ITHEA

International Journal

INFORMATION TECHNOLOGIES & KNOWLEDGE

2011 Volume 5 Number 1

International Journal

INFORMATION TECHNOLOGIES & KNOWLEDGE

Volume 5 / 2011, Number 1

Editor in chief: **Krassimir Markov** (Bulgaria)

Victor Gladun (Ukraine)

Abdelmgeid Amin Ali	(Egypt)	**Koen Vanhoof**	(Belgium)
Adil Timofeev	(Russia)	**Larissa Zaynutdinova**	(Russia)
Aleksey Voloshin	(Ukraine)	**Laura Ciocoiu**	(Romania)
Alexander Kuzemin	(Ukraine)	**Luis F. de Mingo**	(Spain)
Alexander Lounev	(Russia)	**Natalia Ivanova**	(Russia)
Alexander Palagin	(Ukraine)	**Nataliia Kussul**	(Ukraine)
Alexey Petrovskiy	(Russia)	**Nelly Maneva**	(Bulgaria)
Alfredo Milani	(Italy)	**Nikolay Lyutov**	(Bulgaria)
Avram Eskenazi	(Bulgaria)	**Orly Yadid-Pecht**	(Israel)
Axel Lehmann	(Germany)	**Radoslav Pavlov**	(Bulgaria)
Darina Dicheva	(USA)	**Rafael Yusupov**	(Russia)
Ekaterina Solovyova	(Ukraine)	**Rumyana Kirkova**	(Bulgaria)
Eugene Nickolov	(Bulgaria)	**Stefan Dodunekov**	(Bulgaria)
George Totkov	(Bulgaria)	**Stoyan Poryazov**	(Bulgaria)
Hasmik Sahakyan	(Armenia)	**Tatyana Gavrilova**	(Russia)
Ilia Mitov	(Bulgaria)	**Vadim Vagin**	(Russia)
Irina Petrova	(Russia)	**Vasil Sgurev**	(Bulgaria)
Ivan Popchev	(Bulgaria)	**Velina Slavova**	(Bulgaria)
Jeanne Schreurs	(Belgium)	**Vitaliy Lozovskiy**	(Ukraine)
Juan Castellanos	(Spain)	**Vladimir Ryazanov**	(Russia)
Julita Vassileva	(Canada)	**Martin P. Mintchev**	(Canada)
Karola Witschurke	(Germany)	**Zhili Sun**	(UK)

International Journal "INFORMATION TECHNOLOGIES & KNOWLEDGE" (IJ ITK)
is official publisher of the scientific papers of the members of
the ITHEA International Scientific Society

IJ ITK rules for preparing the manuscripts are compulsory.
The **rules for the papers** for IJ ITK as well as the **subscription fees** are given on ***www.foibg.com***.

Responsibility for papers published in IJ ITK belongs to authors.

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

International Journal "INFORMATION TECHNOLOGIES & KNOWLEDGE" Vol.5, Number 1, 2011

Edited by the **Institute of Information Theories and Applications FOI ITHEA**, Bulgaria, in collaboration with:
Institute of Mathematics and Informatics, BAS, Bulgaria,
V.M.Glushkov Institute of Cybernetics of NAS, Ukraine,
Universidad Politécnica de Madrid, Spain.

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

Copyright © 2011 All rights reserved for the publisher and all authors.
® 2007-2011 "Information Technologies and Knowledge" is a trademark of Krassimir Markov

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

DESIGN, IMPLEMENTATION, AND TESTING OF A MINIATURE SELF-STABILIZING CAPSULE ENDOSCOPE WITH WIRELESS IMAGE TRANSMISSION CAPABILITIES.

Dobromir Filip, OrlyYadid-Pecht, Christopher N. Andrews, and Martin P. Mintchev

Abstract*: Video capsule endoscopy (VCE) enables examination of the small intestine. In large-lumen organs of the gastrointestinal (GI) tract (e.g. the stomach and the colon), the capsule tumbles around and therefore cannot image systematically. This limitation underscores the need for a novel approach that allows capsule imaging of these organs without tumbling. This paper describes the design and implementation of a self-stabilizing capsule prototype for colonic imaging. The present design consists of a capsule endoscope (CE) coupled to an expandable stabilizing component comprising a liquid-permeable sac filled with dry superabsorbent polymer granules (swellable material) integrally covered by an outer colon-targeting coating. Once the capsule enters the colon, the outer coating dissolves and this allows the expansion of the swellable material attached to the back side of the capsule endoscope. This volumetric increase of the expandable component provokes peristalsis by activating colonic mass reflex. The expanded end of the capsule stabilizes the entire implement, preventing it from tumbling, and at the same time activating the capsule endoscope. Once activated, the capsule begins to record images and to transmit them to an external receiver, which records the data to a computer. As a safety measure, the expandable component can be electronically separated from the capsule at any time. The capsule is eventually expelled out of the body as a fecal matter. A prototype of the self-stabilizing capsule has been developed conceptually and electronically to fulfill the requirements of image stabilization while coping with the restrictions of commercial CEs. Self-stabilizing CEs and non-stabilized CEs were comparatively tested in laboratory and canine experiments. The self-stabilizing CE eliminated the tumbling effect and demonstrated its potential to greatly improve colonic imaging.*

Keywords*: Colon, Gastrointestinal Tract, Video Capsule Endoscopy, Wireless Image Transmission,*

ACM Classification Keywords*: A.0 General Literature - Conference proceedings; B.7.1. – Advanced Technologies.*

Introduction

A. Capsule Endoscopy

Video capsule endoscopy (VCE) is a new technology that has been developed for investigation of the small bowel [1]. VCE represents a major advance in screening the small intestine because it is simple, safe, non-invasive, reliable, and well tolerated by patients [2].

The size of the small intestine below the duodenum is just right to allow the capsule to be carried through this tubular organ without significant tumbling [3]. Thus, in this part of the small intestine, the entire surface is visualized as the capsule travels through [4]. Captured images are transmitted through the human body to an external receiver [5]. No control of travel, orientation and/or posture is available because the device is conveyed

by natural peristalsis [6]. Therefore, mass lesions located in the proximal small bowel may be missed on the basis of capsule velocity or tumbling [7]. There are several possible reasons that small bowel mass lesions are missed by capsule endoscope (CE). As the capsule passes along the duodenal sweep, it may increase in speed resulting in a missed lesion [8].

The time of passage of a capsule through the entire gastro intestinal (GI) tract is about 8–10 h, with approximately 1 h spent in the stomach, 4 h in the small intestine, and 5 h in the colon [9]. However, GI motility varies widely in the general population. Twenty-four to 48 hours for a CE to go through the entire GI tract as a result of its passive movement were reported in some cases [10]. Generally, in 20% to 30% of the VCE procedures, the capsule does not reach the cecum within recording time, with incomplete imaging of the small bowel, which limits the value of CE [11,12].

Recently, commercial CEs have also been extending their reach into other parts of the GI tract including the esophagus and the colon [13]. However, with the current design approach, the application of VCE is currently limited to small-lumen organs [14].

B. The Problem of Information Transmission in the Colon

The colon measures about 150 cm in length and 6 cm in diameter [15]. It consists of at least two anatomically distinct sections: 1) the sigmoid, which is relatively narrow and typically has one or several sharp angulations or bends, and 2) the more proximal colon which is distinguished by relatively straight sections that progressively widen in luminal calibre. These disparities in the anatomy of the colon create the greatest challenge for VCE imaging. Thus, it is important to emphasize that current capsule endoscope design is not suitable for colonic imaging [16]. Autonomous swallowable video capsule has dimensions (31 mm x 11 mm) that are smaller than the lumen diameter of the colon and the capsule may toss, tumble, turn, and vibrate as it passes through the colon [17,18]. With the tumbling of the capsule in this wider space, important lesions may be missed and generally the entire internal surface cannot be viewed [19]. During the capsule transit, the lens hood can be buried in the mucosa of the colon resulting in an incomplete imaging [20]. The capsule may stagnate in some sections of the colon, may pass quickly over other sections, and may retract for a distance [21]. Therefore, it may be difficult to determine the exact location of the in-vivo device in space, in what section of the colon a point of interest pathology may have been detected or important pathologies could be completely missed [22].

During a quick passage of the tumbling capsule in large lumens, the solid capsule body can create air bubbles in the liquid environment that along with debris can significantly obscure the view of the colonic mucosa. Such obscuring of the visual field may reduce the diagnostic yield. Therefore, it is essential to reduce these potential causes of poor visualization by eliminating the tumbling and optimizing the bowel preparation for capsule endoscopy [23]. In addition, collapsed bowel, large folds, cincturing of bowel by tumors and blood in the lumen may obscure visualization of an underlying mass lesion with the current design of capsule endoscopes [24,25].

The PillCam Colon capsule (Given Imaging, Yoqneam, Israel) is the only CE currently in use for colonic investigation. In the most recent study of 56 patients, colon capsule endoscopy (CCE) was followed by conventional colonoscopy (CSPY). Polyp detection rate (per patient) was 50% (n = 28) for CSPY and 62% (n = 35) for CCE. For relevant polyps (>5 mm) there was a correspondence in the detection rates of both methods ($p < 0.05$). The mean sensitivity was 50% (95% confidence interval [CI], 19 to 81), the mean specificity was 76% (95% CI, 63 to 86), the positive predictive value (PPV) was 20% and the negative predictive value (NPV) was 93% [26]. An European study from 2009 has shown CCE being able to diagnose 64% of total polyps, which were

detected by CSPY [27]. These results indicate the general problem of CE tumbling during its transit in the colon and the need for CE stabilization.

It has been hypothesized that VCE may, by reducing the indignity and discomfort of endoscopic procedures, have the additional benefit of encouraging many who currently forgo medically recommended examinations (particularly colonoscopy) to have it done [28].

The VCE could have a place as a non-invasive screening procedure for colon cancer, which can be performed more safely, more tolerably, and more often than standard colonoscopy. The issue to be solved is improvement of the sensitivity of this modality by eliminating the unwanted tumbling and tossing effects [29].

C. The e-Stool Concept

The e-Stool concept seeks to eliminate the unwanted tumbling and tossing effects of CE traversing the colon by an expandable stabilization apparatus attached to the back-end. The stabilization apparatus can be narrow and short enough to negotiate the narrow portions of the colon, yet it adjusts to the wider luminal calibre of the organ. This design brings a new revolutionary VCE concept to increase the sensitivity and provide systematic visualization of the large intestine.

As the expandable component is fluid-permeable, an obstruction in the small intestines is not created. The risk of stagnation of the capsule in the case of bowel obstruction is eliminated by installing a separation mechanism within the capsule. As the expandable component is activated it can be electronically separated from the capsule at any time.

In addition, the expandable component is biocompatible and is able to disintegrate in the GI tract after a certain period of time.

D. Aim of the Study

The aim of the present study was to design a 1:1 prototype of the e-Stool CE, and to demonstrate the practical feasibility of VCE self-stabilization and separation concepts.

Method

The complete block diagram of the self-stabilizing CE prototype is shown on Fig.1.

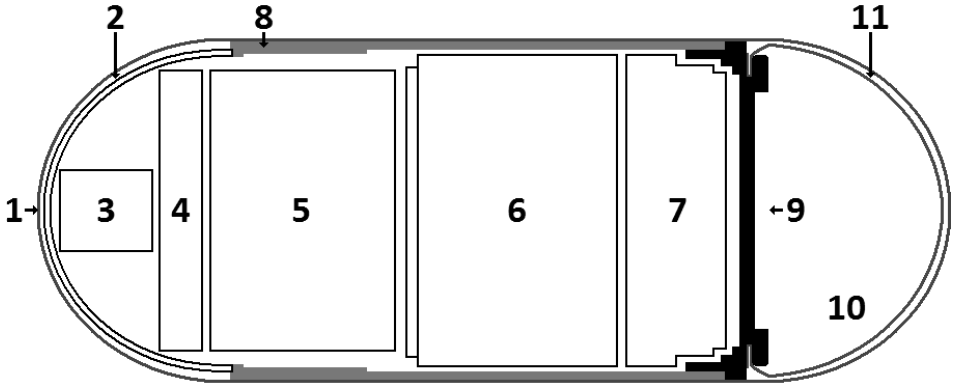


Fig. 1: Self-stabilizing CE Prototype. 1 – outer coating, 2 – optical dome , 3 – longitudinally positioned CMOS color camera, 4 – illumination system (LEDs), 5 – transmitter, 6 – battery, 7 – micro-heater, 8 – tapered cylinder, 9 – sealable lid, 10 – superabsorbent granules, 11 – PGA mesh.

The design consists of an outer coating, capsule endoscope and expandable component, which comprises the superabsorbent polymer granules contained in a porous PGA mesh that is attached to the back-end of the capsule using polydioxanone (PDS) suture. The CE comprises of complementary metal–oxide–semiconductor (CMOS) camera, radio frequency (RF) transmitter, light emitting diodes (LEDs) illumination system, battery, and a microheater. The outer coating is made of the colon-targeting agent Eudragit, a coating resistant to gastric acid that dissolves only in the intestinal tract at a pH range of 5.5-7 [30]. Therefore, when the capsule reaches the lower portion of the small intestine (the ileum) or the colon, the Eudragit coating starts to dissolve due to the change in pH. The electronics inside the capsule is activated by moisture switch and the expansion of the stabilizing material is initiated when the gastric juices start permeating the PGA mesh. The expansion process is continuous and results in volumetric and mass increase of the stabilizing component. In general, the peristaltic reflex can be elicited in the large intestine by local distension of the colonic wall, which involves activation of both ascending excitatory and descending inhibitory neural reflex pathways to the neighboring colonic smooth muscle layers [31]. Therefore, the volumetric increase activates the peristaltic reflex in the colon, which ultimately provokes natural peristalsis that propels the capsule. The expansion is completed before the peristalsis starts, allowing systematic colonic visualization starting in the cecum. As the natural peristalsis moves the capsule smoothly and painlessly through the colon, an intraluminal video is acquired at an appropriate frame rate. This video is subsequently wirelessly transmitted to a PC via an RF link. The expanded implement at the back-end of the capsule stabilizes it and prevents it from tumbling while it slides through the colon towards the rectum.

Some portions of the colon are convoluted, namely the cecum, the hepatic flexure, the transverse colon, the splenic flexure, and the sigmoid. [32]. As a safety layer to eliminate device obstruction in the gut, a separation mechanism was designed. After the expandable component has been activated it can be electronically detached from the capsule at the operator's command by enabling a fast reacting micro-heater placed onto the capsule cap. The stabilizing (expandable) component is affixed compactly to the capsule using a mechanism made of low melting point biodegradable PDS suture. The micro-heater activates the heating of a Ni-Chrome filament in the presence of a magnetic field. The heating filament reaches a suitable temperature to melt the PDS suture (≈107˚C) when a current is fed through it. This results in eliminating the point of attachment between the capsule and the expandable component, leaving the PGA mesh opened. The natural peristaltic movement of the colon completes the separation of the expandable component and the capsule. The superabsorbent polymer granules, PGA mesh, and PDS suture are bio-absorbable and biocompatible, and therefore can safely be disintegrated in the GI tract after a relatively short time period.

A. Design of the Capsule Body and the Optical Dome

The capsule endoscope must be sealed and resistant to decay within the gut to protect the internal components from conditions inside the patient's body. It is composed of three parts, a tapered cylinder, an optical dome window, and a sealable cap. The dimensions of the capsule were determined to allow it to fit inside the outer casing (Torpac #13 3.2 ml; Torpac Inc., Fairfield, NJ, USA), and to leave enough space for the 0.7 ml expandable component required for the stabilization. The design encompassed threaded cap with a recessed groove, which was used to compactly attach the stabilizing component to the body of the CE using PDS suture. The threaded cap provided a water-proof seal for the capsule and at the same time it maintained a point of attachment for the stabilizing component. The cap material that was chosen had to encompass the following properties, high wear resistance, low surface friction, high strength, and great rigidity. Hence aluminum was selected as the cap

material for the prototype. The final step was to seal the cap to the capsule body using an ultra-thin soft polytetrafluoroethylene (PTFE) film.

An optical dome of hemispherical shape was selected, because it is easier for a patient to swallow the capsule, and body fluids cannot remain on the front cover [33]. The sturdy optical dome is in front of the CE to improve the luminous uniformity and to match the lens in the back of the dome. This dome houses the CMOS imager, the lens, and the illumination system. The lumen of the bowel is illuminated through the optical dome by an illumination system. The images are to be obtained as the gut wall sweeps past the optical dome of the capsule during peristalsis. Subsequently, the obtained image is focused by a short focal aspherical lens onto the CMOS imager.

The intensities of reflections from interfaces of the optical dome are minimized when the angle of incidence and the angle of reflection are near zero, in other words, rays of light are incident perpendicular to the plane of the interface [34]. The optical dome has a shape that prevents the light reflected by the dome itself reaching and saturating the imager thus degrading image quality.

The optical dome was designed from a glass which was selected and adjusted for this design at 1-mm thickness. The dome was attached to the tapered part of the cylinder using a medical grade biocompatible epoxy resin EPO-TEK® 301-2 (Epoxy Technology, Inc., Billerica, MA, USA).

Another important efficiency consideration is how much of the light from each LED source can be collected by the optics. Distribution pattern-matched efficiency accounts for how well the optics is expected to direct the light from the LEDs into the desired distribution shape. The pattern-matched efficiency worsens when the design volume is constrained since the optic is relatively small and close to the LEDs. Light collected by any point on the optics will subtend a greater solid angle when compared to a larger optics. Smaller optics has less control over how well light is redirected.

The distance between the lens and the optical dome was chosen to be 5.4 mm. The outside curvature of the optical dome was 6.7 mm, the inside curvature was 5.9 mm, and thickness was 0.8 mm.

The final CE endoscope prototype length was 22 mm and its outer diameter was 14 mm.

B. Design of the Imager

The proposed design uses a CMOS imager (color camera) which captures images of the colon walls and also suits the space constraint in the capsule.

In the front of the CMOS imager is a short focal-length aspherical lens (0.77mm/f3.0<110°>) that is manually focused at the optical dome and a few centimeters beyond, so that the bowel laying against the optical dome would be in focus. Wide-angle lens selection permits more efficient examination of the large intestine. Improvements in image quality can be achieved by careful attention to the positioning of the lens relative to the imager. The position of the entrance pupil of the optical lens should coincide with the center of the curvature of the optical dome window to prevent any of the reflected light from entering the CMOS imager [35, 36]. A special lens holder was used to position miniature lens over the CMOS imager with high degree of accuracy using a microscope. Lens holder kept the lens' optical axis centered on the active area of the imager and perpendicular to the imager surface.

