil$, *then* $\rho_{\mathbf{b}}(t) \geq \min\left\{ \delta^{\mathbf{b}}, \frac{t(V) - c(m_{\mathbf{b}} - 1)}{\widetilde{n}} \right\}$.

Corollary 3: *If* $m_b = \left\lceil \frac{t(V)}{c} \right\rceil$ *and* $\delta^{\mathbf{b}} \leq \frac{t(V) - c(m_{\mathbf{b}} - 1)}{\widetilde{n}}$, *then* $\rho_{\mathbf{b}}(t) = \delta^{\mathbf{b}}$.

Corollary 4: *If b ∈ $B_{opt}(t)$, then* $\rho_{\mathbf{b}}(t) \leq \min\{\delta^{\mathbf{b}}, \max_{i \in \widetilde{V}} t_i\}$.

Corollary 5: *If b ∈ $B_{opt}(t)$, then* $\rho_{\mathbf{b}}(t) \leq \min\{c - \max_{i \in \widetilde{V}} t_j, \max_{i \in \widetilde{V}} t_j\}$.

Corollary 6: *If b ∈ $B_{opt}(t)$* and $\mathbf{b}^{(d)} \in B^{(m_{\mathbf{b}}-1)}$, *then* $\rho_{\mathbf{b}}(t) \leq \min\{\delta^{\mathbf{b}}, \Delta(\mathbf{b}^{(d)})\}$.

Stability of an Optimal Line Balance for Problem SALBP-2

In this section, we consider problem SALBP-2: To find an optimal balance of the assembly line for a fixed number of stations, i.e., to find a feasible assignment of all operations *V* to exactly *m* stations in such a way that the cycle-time *c* is minimal. For problem SALBP-2, line balance $\mathbf{b}_r$ is *optimal* if along with Condition 1 and Condition 2, it has the minimal cycle time. We denote the cycle time for line balance $\mathbf{b}_r$ with the vector *t* of the operation times as $c(b_r, t)$:

$$c(b_r, t) = \max_{k=1}^{m} \sum_{i \in V_k^{b_r}} t_i .$$

For problem SALBP-2, optimality of line balance $\mathbf{b} = \mathbf{b}_s$ with vector *t* of the operation times may be defined via the following condition.

Condition 4: *$c(b_s, t)$ = min{$c(b_r, t)$: b_r ∈ B(t)}, where B(t) = {b_0, b_1, …, b_h} is the set of all line balances.*

For problem SALBP-2, definition of the stability radius of an optimal line balance is introduced as follows.

Definition 2: *The closed ball $\underline{O}_\rho(\widetilde{t})$ in the space $\mathbf{R}^{\widetilde{n}}$ with the radius $\rho \in \mathbf{R}_+^1$ and center $\widetilde{t} \in \mathbf{R}_+^{\widetilde{n}}$ is called a stability ball of the line balance b $\in B(t)$, if for each vector $t^* = (\widetilde{t}^*, \bar{t})$ of the operation times with $\widetilde{t}^* \in \underline{O}_\rho(\widetilde{t}) \cap \mathbf{R}_+^{\widetilde{n}}$ line balance b remains optimal. The maximal value of the radius ρ of stability ball $\underline{O}_\rho(\widetilde{t})$ of the line balance b is called stability radius denoted by $\underline{\rho}_b(t)$.*

In Definition 2, vector $\bar{t} = (t_{\widetilde{n}+1}, t_{\widetilde{n}+2}, \ldots, t_n)$ of the automated operation times and vector $t = (\widetilde{t}, \bar{t}) = (t_1, t_2, \ldots, t_n)$ of all the operation times are fixed, while vector $\widetilde{t}^* = (t_1^*, t_2^*, ..., t_{\widetilde{n}}^*)$ of the manual operation times may vary within the intersection of the closed ball $\underline{O}_\rho(\widetilde{t})$ with the space $\mathbf{R}_+^{\widetilde{n}}$. For each optimal line balance $\mathbf{b}_r$ ∈ $B_{opt}(t)$, we define a set $W(\mathbf{b}_r)$ of all subsets $\widetilde{V}_k^{b_r}$, *k* ∈ {1, 2, …, *m* }, such that

$\sum_{i \in V_k^{b_r}} t_i$ = *c(b, t)*. It should be noted that set *W*(b_r) may include the empty set as its element. In [Sotskov *et al*, 2005], the following claims have been proven.

Theorem 3: *Let inequality* $t_i > 0$ *hold for each manual operation* $i \in \widetilde{V}$ *. Then for line balance* $\mathbf{b}_s \in B(t)$*, equality* $\underline{\rho}_{b_s}(t) = 0$ *holds if and only if there exists a line balance* $\mathbf{b}_r \in B_{opt}(t)$ *such that condition* $W(\mathbf{b}_s) \subseteq W(\mathbf{b}_r)$ *does not hold.*

Corollary 7: *If* $B_{opt}(t) = \{\mathbf{b}_s\}$*, then* $\underline{\rho}_{b_s}(t) > 0$.

If there exists an index $k \in \{1, 2, \ldots, m\}$ such that

$$\sum_{i \in V_k^{b_0}} t_i < c(\mathbf{b}_0, t), \qquad (6)$$

then we set $\lambda(\mathbf{b}_0) = \{c(\mathbf{b}_0, t) - \max\{\sum_{i \in V_k^{b_0}} t_i : \widetilde{V}_k^{b_0} \notin W(\mathbf{b}_0)\}\} \big/ \widetilde{n}$. Due to (6), the strict inequality $\lambda(\mathbf{b}_0) > 0$ must hold.

If $\sum_{i \in V_k^{b_0}} t_i = c(\mathbf{b}_0, t)$ for each index $k \in \{1, 2, \ldots, m\}$, then we set $\lambda(\mathbf{b}_0) = \min\{t_i : i \in \widetilde{V}\}$. We denote

$\Delta = \min\{\Delta(b_s) : b_s \in B \setminus B(t)\}$, where $\Delta(b_s) = \dfrac{c(\mathbf{b}_s, t) - c(\mathbf{b}_0, t)}{\widetilde{n}}$. Theorem 3 implies the following claim.

Corollary 8: *If* $\underline{\rho}_{b_s}(t) > 0$*, then* $\underline{\rho}_{b_s}(t) \geq \min\{\Delta, \lambda(\mathbf{b}_0)\}$.

Calculating exact value of stability radius $\underline{\rho}_b(t)$ is close to calculating stability radius of the optimal schedule for the makespan criterion (see [Sotskov, 1991; Sotskov, *et al.* 1998]).

Conclusion

In the above sections, the recent results on stability analysis of an optimal line balance are presented. We used the notion of stability radius, which is similar to the stability radius of an optimal schedule introduced in [Sotskov, 1991] for shop-scheduling problems. If stability radius of an optimal line balance is strictly positive, then any independent changes of the operation times t_j, $j \in \widetilde{V}$, within the ball with this radius definitely keep the optimality of this line balance. On the other hand, if stability radius of optimal line balance ***b*** is equal to zero (Theorems 1 and 3), then even small changes of the processing times of all or a portion of the manual operations may deprive the optimality of line balance ***b***. It is worth noting that conditions presented in this paper (except Theorem 2, Theorem 3, Corollary 1, and Corollary 8) may be tested in polynomial time, which is important for real-world assembly lines with large numbers of operations. Moreover, for exact value of stability radius, feasibility of line balance ***b***, which is defined by the value $\delta^{\mathbf{b}}$, may be tested in polynomial time due to Theorem 2 as well.

In practice, the tendency at the design stage must be to find optimal line balance for which stability is as much as possible. The common objective is to assign to each station a set of operations with roughly the same total operation time [Bukchin, Tzur, 2000; Erel, Sarin, 1998; Lee, Johnson, 1991; Sarin, Erel, Dar-El, 1999]. Due to the above results, we have to defer stations with manual operations and stations without manual operations. Theorem 1 shows that for the station with manual operations it is desirable to have some slack between cycle time and station time. The larger this slack is the larger stability radius of the line balance may be. On the other hand, for the stations loaded by only automated operations, such a slack may be as small as possible, which gives the possibility to increase slacks for stations with manual operations.

Since stability radius $\rho_b(t)$ cannot be larger than *c*/2, one has to pay special attention to the manual operations with possible variations of the processing times more than *c*/2 (such operations may cause instability of optimality of the line balance being in process). At the design stage, such an operation has to be divided into shorter operations. If line balance will be used for a long time for assembling the same finished product, it is desirable at the design stage, to construct several optimal line balances, and select among them the one with the best stability characteristics. To this end, it is useful to develop algorithms, which construct a set of optimal line balances (instead of only one optimal line balance), in order to carry out a stability analysis for them. Or, better yet, it is useful to include in the branch-and-bound algorithm or other algorithms used for solving problem SALBP-1 and problem SALBP-2 specific rules in order to construct optimal line balance with larger stability radius. In a concrete study, the set $\widetilde{V}$ of manual operations can be reduced (e.g., only critical manual operations may be considered) or, on the contrary, set $\widetilde{V}$ may be completed by some unstable automated operations. By changing the set of operations with variable times, the designer of the assembly line can study the influence of different operations on stability of optimality and feasibility of line balances.