Initially the lens holder was positioned and tightly fixed to the imager to avoid dislocation during capsule transit. The top of the holder contained a thread which allowed screwing in the lens with the same thread type. The final

optical arrangement allowed the tissue to be in focus even if it is in contact with the optical dome window but also to remain in focus over 4cm if the lumen is opened with residual air or fluid. The imager uses a narrow aperture to increase the depth of field.

The selected image sensor for the final prototype was the 1/18" Color CMOS camera MO-B802-105 (RF-LINKS, Toronto, Ontario, Canada). The camera operates at 3V, 18-19mA, its minimum required illumination is 2Lux at f1.2, and number of effective pixels is 320x240 (NTSC). This CMOS imager is very precise and can detect objects (polyps) as small as 2 mm.

C. Illumination System

The challenge in developing LED illumination system for a capsule endoscope is obtaining uniform illumination [37] on the observed object. This is difficult for several reasons:

- ♣ the optical path of an LED light source is not parallel to the optical axis of the adjacent imaging lenses,
- ♣ the light pattern of the LED depends sensitively on the driving current, location, and projection angles,
- ♣ the object plane of the observed intestine is not flat,
- ♣ the luminescence of an LED is different at different angles.

In general, the spatial light distribution is not uniform at the object plane in front of the capsule endoscope, resulting in a reduced image quality from the CMOS sensor. In some regions of the object plane, the luminance is very small and in other regions it could be very large. In order to solve this problem, the positions and the angles of multiple LEDs were arranged to improve the luminous uniformity of the designed capsule endoscope. Due to space constraints in the capsule, the LEDs must be small and yet provide enough luminous intensity for decent images to be obtained. Their luminous intensities should also be able to provide enough light for the worst-case scenario where the lumen of the colon is very large: in the cecum and in the ascending colon. The power required by the LEDs is dependent on the number and the type of LEDs used.

Nichia NSSW156T LEDs (NICHIA Corporation, Tokyo, Japan) were selected based on their luminous intensity, viewing angle, and dimensions (3.0x1.4x0.8mm). Four white flat top, wide-angle, surface-mount LEDs were utilized to minimize the space volume and the current draw, while still providing adequate illumination to cover ≈ 140° range.

The experimental luminous intensity of a single LED is above 130 mcd at a forward current of 1mA. This provides sufficient combined illumination for the CMOS imager to clearly view an image inside large intestine when the four white LEDs are circumferentially positioned at the camera side end of the capsule.

Additionally, to be effective, the illumination needs to provide an adequate amount of uniform light and to be coming from a correct angle, so as to avoid shadow effects. In general, LEDs light distribution characteristic deviates from its center. When the LEDs are arranged circumferentially around the imager, light from all these illuminating units overlaps at a portion that is in front of the imager and this portion is brighter than the portion where there is no overlap [36]. Because of this imbalance in brightness, the image quality of the images captured by the imager degrades. To improve the brightness balance, LEDs were disposed around the CMOS imager in such a manner that the optical axes of the illuminating units did not intersect with the optical axis of the observation unit (the CMOS imager), and the illuminating units overlapped at a substantially central portion of the image capturing area of the CMOS imager. The tilting of the illuminating units also minimized the refraction from the optical dome, thus minimizing the distorting effect and efficiently illuminating a cylindrical area of 7cm x 18cm.

The optical dome and the LEDs placement was thoroughly designed to minimize the internal light reflections and refractions. Fig. 2 shows the designed illumination system with four tilted LEDs.

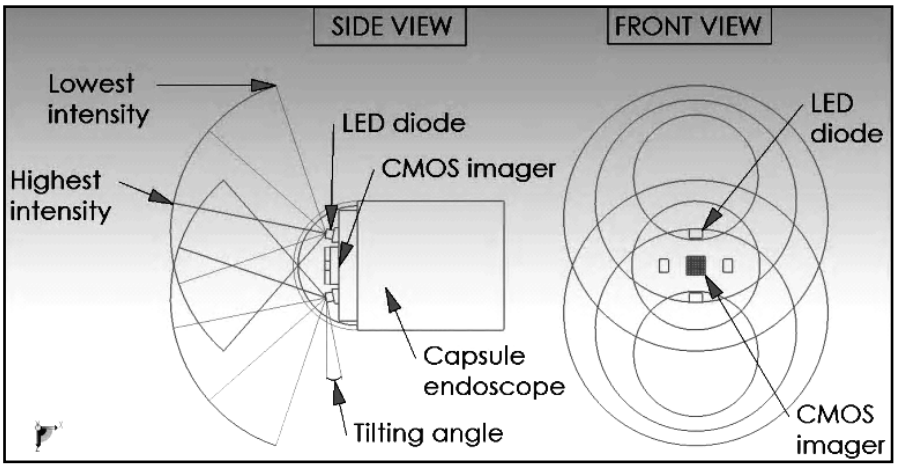


Fig. 2: Tilting of the illuminating LEDs.

The optimal tilt angle which provided adequate illumination for the self-stabilizing capsule endoscope was determined through multiple experiments to be ≈ 11 ± 2°. Three different configurations of LEDs positioning were tested. The configuration of four circumferentially placed LEDs 90 degrees apart proved to be the most effective. We found that the closer to the shell the LEDs were arranged, the lower the total flux was collected by imager. Thus, the distance between each LED and optical axis was chosen to be 5.5mm.

Protel PCB design software (Altium, Sydney, Australia) was used to produce a printed circuit board (PCB) layout of the illumination system. Subsequently, the PCB (copper single-sided board) was milled on the prototyping PCB milling machine and then tinted using a chemical bath. A custom-made ultra-small plastic ring was manufactured to provide a plane for tilting the LEDs, thus ensuring an even tilt angle for all four LEDs. Subsequently, the LEDs were partially placed on the plastic ring and soldered to the PCB with the aid of a microscope.

D. Transmitter, Receiver, and Antenna

The design of a low-power wireless transmission system capable of transmitting through the body was an integral part of this project. The human body is a medium that poses numerous wireless transmission challenges [38]. The RF transmitter must operate in a wide variety of environments and positions that can change with time, because the human body tissue will attenuate the signal variably as the capsule passes through.

There are several critical requirements for the transmitter in a capsule endoscope, including size, power consumption and data rate. Small size is essential, since there is very little room to fit it into the device along with the other components. The hardware needs to be robust in order to withstand the knocks and bumps of normal human body movement. The transmitter layout complements the other electronic components, making the most out of the small space inside the capsule. The antenna connection must be very short to minimize losses and to keep the impedance constant. The length of the antenna must be small enough to fit in the capsule. However, if

the compact antenna in the capsule endoscope is too small, its bandwidth at low frequency will not be wide enough and its radiation efficiency cannot be very high. In addition, in order to transmit the diagnostic real-time high-resolution image data at high speed and because of the effect of human body tissue on signal propagation characteristics, the transmission frequency may slightly shift during operation [39]. Thus, a wide-bandwidth, higher transmission frequency needs to be employed along with a tunable receiver. Unfortunately, an antenna with a higher transmission frequency may cause higher radiation absorption especially in the backmost portion of the intestine and decay the communication link performance in the human body. Since more energy will be absorbed by the human tissue at higher frequencies, many problems exist such as power consumption and the way various positions and orientations of the radiation source influence its characteristics [40]. Consequently, careful selection of transmitter, antenna and frequency for in vivo use is very important. With CMOS technology, the output frequency typically increases as the physical size diminishes. Among the industrial, scientific, and medical (ISM) frequencies, 1.9 GHz is frequently used for transmission, and the smallest transmitters that are commercially available operate in this range.

An RF transmitter chip is used to send data from the CMOS imager wirelessly to be recorded in the computer. The selected transmitter in the final prototype, DX-5A (RF-LINKS, Toronto, Ontario, Canada) operates at 1900MHz (channel frequency), and draws 8mA from a 3V source. This transmitter is ultra-small (4 mm x 5 mm), gain-tunable, and is capable of transmitting a standard NTSC signal.

The transmitter antenna size and shape needed to be modified in order to achieve maximum efficiency in human body environment, fit in the capsule along with all other components, and minimize antenna coupling to the battery. The capsule travels passively in the GI tract, and therefore is randomly oriented. Thus, the direction of transmission radiation is uncontrolled, and to detect the transmitted signal independently of transmitter position, the antenna is required to emit an isotropic radiation pattern [41, 42].

The selected external receiver VRX-2000M (RF-Links, Toronto, Ontario, Canada) outputs the received video signal from the in-vivo sensing device into a video capturing device DVD Xpress DX2 (ADS Tech, Walnut, CA, USA). The small external receiver can monitor the 1900-2100 MHz video bandwidth and the frequency is controlled in 0.25MHz steps, which is entirely sufficient for stable real-time recording and display of the results. The receiver provides very high sensitivity of -86dBm. In addition, it has a built-in special Automatic Frequency Tuning (AFT) control function for best signal receiving stability.

The DX2 outputs the video signal to a PC display via a fast USB 2.0 connection while the CE is inside the patient's body. Using video capture software, the video was captured to computer's hard drive for further analysis.

E. Design of the Self-stabilizing Component

The stabilizing component should be able to deform under external pressure but revert to its original shape when the pressure removed. To provide a maximum safety to patients, necessary requirements for the stabilizing component are: permeability to fluids and gases, compliance characteristics mimicking soft stool, and biodegradability in 2-3 days. The expanded implement should also maintain its consistency and not change state under the influence of water or colonic fluids. In addition, the super-absorbent polymer has to be biocompatible, to swell extensively, to swell in a relatively short period of time, to exert a reasonable swelling pressure on the walls of the lumen, and to withstand the maximum luminal inner pressure in the colon.

The faster the expansion process is completed, the more rapidly the imaging capsule would stabilize in the colon, allowing systematic imaging of the organ once the capsule enters the cecum. Therefore, the expected overall expansion time should be less than 1 min.

The materials that had the desired properties to design the expandable component were categorized according to the preferred mechanism of expansion, osmosis (release of potential energy) and expansion rates. The release of potential energy results in an expansion of the material that possesses the potential energy. Osmosis has proven to be effective in several medical applications such as stents that help relieve pathological obstruction of tubular structures in vascular, urologic and GI systems, as well as in self-expanding prostheses [43].

The swelling behavior of superabsorbent osmosis-based materials depends on pH, temperature, ionic strength, or solvent composition. The rate of expansion in osmosis depends on numerous factors, including the structure of the material, the membrane and the concentration of the solution. On the other hand, the release of potential energy is rapid, almost instantaneous. The superabsorbent material used in this design was cross-linked sodium polyacrylate polymer granules. These granules exhibit strong hydrophilicity and can absorb hundreds of times their weight in water without dissolving upon swelling [44].

The stabilizing component adapts to the inner lumen shape, maximizing the contact surface area between the two. The dynamic viscosity of the swollen superabsorbent (hydrogel) is in the range of 900-7000 centipoises (cPs) [45], which is considerably less than healthy human feces viscosity (semiliquid; <50000cPs) [46]. When the tissue of the wall is stretched, the friction will increase rapidly. This is very important during the CE transit in the colon since it allows for a very smooth movement of the capsule by reducing friction between the expanded implement and the intestinal wall. The selected superabsorbent polymer provides high fluid retention even under pressure.

To overcome sharp colonic turns such as the hepatic and the splenic flexures the expandable component must behave similarly to the way formed stool does, which has been incorporated as an important feature of this design. This bending capability should be uniform up to the base of the expandable component that is attached to the rigid, but relatively small imaging component.

The pressure exerted on the walls of the colon upon expansion should not be harmful. The colon wall of the human body can bear≈ 7.71 pound -force/square inch (398.72 mmHg) [47]. The forces due to the internal pressure of the stabilizing component acting on the colonic walls can potentially harm the walls. As a result, stresses in the colonic wall on the cross-section along the colonic axis and on the cross-section perpendicular to the axis are created. Generally, these stresses are tensile and are known as circumferential (hoop) and longitudinal. The normal colonic wall thickness ranges from 0.1 to 2 mm in colonic segments with a diameter of ≥4-6 cm, from 0.2 to 2.5 mm in colonic segments with a diameter of 3-4 cm, from 0.3 to 4 mm in colonic segments with a diameter of 2-3 cm, and from 0.5 to 5 mm in colonic segments with a diameter of 1-2 cm [48].

Thus, the colon can be regarded as a thin walled cylinder if certain assumptions are made. Firstly, the stress differences across the thickness of the bowel wall should not be considered significant and secondly the weight of any bowel contents should also be considered insignificant as a stress raiser. These are reasonable assumptions in the bowel if the internal diameter is greater or equal to 10 times the thickness of the bowel wall and the bowel wall receives no external support. Generally, it is known that in the particular case of a thin-walled cylinder, the longitudinal stress is half of the hoop stress [49]. Hence, the permissible stress in the colon should be less than the hoop stress. Thus, in a distendable biological tube like the large intestine, the principal stress of interest is the

hoop stress [50]. Fig. 3 illustrates a simple model of the pressure P_(expan.) exerted upon the expansion of the stabilizing component on the colonic walls during peristalsis.

There are two possible scenarios in the colon. In the first scenario only expansion is present but no peristaltic activity. In this case the following assumption is made:

$$\boldsymbol{P_{expan.}} < P_{\boldsymbol{net}} \tag{1.1}$$

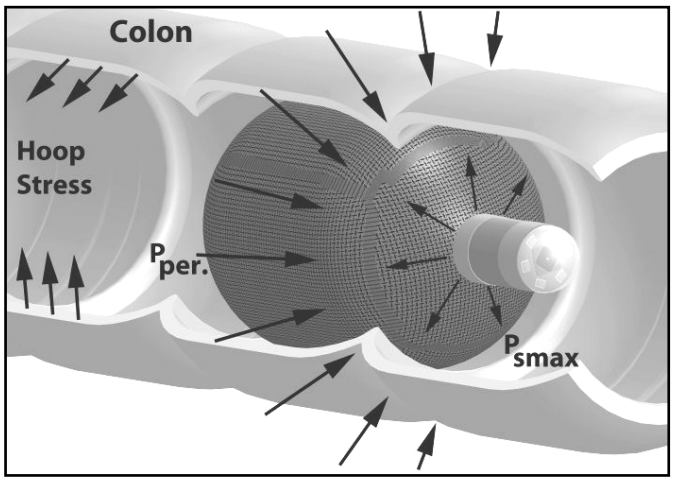


Fig. 3: Pressure in the colon during expansion.

The net pressure that the walls of the colon can bear, $\boldsymbol{P_{net}}$ can be calculated using equation (1.2) **[51]**, assuming that the pressure applied on the walls would mainly result in a hoop (tensile) stress and in a negligible longitudinal stress:

$$\boldsymbol{P_{net}} = \frac{\sigma_{\boldsymbol{h}}\boldsymbol{t}}{\boldsymbol{R}} \tag{1.2}$$

where:

$\boldsymbol{P_{net}}$ = Net pressure that the walls of the colon can bear upon expansion, Pa;

$\boldsymbol{P_{expan.}}$ = Maximum swelling/expanding pressure on the walls of the colon upon expansion, Pa;

$\sigma_{\boldsymbol{h}}$ = Hoop (tensile) stress of the colon, N/m^2;

$\boldsymbol{t}$ = Thickness of the colonic walls, m;

$\boldsymbol{R}$ = Inner radius of the colon, m;

The colon tensile strength is 81 to 139 g/mm^2 **[52].** Using equation (1.2), we estimated the maximum $\boldsymbol{P_{net}}$ of 13.9 kPa (104 mmHg). Previously, it was found that during colonoscopic air insufflations colon perforation can occur at intraluminal pressures greater than 140 mm Hg (or 18.67 kPa) **[53, 54]**. The estimated$\boldsymbol{P_{net}}$ of 104mmHg is well below the reported critical value of 140 mmHg which can perforate the colon.

The second scenario assumes expansion and peristalsis acting simultaneously. Thus, the net pressure applied on the walls of the colon is the difference between the maximum swelling/expanding pressure applied by the stabilizing component, $\boldsymbol{P_{expan.}}$and the pressure exerted on the device by the colonic walls, $\boldsymbol{P_{per.}}$and is shown in

equation (1.3). The difference between the maximum swelling/expanding pressure $P_{expan.}$ and the peristalsis pressure $P_{per.}$should be less than the net pressure, P_{net} so that no harm is caused to the colonic walls:

$$P_{expan.} - P_{per.} < P_{net} \quad (1.3)$$

where:

$P_{per.}$ =peristaltic pressure in the colon, Pa;

The mean peak amplitude of antegrade propagating pressure waves is 41.8+/-2.3 mmHg with a range of 5-169 mmHg **[55]**. It can be seen that the difference between the swelling pressure and the peristaltic pressure should be lower than P_{net}. However, it is important to simultaneously include in this analysis the antegrade propagating wave. Therefore, the stabilizing component was designed to create maximum swelling pressure of 90 ± 15 mmHg, because if this pressure is higher than the antegrade propagating pressure, severe constipation or perforation might result **[56]**. The self-stabilizing capsule before assembly is shown in Fig.4

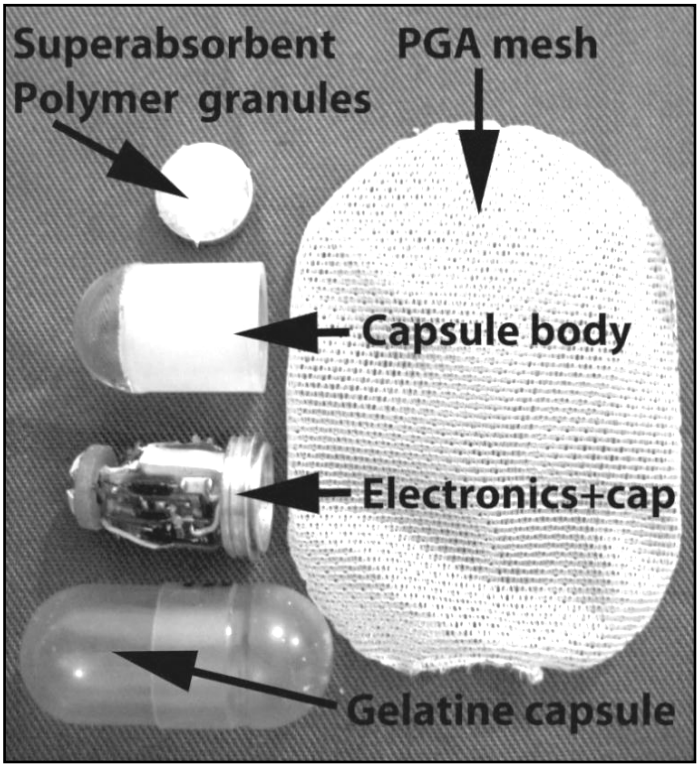


Fig. 4: The self-stabilizing capsules before assembly. The expandable component made from permeable PGA mesh filled with superabsorbent polymer granules was attached to the CE prototype containing the batteries, the imager, and the lens. Subsequently, the entire implement was put in a gelatin capsule.

The self-stabilizing component selected for the final prototype was a woven, bioabsorbable liquid-permeable, flexible polyglactin 910 mesh (Vicryl, Ethicon Inc, Somerville, NJ, USA) filled with superabsorbent polymer granules (Favor PAC, Evonik Industries, Stockhausen, Germany).

Initially, the mesh was thermally processed to obtain oval shape (side length=2cm, radius=2cm) followed by cutting a slit (5mm) through which the polymer (≈ 0.7 ml) was inserted into the oval-shaped mesh. Subsequently,

the mesh was pulled over the end part of the CE through this opening, compactly attached to a specially designed cap using PDS absorbable suture, and covered with an outer casing.

F. Design of the Attaching Mechanism

Due to the geometry of the self-stabilizing CE and the pressure applied in the colon, there is a higher risk of the stabilizing and the imaging components disassociating from each other as a result of shearing. Consequently, the attachment point on the cap has been designed to maintain a strong connection between the CE and the stabilizing component utilizing PDS II (USP 5-0) absorbable suture (Ethicon, Inc. in Somerville, NJ, USA). The attachment of the stabilizing component around the circumference of the sealable cap uniformly distributes the tensional load. This increases the strength of the connection, and prevents the two components from flexing. In addition, this type of attachment normally preserves a common central axis between the stabilizing component and the capsule, while ensuring the necessary flexibility to systematically visualize convoluted portions of the colon.

The PGA mesh was pulled up over the perimeter of the lid, in such a way that it covered the entire circumference of the external groove. Then, looking at the back of the lid, a suture was threaded into the upper left hole and out through the mesh into the groove. It was then wound tightly one full turn clock-wise around the outside of the mesh pulling it into the groove. Next, it was threaded out the other hole to the back of the lid where the two ends of the suture thread were then tied tightly together.

The breaking strength of the stabilizing component should be greater than the maximum possible pressure in the colon if the worst-case scenario is considered. In this scenario, the peristaltic pressure holds one end of the stabilizing device fixed, while random pulling pressure is applied on the rest of the capsule. The random pressure is assumed to have a pulling effect, i.e. a lateral pressure of the same magnitude as the transversally applied random pressure. Figure 5 shows the worst-case scenario where one end is held and the other is pulled laterally and circumferentially.

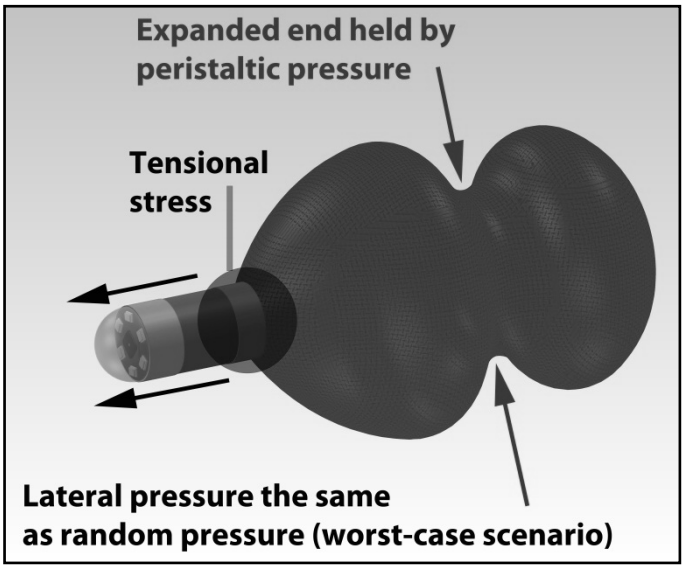


Fig. 5: Worst-case scenario for the stabilizing component where one end is held and the other is pulled by lateral and circumferential forces.

Previously, it was found that the mean peak amplitude of antegrade propagating pressure waves for all colonic regions is significantly greater than that of the retrograde propagating pressure waves [57]. The colonic pressure activity is complex and usually ranges from 25 mmHg up to 400 mmHg [58]. A value of 400 mmHg was used to calculate the worst-case scenario for the attachment between the CE imager and the expandable component.

G. Design of the Separation Component

The separation component was designed utilizing an electronically-controlled microheater. The general requirements for the microheater module can be summarized as follows: excellent temperature uniformity over the sensitive area, very low input power and ultra- small dimensions. For the final prototype, the material selected for the microheater's heating filament was a non-magnetic alloy of nickel and chromium (80% nickel and 20% chromium). Ni-Chrome converts heat into electricity through Joule heating. This alloy has a high melting point of about 1400 °C. The Ni-Chrome based microheater is controlled via two specially disposed miniature magnetic dry reed switches (contact layers: gold, sputtered ruthenium). These switches are of the double-ended type and may be actuated by an electromagnet, a permanent magnet or a combination of both. The connection between these switches is flexible and partially movable to improve reliability and sensitivity of switching. The diameter of the Ni-Chrome filament is very small (0.001mm), which allows it to heat up quickly when exposed to a relatively small current. Unfortunately, this means that the filament is very fragile, and could be severed during assembly when the PDS suture is tensioned on top of it. This situation was avoided by using a diameter of Ni-Chrome wire that was small enough to be partially concealed within the uneven surface features of the sealable cap. Thus, an additional structural support to the Ni-Chrome wire was provided by the sealable cap, which effectively increased the breaking point of the wire. Using a microscope, the PDS suture was positioned on the top of the heating filament (1.4 mm in length) making a very slight contact with it.

The magnetic reed switches activate the heating of Ni-Chrome filament in the presence of a magnetic field, and deactivate it once the field is removed. When a current passes through the filament, it achieves the requisite temperature (melting point of 107°C) to melt the PDS II suture [59], separating the capsule from the expandable material. In the specific design the switching time was less than 100 ms and the separation time was 1.5 seconds using an 89-mA pulse delivered from the power supply. The effective magnetic field of operation was determined by multiple experiments and was in a range of 2-15Ampere-Turns (current in the coil multiplied by the number of turns). Fig. 6 illustrates a CAD-rendering of the microheater module.

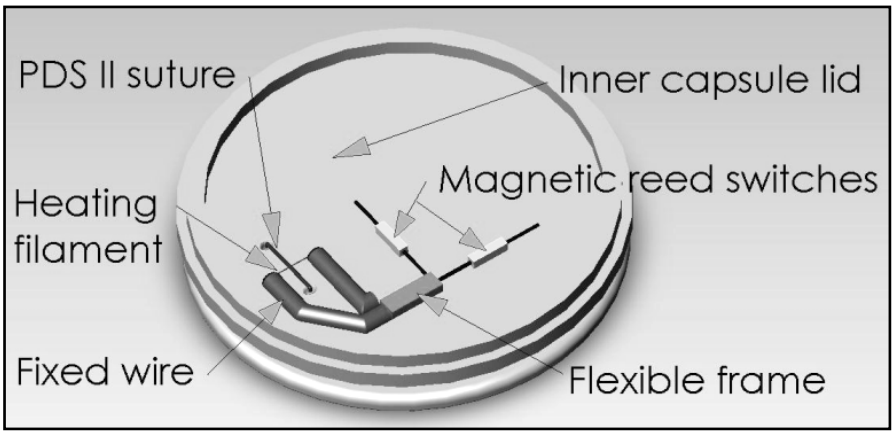


Fig. 6: Microheater module.

H. Power Supply and Encapsulation

Commercial CEs use small silver oxide batteries (usually two), which account for more than half of the weight of the capsule. In all applications of capsule endoscopy it is imperative to consider battery life/performance trade-offs and toxicity of the battery technology. Careful component choice and efficient electrical design are important in extending the life and improving the performance of the capsules.

Since there are only certain types of batteries commercially available that can fit in the limited space of the capsule, the power supply needs to be designed very efficiently. The CR 1/3N battery (Duracell Canada Inc, Mississauga, Ontario, Canada) provides enough continuous current to allow for the reliable operation of all electronics inside the capsule. The operational voltage, continuous discharge current, small weight, and small dimensions were deciding factors when selecting an effective power supply. Another important parameter to consider was the internal resistance as the battery must provide enough continuous discharge current of 28-30mA for the electronics and 400 msec/80mA pulse to trigger the microheater simultaneously. The lower the internal resistance, the less restriction the battery encounters in delivering the needed power spikes. Moreover, the smaller the battery (i.e. its diameter), the higher the internal resistance. Using a DC load test the internal resistance was measured to be in range of 250 mOhms.

For the initial prototype model, the outer casing is a hard-shell gelatin capsule from Torpac, Inc. (Fairfield, NJ, USA) which dissolves very rapidly (2-3 minutes) in water. This capsule is sized for animal testing, and with the design limitations in mind, a capsule of size #13 (1/8 oz. 3.2 ml) was chosen. The gelatin capsule provided an adequate space for the CE containing both the electronics and the self-stabilizing component. In the present prototype, the emphasis is on demonstrating stabilization and separation, and the testing assumes that the outer casing has reached the targeted organ.

In a final design, a colon targeting film cover can be utilized such as the recently proposed coating prepared by mixing Eurylon VII [high amylose maize starch (Roquette, Lestrem, France)] aqueous dispersion and Eudragit S ethanolic solution (Evonik, Darmstadt, Germany) (ratio 3:7 solids) [60].

I. Laboratory Testing

1) Image Transmission Test

A medium approximating human flesh, a fatty beef tripe was selected in order to confirm that an image can be transmitted successfully through. The prototype of the capsule endoscope was placed between two pieces of the tripe (>5cm thickness each) and was fully covered, with the CE being positioned in the center. The receiver antenna was placed 2m from the outer tripe surface to sense the receiving signal. The tripe was manually squeezed both periodically and randomly to simulate the dynamic impedance change of moving flesh, at a distance of 2m from the receiver. The squeezing continued for five minutes, monitoring the PC screen for qualitative changes in the clarity of the images.

2) Vertical Drop Test

A 1.5-liter transparent acrylic tube (4.5 cm diameter, 90 cm length) was sealed at one end, filled with water/oil (1:1), and a non-stabilized or stabilized capsule was set to fall down its entire length. The oil-water mixture reduced the speed at which the capsules fell, and enabled more consistent video comparison.

Agitation of the cylinder was delivered manually, with a displacement of 18 cm from the center of the tube at a frequency of approximately one back and forth oscillation per second. A scaled paper grid was placed on the

floor below the center of the acrylic tube for the purpose of amplitude control while agitation of the tube was performed (Fig. 7).

3) Separation Test

The self-stabilizing CE was activated by placing it into a container filled with saline solution. Once the stabilizing component achieved its final swollen state it was manually removed from the container. Subsequently, a magnetic field was created with intensity in the 5-15Ampere-Turns range to which the expanded self-stabilizing CE was subjected to. The separation time and the needed force that resulted in physical separation between the capsule and stabilizing component were recorded.

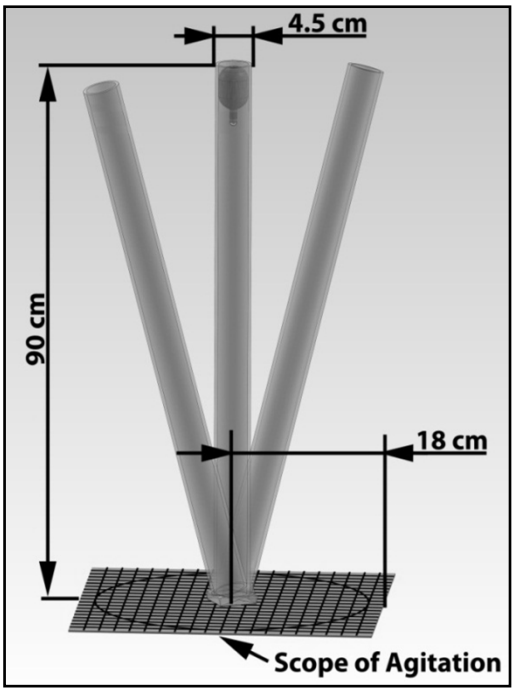


Fig. 7: Self-stabilizing CE-acrylic tube test.

J. Pilot Animal Testing

Two mongrel dogs (1M, 1F, 25.3+/-4.5 kg) were included in an acute testing of the stabilized and the non-stabilized capsules. Colon preparation involved a 48-hour liquid diet and the administration of phosphosoda enema in the morning before the surgery. At the end of the experiment the animals were euthanized with an intravenous injection of Euthanyl, (480mg/4.5kg, Bimeda-MTC Animal Health Inc., Cambridge, ON, Canada). This research was approved by the Life and Environmental Sciences Animal Care Committee, University of Calgary, Calgary, Alberta, Canada.

Each of the animals underwent an induction with an intravenous injection of thiopental (Thiotal 15 mg/kg IV, Vetoquinol Canada, Lavaltrie, QC, Canada) and was subsequently maintained on inhalant isoflurane and oxygen (Halocarbon Laboratories, River Edge, New Jersey, USA) with a vaporizer setting of 1-3%.

A laparotomy was followed by exteriorization of a 15 cm segment of the proximal descending colon while preserving intact the mesenteric innervations and blood supply. A 1-cm transverse full thickness incision was made in the colon 2 cm proximal to the beginning of the chosen colon segment.

A single CE, either with stabilizing component or without stabilizing component was inserted using forceps through the incision into the lumen of the colon. Subsequently, the incision was temporarily closed with 4-0 Dexon

suture (Covidien, Mansfield, MA) and 200ml saline was injected at the level of the capsule. Each animal was then administered an intravenous injection of Neostigmine (0.04mg/kg, APP Pharmaceuticals, Schaumburg, Illinois, USA) to speed up the transit time by inducing pharmacologically colonic peristalsis. The inserted capsule was propelled distally through the colon and expelled naturally through the anus as a result of the Neostigmine-induced colonic peristalsis. Upon the expulsion of the capsule through the anus, the video data which was collected by the RF receiver were transferred to the PC for further image processing. After the clearance of Neostigmine from the circulation, another capsule was administered in a similar manner. Ultimately, 2 such administrations per animal were sequentially performed, two with the non-stabilized and two with the stabilized capsules.

Results

A. Capsule Dimensions and Performance

The volume of the expandable component (polymer and mesh) was 0.9±0.5 ml before expanding and 67±7 ml after the expansion in water. The fully expanded volume in saline decreased to 54±5 ml. This was because the saline solution contained heavy ions, which slowed down the osmosis process and lowered the volume of the fully swollen expandable component The total volume of the self-stabilizing unit and the dry expandable stabilizing component) was in the 3ml range, which is very close to the volume of 2.7 ml of the Given's PillCam Colon capsule.

Prior to swelling each mesh had an oval shape of 5.5 cm length and a diameter of 4 cm. The expansion time was less than 1 min in water and less than 2 minutes in a saline solution.

The average maximum swelling pressure that the expandable component exerted on the colonic walls was measured to be 95±18 mmHg at 136 gram force per square cm, gf/cm2. This value is well within the pre-design constraints. The tensile strength of the mesh was 341±8 mmHg.

The minimized overall weight (5.1g) of the designed capsule endoscope allowed a common centroid axis for the capsule endoscope and the expanded implement.

The average operating time of the capsule endoscope prototype was 76 minutes and its total power consumption was 84 mW.The fully expanded capsule is shown in Fig.8.

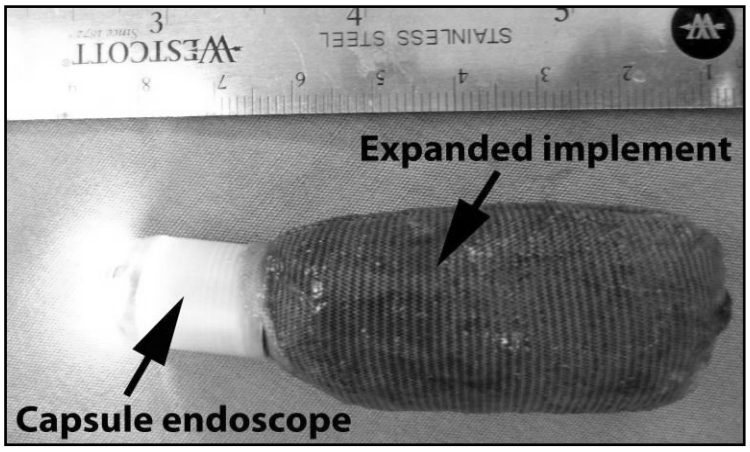


Fig. 8: Fully expanded self-stabilizing capsule endoscope. Due to its permeability, the stool-like tail swelled immediately after the degradation of the gelatin cover to provide stabilization to the imaging module.

B. Laboratory Test Results

1) Image Transmission Test

Continuous images were successfully transmitted from the capsule and were easily captured using wireless receiver without decrease in video quality.

2) Vertical Drop Test

A total of 30 laboratory video footages for each modality (stabilized and unstabilized) were collected for the vertical drop, resulting in 36 minutes of recorded video for the stabilized case and 22 minutes for the unstabilized.

The expandable portion of the stabilized capsule absorbed and attenuated the manually delivered agitation disturbances, which resulted in a minimal CE lateral movement during the agitation process.

It was clearly seen that that no tumbles have occurred in the stabilized case. Since the connection between the expandable stabilizing component and the capsule was rigid, the completed assembly was unable to deform permanently, and the CE self-stabilized itself during the agitation process.

In addition, the stabilized capsule showed significant improvement in image quality versus the non-stabilized capsule.

At no time in this testing did the stabilized capsule faced closely the cylinder wall because both the combined length of the capsule and the stabilization mechanism exceeded the diameter of the acrylic tube.

3) Separation Test

In each of the successful trials that were conducted, the expanded part separated from the capsule cap by means of gravity in less than 1.5 seconds after entering the created magnetic field. Ten trials were conducted with the complete separation mechanism, and the only failures of the system were identified to be due to poor workmanship, which was dramatically improved with the aid of a microscope. The initial difficulty was keeping close proximity between the heating filament and the PDS suture, without which the heat generated by the filament was able to dissipate through the small volume of air within the capsule. The microheater was determined to be entirely suitable for this application, with the only potential design challenge left being the reliability of its manufacturing. Fig. 9 shows the CE and the stabilizing component after the physical separation.