In [Sotskov, Dolgui, 2001], slightly different definition of stability radius of an optimal line balance has been used for problem SALBP-1. Namely, it was assumed that $O_{\rho_b(t)} \subset \boldsymbol{R}_+^{\tilde{n}}$ (as a result, such a stability radius cannot be greater than $\min\{t_i : i \in \widetilde{V}\}$). The above Definition 1 is more appropriate for practical assembly lines. As a subject for future research, it is important to consider stability characteristics of an assembly line balance provided that for each operation time an interval of possible variation is known in advance. Such an assumption seems to be realistic for many real-word assembly lines.

This research was partially supported by ISTC (Project B-986). The author would like to thank Prof. Alexandre Dolgui, Prof. Marie-Claude Portmann, and Prof. Frank Werner for joint research in stability analysis.

Bibliography

[Bukchin, Tzur, 2000] J. Bukchin, M. Tzur, Design of flexible assembly line to minimize equipment cost, *IIE Transactions* 32 (2000) 585-598.

[Erel, Sarin, 1998] E. Erel, S.C. Sarin, A survey of the assembly line balancing procedures, *Production Planning & Control* 9 (5) (1998) 414-434.

[Garey, Johnson, 1979] M.R. Garey, D.S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness.* Freeman, San Francisco, USA, 1979.

[Lee, Johnson, 1991] H.F. Lee, R.V. Johnson, A line-balancing strategy for designing flexible assembly systems, *The International Journal of Flexible Manufacturing Systems* 3 (1991) 91-120.

[Sarin, Erel, Dar-El, 1999] S.C. Sarin, E. Erel, E.M. Dar-El, A methodology for solving single-model, stochastic assembly line balancing problem, *OMEGA – International Journal of Management Science* 27 (1999) 525-535.

[Scholl, 1999] A. Scholl, *Balancing and Sequencing of Assembly Lines*, Heidelberg: Physica-Verlag, Germany,1999.

[Sotskov, 1991] Yu.N. Sotskov, Stability of an optimal schedule, *European Journal of Operational Research* 55 (1991) 91-102.

[Sotskov, Dolgui, 2001] Yu.N. Sotskov, A. Dolgui, Stability radius of the optimal assembly line balance with fixed cycle time. In: *Proceedings of the IEEE Conference ETFA'2001* (2002) 623-628.

[Sotskov, Dolgui, Portmann, 2006] Yu.N. Sotskov, A. Dolgui, M-C. Portmann, Stability of an optimal balance for an assembly line with fixed cycle time, *European Journal of Operational Research* 168 (2006) 783-797.

[Sotskov, *et al.* 2005] Yu.N. Sotskov, A. Dolgui, N. Sotskova, F. Werner, Stability of optimal line balance with given station set. In: *Supply Chain Optimisation* (2005) 135-149.

[Sotskov, *et al.* 1998] Yu.N. Sotskov, V.S. Tanaev, F. Werner, Stability radius of an optimal schedule: A survey and recent developments, Chapter in: G. Yu (Ed.) *"Industrial Applications of Combinatorial Optimization"* **16**, Kluwer Academic Publishers, Boston, USA, 1998, 72-108.

Author's Information

Yuri N. Sotskov – Professor, DSc, United Institute of Informatics Problems of the National Academy of Sciences of Belarus, Surganova str., 6, Minsk, Belarus; e-mail: sotskov@newman.bas-net.by

REALIZATION OF AN OPTIMAL SCHEDULE FOR THE TWO-MACHINE FLOW-SHOP WITH INTERVAL JOB PROCESSING TIMES

Natalja Leshchenko, Yuri Sotskov

Abstract: Non-preemptive two-machine flow-shop scheduling problem with uncertain processing times of n jobs is studied. In an uncertain version of a scheduling problem, there may not exist a unique schedule that remains optimal for all possible realizations of the job processing times. We find necessary and sufficient conditions (Theorem 1) when there exists a dominant permutation that is optimal for all possible realizations of the job processing times. Our computational studies show the percentage of the problems solvable under these conditions for the cases of randomly generated instances with $n \le 100$. *We also show how to use additional information about the processing times of the completed jobs during optimal realization of a schedule (Theorems 2 – 4). Computational studies for randomly generated instances with* $n \le 50$ *show the percentage of the two-machine flow-shop scheduling problems solvable under the sufficient conditions given in Theorems 2 – 4.*

Keywords: Scheduling, flow-shop, makespan, uncertainty.

ACM Classification Keywords: F.2.2 Non-numerical algorithms and problems: Sequencing and scheduling

Introduction

In scheduling theory, it is usually assumed that the job processing times are known exactly before scheduling. However, the real-world scheduling problems usually are not deterministic: Machines may break down, activities may take longer time to be executed than it is expected before scheduling, jobs may be added or canceled, etc. In operations research literature, there are different approaches concerning management of uncertainty in scheduling (see surveys [Aytug et al., 2005; Davenport, Beck, 2000; Gupta, Stafford, 2006]).

The *stochastic method* [Elmaghraby, Thoney, 2000; Pinedo, 1995] for dealing with uncertainty is useful when the process has enough prior information to characterize the probability distributions of the random processing times and there are a lot of realizations of a similar process. In the particular case of the stochastic scheduling problem, random processing times may be *controllable*, and the objective is to choose the optimal processing times (which are under control of a decision-maker) and the optimal schedule with the chosen job processing times. For such a problem, the objective function depends on both the job processing times and the job completion times (see, e.g., [Jansen, Mastrolilli, Solis-Oba, 2005]). The current trends in the field of scheduling under the *fuzziness* notion have been presented in [Slowinski, Hapke, 1999]. In the field of operations research for the problems under uncertainty auxiliary criteria are often used. The most popular auxiliary criteria are criteria introduced by Wald, Hurwicz, and Savage (see [Shafransky, 2005] for a brief survey).

In spite of several developments, flow-shop scheduling problem with uncertain job processing times remains unsolved (see [Gupta, Stafford, 2006]). In the most of these developments, Johnson's rule and analysis methods play a significant role. In this paper, we consider a two-machine flow-shop scheduling problem with interval job processing times. A scheduling problem with interval job processing times is rather general, since most events that are uncertain before scheduling may be considered as factors that vary the job processing times. The processing time may depend on the distance between machines, the type of transport used, traffic conditions, intervals of availability of machines, possible machine breakdowns, emergence of new unexpected jobs with high priority, early or late arrival of raw materials, etc. In [Gupta, Stafford, 2006], there were discussed 21 restrictions involved in the classical flow-shop problem (denoted as $F \| C_{\max}$) with the fixed job processing times, where criterion $C_{\max}$ denotes minimization of a schedule length. Nine of these restrictions addressed the criterion and type of the processing system, but all the remaining restrictions may be overcome by using suitable intervals for possible variations of the job processing times.

Problem Setting

We consider the non-preemptive flow-shop scheduling problem with two machines and random bounded job processing times (only lower and upper bounds of the job processing times are assumed to be given before scheduling). This problem is denoted as $F2 \mid t_{jm}^{L} \le t_{jm} \le t_{jm}^{U} \mid C_{\max}$.

Two machines $M = \{M_1, M_2\}$ have to process set of *n* jobs $J = \{1, 2, ..., n\}$ with the same machine route (M_1, M_2). All the *n* jobs are available to be processed from time $\tau = 0$. In contrast to deterministic scheduling problem, it is assumed that processing time t_{jm} of job $j \in J$ on machine $M_m \in M$ is not fixed before scheduling. In a realization of the process, t_{jm} may be equal to any real value between lower bound $t_{jm}^{L} \ge 0$ and upper bound $t_{jm}^{U} \ge t_{jm}^{L}$ being given before scheduling. The probability distribution of the random job processing time is unknown.

Thus, we address the stochastic flow-shop scheduling problem for the case when it is hard to obtain exact probability distributions for random bounded job processing times, and when assuming a specific probability distribution is not realistic. It has been observed that, although the exact probability distribution of the job processing times may be unknown in advance, upper and lower bounds on the job processing times are easy to obtain in many practical cases. In such a case there may not exist a unique schedule that remains optimal for all possible realizations of the job processing times and this question is considered in detail in the next section.

If equality $t_{jm}^{L} = t_{jm}^{U}$ holds for each job $j \in J$ and each machine $M_m \in M$, then problem $F2 \mid t_{jm}^{L} \le t_{jm} \le t_{jm}^{U} \mid C_{\max}$ turns into a *deterministic* flow-shop problem (denoted as $F2 \,||\, C_{\max}$) that is polynomially solvable due to Johnson’s algorithm [Johnson, 1954]. Permutation that is constructed by Johnson’s algorithm is called a *Johnson's permutation*. At least one optimal permutation for problem $F2 \,||\, C_{\max}$ is a Johnson’s permutation. (It should be noted however, that for the problem $F2 \,||\, C_{\max}$, an optimal schedule may also be defined by permutation that is not a Johnson’s permutation.)

In contrast to deterministic problem $F2 \,||\, C_{\max}$ we call problem $F2 \mid t_{jm}^{L} \le t_{jm} \le t_{jm}^{U} \mid C_{\max}$ as an *uncertain* scheduling problem.

Existence of a Dominant Johnson’s Permutation for the Uncertain Flow-Shop Problem

Let *T* denote a set of feasible vectors $t = (t_{1,1}, t_{1,2}, ..., t_{n1}, t_{n2})$ of the job processing times:

$$T = \left\{ t \mid t_{jm}^{L} \le t_{jm} \le t_{jm}^{U}, j \in J, m \in M \right\}.$$

The set *S* of all feasible permutations (schedules) has cardinality $|S| = n!$. Permutation $\pi_i \in S$ *dominates* each other permutation $\pi_k \in S$ with $k \ne i$ if inequality $C_{\max}(\pi_i, t) \le C_{\max}(\pi_k, t)$ holds for each permutation $\pi_k \in S$, where $C_{\max}(\pi_i, t)$ denotes objective value $C_{\max}$ (length of a schedule) to the deterministic problem $F2 \,||\, C_{\max}$ with the fixed vector $t \in T$ of the job processing times.