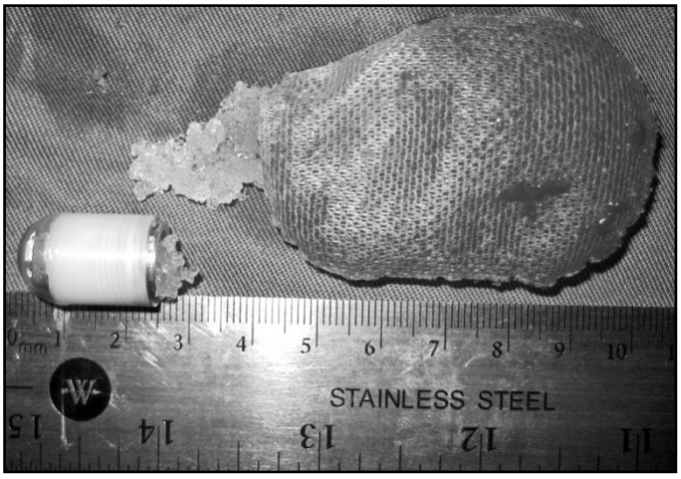


Fig. 9: Separated CE and stabilizing component.

Once the capsule was electronically separated, the average force that would result in physical separation of the device was in the range of 80gf. Thus, the peristalsis can complete the physical separation process after the electronic separation.

C. Results from the Pilot Animal Tests

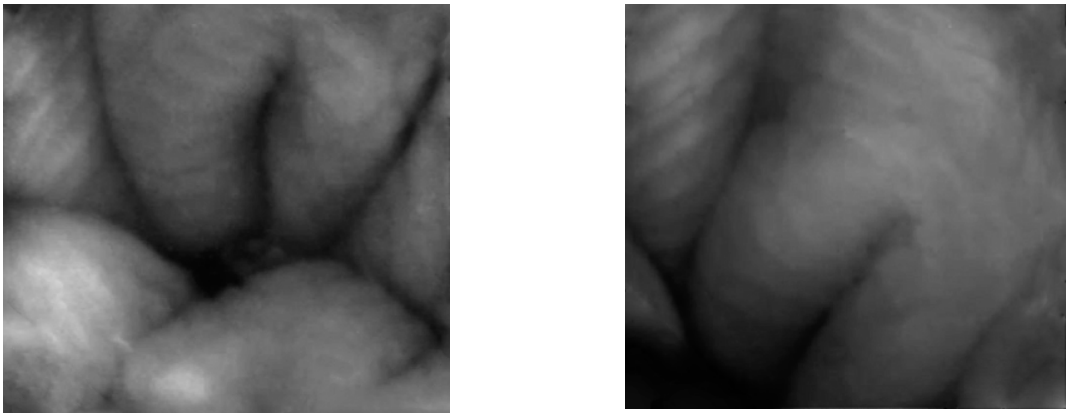
Totally, there were 4 video footages for each of the stabilized and the unstabilized cases, resulting in 28 minutes of recorded video for the former and 42 minutes for the latter.

The stabilized capsule did not miss vital intraluminal visualization of the colon during its distal peristaltic transit.

The unstabilized CE was rapidly changing its position and rendered some parts of the video completely unusable. In addition, the sensitivity of the CMOS imager was greatly decreased by this effect during peristalsis. It was evident that during rapid capsule repositioning (e.g. during rapid persistaltic activity) the stabilized capsule reacted quickly by stabilizing itself towards the centroid of the lumen.

In addition, it was observed that the unstabilized capsule faced directly the colonic wall during its tumbling movement. Thus, the unstabilized capsule could not systematically visualize the entire portion of the colon it was traversing. This inevitably resulted in missed areas.

A comparative example of images for the self-stabilizing capsule (a) and the non-stabilized CE (b) is shown in Figure 10.



(a) (b)

Fig. 10: Example of images for both modalities, (a) self-stabilizing and (b) non-stabilized CE)

At no time in this study did the self-stabilizing capsule tumbled within the colon. The simple reason for this was that both the combined length of the capsule and the stabilization mechanism exceeded the diameter of the lumen. Because the connection between the two was flexible, the completed assembly was unable to deform permanently, and the scenario in which the self-stabilizing CE could closely be facing the colonic wall became a geometric impossibility with the fully expanded stabilizing component.

Conclusion

The feasibility of self-stabilized capsule endoscopy has been demonstrated in laboratory and acute canine experiments. New self-stabilized capsule design was described. The goal of this study was the improvement of sensitivity of VCE in the colon. The proposed device has the capabilities to provide colonic images of quality comparable to the ones received from the small intestines using traditional VCE. In comparison with the conventional CE design, the proposed self-stabilizing CE delivered a significant improvement in canine colonic imaging. It has the potential to greatly improve quality of colonic imaging. Self-stabilized CE could be an alternative diagnostic modality for colon neoplasia screening. Further larger-scale clinical trials are needed to confirm and elucidate further the findings of this pilot study.

Acknowledgement

This research was supported by Sandhill Scientific, USA.

References

[1] T. Nakamura, A. Terano, “Capsule endoscopy: past, present, and future”, J Gastroenterol., vol. 43, no. 2, pp. 93-9, 2008

[2] O. Barkay, M. Moshkowitz, Z. Fireman et al., “Initial experience of video capsule endoscopy for diagnosing small bowel tumors in patients with GI polyposis syndromes”, Gastrointest.Endosc., vol. 62, no. 3, pp. 448-52, 2005

[3] N. Pallotta, F. Baccini, E. Corazziari, “Contrast ultrasonography of the normal small bowel”, Ultrasound Med Biol., vol. 25, no. 9, pp. 1335-40, 1999

[4] B.R. Davis, H. Harris, G.C. Vitale, “The evolution of endoscopy: wireless capsule cameras for the diagnosis of occult gastrointestinal bleeding and inflammatory bowel disease”, SurgInnov., vol. 12, no. 2, pp. 129-33, 2005

[5] Z. Fireman, “Capsule endoscopy: Future horizons”, World J GastrointestEndosc., vol. 2, no. 9, pp. 305-7, 2005

[6] G. Bresci, G. Parisi, M. Bertoni, E. Tumino, A. Capria, “The role of video capsule endoscopy for evaluating obscure gastrointestinal bleeding: usefulness of early use”, J Gastroenterol., vol. 40, no. 3, pp. 256-9, 2005

[7] A. Ross, S. Mehdizadeh et al., “Double balloon enteroscopy detects small bowel mass lesions missed by capsule endoscopy”, Dig Dis Sci., vol. 53, no. 8, pp. 2140-3, 2008

[8] H. Goto, “Diagnosis and treatment of small bowel diseases are advanced by capsule endoscopy and double-balloon enteroscopy”, Clinical J of Gastroenterology, vol. 3, no. 5, pp. 219-225, 2010

[9] J.L. Toennies, G. Tortora, M. Simi, P. Valdastri, and R. J. Webster III, “Swallowable medical devices for diagnosis and surgery: The state of the art”, J of Mechanical Engineering Science, 2010, vol. 224, no. 7, pp. 1397-1414, 2010

[10] W. El-Matary, “Wireless capsule endoscopy: indications, limitations, and future challenges”, J PediatrGastroenterolNutr, vol. 46, no. 1, pp. 4-12, 2008.

[11] Moglia, A., Menciassi, A., Dario, P., and Cuschieri, A. “Capsule endoscopy: progress update and challenges ahead”, Nat. Rev. Gastroenterol. Hepatol., 2009, vol. 6,no. 6, pp. 353-362, 2009.

[12] J. Westerhof, R.K. Weersma, J.J. Koornstra, “Risk factors for incomplete small-bowel capsule endoscopy”, GastrointestEndosc., vol. 69, no. 1, pp.74-80, 2009

[13] R. Eliakim, K. Yassin, I. Shlomi, “A novel diagnostic tool for detecting oesophageal pathology: the PillCamoesophageal video capsule”, Aliment PharmacolTher, vol. 20, no. 10, pp. 1083-9, 2004.

[14] ASGE Technology Commitee, “Capsule endoscopy of the colon”, J GastrointestEndosc, vol. 68, no. 4, pp. 621-623, 2008.

[15] G.Hounnou, C. Destrieux, J. Desme, P. Bertrand, and S. Velut, “Anatomical study of the length of the human intestine”, Surg. and Radiol. Anat., vol. 24, no. 5, pp. 290-294, 2002.

[16] C. Schmidt, “Capsule Endoscopy To Screen for Colon Cancer Scores Low on Sensitivity, High on Controversy”, J Natl Cancer Inst, vol. 101, no. 21, pp. 1444-1445, 2009.

[17] C. Spada, M.E. Riccioni, C. Hassan, G. Costamagna, “PillCam capsule endoscopy for the diagnosis of colonic diseases”, RecentiProg Med., vol. 101, no. 6, pp. 227-31, 2010

[18] E.Horn, D. Gat, R. Rabinovitz, Y. Nitzan.“System and method for modeling a tracking curve of an in vivo device”, U.S. Patent 7 761 134, Jul. 20, 2010.

[19] K.N.C. Hin, O. Yadid-Pecht, M. Mintchev, “e-Stool: Self-Stabilizng Capsule for Colonic Imaging”, in Proceedings of the 20th Int. Symposium on Neurogastroenterology and Motility; Jul 3-6 2005, pp. 86-90.

[20] M. Classen, G. N. J. Tytgat, and C. Lightdale, “Wireless video capsule endoscopy”, in Gastroenterological Endoscopy , 2nd ed., D. Cave, Ed. Notzingen, Germany: Thieme, pp. 127-128, 2010.

[21] H.M. Kim, Y.J. Kim, H.J. Kim, S. Park, et al. ”A Pilot Study of Sequential Capsule Endoscopy Using MiroCam and PillCam SB Devices with Different Transmission Technologies”, J Gut and Liver, vol. 4, no. 2, pp. 192-200, 2010.

[22] J.D. Waye, D.K. Rex, C.B. Williams, “Reports and Imaging”, in Colonoscopy: Principles and Practice, 2nd ed., C.B. Williams, Ed. Oxford: Wiley Blackwell, 2003, pp. 251-252.

[23] Y.H. Fang, C.X. Chen, B.L. Zhang, “Effect of small bowel preparation with simethicone on capsule endoscopy”, J Zhejiang UnivSci B., vol. 10, no. 1, pp. 46-51, 2009

[24] A. Ross, S. Mehdizadeh, J. Tokar, et al., “Double- balloon enteroscopy detects small bowel mass lesions missed by capsule endoscopy”, Dig Dis Sci. 2008; vol.53, no. 8, pp. 2140–2143, 2008.

[25] A. Postgate, E. Despott, D. Burling, et al. ”Significant small-bowel lesions detected by alternative diagnostic modalities after negative capsule endoscopy”, GastrointestEndosc, vol. 68, no. 6, pp. 1209-14, 2008.

[26] J.B. Pilz, S. Portmann, P. Shajan, C. Beglinger, L. Degen, “Colon Capsule Endoscopy Compared to Conventional Colonoscopy under Routine Screening Conditions”, J BMC Gastroenterology, vol. 10, no. 66, 2010.[27]A. Van Gossum, M. Munoz-Navas and I. Fernandez-Urien et al., Capsule endoscopy versus colonoscopy for the detection of polyps and cancer, N Engl J Med, vol. 361, no. 12, pp. 264-270, 2009.

[28] P. Valdastri, R.J. Webster, C. Quaglia, M. Quirini, A. Menciassi, and P. Dario, “A new mechanism for mesoscale legged locomotion in compliant tubular environments”, IEEE Trans. Robot., vol. 25, no. 5, pp. 1047-1057, 2009.

[29] R.H. Fletcher, “ACP Journal Club: Capsule endoscopy had low sensitivity for detecting colonic lesions and high specificity for large lesions”, N Engl J Med., vol. 361, no. 3, pp. 264-70, 2009.

[30] N. Otterbeck, “Pellet formulation for the treatment of the intestinal tract”, U.S. Patent Application 0017117, Jan. 1, 2009.

[31] S. Nicholas, N.J. Spencer, “Peristalsis and fecal pellet propulsion do not require nicotinic, purinergic, 5-HT3, or NK3 receptors in isolated guinea pig distal colon”, Am J PhysiolGastrointest Liver Physiol, vol. 298, no. 6, pp. G952-61, 2010.

[32] A.L. Baert, P. Lefere and S. Gryspeerdt, “How to avoid pitfalls in imaging: Causes and solutions to overcome false-negative and false-positive findings”, in Virtual Colonoscopy: A Practical Guide, 2nd edition; Ed. Germany, Berlin: Springer Verlag; 2010, pp. 113-114.

[33] S. Kimoto, N. Fujimori, et al., “Capsule endoscope and a capsule endoscope system”, U.S. Patent 7 316 647, Aug. 8, 2008.

[34] E.S. Bennett, B.A. Weissman, “Basic topics”, in Clinical Contact Lens Practice, B.A. Weissman, Ed. Philadelphia: Lippincott Williams&Wilkins; 2005, pp. 111-112.

[35] P. Swain, “At a watershed?Technical developments in wireless capsule endoscopy”, J Dig Dis., vol. 11, no. 5, pp. 259-65, 2010.

[36] T. Yokoi, et al., “Capsule endoscope”, U.S. Patent 7 505 802, Mar. 17, 2009.

[37] O.Y. Mang, at el., “Multiple LEDs luminous system in capsule endoscope”, in Proceedings SPIE the International Society for Optical Engineering; Feb. 6 2008, vol. 8, no. 7

[38] H. Yu, G.S. Irby, D.M. Peterson, M.-T. Nguyen, G. Flores, N. Euliano and R. Bashirullah, "A Printed Capsule Antenna for Medication Compliance Monitoring", IEEE Electronics Letters, vol. 43, no.: 22, pp. 1179–1181, 2007

[39] S. Kwak, K. Chang, and Y. J. Yoon, “Ultra-wide band spiral shaped small antenna for the biomedical telemetry”, in Proc. Asia-PactificMicrow. Conf., 2005, vol. 1, pp. 241-244.

[40] L.S. Xu, Max Q.H. Meng and H.L. Ren, “Electromagnetic Radiation from Ingested Sources in the Human Intestine at the Frequency of 2.4GHz”, in Proc. of the 23rd Progress In Electromagnetics Research Symposium; Mar. 24 2008, pp. 176-181.

[41] J. Kim, and Y. Rahmat-Samii, “Implanted Antennas Inside a Human Body: Simulations, Design, and Characterization”, IEEE Transactions on Microwave theory and techniques, vol. 52, no. 8, pp. 1934-1943, 2004.

[42] L. .C. Chirwa, P.A. Hammond, S. Roy, and D.R.S. Cumming, “Radiation from ingested wireless devices in biomedical telemetry”, IEEE Electronic Letters, vol. 39, no. 2, pp. 178-179, 2003.

[43] A. Lauto, M. Ohebshalom, M. Esposito, J. Mingin, P.S. Li, D. Felsen, M. Goldstein, D.P. Poppas, “Self-expandable chitosan stents: design and preparation”, J Biomaterials, vol. 22, no. 13, pp. 1869-1874, 2001.

[44] J. Haselbach, S. Hey, T. Berner, “Short-Term Oral Toxicity Study of FAVOR PAC in Rats”, J Regulatory Toxicology and Pharmacology, vol. 32, no. 3, pp. 310-316, 2000.

[45] J.A. Ji, J. Liu, S.J. Shire, T.J. Kamerzell, S. Hong, K. Billeci, Y. Shen and Y.J. Wang, “Characteristics of rhVEGF release from topical hydrogel formulations”, Pharm Res, vol. 27, no. 4, pp. 644-54, 2010.

[46] H.H. Wenzl, K.D. Fine, L.R. Schiller, J.S. Fordtran, ”Determinants of decreased fecal consistency in patients with diarrhea,”, Gastroenterology, vol. 108, no. 6, pp. 1729-1738, 1995.

[47] T. McCracken, R. Walker, “The digestive system”, in New Atlas of Human Anatomy, Ed. New York: Barnes & Noble Books, 1999, pp. 134-135.

[48] W. Wiesner, K.J. Mortele, H. Ji, P.R. Ros, “Normal colonic wall thickness at CT and its relation to colonic distension”, J Comput Assist Tomogr, vol. 26, no. 1, pp. 102-6, 2002.

[49] W. Sapsford, “Aeromedical evacuation following abdominal surgery”, J R Army Med Corps, vol. 148, no. 3, pp. 248-54, 2002.

[50] C. Gao, H. Gregersen, “Biomechanical and morphological properties in rat large intestine”, J Biomech, vol. 33, no.9, pp. 1089-1097, 2000.

[51] P.R. Akshoy, R. Pijush, M. Sanchayan, “Stresses in thin walled pressure vessels” in Mechanical sciences: engineering mechanics and strength of materials, Ed. New Delphi: Prentice Hall of India, 2005, pp. 525-533.

[52] D.A. Watters, A.N. Smith, M.A. Eastwood, K.C. Anderson, R.A. Elton, J.W. Mugerwa, “Mechanical properties of the colon: comparison of the features of the African and European colon in vitro”, Gut, vol. 26, no. 4, pp. 384-92, 1985.

[53] S.S. Rao, P. Sadeghi, J. Beaty, R. Kavlock, K. Ackerson, “Ambulatory 24-h colonic manometry in healthy humans” Am J PhysiolGastrointest Liver Physiol, vol. 280, no. 4, pp. 629-639, 2001.

[54] R.A. Kozarek, R.A. Sanowski, “Use of pressure release valve to prevent colonic injury during colonoscopy”, GastrointestEndosc, vol.26, no.4, pp. 139-142, 1980.

[55] I.J. Cook, Y. Furukawa, V. Panagopoulos, P.J. Collins, J. Dent, “Relationships between spatial patterns of colonic pressure and individual movements of content”, Am J PhysiolGastrointest Liver Physiol, vol. 278, no. 2, pp. G329-G341, 2000.

[56] M.E. Pezim, J.H. Pemberton, K.E. Levin, W.J. Litchy, S.F. Phillips, "Parameters of anorectal and colonic motility in health and in severe constipation", Dis Colon Rectum, vol. 36, no. 5, pp. 484-491, 1993.

[57] P.G. Dinning, M.M. Szczesniak, I.J. Cook, "Proximal colonic propagating pressure waves sequences and their relationship with movements of content in the proximal human colon", Neurograstroenterol Mot, vol. 20, no. 5, pp. 512-520, 2008.

[58] S.S. Rao, P. Sadeghi, J. Beaty, R. Kavlock, K. Ackerson, "Ambulatory 24-h colonic manometry in healthy humans" Am J PhysiolGastrointest Liver Physiol, vol. 280, no. 4, pp. 629-639, 2001.

[59] A.M. Greenberg, J. Prein, "Bioresorbable materials for bone fixation: Review of biological concepts and mechanical aspects", in Craniomaxillofacial Reconstructive and Corrective Bone Surgery Principles of Internal Fixation Using AO/ASIF Technique"; R. Suuronen, C. Lindqvist ,Ed. New York: Springer-Verlag; 2002, pp.113-114.

[60]V.C. Ibekwe, M.K. Khela, D.F. Evans, A.W. Basit, "A new concept in colonic drug targeting: a combined pH-responsive and bacterially-triggered drug delivery technology", Aliment PharmacolTher, vol. 28, no.7, pp. 911-916, 2008.

Autors Information

D. Filip *is with the Department of Electrical and Computer Engineering, University of Calgary, Calgary, AB T2N 1N4, Canada.*

O. Yadid-Pecht *is with the Department of Electrical and Computer Engineering, University of Calgary, Calgary, AB T2N 1N4, Canada.*

C.N. Andrews *is with the Department of Medicine, University of Calgary, Calgary, AB, T2N 1N4, Canada.*

**M. P. Mintchev* *is with the Department of Electrical and Computer Engineering, University of Calgary, 2500 University Dr. NW, Calgary, AB, T2N 1N4, Canada (e-mail: mintchev@enel.ucalgary.ca).*

EVOLVING CASCADE NEURAL NETWORK BASED ON MULTIDIMESNIONAL EPANECHNIKOV'S KERNELS AND ITS LEARNING ALGORITHM

Yevgeniy Bodyanskiy, Paul Grimm, Nataliya Teslenko

***Abstract**: At present time neural networks based on Group Method of Data Handling (GMDH-NN), nodes of which are two-input N-Adalines, is well-known. Each of N-Adalines contains the set of adjustable synaptic weights that are estimated using standard least squares method and provides quadratic approximation of restoring nonlinear mapping. On the other hand, for needed approximation quality ensuring this NN can require considerable number of hidden layers. Approximating properties of GMDH-NN can be improved by uniting the approaches based on Group Method of Data Handling and Radial-Basis-Functions Networks that have only one hidden layer, formed by, so-called, R-neurons. Such networks learning reduces, as a rule, to the tuning of synaptic weights of output layer that are formed by adaptive linear associators. In contrast to neurons of multilayer structures with polynomial or sigmoidal activation functions R-neurons have bell-shaped activation functions. In this paper as activation functions multidimensional Epanechnikov's kernels are used. The advantage of activation function is that its derivatives are linear according all the parameters that allows to adjust sufficiently simply not only synaptic weights but also centers with receptive fields. Proposed network combines Group Method of Data Handling, Radial-Basis-Functions Networks and cascade networks and isn't inclined to the "curse of dimensionality", is able to real time mode information processing by adapting its parameters and structure to problem conditions. The multidimensional Epanechnikov's kernels were used as activation functions, that allowed to introduce numerically simple learning algorithms, which are characterized by high speed.*

***Keywords**: evolving neural network, cascade networks, radial-basis neural network, Group Method of Data Handling, multidimensional Epanechnikov's kernels.*

***ACM Classification Keywords**: F.1 Computation by abstract devices – Self-modifying machines (e.g., neural networks), I.2.6 Learning – Connectionism and neural nets, G.1.2 Approximation – Nonlinear approximation.*

Introduction

At present artificial neural networks have gotten a wide spread for extensive class of pattern recognition, identification, emulation, intelligent control, time series prediction etc. problems due to universal approximating properties and abilities to learn. As far as when a number of practical tasks solving the volume of learning sample is limited, then to the foreground learning rate factor comes. At the same time not all the neural networks are able to overcome arising problems and, first of all, so-called "overfitting". As one of the most high-performance networks that are learned based on optimization procedures of second order with high convergence rate is Radial Basis Functions Neural Network (RBFN). Output signal of this network linearly depends on adjusting synaptic weights. At the same time these networks are inclined to so-called "curse of dimensionality", when the number of radial-basis neurons of hidden layer (R-neurons) exponentially grows while input signals state grows.

It is possible to overcome this problem by dividing the initial task in one or another way to a number of subtasks of low dimensionality and grouping obtained solutions to get required result. From computational point of view the

most convenient in this case is Group Method of Data Handling (GMDH) [Ivakhnenko, 1969; Ivakhnenko, 1975; Ivakhnenko, Stepashko, 1975; Ivakhnenko, Madala, 1994] that demonstrated its efficiency when solving a great number of practical tasks.

In [Pham, Liu, 1995] multi-layered GMDH-neural network was considered. It has two-inputs N-adalines as a nodes and output of each node is quadratic function of input signals. At the same time the synaptic weights of each neuron are defined in patch mode using standard least squares method. It can be needed some more quantity of hidden layers to provide necessary approximation quality. That is why on-line learning becomes impossible.

Hybrid architecture of artificial neural network based on ideas of GMDH and consequent forming of cascade neural networks [Avedyan, Barkan, Levin, 1999]. Nodes of this network are compartmental R-neurons with activation functions like multidimensional Epanechnikov's kernels [Epanechnikov, 1969; Friedman, Hastie, Tibshirani, 2003; Bodyanskiy, Chaplanov, Kolodyazhniy, Otto, 2002], that have large degree of freedom, and thus improved approximating properties in comparison with conventional Gaussians.

Compartmental R-neuron with Multidimensional Epanechnikov Kernels and Its Learning Algorithm

Let introduce the structure of compartmental R-neuron presented on fig.1 and concurring with simplified architecture of conventional Radial Basis Functions Neural Network with two inputs x_i and x_j, $i, j = 1, 2, \ldots, n$, where n – dimensionality of input space.

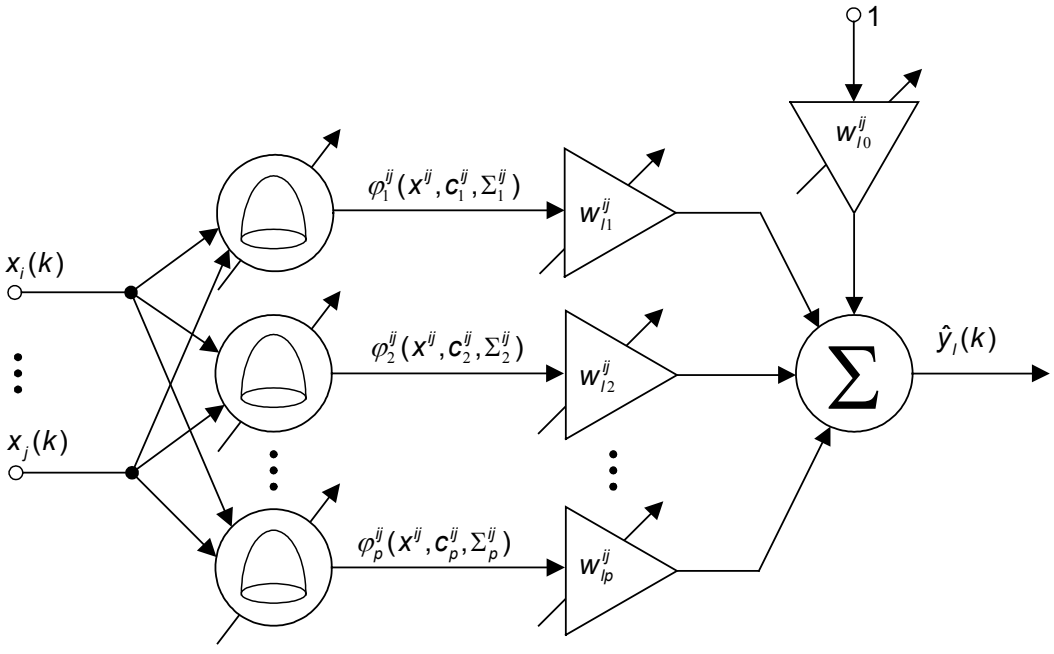


Fig. 1 – Compartmental R-neuron

Compartmental R-neuron contains p activation functions (conventionally in RBFN multidimensional Gaussians or other bell-shaped functions are used) $\varphi_h^{ij}(x^{ij}, c_h^{ij}, \Sigma_h^{ij})$, p+1 synaptic weights that are united to vector $w_l^{ij} = (w_{l0}^{ij}, w_{l1}^{ij}, \ldots, w_{lp}^{ij})$, p two-dimensional vector of centers $c_h^{ij} = (c_h^i, c_h^j)^{\mathrm{T}}$, p (2×2) – matrices of receptive

fields of activation functions Σ_h^{ij},two-dimensional inputs vector $x^{ij} = (x_i, x_j)^{\mathrm{T}}$, one output $\hat{y}_l$; $l = 1,2,\ldots,p$; $k = 1,2,\ldots,N$ – number of observation in processing sample or index of current discrete time.

Multidimensional Epanechnikov's kernels are used as activation functions $\varphi_h^{ij}(x^{ij}, c_h^{ij}, \Sigma_h^{ij})$

$$\varphi_h^{ij}(x^{ij}, c_h^{ij}, \Sigma_h^{ij}) = 1 - \left\| x^{ij} - c_h^{ij} \right\|^2_{\left(\Sigma_h^{ij}\right)^{-1}}, \tag{1}$$

that have bell-shaped by positive definite matrix of receptive field Σ_h^{ij}. The advantage of activation function (1) in comparison with conventional ones is in linearity of its derivatives with respect to all the parameters that allows to adjust not only synaptic weights but also centers with receptive fields sufficiently easy.

At the same time, transformation that is realized by compartmental R-neural has a form

$$\hat{y}_l = w_{l0}^{ij} + \sum_{h=1}^{p} w_{lh}^{ij} \varphi_h^{ij}(x^{ij}, c_h^{ij}, \Sigma_h^{ij}) = w_{l0}^{ij} + \sum_{h=1}^{p} w_{lh}^{ij} \left(1 - \left\| x^{ij} - c_h^{ij} \right\|^2_{\left(\Sigma_h^{ij}\right)^{-1}} \right).$$

Usually RBFN learning comes to synaptic weights w_{lh}^{ij} adjusting, but centers and characteristics of receptive fields are defined priory. At the same time for two-dimensional case it is sufficiently easy to locate the centers at the regular lattice nodes and define receptive field as circles. Learning process itself consists of synaptic weights vector w_l^{ij} estimating by learning sample containing N observations $x^{ij}(k) = (x_i(k), x_j(k))^{\mathrm{T}}$, $y(k)$, $k = 1,2,\ldots,N$, where $y(k)$ - external learning signal.

By introducing to the consideration $(p+1)\times 1$-vector of activation functions $\varphi^{ij}(k) = (1, \varphi_1^{ij}(x^{ij}(k), c_1^{ij}, \Sigma_1^{ij}), \ldots, \varphi_p^{ij}(x^{ij}(k), c_p^{ij}, \Sigma_p^{ij}))^{\mathrm{T}}$ and learning criterion

$$E_l^N = \sum_{k=1}^{N} (y(k) - \hat{y}_l(k))^2 = \sum_{k=1}^{N} e_l^2(k) = \sum_{k=1}^{N} (y(k) - (w_l^{ij})^{\mathrm{T}} \varphi^{ij}(k))^2, \tag{2}$$

using standard least squares method it is easy to obtain required solution in the form

$$w_l^{ij} = \left(\sum_{k=1}^{N} \varphi^{ij}(k)(\varphi^{ij}(k))^{\mathrm{T}} \right)^{+} \sum_{k=1}^{N} \varphi^{ij}(k) y(k), \tag{3}$$

where $(\bullet)^{+}$ – symbol of inversion by Moore-Penrose.

If the data are fed to the processing consequently in on-line mode, then instead of (3) can be used its recurrent variant in the form

$$\begin{cases} w_l^{ij}(k) = w_l^{ij}(k-1) + \dfrac{P_{ij}(k-1)(y(k) - (w_l^{ij}(k-1))^{\mathrm{T}} \varphi^{ij}(k))}{1 + (\varphi^{ij}(k))^{\mathrm{T}} P_{ij}(k-1) \varphi^{ij}(k)} \varphi^{ij}(k), \\ P_{ij}(k) = P_{ij}(k-1) - \dfrac{P_{ij}(k-1)\varphi^{ij}(k)(\varphi^{ij}(k))^{\mathrm{T}} P_{ij}(k-1)}{1 + (\varphi^{ij}(k))^{\mathrm{T}} P_{ij}(k-1) \varphi^{ij}(k)}, \quad P_{ij}(0) = \gamma \mathrm{I}, \quad \gamma >> 0. \end{cases} \tag{4}$$

Algorithms (3) and (4) are effective only in the cases, when required solution is stationary, that is, optimal values of synaptic weights aren't variable in time. Since in many practical tasks it is not so, for example, adaptive identification of non-stationary objects or non-stationary time series prediction, then high-performance adaptive learning algorithm having tracking and filtering properties can be used [Bodyanskiy, Kolodyazhniy, Stephan, 2001]:

$$\begin{cases} w_l^{ij}(k) = w_l^{ij}(k-1) + \eta_w(k)(y(k) - (w_l^{ij}(k-1))^{\mathrm{T}}\varphi^{ij}(k))\varphi^{ij}(k) = \\ = w_l^{ij}(k-1) + \eta_w(k)e_l(k)\varphi^{ij}(k), \\ \eta_w^{-1}(k) = r_w(k) = \alpha r_w(k-1) + \left\|\varphi^{ij}(k)\right\|^2, \quad 0 \le \alpha \le 1, \end{cases}$$

where α – smoothing parameter that defines compromise between filtering and tracking properties.

For the purpose of compartmental R-neuron approximating properties improving not only synaptic weights but also centers with receptive fields can be adjusted. At the same owing to Epanechnikov's kernels using learning algorithms have sufficiently simple form.

By using gradient procedure of criterion (2) minimization and its rate optimization technique [Otto, Bodyanskiy, Kolodyazhniy, 2003] we obtain compartmental R-neuron learning algorithm in the form

$$\begin{cases} w_l^{ij}(k) = w_l^{ij}(k-1) + \eta_w(k)e_l(k)\varphi^{ij}(k), \\ \eta_w^{-1}(k) = r_w(k) = \alpha r_w(k-1) + \left\|\varphi^{ij}(k)\right\|^2, \\ c_h^{ij}(k) = c_h^{ij}(k-1) + \eta_c(k)e_l(k)w_l^{ij}(k)\left(\Sigma_h^{ij}(k-1)\right)^{-1}(x^{ij}(k) - c_h^{ij}(k-1)) = \\ = c_h^{ij}(k-1) + \eta_c(k)e_l(k)g_h(k), \\ \eta_c^{-1}(k) = r_c(k) = \alpha r_c(k-1) + \left\|g_h(k)\right\|^2, \\ \left(\Sigma_h^{ij}(k)\right)^{-1} = \left(\Sigma_h^{ij}(k-1)\right)^{-1} - \eta_\Sigma(k)e_l(k)w_l^{ij}(k)(x^{ij}(k) - c_h^{ij}(k))(x^{ij}(k) - c_h^{ij}(k))^{\mathrm{T}} = \\ = \left(\Sigma_h^{ij}(k-1)\right)^{-1} - \eta_\Sigma(k)e_l(k)G_h(k), \\ \eta_\Sigma^{-1}(k) = \Gamma_\Sigma(k) = \alpha\Gamma_\Sigma(k-1) + Tr(G_h(k)G_h^{\mathrm{T}}(k)). \end{cases}$$

Evolving Cascade Neural Network

Uniting of GMDH and cascade neural networks ideas leads to architecture presented on fig. 2.

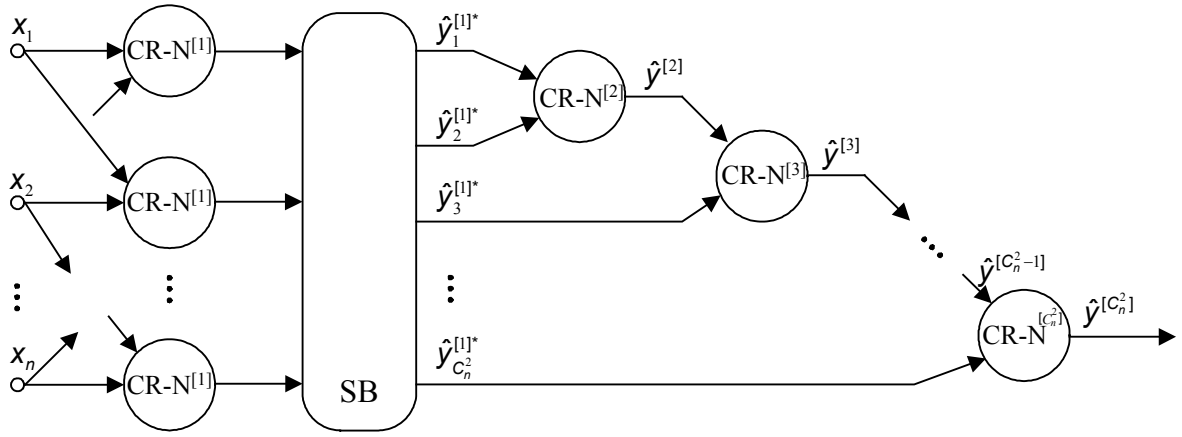


Fig. 2 – Evolving cascade neural network

The first hidden layer of the network is formed similarly to the first hidden layer of GMDH neural network [Pham, Liu, 1995] and contains the number of neurons equal to quantity of combinations of *n* in 2, that is C_n^2. Selection

block SB executes sorting by accuracy, for example, in the sense of variations, of all output signals $\hat{y}_l^{[1]}$ so that the most accurate signal is $\hat{y}_1^{[1]*}$, then $\hat{y}_2^{[1]*}$ and the worst is $\hat{y}_{C_n^2}^{[1]*}$. Outputs of SB $\hat{y}_1^{[1]*}$ and $\hat{y}_2^{[1]*}$ then are fed to the one neuron of the second layer-cascade CR-N[2] that computes signal $\hat{y}^{[2]}$ which in the third cascade is combined with $\hat{y}_3^{[1]*}$. The process of cascades increasing lasts till required accuracy obtaining, at the same time, maximal neurons number of this network is restricted by the value $2C_n^2 - 1$. Thus, neural network is able to process information that are fed in real time by readjusting both its parameters and its architecture in time [Kasabov, 2003] and by adapting to the conditions of the solving task.

Conclusion

Architecture of evolving cascade radial-basis neural network was proposed in this paper. It is formed based on the idea of combining GMDH and cascade networks. Also this network is not disposed to the "curse of dimensionality" and is able to process information in real time by adapting its parameters and structure to the solving task conditions. Using of multidimensional Epanechnikov's kernels as activation functions allowed to introduce numerically simple learning algorithms that are characterized by high-performance.

Acknowledgements

The paper is published with financial support by the project ITHEA XXI of the Institute of Information Theories and Applications FOI ITHEA (www.ithea.org) and the Association of Developers and Users of Intelligent Systems ADUIS Ukraine (www.aduis.com.ua).

Bibliography

[Ivakhnenko, 1969] Ivakhnenko A.G. Self-learning systems for recognition and automatic control. Technika, Kiev, 1969 (in Russian).