We call permutation $\pi_i \in S$ a *dominant Johnson’s permutation* to the uncertain problem $F2 \mid t_{jm}^{L} \le t_{jm} \le t_{jm}^{U} \mid C_{\max}$ if for any feasible vector $t \in T$ of the job processing times permutation π_i is a Johnson's permutation (so permutation π_i is optimal) for the deterministic problem $F2 \,||\, C_{\max}$ with this vector $t \in T$ of the job processing times.

We consider the case when inequality $t_{jm}^{L} < t_{jm}^{U}$ holds for each job $j \in J$ and each machine $M_m \in M$. For this case the following theorem has been proven in [Leshchenko, Sotskov, 2005].

Theorem 1 *Let* $t^L_{jm} < t^U_{jm}$, $j \in J$, $M_m \in M$. *Then there exists a dominant Johnson's permutation* $\pi_i \in S$ *to the uncertain problem* $F2 \mid t^L_{jm} \le t_{jm} \le t^U_{jm} \mid C_{\max}$ *if and only if:*

a) For any pair of jobs i and j with $t^U_{k1} \le t^L_{k2}$, *k = i, j (with* $t^U_{k2} \le t^L_{k1}$, *k = i, j, respectively) either* $t^U_{i1} \le t^L_{j1}$ *or* $t^U_{j1} \le t^L_{i1}$ *(either* $t^U_{i2} \le t^L_{j2}$ *or* $t^U_{j2} \le t^L_{i2}$*);*

b) There exists at most one job j^* *such that* $t^L_{j1} < t^U_{j2}, t^L_{j2} < t^U_{j1}$, *and the following inequalities hold:* $t^L_{j^*1} \ge \max\{t^U_{i1} \mid t^U_{i1} \le t^L_{i2}\}$, $t^L_{j^*2} \ge \max\{t^U_{i2} \mid t^U_{i2} \le t^L_{i1}\}$.

Computational Results for Necessary and Sufficient Conditions of Theorem 1

In this section, we consider randomly generated uncertain flow-shop problems $F2 \mid t^L_{jm} \le t_{jm} \le t^U_{jm} \mid C_{\max}$ and answer (by experiments on PC) the question of how many uncertain instances have a Johnson's permutation that is optimal for all corresponding deterministic problems $F2 \,\|\, C_{\max}$ with feasible vectors $t \in T$ of the job processing times. Namely, for each randomly generated instance $F2 \mid t^L_{jm} \le t_{jm} \le t^U_{jm} \mid C_{\max}$ under consideration we tested whether condition of Theorem 1 hold.

The computational algorithm was coded in C++. For the experiments, we used an AMD 3000 MHz processor with 1024 MB main memory. For each job, the lower bound of job processing time was randomly generated in the range [1,1000] and the upper bound of job processing time was computed as follows: $t^U_{jm} = t^L_{jm}(1 + L\% / 100\%)$.

In each series we generated and tested 1000 instances (for each combination of *n* and *L* under consideration). Random instances have been generated as follows. We tested two-machine flow-shop scheduling problems with *n* ∈ {5, 10, 15, …, 100} jobs, integer processing times uniformly distributed in the range [1,1000], and *L* ∈ {1, 2, …, 10}.

Table 1. Percentage of instances solvable due to Theorem 1

L\n	5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
1	94.5	85.1	70	53.5	36.9	24.9	14.7	8.2	4.8	1.9	1.3	0.6	0.3	0	0	0	0	0	0	0
2	91.2	69.3	45.6	24.2	11.8	4.2	1.1	0.6	0.1	0	0	0	0	0	0	0	0	0	0	0
3	87.7	58.4	28.3	10.3	3.8	1.3	0.2	0	0	0	0	0	0	0	0	0	0	0	0	0
4	82.4	47.4	18.8	5.5	0.8	0.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	75.5	37.4	11.3	2.2	0.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	72.5	32.4	7.8	1.7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	66.9	25	5.4	0.7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	65.8	19.9	2.1	0.4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	63.4	18	2.4	0.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	59.2	14.4	2	0.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 1 presents the percentage of instances in the series for which conditions of Theorem 1 hold. From our experiment, it follows that increasing simultaneously both numbers *n* and *L* decreases the number of instances solvable due to Theorem 1.

Optimal Realization of a Schedule

In Table 1, there are many cases for which percentage of instances solvable due to Theorem 1 is equal to 0, i.e., unique dominant permutation does not exist in each of such cases. For such instances, we propose to use modified Johnson's algorithm developed in [Leshchenko, Sotskov, 2005] to construct partial strong order of jobs *J* (we considered only the cases when strong inequality $t_{jm}^L < t_{jm}^U$ holds for each job $j \in J$ and each machine $M_m \in M$).

Theorem 2 *Let inequality $t_{jm}^L < t_{jm}^U$ hold for each job $j \in J$ and each machine $M_m \in M$. The order $j \to w$ of jobs $j \in J$ and $w \in J$ is optimal for processing jobs of set J if and only if at least one of the following three conditions holds:*

$$t_{w2}^U \le t_{w1}^L \text{ and } t_{j1}^U \le t_{j2}^L;$$

$$t_{j1}^U \le t_{w1}^L \text{ and } t_{j1}^U \le t_{j2}^L;$$

$$t_{w2}^U \le t_{w1}^L \text{ and } t_{w2}^U \le t_{j2}^L.$$

Let us consider instances with graph of the above partial strong order when no more than two jobs are not in this order at any time moment (see Fig. 1 for example).

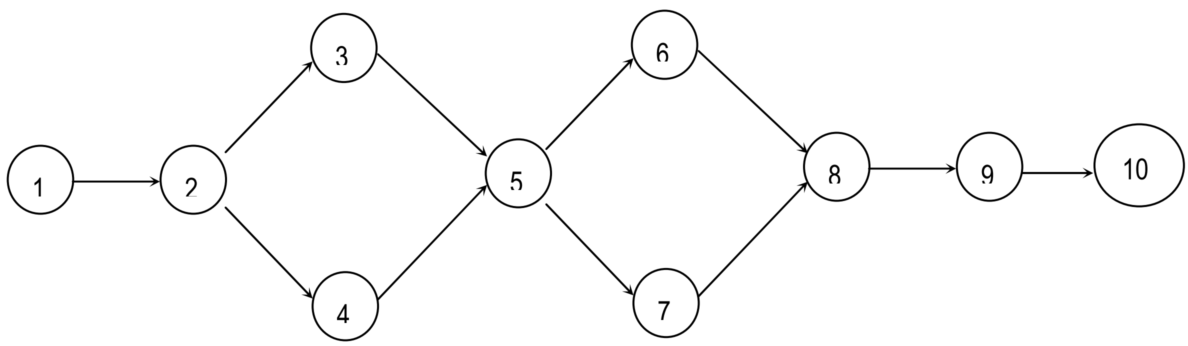


Figure 1. A graph reduction of the partial strong order constructed due to modified Johnson's algorithm

To answer the question of how many instances of the uncertain two-machine flow-shop scheduling problem have such form of the partial strong order (constructed due to Theorem 2), we made experiments on PC analogous to those presented in the previous section.

We call pair of jobs *conflict* pair of jobs if they are ready for processing but their optimal order is not fixed in the above partial strong order (see pair of jobs 3 and 4, and pair of jobs 6 and 7 in Fig. 1). Table 2 presents the percentage of instances in the series where no more than one conflict pair of jobs exists at any time. As we see from Table 2, increasing both numbers *n* and *L* decreases the number of instances with no more than one conflict pair of jobs existing at any time.

For each series we generated and tested 100 instances (for each combination of *n* and *L* under consideration). Random instances have been generated as follows. We tested problems with $n \in \{5, 10, 15, \ldots, 50\}$ jobs, integer job processing times uniformly distributed in the range [10,1000], and $L \in \{1, 2, 3, \ldots, 15\}$. For each job, the lower bound of the job processing times was randomly generated in the range [10,1000] and the upper bound of the job processing times was computed as follows: $t_{jm}^U = t_{jm}^L(1 + L\%/100\%)$. We restricted the experiments by $n \le 50$ due to the results presented in Table 2.

Table 2. Percentage of instances for which no more than one conflict pair of jobs exists at any time (the lower bounds of the processing times are uniformly distributed in the range [10, 1000] with $t_{jm}^U = t_{jm}^L(1 + L\%/100\%)$)

n\ L	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
5	5	8	8	16	18	25	26	20	25	29	37	40	38	36	46
10	15	29	28	37	51	50	54	52	52	63	60	45	55	53	42
15	32	44	48	59	63	55	48	40	45	37	38	33	33	19	23
20	41	65	47	48	45	47	34	36	22	17	13	6	10	5	4
25	52	63	60	42	21	18	14	8	9	8	6	5	2	1	1
30	67	53	50	20	20	12	8	3	3	0	0	0	0	0	0
35	68	51	33	24	10	2	1	0	0	0	0	0	0	0	0
40	69	34	22	5	4	0	0	0	0	0	0	0	0	0	0
45	57	35	18	9	1	0	0	0	0	0	0	0	0	0	0
50	50	21	10	0	0	0	0	0	0	0	0	0	0	0	0

For a conflict pair of jobs, the following sufficient conditions for constructing an optimal order for processing two jobs have been proven. If in the above partial strong order of jobs (with no more than one conflict pair of jobs at any time) two first jobs make conflict pair, then one can use the following sufficient conditions for optimal ordering of the conflict jobs.

Theorem 3 *Let jobs 1 and 2 make first conflict pair of jobs and let job 3 must be processed optimally after them. Then the order* $1 \to 2$ *of jobs* $1 \in J$ *and* $2 \in J$ *is optimal for processing jobs from set J if at least one of the following three conditions holds:*

$$t_{1,1}^U + t_{2,2}^U + \max\{t_{1,2}^U; t_{2,1}^U\} \le t_{1,2}^L + t_{2,1}^L + \max\{t_{1,1}^L; t_{2,2}^L\};$$

$$t_{2,2}^U + \max\{0; t_{1,2}^U - t_{2,1}^L\} \le t_{3,1}^L;$$

$$t_{1,1}^U + t_{2,2}^U + \max\{t_{1,2}^U; t_{2,1}^U\} \le t_{1,1}^L + t_{2,1}^L + t_{3,1}^L.$$

Let processing set of jobs *J* start at time τ_0 = 0, and at time $\tau_1 > \tau_0$ machine M_1 completed all the jobs before the next conflict pair of jobs. Without lose of generality, we assume that jobs (1, 2, …, *j* –1) are completed on machine M_1 in this order, jobs *j* and *j* +1 make next conflict pair of jobs, and job *j* +2 has to be processed optimally after this conflict pair of jobs. At time τ_1 we need to decide which job *j* or *j* +1 has to be processed next on machine M_1 in order to obtain an optimal schedule. It is clear that at time τ_1 the exact processing times of jobs (1, 2, …, *j* –1): $t_{1,1}, t_{2,1}, ..., t_{j-1,1}$, on machine M_1 are already known. We assume also that at time τ_1 the processing times of jobs {1, 2, …, *j* –1} on machine M_2 are also known. Thus, at time τ_1 the initial part $(t_{1,1}, t_{1,2}, ..., t_{j-1,1}, t_{j-1,2})$ of the vector $t \in T$ of the job processing times is already known.

Let c_1 denote the completion time of all jobs (1, 2, …, *j* –1) on machine M_1 and c_2 denote the completion time of all jobs (1, 2, …, *j* –1) on machine M_2.

Theorem 4 The o*rder* $j \to j+1$ *of conflict pair of jobs* $j \in J$ *and* $(j+1) \in J$ *is optimal for processing jobs of set J if at least one of the following seven conditions holds:*

$c_1 + t_{j1}^U \le c_2$; $c_1 + t_{j1}^U + t_{j+1,1}^U \le c_2 + t_{j2}^L$; (1)

$c_1 + t_{j1}^L > c_2$; $t_{j+1,1}^L + t_{j+2,1}^L \ge t_{j2}^U + t_{j+1,2}^U$; $t_{j+1,1}^U \le t_{j2}^L$; (2)

$t_{j+1,1}^L > t_{j2}^U$; $c_1 + t_{j1}^L + t_{j+1,1}^L \ge c_2 + t_{j2}^U$; $t_{j+2,1}^L \ge t_{j+1,2}^U$; (3)

$c_1 + t_{j1}^L \le c_2$; $c_1 + t_{j1}^U > c_2$; $t_{j+1,1}^U \le t_{j2}^L$; $t_{j+1,1}^L + t_{j+2,1}^L \ge t_{j2}^U + t_{j+1,2}^U$; (4)

$$c_1 + t^L_{j1} + t^L_{j+1,1} \le c_2 + t^U_{j2}\,; \qquad c_1 + t^U_{j1} + t^U_{j+1,1} > c_2 + t^L_{j2}\,; \qquad t^L_{j+1,1} > t^U_{j2}\,; \qquad t^L_{j+2,1} \ge t^U_{j+1,2}\,; \tag{5}$$

$$c_1 + t^L_{j1} \le c_2\,; \qquad c_1 + t^L_{j1} + t^L_{j+1,1} \le c_2 + t^U_{j2}\,; \qquad c_1 + t^U_{j1} > c_2\,; \tag{6}$$

$$t^U_{j+1,1} > t^L_{j2}\,; \qquad t^L_{j+1,1} + t^L_{j+2,1} \ge t^U_{j2} + t^U_{j+1,2}\,; \qquad t^L_{j+2,1} \ge t^U_{j+1,2}\,;$$

$$c_1 + t^L_{j1} > c_2\,; \qquad t^U_{j+1,1} > t^L_{j2}\,; \qquad t^L_{j+1,1} \le t^U_{j2}\,; \qquad t^L_{j+1,1} + t^L_{j+2,1} \ge t^U_{j2} + t^U_{j+1,2}\,. \tag{7}$$

The experiments on PC have been realized for calculating percentage of the number of conflict pair of jobs resolved due to Theorem 3 and Theorem 4 (Table 3) and percentage of instances solvable exactly due to Theorems 3 and 4 (Table 4). It is clear that only serious of instances which has non-zero value in Table 2 were considered in Tables 3 and 4 (each serious with zero value in Table 2 is indicated by symbol **–** in Table 3 and in Table 4). For each series we generated and tested 1000 instances with no more than one conflict pair of jobs at any time (for each combination of *n* and *L* under consideration). To this end, we used conditions of Theorem 2.

We tested problems with $n \in$ {5, 10, 15, …, 50} jobs, integer processing times with lower and upper bounds uniformly distributed in the range [10,1000], and $L \in$ {1, 2, 3, …, 15}. We restricted the experiments by $n \le 50$ due to the results given in Table 2. For each job, the lower bound of the job processing times was randomly generated in the range [10,1000] and the upper bound of the job processing times was computed as follows: $t^U_{jm} = t^L_{jm}(1 + L\% / 100\%)$.

Table 3. Percentage of the number of conflict resolutions due to Theorem 3 and Theorem 4

n\L	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
5	73.53	71.21	72.89	70.73	69.91	69.19	69.03	70.32	68.88	67.25	68.28	65.96	66.16	65.17	63.47
10	92.32	91.8	92.76	91.46	91.96	93.33	90.83	91.08	91.38	90.64	90.67	90.54	90.71	90.07	91
15	96.97	97.41	96.85	96.13	97.12	96.89	96.31	97.12	96.8	97.04	97.31	96.14	96.62	95.96	96.07
20	98.3	97.78	98.17	98.13	97.52	97.86	98.17	98.19	98.18	98.08	97.84	97.65	98.17	97.67	97.56
25	98.55	99.04	99.07	98.75	98.73	98.57	98.5	98.59	99	98.73	98.63	98.55	98.68	98.52	98.19
30	99.2	99.18	99.03	99.11	99.01	99.03	99.13	99.11	98.82	-	-	-	-	-	-
35	98.76	99.22	99.34	99.28	99.34	99.48	99.2	-	-	-	-	-	-	-	-
40	99.36	99.45	99.57	99.63	99.47	-	-	-	-	-	-	-	-	-	-
45	99.42	99.72	99.57	99.57	99.42	-	-	-	-	-	-	-	-	-	-
50	99.3	99.51	99.49	-	-	-	-	-	-	-	-	-	-	-	-

Table 4. Percentage of the instances with all conflicts being resolved due to Theorem 3 and Theorem 4

n \ L	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
5	78.3	75.6	77.5	75.5	74.8	73.8	73.6	74.4	74	72.3	73	70.3	71.6	70	69
10	92.4	91.6	92.4	90.7	90.9	92.2	88.9	89	89.2	88	88.1	86.8	86.6	85.5	87.9
15	96.7	96.7	95.8	94.5	95.1	94.6	93.3	94.2	92.7	93.6	93.6	90.5	91.6	89.1	88.9
20	97.8	96.5	96.6	95.9	94.1	94	94.7	94.8	93.6	93	92.4	90.5	92.1	89.9	90.2
25	97.9	98.1	97.6	96.3	95.6	94.3	93.3	93.4	94.9	93.3	92.2	91.5	92	90.7	88.1
30	98.4	98	97.2	96.1	95.7	94.9	94.4	94.8	94	91.5	-	-	-	-	-
35	97.4	97.4	97.2	96.3	95.9	96.4	94	-	-	-	-	-	-	-	-
40	98.4	97.7	97.7	97.6	96	-	-	-	-	-	-	-	-	-	-
45	98.2	98.6	97.1	96.5	94.5	-	-	-	-	-	-	-	-	-	-
50	97.5	97	95.9	-	-	-	-	-	-	-	-	-	-	-	-

From Tables 3 and 4 it follows that the most randomly generated instances of the uncertain problem $F2 \mid t_{jm}^L \le t_{jm} \le t_{jm}^U \mid C_{\max}$ (each of which has no more than one conflict pair of jobs at any time) are solved exactly during realization of the process due to sufficient conditions given in Theorems 3 and 4 (in spite of the uncertain job processing times before scheduling).

Conclusion

It is clear that in spite of uncertainty of the job processing times it is necessary to choose only one schedule for practical realization of the process. Theorem 1 allows obtain a dominant permutation for problem $F2 \mid t_{jm}^L \le t_{jm} \le t_{jm}^U \mid C_{\max}$ before realization of the process. Indeed, such a dominant schedule (if any) is the best one for any feasible realization of the job processing times. If condition of Theorem 1 does not hold, then one can use Theorem 2, Theorem 3, or Theorem 4 for constructing optimal schedule during realization of the process under consideration.