[Ivakhnenko, 1975] Ivakhnenko A.G. Long-term prediction and complex systems control. Technika, Kiev, 1975 (in Russian).

[Ivakhnenko, Stepashko, 1975] Ivakhnenko A.G., Stepashko V.S. Modeling noise-immunity. Naukova Dumka, Kiev, 1975 (in Russian).

[Ivakhnenko, Madala, 1994] Ivakhnenko A.G., Madala H.R. Inductive Learning Algorithms for Complex Systems Modeling. CRC Press, London-Tokio,1994.

[Pham, Liu, 1995] Pham D.J., Liu X. Neural Networks for Identification, Prediction and Control. Springer-Verlag, London, 1995.

[Avedyan, Barkan, Levin, 1999] Avedyan E. D., Barkan G.V., Levin I.K. Cascade neural networks // Avtomatika i Telemekhanika. – 1999. – №3. – P. 38-55. (in Russian).

[Epanechnikov, 1969] Epanechnikov V.A. Non-parametric estimate of multidimensional probability density // Probability theory and its application. – 1969. – 14. – №1. – P. 156-161. (in Russian).

[Friedman, Hastie, Tibshirani, 2003] Friedman J., Hastie T., Tibshirani R. The Elements of Statistical Learning. Data Mining, Inference and Prediction. Springer, Berlin, 2003.

[Bodyanskiy, Chaplanov, Kolodyazhniy, Otto, 2002] Bodyanskiy Ye., Chaplanov O., Kolodyazhniy V., Otto P. Adaptive quadratic radial basis function network for time series forecasting // Proc. East West Fuzzy Coll. 2002.– Zittau/Goerlitz: HS, 2002.– P.164-172.

[Bodyanskiy, Kolodyazhniy, Stephan, 2001] Bodyanskiy Ye., Kolodyazhniy V., Stephan A. An adaptive learning algorithm for a neuro-fuzzy network / Ed. by B.Reusch "Computational Intelligence. Theory and Applications". – Berlin-Heidelberg-New York: Springer, 2001.– P.68-75.

[Otto, Bodyanskiy, Kolodyazhniy, 2003] Otto P., Bodyanskiy Ye., Kolodyazhniy V. A new learning algorithm for forecasting neuro-fuzzy network// Integrated Computer-Aided Engineering. – 2003. – 10. – №4. – P. 399-409.

[Kasabov, 2003] Kasabov N. Evolving Connectionist Systems: Methods and Applications in Bioinformatics. Springer–Verlag, London, 2003.

Authors' Information

Yevgeniy Bodyanskiy *– Professor, Dr.-Ing. habil. Scientific head of Control Systems Research Laboratory, KhNURE, Senior member of IEEE, professor of Artificial Intelligence Department of KhNURE.*

e-mail: bodya@kture.kharkov.ua

Major Fields of Scientific Research: hybrid systems of computational intelligence.

Paul Grimm *– Professor, Computer Graphics Department, Erfurt University of Applied Sciences.*

e-mail: grimm@fh-erfurt.de

Major Fields of Scientific Research: image processing using intelligent computer technologies.

Nataliya Teslenko *– Ph.D., Senior researcher of Control Systems Research Laboratory, KhNURE.*

e-mail: ntntp@ukr.net

Major Fields of Scientific Research: self-learning evolving neuro-fuzzy models and systems in the intelligent data analysis tasks

REGIONS OF SUFFICIENCY FOR METRICAL DATA RETRIEVAL

Vladimir Mashtalir, Konstantin Shcherbinin, Vladislav Shlyakhov, Elena Yegorova

***Abstract**: In this paper the fast metrical search for large scaled and poor structured databases using objects eliminating from the consideration without calculating the distance between them and query is considered and grounded. This search is based on the pre-calculated distances between pivot points and database objects, and triangular inequality as the base for the objects elimination without calculating distances between them. For that the sufficient and necessary conditions of the objects elimination are explored and mathematical foundation for the sufficiency and necessity regions in the metric space are given.*

***Keywords**: Metrical Retrieval, pivot point, sufficiency*

***ACM Classification Keywords**: H3.3.3 Information Search and retrieval.*

Introduction

Nowadays many application require process and retrieve data which is more and more characterized with much less structure and queries precision, as for instance multimedia or video database systems [Geetha, 2008], [Natsev, 2006]. The efficiency of such systems is mainly measured by the search speed and quality. Many ordinary retrieval techniques appear to be inadequate for modern large-scale databases [Snoek, 2008].

The need for more suitable and effective tools gave a push to develop retrieval methods based upon the distance, i.e. performing the search in the metric space [Barsi, 2011], [Kinoshenko, 2010]. Allowing the similarity queries execution this approach was used in a lot of techniques and methods of creating indexing structures. In general the approaches are divided into tree-based (M-tree algorithms, Slim-tree, Slim-Down algorithm, Multi-Way Insertion algorithm) and hash-based (D-index, Insertion and Search strategies) metric indexing [Dohnal, 2003], [Liu, 2007], [Zezula, 2006]. The first ones traverse trees and visit nodes which reside within the query region. In the best case it means the logarithmic search cost. Second ones provide a direct access to searched regions with no additional traversals of the underlying structure, but still the costs increase linearly with the growth of the dataset. It means that with the dataset growth the ability to perform the reasonable time search is rather limited.

Expansion of CBIR, CBVR, CBMR (Content Based {Image, Video, Multimedia} Retrieval) systems led to developing a wide range of algorithms using off-line processing, what allowed to narrow the search to the database objects distance matrices analysis [Liu, 2007], [Kinoshenko[1], 2010]. Still under very large data volumes using of the complete matrix becomes not effective enough and arises the need to analyze sparse matrices which practically represent inner and intra-cluster distances.

One more key aspect of increasing the search speed in the database metrical search is objects eliminating from the consideration without calculating the distance between them and query. This paradigm – eliminate what is not needed and perform the search among what is left – provides essential reduction of the potentially tolerant objects set cardinality and also allows to rank the objects based on their similarity [Zezula, 2006]. At that it is not

substantial if the decision is made based on the law level features or actively using semantics. This search is based on the pivot points, the distance to them is preliminary calculated distances between the database objects and triangular inequality is the base for the objects elimination without calculating distances between them. Obviously the more a priori information is considered the bigger speed increase can be expected.

The next paper is organized as following: in section 2 the mathematical premises for metrical search sufficiency and necessity regions are defined, in section 3 and 4 sufficiency conditions analysis and the grounds of sufficiency region construction using one pivot point is conducted, section 5 describes sufficiency region construction using n pivot points.

Sufficiency and necessity regions for metric search

In some phase space Φ we shall consider a finite points configuration $K \subseteq \Phi$, for which the distance matrix $\mathcal{P}$ is known, i.e. a set of point $x_1,\ldots,x_n \in \Phi$ actually defines a set of objects in the database (in this case can be video strings, characterized by segments features in time and/or space, key frames, events, etc) among which a subset of most similar to $y \in \Phi$ query will be selected. Thus, a symmetric (it is essential for the objects storing and indexing) $(n \times n)$ of-line calculated matrix is given

$$\mathcal{P} = \left(\rho(x_i, x_j)\right)_{i,j=1}^{n} \tag{1}$$

where $\rho(x_i, x_j)$ is a distance between points x_i, x_j.

For some query i.e. point $y \in \Phi$, we shall define the distance to the pivot point $x* \in K$ (we shall not discuss here the choice of one or few point like that, as well as the optimal number of those)

$$\delta = \rho(y, x*). \tag{2}$$

We shall introduce the objects similarity threshold Δ which characterizes similarity measure of points in metrical space Φ with the current point y. Thus if

$$\rho(y, x_k) > \Delta \tag{3}$$

Then point x_k so much differs from y, that it is automatically eliminated from the further analysis.

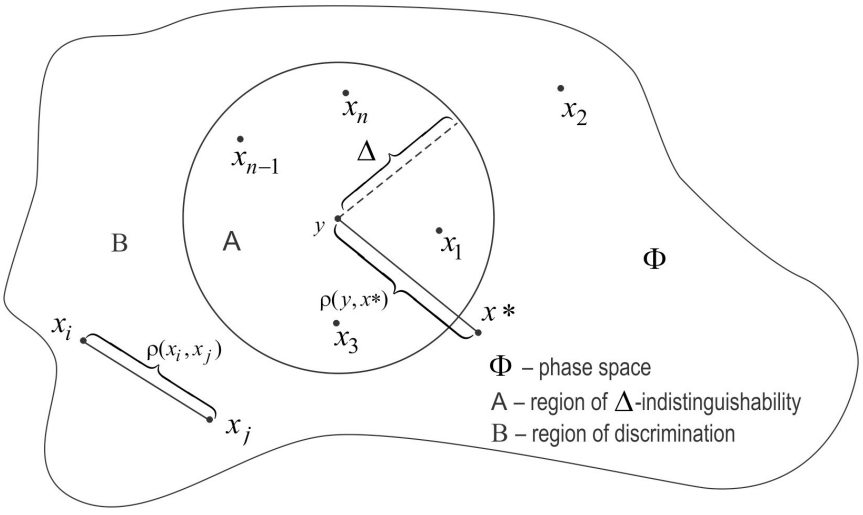


Fig. 1. Geometrical interpretation of metrical search conditions

Based on the stated above, we shall aim to create a procedure of defining the closest points of the K configuration to the given point $y \in \Phi$. At that we consider a distance matrix $\mathcal{P}$, formed according to (1), similarity measure Δ and known distance (2) between current point y and pivot point $x* \in K$. It should be emphasized that this task is first of all conditioned with the complexity of computing $\rho(y, x_k)$ distances, what under large cardinalities of configurations $K \subseteq \Phi$ can lead to intolerable time expenditures. It is clear, that the search result (if exists, as not all points satisfy (3)) can be the closest according the given metric point or m points ordered ascending on the distance to the query. In other words, the result of (Δ, k)- search on query $y \in \Phi$ are elements of set $K^m = \{x_i, x_i, \ldots, x_i\} \subseteq K$, $0 < m \le k \le n$, for which

$$\begin{aligned} &\forall x_{ij} \in K^m, \forall (x \in K) \setminus K^m, \forall y \in \Phi \quad y\ x_{ij} \le \Delta\ \Delta \ge 0, \\ &\rho(y, x_{ij}) \le \rho(y, x), \rho(y, x_{ij}) \le \rho(y, x_{ij+1}), j = \overline{1, m-1}. \end{aligned} \tag{4}$$

Let us now analyze sufficiency and necessity regions for search (4). We shall fix two points of given configuration $x_1, x_2 \in K$. Then naturally constructing the procedure of choosing the closest among two objects in the database without calculating the distance between then and query y allows to solve the stated task in the most general case. At that it is assumed that analysis (choice of the closest one) of any pairs of points of set K (total number of variants for the analysis is C_n^2) in sense of time and material costs in view of measuring the distance $\rho(y, x_1)$ and $\rho(y, x_2)$ is quite insignificant value. Such an assumption is common for the video frames processing tasks as searching for most similar frames or their sequences to the query commonly needs to be carried out within rough time restrictions, and distance calculations in a phase space Φ is usually computationally expensive. Let us introduce the next definitions.

Definition 1. A sufficiency region S we shall call a region from $\mathbb{R}^n$ with following properties:

– a function $\varphi : K \times K \to \mathbb{R}^n$ exists, i.e. $\varphi(x_1, x_2) = (\alpha_1, \alpha_2, \ldots, \alpha_n)$;

– for this function is true that if $\varphi(x_1, x_2) \in S$, then it is sufficient to find $min\ \{\rho(y, x_1), \rho(y, x_2)\}$.

Shall we point out that the values of distances $\rho(y, x_1), \rho(y, x_2)$ are not known, but we can indicate the minimal one. Namely this property gives the main premise for constructing the fast-operating elimination procedures.

Definition 2. The region of necessity N we shall call a region from $\mathbb{R}^n$ which has the following properties:

– a function $\psi : K \times K \to \mathbb{R}^n$ exists, i.e. $\psi(x_1, x_2) = (\alpha_1, \alpha_2, \ldots, \alpha_n)$;

– for this function is true that if one can indicate a proper inequality from the set

$$\begin{cases} \rho(y, x_1) \ge \rho(y, x_2); \\ \rho(y, x_2) \ge \rho(y, x_1), \end{cases}$$

then condition $\psi(x_1, x_2) \in N$ should be fulfilled.

The core of these definitions is that dimensionality of the sufficiency and necessity is defined by the fact how fully we use the input information (the more fully we use it the larger dimension we shall get). From the other hand if $\varphi(x_1, x_2)$ is in region S it is enough to choose the point most close to query y. On the contrary, if we know the

point most close to y, it is necessary that membership $\psi(x_1, x_2) \in N$ is fulfilled. It is clear that knowing of the sufficiency region in general allows to solve the given task. And ability to choose the closest point from an arbitrary pair allows to sequentially find the most close one from the K set when following for instance the next procedure:

– choose the closest to query y point from x' and x'';

– suppose that it is point x', then we eliminate x'' and move to the next step;

– choose the next (is matching has not finished) point x''', and then the closest one from x' and x''' and go back to the previous step.

Finally we end up with the point from K set which is the closest one to the query. At that the matching operations number would be equal to $card(K) - 1$. All said above can be formulated in next statement.

Proposition 1. Knowledge of sufficiency region S for arbitrary pair of point $x_1, x_2 \in K$ allows to find the point closest to query y.

Now let us analyze the necessity region. For that we should first consider some properties of the sufficiency and necessity regions.

Regions S and N are parted into two subregions S', S'' (N', N''). Region S' satisfies the situation when

$$\rho(y, x_1) \geq \rho(y, x_2) \tag{5}$$

and S'' is for

$$\rho(y, x_2) \geq \rho(y, x_1). \tag{6}$$

Then $S = S' \cup S''$ and $S_1 \cap S'' = \varnothing$ except for the situation when $\rho(y, x_1) = \rho(y, x_2)$. It is clear that similar property is also fulfilled for N.

Let us assume now that it is a priori known which of the inequalities (5) or (6) is fulfilled. Then if (5) is fulfilled, it should be $\psi(x_1, x_2) \in N'$, and if (6) is fulfilled – $\psi(x_1, x_2) \in N''$.

Suppose that

$$\psi(x_1, x_2) \notin N' \tag{7}$$

then (5) is not fulfilled and so the (6) is true. In other words from $\psi(x_1, x_2) \notin N'$ follows condition (6), i.e. it is sufficient for (6) or $\psi(x_1, x_2) \notin N' \subset S''$.

From the other hand if (7) is true then (6) is fulfilled and so $\psi(x_1, x_2) \in N''$. Thus for the regions exists an enclosures chain: $N' \subset \psi(x_1, x_2) \notin N' \subset S''$, i.e. $N'' \subset S''$. But D'' is a sufficiency region of inequality (6), from which follows the sufficiency region N'', and so $S'' \subset N''$. From here it follows that $S'' = N''$. Similarly we can show that $S' = N'$. Therefore $N = S$. We can formulate that as following.

Proposition 2. Under the metrical search based on alternatives (5), (6) the sufficiency region coincides the necessity region.

Thus, the conditions which characterize regions of sufficiency and necessity are necessary and sufficient. From proposition 2 it follows that if one can construct the necessity region, it fully solves the given task. But any sufficiency subregion is also of the same kind, thus the sufficiency region can be narrowed, which is not possible

for the necessity region. Practically we state that the maximal sufficiency region coincide the necessity region. At that regions of sufficiency and necessity should satisfy the $S \subset N$ enclosure and only $S_{max} = N$.

When stating the above we certainly assume that S', S'' are the maximal regions of sufficiency for fulfillment of inequalities (5) and (6) accordingly.

It should be particularly noted that when solving the non-alternative search tasks the regions S and N do not have to satisfy the Proposition 2.

Summarizing the above we can introduce some generalized concept of aggregated phase space.

Definition 3. Cumulative phase space $\widetilde{\Phi}$ we shall call a couple $\widetilde{\Phi} = \langle \Phi, U \rangle$, where Φ is a phase space to which belong the current point y and points configuration K, and U is a set of input data used under the objects elimination.

All so far obtained results are true also for the cumulated phase space $\widetilde{\Phi}$. It relates to the fact that considering the conditions of sufficiency and necessity we did not take into consideration the initial data set (1) – (3) as set

$$U = \{\mathcal{P}, \delta, \Delta\}. \tag{8}$$

Further we shall consider finding the sufficiency regions for pairs of points under different concrete realizations of U conditions.

Sufficiency conditions analysis when having one pivot point

We shall start from the situation when conditions have form (8) and configuration K consist of 3 points $K = \{x*, x_1, x_2\}$, where the pivot point $x*$ is separated and for it $\rho(y, x*) = \delta$ is known. Now we can concretize the elements of U as

$$\mathcal{P} = \begin{pmatrix} \rho(x*, x*) & \rho(x*, x_1) & \rho(x*, x_2) \\ \rho(x*, x_1) & \rho(x_1, x_1) & \rho(x_1, x_2) \\ \rho(x*, x_2) & \rho(x_1, x_2) & \rho(x_2, x_2) \end{pmatrix}, \rho(y, x*) = \delta > 0 \tag{9}$$

and point for which the condition of similarity degree condition $\rho(y, x*) < \Delta$ is fulfilled are considered.

We should notice that taken into consideration points configuration K of phase space Φ automatically satisfies the last condition as in contrary the closest point search algorithm is obvious or leads to the confluent situation. Indeed if at least one point of configuration $K = \{x*, x_1, x_2\} \in \Phi$ does not lie in the neighborhood $U_\Delta(y)$ of radius Δ of the current point $y \in \Phi$, then other variants are possible.

Let $x* \notin U_\Delta(y)$. Then the chosen pivot point in the configuration conditionally is “not quite similar” to the current one. So in this case the pivot point choice is not successful and another point should be chosen

Assume that one of the points x_1 or x_2 does not belong to neighborhood $U_\Delta(y)$. If it is x_1, then obviously x_2 is the closest one. If both of then do not belong, these is no sense of searching the closest one.

It has to be mentioned that as opposed to the pivot point $x*$, it is not possible to define membership of configuration K points in neighborhood $U_\Delta(y)$, as distances $\rho(y, x_1)$ and $\rho(y, x_2)$ are not known. But it is meant that when searching for the sufficiency region this condition should be automatically fulfilled. If there is not

need in similarity thresholds or there is not possible to choose them, it is assumed that $\Delta = \infty$, and according to (4) one search k points ranged according to their similarity degree.