Clearly, this approach is useful if the level of uncertainty is low enough (the best results were obtained for the non-zero cases presented in Table 2). If level of uncertainty exceeds a certain threshold, then others approaches to problem $F2 \mid t_{jm}^L \le t_{jm} \le t_{jm}^U \mid C_{\max}$ outperform the approach based on Theorems 1 – 4. If there is no possibility to construct one dominant schedule for problem $F2 \mid t_{jm}^L \le t_{jm} \le t_{jm}^U \mid C_{\max}$ (i.e., conditions of Theorem 1 do not hold), it may be fruitful to construct more general schedule form on the basis of a partial strong order of jobs *J* (Theorem 2). Then one can consider the realization stage of a schedule for the flow-shop when a part of the schedule is already realized. Theorems 3 and 4 show how to use additional information about realized operations in order to obtain better solution than that constructed before scheduling. In such a case, a realistic solution process can be seen as consisting of *static* and *dynamic* phases. At the static phase, a scheduler can construct a family of the dominant permutations. At the dynamic phase of the decision-making, a scheduler has to select an appropriate schedule from such a family of the dominant permutations to react in real-time to the actual processing times of the already completed jobs.

Thus, our approach falls into the category of predictive-reactive scheduling. The static phase (based on Theorems 1 and 2) may be considered as predictive scheduling and dynamic phase (based on Theorems 3 and 4) may be considered as reactive scheduling (see [Aytud *et al.*, 2005; Gupta, Stafford, 2006]).

It is interesting to find sufficient conditions for choosing the unique permutation that is optimal for any feasible processing times of the remaining operations. If a scheduler cannot make right decision at time $\tau > 0$, he (she) has to use one of the solution policies, which does not guarantee to find an optimal schedule for any realization of the remaining job processing times. The solution policy may be optimistic or pessimistic (see [Aytud *et al.*, 2005; Shafransky, 2005]), or a scheduler can minimize objective function in average. For an uncertain problem it may be necessary to look for an optimal scheduling policy that stochastically minimizes the makespan [Ku, Niu, 1986; Pinedo, 1995]. To this end, it is necessary to obtain the reliable probability distributions for the random processing times. In general case, the choice of the job may also be based on minimization of possible loss of the objective function value.

The research was partially supported by INTAS (project 03-51-5501) and ISTC (project B-986). The authors would like to thank Ms. Natalja G. Egorova for the help in computational experiments.

Bibliography

[Aytug *et al.*, 2005] H. Aytug, M.A. Lawley, K. McKay, S. Mohan, and R. Uzsoy. Executing production schedules in the face of uncertainties: A review and some future directions. European Journal of Operational Research. 161, 86-110, 2005.

[Davenport, Beck, 2000] A.J. Davenport, and J.C. Beck A survey for techniques for scheduling with uncertainty. Unpublished, available from http://www.eil.toronto.ca/profiles/chris/chris.papers.html, 2000.

[Elmaghraby, Thoney, 2000] S. Elmaghraby, and K.A. Thoney. Two-machine flowshop problem with arbitrary processing time distributions. IIE Transactions, 31, 467-477, 2000.

[Gupta, Stafford, 2006] J.N.D. Gupta, and E.F. Stafford Jr. Flowshop scheduling research after five decades. European Journal of Operational Research. 169, 699-711, 2006.

[Jackson, 1956] J.R. Jackson. An extension of Johnson's results on job lot scheduling. Naval Research Logistic Quarterly, 3, 201-203, 1956.

[Jansen, Mastrolilli, Solis-Oba, 2005] K. Jansen, M. Mastrolilli, R. Solis-Oba. Approximation schemes for job shop scheduling problems with controllable processing times. European Journal of Operational Research. 167, 297-319, 2005.

[Johnson, 1954] S.M. Johnson. Optimal two- and three-stage production schedules with setup times included. Naval Research Logistics Quarterly, 1, 61—68, 1954.

[Ku, Niu, 1986] P.S. Ku, and S.C. Niu. On Johnson's two-machine flow-shop with random processing times. Operations Research. 34, 130-136, 1986.

[Leshchenko, Sotskov, 2005] N.M. Leshchenko, and Yu.N. Sotskov. Two-machine minimum-length shop-scheduling problems with uncertain processing times. In: Proceedings of XI International Conference "Knowledge-Dialogue-Solution", June 20-24, Varna, Bulgaria, 375-381, 2005.

[Pinedo, 1995] M. Pinedo. Scheduling: Theory, Algorithms, and Systems. Prentice-Hall, Englewood Cliffs. 1995.

[Slowinski, Hapke, 1999] R. Slowinski, and M. Hapke. Scheduling Under Fuzziness. Physical-Verlag, Heidelberg, New York. 1999.

[Shafransky, 2005] Ya.M. Shafransky. Scheduling problems with uncertain parameters: research directions and some results. Informatika. 3, 5-15, 2005 (in Russian).

Authors' Information

Natalja M. Leshchenko – United Institute of Informatics Problems of National Academy of Sciences of Belarus, Surganova str., 6, 220012, Minsk, Belarus; e-mail: leshchenko@newman.bas-net.by

Yuri N. Sotskov – Prof., DSc. United Institute of Informatics Problems of National Academy of Sciences of Belarus, Surganova str., 6, 220012, Minsk, Belarus; e-mail: sotskov@newman.bas-net.by

ACCESS RIGHTS INHERITANCE IN INFORMATION SYSTEMS CONTROLLED BY METADATA

Mariya Chichagova, Ludmila Lyadova

Abstract: All information systems have to be protected. As the number of information objects and the number of users increase the task of information system's protection becomes more difficult. One of the most difficult problems is access rights assignment. This paper describes the graph model of access rights inheritance. This model takes into account relations and dependences between different objects and between different users. The model can be implemented in the information systems controlled by the metadata, describing information objects and connections between them, such as the systems based on CASE-technology METAS.

Keywords: access control mechanisms, graph model, metadata, CASE-technology.

ACM Classification Keywords: D.2 Software engineering: D.2.0 General - Protection mechanisms.

Introduction

As information systems become larger and more complex, and as the number of their users increase, there are growing needs for methods that can simplify and even partly automate the process of access rights assignment.

The main problem of traditional access control mechanisms is that they don't take into account the relations between information objects.

The technology METAS [Lyadova, 2003], [Ryzhkov, 2002] allows to create dynamically adjusted information systems. Means of adaptation are based on use of the metadata describing information objects and connections between them from the various points of view. Functioning of information system is carried out through interpretation metadata by MDK METAS The metadata can form a basis for realization of mechanisms of the rights equivalence, in particular, the mechanism of the access rights inheritance that allows to lower labour input of assignment of the rights the manager of system. At the same time there are problems at definition of the rights of access on the objects accessible on several relations from parental objects to which various rights are appointed.

The model proposed in this paper take such relations in consideration and the rules for it are formulated.

To define the inheritance mechanism we shall formally describe model of distribution of the access rights. As basic elements of the model are used access graph and rules of its transformation.

Graph Model

An information system consists of objects and subjects. Access control describes whether specific subject can access specific object.

Let *O* is a set of objects and *S* is a set of subjects. $G = (V, E)$ is a finite directed labelled graph, where $V = O \cup S$ is a set of *nodes* and *E* is a set of *arcs*.

Notation $v_i \to v_j$ means that there is an arc $(v_i, v_j) \in E$ in graph *G*. A node $s_i \in S \subseteq V$ is called *subject-node* and a node $o_i \in O \subseteq V$ is called *object-node*.

If $v_i \to v_j$ and both of these nodes are subject-nodes (or both of them are object-nodes) node v_i is called *parent* of node v_j and node v_j is called *child* of node v_i.

Subject-nodes which have not got any children are called *users* all other subject-nodes are called *groups*.

Object-nodes which have not got any children are called *leaves* all other subject-nodes are called *roots*.

Arcs between objects present the relations between them. The arc's direction depends on type of relationship between objects:

- 1 : 0..1 ⇒ arc from "0..1" to "1";
- 1 : M ⇒ arc from "1" to "M";
- 0 : 1..M ⇒ arc from "0" to "1..M";
- M : M ⇒ bidirectional arc.

For example, for part of the database scheme which is shown on fig. 1 the subgraph on fig. 2 is fitted.

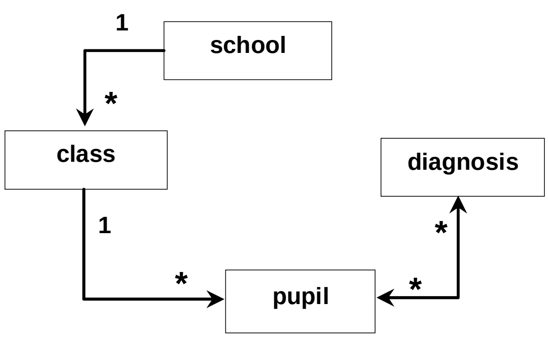


Figure 1. Database Scheme Fragment

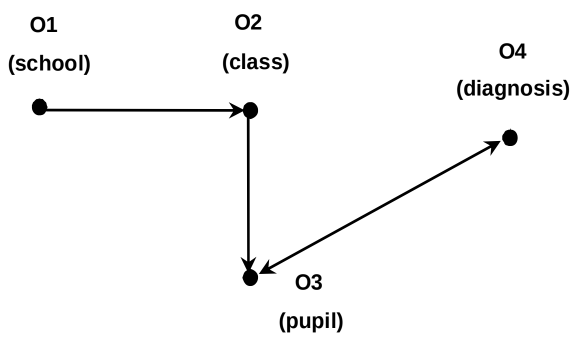


Figure 2. Object-nodes Subgraph

Each of the subjects must be connected by the arc to each of the objects. An arc between subject-node and object-node is called *access arc*.

An access arc's label determines the assigned access right of this subject to the object. *Assigned access right* is determined by the information system's administrator and it can deny or allow access to information objects.

Access rights that take into account relations between subjects and objects are called *actual access rights*.

Subject-nodes

In this section relations between subjects are considered and the rules that can take them into account are formulated. Access rights that allow for relations only between subjects are called *actual subject's access rights*.

Let $parent(s_i) = \{s_j \in S \mid \exists\, (s_j, s_i) \in E\}$ – is a set of parents for subject-node $s_i \in S$.

Let access arcs have following labels:

- $right(s_i, o_j) \in \{0, 1, 2, 3\}$ is an *assigned access right* of subject $s_i \in S$ to object $o_j \in O$, where 0 means that right is not assigned, 1 – access is denied, 2 – subject has a partial access and 3 – access is allowed. *Partial access* means that access is allowed only if certain conditions are fulfilled. These conditions are determined by administrator.
- $a\_right(s_i, o_j) \in \{1, 2, 3\}$ is an *actual subject's access right* of subject $s_i \in S$ to object $o_j \in O$, where 1 means that access is denied, 2 – subject has a partial access and 3 – access is allowed.

The subject's access rights can depend on its parents' rights. The process of determination the actual subject's rights is called *subject's rights inheritance*.

Let $s_i \in S$. The inheritance should be done according to the following two rules.

Rule 1. If subject has an assigned right $right(s_i, o_j) \neq 0$ to object $o_j \in O$, the actual subject's right is determined as $right(s_i, o_j)$, i.e.

$$a\_right(s_i, o_j) = right(s_i, o_j) \quad (1)$$

The main idea of this rule is that explicit assignment is more significant than inheritance.

Rule 2. If an access right to object $o_j \in O$ isn't assigned, i.e. $right(s_i, o_j) = 0$, the actual subject's right is determined as maximum of its parents' actual rights :

$$a\_right(s_i, o_j) = \max_{s_k \in parent(s_i)} (a\_right(s_k, o_j)) \quad (2)$$

If an access right to object $o_j \in O$ isn't assigned and the subject hasn't got any parents its actual right is determined as 1. It means that access is denied.

Note that by definition the maximum value for *a_right* is 3 (that means that access is allowed).

For finding actual subject's rights the above two rules are recursively applied.

Object-nodes

In this section relations between objects are considered and the rules that can take them into account are formulated. Access rights that allow for relations only between subjects are called *actual object's access rights*.

Note that the same object can be connected with different objects and rights can depend on the object from which the access is done.

As access rights depend on access context let define $context(s_i, o_j)$ as a current *access context*, i.e. list of object-nodes (path from one of the roots to current object-node).

Parent from context is an object-node from the access context which is the parent for node $o_j \in O$. Let $c\_parent(o_j)$ is a parent for object-node $o_j \in O$ from context.

Let access arcs also have following labels:

- $o\_right(s_i, o_j) \in \{1, 2, 3\}$ is an *actual* o*bject's access right* of subject $s_i \in S$ to object $o_j \in O$, where 1 means that access is denied, 2 – subject has a partial access and 3 – access is allowed.

Let arcs between object-nodes have the following labels:

– *inherit*(o_k, o_j) $\in$ {*true*, *false*} shows the possibility of inheritance. Let *inherit*($\emptyset$, o_j) = *false* that means that inheritance from empty object is forbidden.

The process of determination the actual object's rights is called *object's rights inheritance*.

Let $s_i \in S$. The inheritance should be done according to the following three rules.

Rule 3. If subject has an assigned right *right*(s_i, o_j) $\neq 0$ to object $o_j \in O$, the actual object's right is determined as *right*(s_i, o_j), i.e.

$$o\_right(s_i, o_j) = right(s_i, o_j) \quad (3)$$

This rule is the same as the rule 1 for subjects' rights inheritance.

Rule 4. If an access right to object $o_j \in O$ isn't assigned, i.e. *right*(s_i, o_j) = 0, and *inherit*(o_k, o_j) = *false* where o_k = *c_parent*(o_j) the actual object's right is determined as prohibition of access, i.e.

$$o\_right(s_i, o_j) = 1 \quad (4)$$

This rule means that if the inheritance is forbidden in current context and there are no assigned rights the access is denied.

Rule 5. If an access right to object $o_j \in O$ isn't assigned, i.e. *right*(s_i, o_j)=0, and *inherit*(o_k, o_j)=*true* where o_k=*c_parent*(o_j) the actual object's right is determined as follows:

$$o\_right(s_i, o_j) = o\_rigth(s_i, c\_parent(o_j)) \quad (5)$$

In order to determine actual access rights in rules 3, 4, 5 we should use *a_right* instead *right*.

Conclusion

Using of access rights inheritance allows to simplify the access rights assignment by automatic taking into account relations between object and subject. This method also permits to avoid some kind of mistakes which can be made by information system's administrator.

In addition to the means described in the given article means of the control of correctness of obvious assignment of the access rights for objects and their attributes are offered also [Mikov, 2003].

The architecture of the CASE-system METAS, the models of the metadata used in this system, and principles of its functioning are described in several articles [Lyadova, 2003], [Ryzhkov, 2002].

The system METAS is developed on the .NET platform. The technology ADO.NET, providing independence from DBCS, is used for access to the objects in database.

Bibliography

[Lyadova, 2003] L.N. Lyadova, S.A. Ryzhkov. CASE-technology METAS. In: Scientific Articles Collection “Mathematics of Program Systems”. Perm University, Russia, 2003, pp. 4-18.

[Mikov, 2003] A.I. Mikov, M.V. Chichagova. The Control over Rights Assignment. In: Scientific Articles Collection “Mathematics of Program Systems”. Perm University, Russia, 2003, pp. 207-212.

[Ryzhkov, 2002] S.A. Ryzhkov. The Concept of the Metadata in the Development of Information Systems. In: Scientific Articles Collection “Mathematics of Program Systems”. Perm University, Russia, 2002, pp. 25-35.

Authors' Information

Mariya Chichagova – Perm State University, Assistant of the Department of Computer Science; PSU, 15, Bukirev St., Perm, 614990, Russia; e-mail: chichagova@dom.raid.ru

Ludmila Lyadova - Institute of Computing, Deputy Director; 19/2-38, Podlesnaya St., Perm, 614097, Russia; e-mail: lnlyadova@mail.ru

THE COMPUTER SYSTEM FOR THE ESTIMATION OF THE REALIZABILITY OF THE CREW ACTIVITY ALGORITHMS

Boris Fedunov, Denis Vidruk

Abstract. When designing specification on-board algorithm (the algorithm, realized on on-board digital computing machine, and algorithm to activity of the crew necessary to conduct the estimation their realizing. Presented computer system allows in interactive mode with user to value the temporary expenseses of the operator on processes decision making and their realizing, participations it in process of the spying.

Key words: the algorithms to activity of the crew, designing algorithm, realizing estimation.

ACM Classification Keywords: B.8.2 Performance Analysis and Design Aids; D.2.2 Computer-aided software engineering.

Introduction

During system designing of anthropocentric object onboard intelligence algorithms specifications, designers face with necessity of defined algorithms structure realization opportunity check. The algorithms of onboard intelligence intended for realization through algorithms of a crew member activity (CAA – crew activity algorithms) are estimated through their performance time interdictions. Correlation of the defined expenses with ones allowable on external conditions projected anthropocentric object will function, allows to determine the realizability of CCA designed structure. During system designing of onboard intelligence algorithms, operator activity algorithms are represented as the earl (ODE – operator decisions earl), which apexes are decisions accepted by the operator, the beginning and the end of tracking stages, and arcs – causal relation of apexes [1,2]. Computer system «GRO-otsenka» is developed for automation CAA structure estimation process. It is intended for use in designing system of anthropocentric object onboard algorithms of system-forming kernel specifications at a stage of development of their specification.

The technology onboard algorithms specifications development includes the following stages:

A) Development of naturally language technical document «crew – the onboard equipment» system logic». The text of the document is usually structured on typical situations (TS) of functioning of projected anthropocentric object and their problem subsituations. [3];

B) Onboard algorithms specifications development, including the algorithms intended for realization on onboard digital computers (ODCA), and algorithms of crew activity (CAA). The design stage is supported by computer system "Bort" [4];

C) An estimation of onboard algorithms designed specification realizability:

ODCA - by the onboard digital computing system.

CAA - by crew for set time.

The stage is provided with «GRO-otsenka» and «BTsVS - otsenka» computer systems

D) Estimation of onboard algorithms developed specification efficiency.

The stage is provided with system of computer-imitating mathematical models [5] of anthropocentric objects functioning typical situations (see for example [6,7]) and imitating mathematical model of algorithms of aim prediction operative level [1,5] in which the crew is necessarily involved.