Let us denote $\rho(x*, x_1) = \alpha, \rho(x*, x_2) = \beta$ and $\rho(x_1, x_2) = \gamma$. Then distance matrix $\mathcal{P}$ with consideration of (9) and metric properties takes form

$$\mathcal{P} = \begin{pmatrix} 0 & \alpha & \beta \\ \alpha & 0 & \gamma \\ \beta & \gamma & 0 \end{pmatrix}. \tag{10}$$

Based on the distance matrix $\mathcal{P}$ (actually assemblage of numbers (α, β, γ)) the sufficiency region can be constructed in $\mathbb{R}^3$. Using the definition from Section 2 we have function $\varphi : K \times K \to \mathbb{R}^3$ defined by the distance matrix $\mathcal{P}$ and equal to

$$\varphi(x_1, x_2) = (\alpha, \beta, \gamma) \in \mathbb{R}^3. \tag{11}$$

Thus consideration of the metric properties, and first of all the triangular inequality creates the premises for sufficiency region constructing.

Sufficiency region construction under one pivot point

Let us assume that we have a situation according to the schema on the Figure 1 and function $\varphi(x_1, x_2)$ given by equality (11). Then sufficiency region in $\mathbb{R}^3$ under parameters (α, β, γ) is defined as following:

– restrictions introduced by virtue of α, β, γ being distances, i.e.

$$\alpha, \beta, \gamma > 0; \tag{12}$$

– restrictions which follow from the triangular inequality according to Figure 2.

$$\begin{cases} |\rho - \alpha| \le \rho(y, x_1) \le \rho + \alpha; \\ |\rho - \beta| \le \rho(y, x_2) \le \rho + \beta \end{cases} \tag{13}$$

are inequalities corresponding to the triangular with current point y, but there is a set of inequalities corresponding to the configuration K:

$$\begin{cases} \alpha + \beta \ge \gamma; \\ \alpha + \gamma \ge \beta; \\ \gamma + \beta \ge \alpha; \end{cases} \tag{14}$$

– condition which follows from the similarity criterion, i.e. restrictions of configuration $K \in U_\Delta(y)$ or

$$\begin{cases} \rho(y, x*) = \rho \le \Delta; \\ \rho(y, x_1) \le \Delta; \\ \rho(y, x_2) \le \Delta. \end{cases} \tag{15}$$

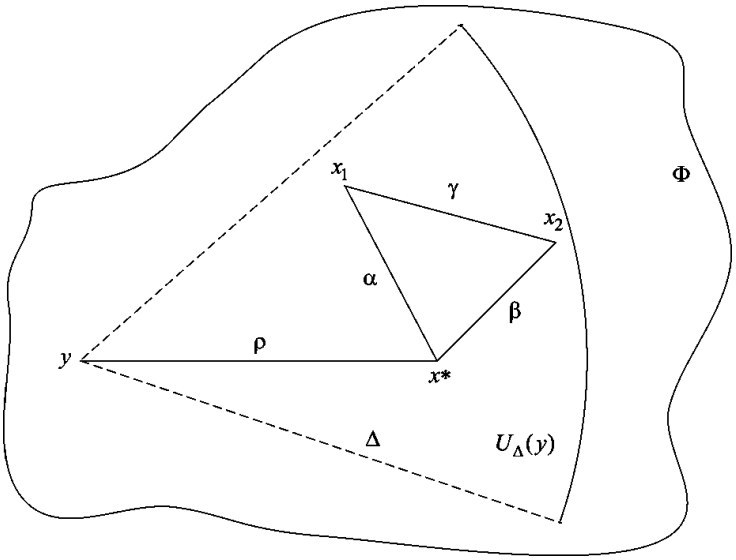


Fig. 2. Geometrical interpretation of search conditions in $\mathbb{R}^3$

In the result it is easy to understand that the sufficiency region will be a region where at least one of the lower bounds of inequalities (13) ranks over the upper bound of the same inequalities. Formally it can be written as

$$|\rho - \alpha| \geq \rho + \beta \tag{16}$$

or

$$|\rho - \beta| \geq \rho + \alpha \ . \tag{17}$$

One can see that from (16) follows $\rho(y, x_1) \geq \rho(y, x_2)$, i.e. point x_2 is closer to current point y. And vice versa, from (17) it follows that $\rho(y, x_2) \geq \rho(y, x_1)$, i.e. point x_1 is closer to current point y.

Taking into account all inequalities (12) – (15) the sufficiency region in $\mathbb{R}^3$ satisfies the conditions set

$$\begin{cases} \alpha, \beta, \gamma > 0; \rho, \Delta > 0; \\ \begin{cases} |\rho - \alpha| \geq \rho + \beta; \\ |\rho - \beta| \geq \rho + \alpha; \end{cases} \\ \alpha + \beta \geq \gamma; \\ \alpha + \gamma \geq \beta; \\ \gamma + \beta \geq \alpha; \\ \rho \leq \Delta; \\ \rho + \alpha \leq \Delta; \\ \rho + \beta \leq \Delta. \end{cases} \tag{18}$$

One should note that the last three inequalities of system (18) provide the membership of configuration $K \in U_\Delta(y)$. Indeed, $\rho \leq \Delta$ signifies that $x* \in U_\Delta(y)$. And if $\rho + \alpha \leq \Delta$, then from (13) it follows

$$\rho(y, x_1) \leq \rho + \alpha \leq \Delta$$

or $x_1 \in U_\Delta(y)$. Similarly from $\rho + \beta \le \Delta$ follows that $x_2 \in U_\Delta(y)$.

Thus the next statement is proved.

Proposition 3. If elements α, β, γ of distance matrix (10) satisfies the system of conditions (9), then the choice of the closest point of configuration K to the query y can be made without calculating the distance to it.

As the considered choice situation is limited by $\mathbb{R}^3$, there is a possibility to give a geometrical interpretation of the system (9) or explicitly point out the sufficiency region in the $\mathbb{R}^3$ space as a region restrained by planes. We shall start with the inequality of type

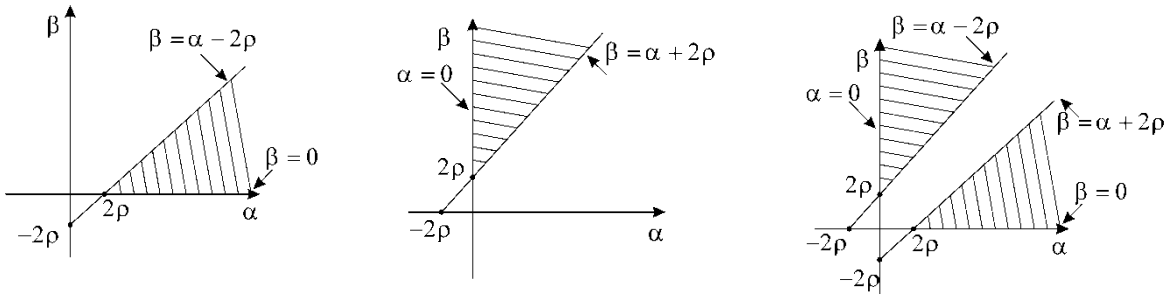
$$\begin{cases} |\rho - \alpha| \ge \rho + \beta; \\ |\rho - \beta| \ge \rho + \alpha \end{cases} \tag{19}$$

Note that this set of conditions is the most essential one, as it allows to choose the closest point (otherwise the algorithm of choosing can be terminated). Practically below we shall construct a sufficiency region projection on the plane of variables (α, β) or plane $\gamma = 0$. System (19) can be put in form of

$$\begin{cases} \begin{cases} \rho - \alpha \ge \rho + \beta; \\ \rho - \alpha \le -\rho - \beta; \end{cases} \\ \begin{cases} \rho - \beta \ge \rho + \alpha; \\ \rho - \beta \le -\rho - \alpha. \end{cases} \end{cases} \tag{20}$$

We shall consider the first part of system (20), i.e. union of first two inequalities. It is easy to see that the first one $\rho - \alpha \ge \rho + \beta$ leads to inequality $\alpha + \beta \le 0$ and taking into account $\alpha, \beta > 0$ is an empty set. It means that now we have only the second one $\rho - \alpha \le -\rho - \beta$, which is equivalent to inequality $\alpha - \beta \ge 2\rho$, which defines the region of the first quadrant, restricted by lines $\beta = 0$ and $\beta = \alpha - 2\rho$, what is shown as a hatched part on Figure 3.a.

In similar way the second two unity conditions are shown as hatched region of Figure 3.b., and Figure 3c shows all inequalities of (20) in $\mathbb{R}^2$ variant. The $\mathbb{R}^3$ variant is shown on Figure 4.



a) b) c)

Fig. 3. Graphical interpretation of the a) first pair of inequalities (20); second pair of inequalities (20); all inequalities (20) in $\mathbb{R}^2$

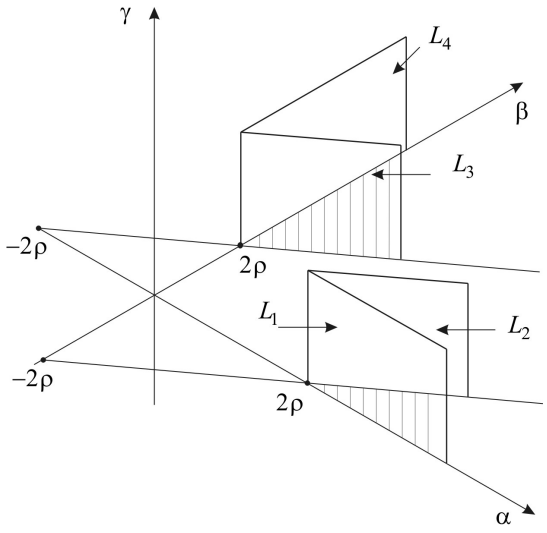


Fig 4. Graphical interpretation of inequalities (20) in $\mathbb{R}^3$

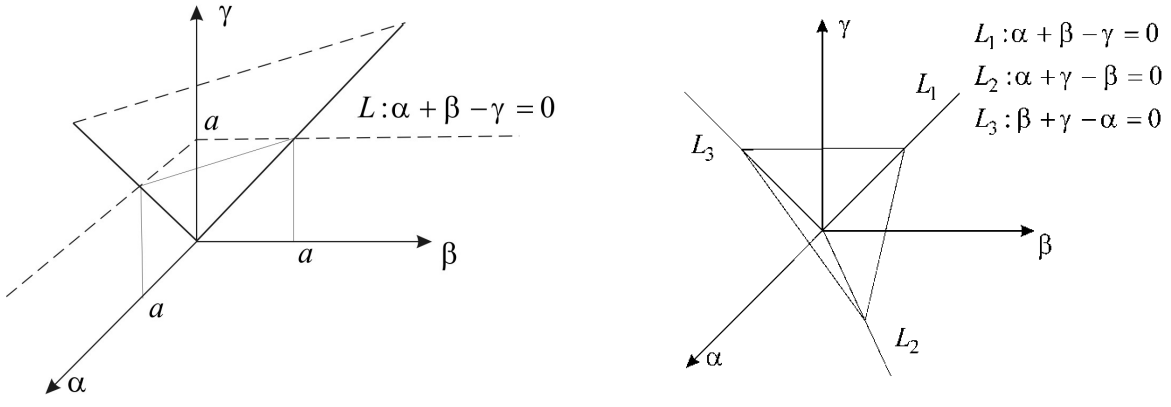
Here the region corresponding to (20), is restrained below by plane $\gamma = 0$ and side planes L_1, L_2, L_3, L_4, the inequalities of which have form

$$L_1 : \beta = 0,$$
$$L_2 : \alpha - \beta - 2\rho = 0,$$
$$L_3 : \alpha - \beta + 2\rho = 0,$$
$$L_4 : \alpha = 0.$$

Now consider the part of system (18) in form

$$\begin{cases} \alpha + \beta \geq \gamma \\ \alpha + \gamma \geq \beta \\ \gamma + \beta \geq \alpha \end{cases} \tag{21}$$

It is clear to see that in the first inequality its boundary is plane $\alpha + \beta - \gamma = 0$, and region satisfying condition $\alpha + \beta \geq \gamma$ is under plane L (Figure 5a). Naturally the same will correspond (regarding to other axes) to the two other inequalities of system (21). In whole the searched region can be geometrically represented (Figure 5b) as inner part of an infinite regular triangular pyramid.



a) b)

Fig. 5. Geometrical interpretation of the a) first inequality of system (21); b) inequalities system (21)

Finally consider the last three inequalities of system (18)

$$\begin{cases} \rho \le \Delta; \\ \rho + \alpha \le \Delta; \\ \rho + \beta \le \Delta. \end{cases}$$

First of it is practically the condition which should be fulfilled for all parameters of system (18) (in this system α, β, γ are variables in regard to which the sufficiency region is constructed, and ρ and Δ are the system's parameters), so the sufficiency region would not be empty. And the other two inequalities from geometrical point of view define the inner part of an infinite parallelepiped refined by planes $\alpha = 0, \beta = 0, \gamma = 0, \alpha = \Delta - \rho, \beta = \Delta - \rho$ (Figure 6).

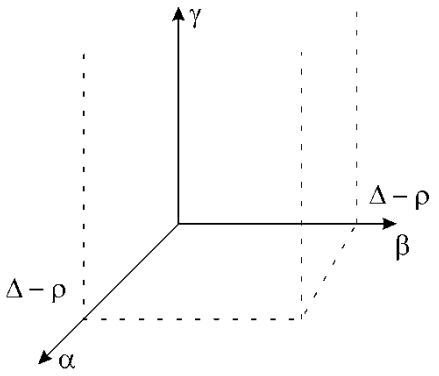


Fig. 6. Geometrical interpretation of threshold restrictions in (18)

Result of intersection of regions shown on Figures 5 and 6 gives us the sufficiency region. At that in order of the sufficiency region not to be empty inequality $\Delta - \rho > 2\rho$ or $\Delta > 3\rho$ should be fulfilled. It finally means that parameters of system (18) ρ and Δ should satisfy more strict condition than $\rho \le \Delta$, namely $\rho < \Delta/3$.

Sufficiency region construction on *n* pivot points

We shall consider the most general situation when the procedure of defining the closest configuration point is carried out by using n pivot points. Under our approach it means that configuration K consists of $n+2$ points, as to the pivots 2 compared points are added, and taking into account the current point y it can be represented as shown on Figure 7, notations of which will be explained below.

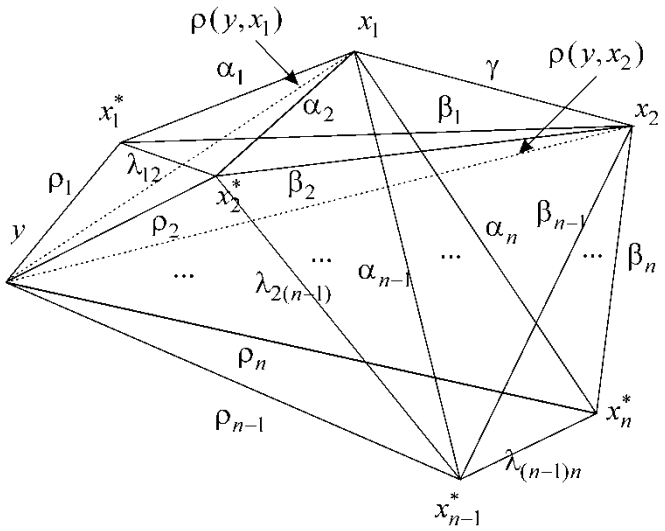


Fig. 7. Geometrical interpretation of search conditions with *n* pivot points

The notations are based on the following: 1) compared distances are denoted as $\rho(y, x_1)$ and $\rho(y, x_2)$; 2) $\rho(x_1, x_2) = \gamma$; 3) $\rho(x_i^*, x_1) = \alpha_i$, $\rho(x_i^*, x_2) = \beta_i$ $i = 1,...,n$ 4) $\rho(y, x_i^*) = \rho_i$; 5) $\rho(x_i^*, x_j^*) = \lambda_{ij}, i \neq j, \lambda_{ij} = \lambda_{ji}$. Here the distance matrix $\mathcal{P}$ will have the order $(n+2)\times(n+2)$, as configuration K contains $n+2$ points and has a form

$$\mathcal{P} = \begin{pmatrix} 0 & \gamma & \alpha_1 & \dots & \alpha_i & \dots & \alpha_n \\ \gamma & 0 & \beta_1 & \dots & \beta_i & \dots & \beta_n \\ \alpha_1 & \beta_1 & 0 & \dots & \lambda_{1i} & \dots & \lambda_{1n} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \alpha_i & \beta_i & \lambda_{i1} & \dots & 0 & \dots & \lambda_{in} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \alpha_n & \beta_n & \lambda_{n1} & \dots & \lambda_{ni} & \dots & 0 \end{pmatrix} \begin{matrix} x_1 \\ x_2 \\ x_1^* \\ \dots \\ x_i^* \\ \dots \\ x_n^* \end{matrix} \quad . \qquad (22)$$
$$\begin{matrix} x_1 & x_2 & x_1^* & \dots & x_i^* & \dots & x_n^* \end{matrix}$$

From here it follows that matrix (22) contains $(n+2)^2 - (n+2)/2$ or $r = \dfrac{(n+2)(n+1)}{2}$ parameters, what is $\psi(x_1, x_2) \in \mathbb{R}^r$, where we should construct the sufficiency region.

As in the previous situation, the main part of parameters restrictions which provide the sufficiency region constructing are those which contain unknown compared distances $\rho(y, x_1)$ and $\rho(y, x_2)$, and if in case with

one pivot point there were two of them (12), then for n pivots their number will be $2n$ and they will take form

$$\begin{cases} |\rho_1 - \alpha_1| \le \rho(y, x_1) \le \rho_1 + \alpha_1; \\ |\rho_1 - \beta_1| \le \rho(y, x_2) \le \rho_1 + \beta_1; \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ |\rho_i - \alpha_i| \le \rho(y, x_1) \le \rho_i + \alpha_i; \\ |\rho_i - \beta_i| \le \rho(y, x_2) \le \rho_i + \beta_i; \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ |\rho_n - \alpha_n| \le \rho(y, x_1) \le \rho_n + \alpha_n; \\ |\rho_n - \beta_n| \le \rho(y, x_2) \le \rho_n + \beta_n. \end{cases}$$

Having grouped them in a different way, namely

$$\left\{ |\rho_i - \alpha_i| \le \rho(y, x_1) \le \rho_i + \alpha_i \right., \tag{23}$$

where $i = \overline{1,n}$ and

$$\left\{ |\rho_i - \beta_i| \le \rho(y, x_2) \le \rho_i + \beta_i \right. \tag{24}$$

Where $i = \overline{1,n}$, then we can make a choice of the closest point among x_1 and x_2 to the current point y and define the main part of the sufficiency region if at least one of the inequality system (3) lower bounds exceeds at least one of the inequality system (4) upper bounds and vice versa. To be more precise it means that the sufficiency region has a projection in coordinate subspace of parameters $\{\alpha_1,\ldots,\alpha_n,\beta_1,\ldots,\beta_n\}$ of the main space $\mathbb{R}^r$, yet without considering other restrictions. This projection is defined by union of two inequalities systems: the first one define x_1 as the closest point and has a form

$$\left\{ \rho_i + \alpha_i \le |\rho_j - \beta_j| \right. \tag{25}$$

where the union is fulfilled for all parameters with repetition from the index set $\{1,2,\ldots,n\}$, т.е. $i, j = \overline{1,n}$. The number of all these pairs is equal to n^2.