Described computer system «GRO-otsenka» works in coordination with the mentioned above computer systems. It receives the information from computer system "Bort"in the coordinated format and prepares the initial data [8] for designed CAA interpretation in the computer-imitating mathematical models [5] of anthropocentric objects functioning typical situations.

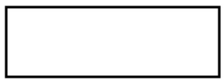
«GRO-otsenka» system represents the computer integrated environment of an estimation operators decisions earl, containing the interface of the designer, dialogue procedures for an estimation, completeness check procedures and a coordination of the data, procedures of an realizability estimation, estimation process support procedures.

Structure of the activity of operator in the technical anthropocentric object

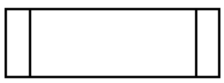
In the anthropocentric object the operator makes decisions by the operationally appearing problem, it realizes the decisions accepted and participates in different operations of tracking [9 - 13]. All decisions of operator are classified as the π-decisions (perceptive- identification), the ρ-decisions (speech-thought) and the π-ρ-decisions (heuristic) [9,13]. The class of the π-decisions is characterized only by search time, perception (quantity of operational units of perception (OUP)) and comprehension by the operator of necessary information. The process of decision making occurs instantly. Each ρ-decision with the design of the activity of operator is described through the composition of information, on the basis of which the operator must make this decision, also, through the algorithm of decision making. Each decision is described:

- by the input information: the composition of information in the information-control field of cab, on which the operator must make this decision; composition and the duration of voice communication, which is transferred to operator by cabin vocal informant and which is used with making of this decision;
- by the structure of the decision: a quantity of operational units of perception (OUP), composition and the sequence of the elementary acts of decisions working out, represented by the indicative symbolism on the personnel cabin indicators;
- by output information: by composition and by the sequence of manual operations, necessary for the realization of the accepted decision.

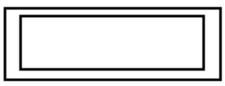
With the fulfillment of the operations of tracking it is assumed, what operator works in discretely - continuous regime, being distracted from the operation of tracking to the period of adoption and realization of the decision (decisions). After abstraction the operator again returns to the process of tracking. In this case the composition and the description of dynamic is sectional this servo system in the stage of design in question they are absent. There is only an idea about the dependence of the time of finalizing by the operator (τ-tracking) of that accumulating in the time of its abstraction (τ-otv) of following error. According to the developed requirements each apex of ODE represented for the estimation can belong to one of the established types (table 1):



Rectangles designate the algorithms of decision making with the indication of the type of decision (the π-decision, the ρ-decision, π-ρ-the decision). Each CAA-decision is always accompanied by the fragment of sequence information on the control indicators (it is possible vocal information, on which decision is chosen).



The apex with the appropriate number, designating the accepted decision algorithms of realization by the operator. Each CAA-realization is compulsorily accompanied by indication of the utilized elements of control and their arrangement or by indication of their generalized description.



Place of the replacement of the conceptual model of the operator behavior

Types of apexes are given below do not contribute time expenditures, in the general time estimation.

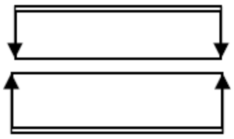
	The apex designating branching in ODE. It contains logical conditions for the actions of operator. It has two outputs: truth and lie.
	Apexes designating algorithms participations of operator in the work of tracking system, accordingly beginning and the end of the tracking.
(+)	Marker is placed on the arcs of earl in those places, where in the opinion of the developer of earl, operator can be distracted to the realization of the operation of tracking.

Table 1: Apexes types.

The apexes of earl CAA-decision and CAA-realization are characterized by time, operator spent for perception, comprehension, the production of decisions and fulfillment of manual operations. Apex, corresponding to the replacement of the conceptual model of the operator behavior, is characterized by the time of the replacement of this model. Between the apexes there are causal relationships (on the earl depicted as arcs). The beginning of the operator actions must be an operation of tracking, located in the root of earl; the activity of operator is turned from the root apex to the final (operation of tracking). The earl is represented with the treelike structure (one beginning, many ends).

The ODE "+" marker indicates the places of the possible (on semantics) distraction of operator to the process of tracking.

All apexes of the CAA type must be accompanied by information, on which with the aid of the experimental data, existing in the system «GRO-otsenka», the time of apex is determined. It is necessary to afford screen of control indicators. on the basis of which data, necessary for the description CAA will be compiled. Apex, designating the beginning of tracking, must be accompanied by table or earl of the dependence of the correction time on the distraction time

$$(\tau_{cor} = f(\tau_{dist})).$$

System Characteristic

«GRO-otsenka» system is intended for the automation of the evaluation process of the operator decisions earl realization. On the basis of data, entered manually by design engineer or exported from the system "Bort", «GRO-otsenka» system calculates ODE in the dialogue with design engineer:

- time, which operator will expend on decision making (CAA apexes), on its realization (CAA-R apexes), the replacement of the conceptual model of behavior;
- the arrangement (among the permissible points) of the points of the operator optimum switch to the tracking regime.

According to this information «GRO-otsenka» system calculates time (of each branch of the introduced operator decisions earl) spent by operator on all forms of its activity, indicated on this branch, and the mean-square deviation of this time. Such calculations are done for each branch of earl and on any selected sections inside any branch.

System automatically determines the ODE bottlenecks and exports the estimation results of the introduced earl for further processing in other programs.

Regimes of the design engineer work in the «GRO-otsenka» system.

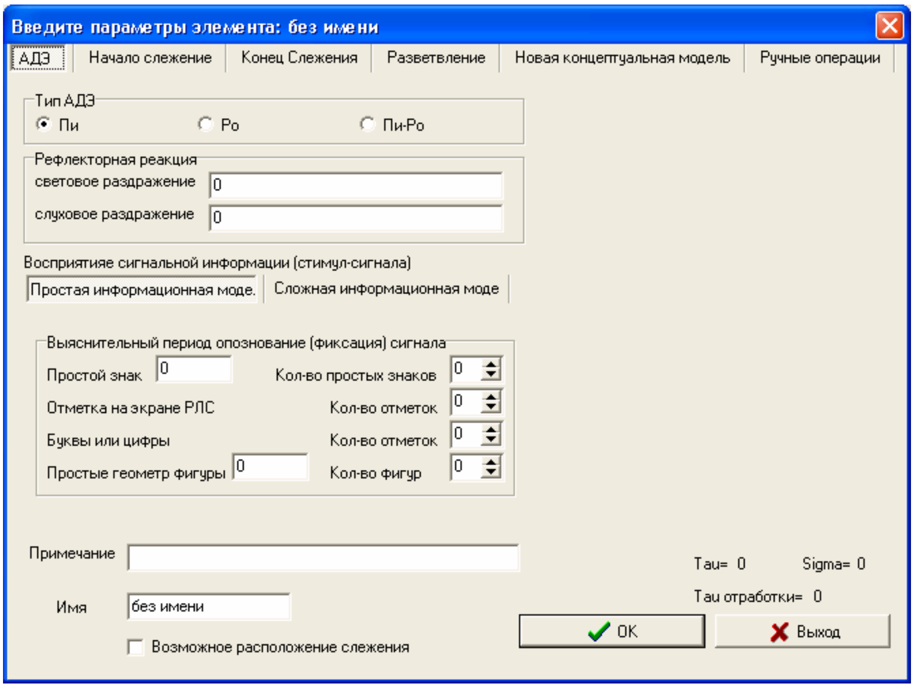
After the entering ODE into «GRO-otsenka» system the design and planning engineer fixes his attention in the concrete apex of ODE and selects the type of this apex from the following:

1. «АДЭ» – this type of apex is used for indication of the of crew activity algorithm.
2. «Начало слежения» - this type of apex is used for the indication of the beginning of the tracking operation.
3. «Конец слежения» - this type of apex is used for the indication of the ending of the tracking operation.
4. «Разветвление» - this type of apex is used for indication of branching.
5. «Новая концептуальная модель» - this type of apex is used for the indication of a conceptual model change.
6. «Ручные операции» - this type of apex is used for the indication of manual actions realization of the crew CAA-R.

Depending on the type of the selected apex, the corresponding supplementary sheet with its parameters is opened. There are parameters common for all types of earl apexes: "Имя" – name of the apex, "Примечание" - text field for any note for this apex, and the "Возможное расположение слежения" checkbox, which is placed in such a case of finalizing the deviation of the tracking parameters is possible after current of apex.

«АДЭ» apexes type.

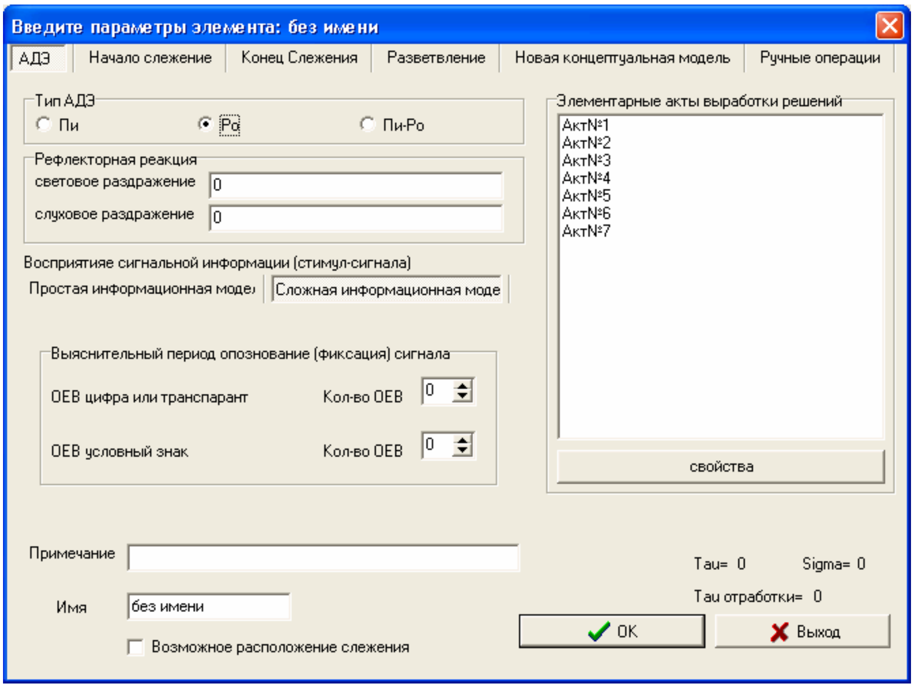
If «АДЭ» apex is selected, design engineer should select the type of the solution is used in this case. There are three accessible types: "π - decision", "ρ - decision" and " π-ρ decision" (page 1).



Page 1: Properties window of earl apexes.

When π-decision is selected, the edit fields for the data being entered according the estimation procedure of operator time expenditures [8] are available.

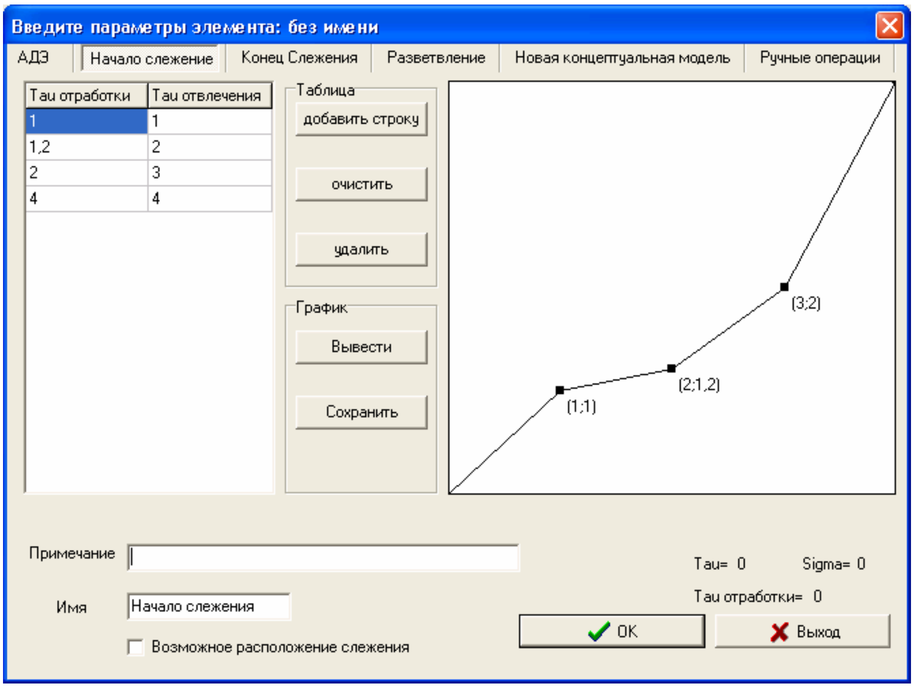
When ρ-decision is selected, the edit fields for the data being entered according the estimation procedure of operator time expenditures [8] are available (page 2).



Page 2: Estimation of CAA ρ -decision.

"Elementary reports of working out decisions" data sheet is also available.

When π-ρ-decision is selected the edit field for entering the time, spent on this decision, which is taken from the appropriate half-scale simulation.



Page 3: The beginning of the tracking operation.

«Начало слежения» apexes type.

When this type of apex is selected, the edit table for entering the dependence of the finalizing time on the time of abstraction for the tracking is available (page 3).

With the "добавить строку" button it is possible to add new line into the table. Button "удалить" makes it possible to remove line from the table. Button "очистить" removes all lines from the table.

Also there is a button "Вывести", which makes it possible to represent earl and button "сохранить" which allows to save earl as a graphic file.

«Конец слежения» apexes type.

This type of apex indicates that the tracking is finished. This type of ODE apex is logical and influences only the calculation of branches, without making changes in the results of time expenditures calculations.

«Разветвление» apexes type.

This type of apex indicates branching of the earl. This type of ODE apex is logical and influences only the calculation of branches, without making changes in the results of time expenditures calculations.

«Новая концептуальная модель» apexes type.

This type of apex indicates the change of conceptual model. This type of apex does not have parameters influencing the calculating results. The time of the change of conceptual model of operator behavior is equal of 1.2sek by default [13].

«Ручные операции» (CAA-R) apexes type.

This apex indicates realization of the accepted decision. Design engineer should manually fulfill the corresponding table (name of each manual operation and the time needed for it).

«Новая операция» button adds a new operation string into the table. «Удалить» button deletes selected operation string. «Очистить» button clears all the table.

Calculation of the earl parameters.

It is necessary to be convinced that the earl is built correctly. This checking is produced by looking through entered ODE with visual monitoring of the following positions:

1. The first apex of earl must have a type "Новое слежение".
2. All apexes with the type "Новое слежение" must have the filled table of tracking.
3. The last apex of earl must have a type "Конец слежения".
4. Cycles are not admitted in the earl - they must be opened to corresponding branches.
5. If the apex of graph is branching off and does not have the «Разветвление» type all apexes on the branch "no" they would not be considered in the calculations.
6. If the apex of graph has the «Разветвление» type and does not have the appropriate branches only existing branches ODE would be considered in the calculations.

If some of requirements for the earl are not fulfilled the system will give out the corresponding warning. For calculating of the earl parameters the «Пересчитать ветки» button should be pressed. ODE branches will be calculated with the time needed for their fulfillment (including the time spent on finalizing of tracking) and its mean-square deviation. Parameters of branch can be examined after selecting it in the menu of the «Параметры ветки» division. System also provides the verification of the assigned time limitations in the section of ODE branch.

Conclusion

The "GRO-otsenka" computer system allows to solve the problems of the realizability evaluating of designed ODE in the early stage of the onboard algorithms design. It significantly facilitates the task of ODE realizability evaluating, it makes it possible to obtain the result of estimation and the corresponding information immediately. The "GRO-otsenka" system has comfortable interface and sufficient execution speed (for the solution of the problems of this level). It is oriented as to manual input of initial data, so to obtaining of them directly from the "Bort" system.

Bibliography

1. Fedunov B.E. The Constructive semantics антропоцентрических systems for development and analysis specification algorithm of the on-board intellect. Izd. THE WOUNDS. T and SU. 1998г. 5
2. Poznyakov P.V., Fedunov B.E. The Bases to information integration of the on-board equipment. M.: Izd-in MAY, 1993г.
3. Nevzorova O.A., Fedunov B.E., System of the analysis technical text "LOT": the main concepts and design resheniya. // Izv. THE WOUNDS. TISU. - 2001. - 3.- S. 138-149.
4. Kondrikova T.A. Fedunov B.E. "Board" - a Computer system of support of the process of the designing on-board intellect specification. Izd. THE WOUNDS. T and SU. 1999г
5. Fedunov B.E. The Estimation to efficiency specification algorithm of the on-board intellect антропоцентрических ob'ektov. - M., Mehatronika, automation, management. 2002, 5.
6. Tkachenko A.P., Fedunov B.E. The Estimation to efficiency specification algorithm of the on-board intellect антропоцентрического ob'ekta. // Izv. THE WOUNDS. THE YEW. 2002. 5
7. Turnovskiy M.B., Fedunov B.E. The Comparison and estimation specification algorithm of the on-board intellect of the plane - a fighter in дуэльной of the situations distant air boa. Izv.THE WOUNDS. TISU. .6, 2004. Str.128-141.
8. Fedunov B.E. The Methods of the estimation реализуемости column of the decisions of the operator антропоцентрического object at development algorithm on-board intellect. 2002г., Moscow, GOSNIIAS.
9. The Bases to human engineering / Under ред. B.F. Lomova. M.: Vyssh. shk., 1977г.
10. Introduction to эргономику / Under ред. V.P. Zinchenko. M.: Sov. radio, 1974г.
11. Zarakovskiy G.M. The Psychological analysis to labor activity. M.: Science, 1967г.
12. Cibulevskiy I.E. The Persons as section watching systems. M.: Science, 1981г.
13. The Guide to human engineering / Under ред. B.F. Lomova. M.: Machine building, 1982г.

Autors' Information

Boris E. Fedunov – doctor of technical sciences, professor, graduated the Moscow Aviation Institute (State University of Aerospace Technologies) in 1960 as aircraft designing engineer, Moscow State University in 1965 as mathematician. Now he works at the State Research Institute of Aviation Systems and delivers a lectures in Moscow Aviation Institute on syntheses of on-board algorithms systems of piloting aircrafts и applied systems analysis. He is member scientific council of the Russian Association of an Artificial Intellect (collective member of European Coordination Committee on the Artificial Intellect).

State Scientific Research Institute for Aviation System (GosNIIAS),
ul. Viktorenko 7, Moscow, 125319 Russia; e-mail boris_fed@gosniias.ru

Denis G. Vidruk – graduate student, graduated the Moscow Aviation Institute (State University of Aerospace Technologies) in 2005.

TABLE OF CONTENTS OF VOLUME 14, NUMBER 2