Similar system in form

$$\left\{ \rho_j + \beta_j \le |\rho_i - \alpha_i| \right.$$

provides the choice of x_2 as the closest one. In the result the sufficiency region projection (without the restrictions which shall be defined later in the coordinate space $\{\alpha_1,\ldots,\alpha_n,\beta_1,\ldots,\beta_n\}$) is a union of $2n^2$ inequalities which have a form

$$\begin{cases} \rho_i + \alpha_i \le |\rho_j - \beta_j|; \\ \rho_j + \beta_j \le |\rho_i - \alpha_i| \end{cases} \tag{26}$$

where $i, j = \overline{1,n}$.

Let us now consider other restrictions which arise when using n pivot points. First of all it is related to the triangulars which describe one or two pivot points and one or two of the compared ones. Formally there is a set of pivot points $\{x_1^*,\ldots,x_n^*\}$ and set of compared ones $\{x_1, x_2\}$, then it is clear that number of the triangulars with

vertexes from these two sets can be found as next: if there is one pivot point, their number is n, namely $\Delta(x_i^*, x_1, x_2)$, where $i = \overline{1,n}$; if there are two pivot points, then the number is $2C_n^2$, as from the set $\{x_1, x_2\}$ we can choose one point by two means, and from set $\{x_1^*,...,x_n^*\}$ choosing two points is by C_n^2 choosing means. Thus we get triangulars of type $\Delta(x_i^*, x_j^*, x_k)$, where $i \neq j$, $i, j = \overline{1,n}$ and $k = 1,2.$

From the triangular inequality to the first type of triangulars $\Delta(x_i^*, x_1, x_2)$ correspond n restrictions of type

$$\begin{cases} \alpha_i + \beta_i \geq \gamma; \\ \alpha_i + \gamma \geq \beta_i; \\ \gamma + \beta_i \geq \alpha_i, \end{cases} \tag{27}$$

($i = \overline{1,n}$), to the second type of triangulars $\Delta(x_i^*, x_j^*, x_k)$ correspond C_n^2 of restrictions of type

$$\begin{cases} \alpha_i + \alpha_j \geq \lambda_{ij}; \\ \alpha_i + \lambda_{ij} \geq \alpha_j; \\ \lambda_{ij} + \alpha_j \geq \alpha_i, \end{cases} \tag{28}$$

when $k = 1$, i.e. $\Delta(x_i^*, x_j^*, x_1)$, and when $k = 2$, i.e. $\Delta(x_i^*, x_j^*, x_2)$ restrictions will be as following

$$\begin{cases} \beta_i + \beta_j \geq \lambda_{ij}; \\ \beta_i + \lambda_{ij} \geq \beta_j; \\ \lambda_{ij} + \beta_j \geq \beta_i, \end{cases} \tag{29}$$

at that $i, j = \overline{1,n}$ and $i \neq j$ for systems (27) and (28).

It should be mentioned that starting with three pivot points additional restrictions arise. They are related to the triangulars which vertexes contain only pivot points (in case that we have at least three pivot points). The number of such triangulars is C_n^3 and in our notations they have a form $\Delta(x_i^*, x_j^*, x_l^*)$ and correspond to the inequalities systems

$$\begin{cases} \lambda_{ij} + \lambda_{il} \geq \lambda_{jl}; \\ \lambda_{ij} + \lambda_{jl} \geq \lambda_{il}; \\ \lambda_{jl} + \lambda_{il} \geq \lambda_{ij}, \end{cases} \tag{30}$$

where $i, j, l = \overline{1,n}$ и $i \neq j, i \neq l, l \neq j$.

Finally the last type of restrictions is connected with the triabgular inequality when one of the triangular's vertexes is the current point y, and two others are pivots x_i, x_j, where $i \neq j$ and $i, j = \overline{1,n}$. These are triangulars of type $\Delta(y, x_i^*, x_j^*)$. The number of those is C_n^2, and they correspond to the inequalities system

$$\begin{cases} \rho_i + \rho_j \geq \lambda_{ij}; \\ \rho_i + \lambda_{ij} \geq \rho_j; \\ \lambda_{ij} + \rho_j \geq \rho_i, \end{cases} \tag{31}$$

where $i \neq j$ and $i, j = \overline{1,n}$.

But only one parameter λ_{ij} of the distance matrix takes part in these systems set, so it is expedient to substitute system (31) with the following one

$$\left\{\left|\rho_i - \rho_j\right| \leq \lambda_{ij} \leq \rho_i + \rho_j, \right. \tag{32}$$

where $i \neq j$ and $i, j = \overline{1,n}$.

Thus we can formulate a statement as for the type of the sufficiency region in the case of n pivots.

Proposition 4. If elements of distance matrix (22) удовлетворяют системам условий (26) – (30), (32), all are positive (situated in the first quadrant) and belong to Δ-neighborhood of query y, i.e. satisfy the restrictions

$$\begin{cases} \{\rho_i + \alpha_i < \Delta; \\ \{\rho_i + \beta_i < \Delta \end{cases} \tag{32}$$

where $0 < \rho_i < \Delta$, then it is enough to choose the closest point of configuration K to the query y without calculating the distance to it.

Conclusion

The sufficient conditions for the database search based on the pre-calculated distances between pivot points and database objects, and triangular inequality as the base for the objects elimination without calculating distances between them are obtained and grounded. There is still a problem of choosing the pivot points and their quantity which was not a part of the current exploration.

Acknowledgements

*The paper is published with financial support by the project ITHEA XXI of the Institute of Information Theories and Applications FOI ITHEA (*www.ithea.org*) and the Association of Developers and Users of Intelligent Systems ADUIS (*www.aduis.com.ua*).*

Bibliography

[Barsi, 2011] R. Barsi, T. Hassner, L. Zelnik-Manor. Approximate Nearest Subspace Search. In: IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 33, No 2, 2011.

[Dohnal, 2003] V. Dohnal, C. Gennaro, P. Savino, P. Zezula. D-index: Distance Searching Index for Metric Data Sets. In: Multimedia Tools and Applications, Vol. 21, No 1, 2003.

[Geetha, 2008] P. Geetha, V. Narayanan. A Survey of Content-Based Video Retrieval. In: Journal of Computer Science, Vol.4, No 6, 2008.

[Kinoshenko, 2010] D. Kinoshenko, V. Mashtalir, V. Shlyakhov, E. Yegorova. Nested Partitions Properties for Spatial Content Image Retrieval. In: International Journal of Digital Library Systems, Vol. 1, No 3, 2010.

[Kinoshenko[1], 2010] D. Kinoshenko, V. Mashtalir, E. Yegorova. Block-Diagonal Forms of Distance Matrices for Partition Based Image Retrieval. In: Pattern Recognition, Recent Advances, A. Herout (ed). Vucovar: In-Thech, 2010.

[Liu, 2007] Y.Liu, D. Zhanga, G. Lua, W.-Y. Ma. A Survey of Content-Based Image Retrieval with High-Level Semantics. In:

Pattern Recognition. Vol. 40, No. 1, 2007.

[Natsev, 2006] A. (P). Natsev. Multimodal Search for Effective Video Retrieval. In: Image and Video Retrieval. Berlin-Heidelberg: Springer-Verlag, Lecture Notes in Computer Science, Vol. 4071, 2006.

[Snoek, 2008] C.G.M. Snoek, M. Worring. Concept-Based Video Retrieval. In: Foundations and Trends in Information Retrieval, Vol. 2, No. 4. 2008.

[Zezula, 2006] A.P. Zezula, G. Amato, V. Dohnal, M. Batko. Similarity Search. The Metric Space Approach. In: Advances in: Database Systems, Vol.32. N.Y.: Springer Science+Business Media, Inc, 2006.

Authors' Information

Vladimir Mashtalir *– Professor, Dean of Computer Science Faculty, Kharkov National University of Radio Electronics, Lenin ave 14, 61166, Kharkov, Ukraine, email: mashtalir@kture.kharkov.ua*

Major Fields of Scientific Research: Pattern recognition, information retrieval

Konstantin Shcherbinin *– Ph.D. student, Kharkov National University of Radio Electronics, Lenin ave 14, 61166, Kharkov, Ukraine, email: shcherbinin@kture.kharkov.ua*

Major Fields of Scientific Research: Video segmentation and retrieval

Vladislav Shlyakhov *– Leading Researcher, Kharkov National University of Radio Electronics, Lenin ave 14, 61166, Kharkov, Ukraine.*

Major Fields of Scientific Research: Artificial intelligence, multialgebraic systems

Elena Yegorova *– Senior Researcher, Kharkov National University of Radio Electronics, Lenin ave 14, 61166, Kharkov, Ukraine, email: elena.yegorova@gmail.com*

Major Fields of Scientific Research: Image segmentation and retrieval

METHOD FOR EVALUATING OF DISCREPANCY BETWEEN REGULARITIES SYSTEMS IN DIFFERENT GROUPS

Oleg Senko, Anna Kuznetsova, Natalia Malygina, Irina Kostomarova

Abstract: *A new method of data analysis is discussed. Goal of represented techniques is complete and statistically valid comparing of regularities existing in two different groups of objects. It is supposed that regularities that tie levels of forecasted and explanatory variables are searched with the help of optimal partitioning technique. The developed technique was applied for analysis of genetic factors impact on severity of discirculatory encephalopathy (DEP). At the first stage computer method for evaluating of DEP severity was developed with the help of pattern recognition techniques. It was revealed that computer estimates of severity rather strongly correlate with DD variant of gene coding angiotensin-converting enzyme (ACE). Systems of regularities that ties various clinical or biochemical factors with computer estimates of severity were found with the help of optimal valid partitioning (OVP) method in groups of patients with three different variants of gene coding ACE. Statistically significant discrepancies were found between revealed regularities systems with the help of developed methods of comparing.*

Keywords: *Optimal partitioning, statistical validity, permutation test, regularities, explanatory variables effect,pattern recognition, discirculatory encephalopathy, genetic factors*

ACM Classification Keywords: *H.2.8 Database Applications - Data mining, G.3 Probability and Statistics - Nonparametric statistics, Probabilistic algorithms*

Introduction

Goal of this paper is development of method for statistically valid comparing of regularities that were found in two sets of objects. It is supposed that objects from both sets are described by the same variables. Such task arises in particular when we try to evaluate effect of gene's variant on biological processes related to severity of some disease. In present paper developed technique is used to evaluate influence polymorphous variants of genes coding angiotensin-converting enzyme (ACE), lipoprotein lipase (LPL) and cholesterol ester transfer protein (CETP) on biological processes related to severity of dyscirculatory encephalopathy [Vodolagina et al,2010]. Previously developed in works [Senko and Kuznetsova,1998], [Senko et al, 2003], [Senko and Kuznetsova, 2006] optimal valid partitioning (OVP) technique was used for regularities searching. OVP technique belongs to family of data analysis methods aimed to study regularities that may be described as subregions in explanatory variables space where levels of forecasted variable Y significantly differ from Y levels in neigbouring subregions . Method of logical regularities searching [Ryazanov,2003] or various models of classification or regression trees may be mentioned thereupon [Kuncheva, 2004].

OVP method. OVP technique implements partitioning of explanatory variables $X_1,\ldots,X_n$ space that allows to achieve possibly separation of forecasted variable Y levels. Forecasted variable Y in particular may be binary variable indicating two dichotomous groups. Optimal partitions are searched by initial dataset

$\tilde{S}_* = \{(y_1, \mathbf{x}_1), \ldots, (y_r, \mathbf{x}_r)\}$ where y_j is value of Y and $\mathbf{x}_j$ is vector of $X-$ variables for object s_j. Partition with the maximal value of special quality functional F_Q is searched inside several partitions families that differ from each other by dimensions, complexities of geometrical forms. Revealed regularities in OVP method are verified with the help of permutation tests [Gorman, 2001]. Permutation test implement verification by comparing maximal value of F_Q that was achieved at initial true dataset with maximal F_Q values that were calculated at variety of artificial datasets. At that artificial datasets are generated from initial dataset by random permutation of Y values relatively fixed position of $\mathbf{x}$ descriptions. Statistical validity of regularity (p-value) is estimated as fraction of random permutations for which maximal F_Q at artificial dataset exceeds maximal F_Q at initial dataset. Besides functional F_Q and p-value additional validity measure F_{2Q} is used that is ratio of maximal value of F_Q that was achieved at random dataset to optimal F_Q value at initial dataset. In case of more complicated regularities validity is evaluated with the help of additional second variant of permutation test. Instead of testing null hypothesis that Y is completely independent on $X-$ variables second variant implement testing of null hypotheses that Y is independent on $X-$ variables inside subregions of $X-$ space related to more simple regularities that were previously revealed for the same variables. Let r^{ij} is two-dimensional regularity with one boundary for each variable from pair (X_i, X_j). Let statistically valid simplest regularities exist for variables related to two-dimensional regularity r^{ij}. Then two p-values and two values of F_{2Q} may be calculated for r^{ij}: p_1 and F_{2Q}^1 are calculated from null hypothesis that is independent on (X_i, X_j) inside two subregions formed by partition of X_i range, p_2 and F_{2Q}^2 are calculated from null hypothesis that is independent on (X_i, X_j) inside two subregions formed by partition of X_j range. In case there is no simple valid regularity for one or for both variables from (X_i, X_j) values p_1, F_{2Q}^1 and/or p_2, F_{2Q}^2 are calculated from null hypothesis that Y is completely independent on (X_i, X_j). Using p-values and functionals p_1, F_{2Q}^1 and p_2, F_{2Q}^2 allows to evaluate contribution of each explanatory variable to two-dimensional regularity

OVP method was rather successfully used in several biomedical tasks [Kuznetsova et al, 2000]. However there is problem arising during OVP using that is related to large complexity of regularities systems in high-dimensional tasks where number of regularities may achieve several hundreds. Some additional mathematical tools are necessary to help researcher to evaluate regularity system. One of possible approaches that is associated with classification of regularities was suggested in [Senko et al, 2010]. Aanother method is suggested here that allows evaluating importance of each explanatory variable by calculating corresponding index. Importance of explanatory variables X is calculated as sum of contributions of X to various regularities. Let $\tilde{R} = \{r^{ij}\}$ is found regularities system. Index $\gamma(X_i)$ characterizing importance of X_i may be calculated as sum

$$\gamma(X_i, \tilde{R}) = \sum_{r_*^{ij} \in \tilde{R}} F_{2Q}^1(r_*^{ij}) + \sum_{r_*^{ji} \in \tilde{R}} F_{2Q}^2(r_*^{ji}).$$

Another problem that arises particularly in genetic researches is necessity of comparing of regularities that are found for the same explanatory variables in different groups of patients.

Evaluating of genetic factors effect in patients with discirculatory encephalopathy

Discirculatory encephalopathy is form of cerebral and vascular chronic insufficiency that is caused by worsening of blood supply in brain tissue [Maksudov, 1975].

Clinical data. Database developed at neurology department of "Scientific and clinical center (SCC) gerontology" was used in represented researches. This database includes information about variety of anamnestic, clinical, laboratory indicators in population of 358 patients with range in age from 35 to 102. Database also includes information about results of instrumental method (Magnetic Resonance Tomography, Ultrasonic Imaging, Electrocardiography and others) and results of molecular and genetic screening of genes coding ACE, LPL and CETP that was received in laboratory of age specific population genetics of "SCC gerontology". Levels of many indicators in group with age over 89 deviate significantly from mean levels in group under 89. So group over 89 was deleted from analysis. Population under 89 includes three groups of patients with different severity stages: 43 patients with first stage, 145 patients with second stage and 47 patients with third stage.

Computer method of severity evaluating. Influence of explanatory variables on severity may be estimated by using or standard statistical tests or OVP technique in groups with first or third severity stages. However full number of patients with first and third stages (90) is too small to evaluate effect of explanatory variables inside groups with fixed gene variant**.** At that number of patients with second intermediate severity stage significantly exceeds number of patients with first and third stages. It is possible to calculate computer estimates of severity by valid clinical or laboratory indicators for all patients using algorithm previously trained at dataset including first and third severity groups.

Performance of several pattern recognition methods from program system «RECOGNITION» was evaluated with the help of cross validation technique. «RECOGNITION» is program system for intellectual data analysis including collection of pattern recognition techniques that are based on different approaches: standard statistical methods, methods based on voting by systems of regularities, logical and combinatorial methods, neural networks, linear divisors, support vector machine, classifying trees and others. Program system also includes tools for calculating collective solutions. Goal of our researches was evaluating of relationship between estimates of severity and genetic factors. So only those explanatory variables were used that corresponds to anamnestic, clinical, laboratory indicators and results of instrumental methods. Results of molecular and genetic screening were not used as predicting factor at this stage of studies. It must be noted that one of explanatory variable indicating that patient underwent acute stroke strongly correlate with cerebral encephalopathy severity and its contribution to solving rules may significantly exceed contributions of another variables. So it is interesting to evaluate performance of pattern recognition methods without indicator of acute cerebral disorders. Correct recognition rates by full set of variables and by set without indicator of acute cerebral disorders (ACD) are given in table 1.

It is seen that forecasting ability of recognition algorithms without indicator of acute cerebral disorders is somewhat lower forecasting ability of whole set of variable. Algorithms trained at groups with first and third severity stages were then used for computer diagnostics of second group. Besides single algorithms collective rules were also studied that ascribe recognized object to a class where it is put by majority of algorithms. Calculated estimate for groups with first or third severity stages will be further referred to as calculated severity estimates. Calculated severity estimates were compared with genetic factors indicating variants of genes coding

ACE, LPL or CETP . Statistical validity of relationship between genetic factors and calculated severity estimates was evaluated with the help of standard statistical criteria Wilcoxon-Mann-Whitney (WMW), Chi-square(X^2).

Table1. Diagnostic abilites of different pattern recognition models by full set of explanatory variable and by set of explanatory variables without indicator of acute stroke evaluated with the help of cross validation technique.

Pattern recognition models	Full set 125 explanatory variables	Without indicator of acute stroke
Linear Fisher discriminant (LDF)	77,00%	74,50%
k-nearest neighbors (k-NN)	67,80%	74,30%
Statistically weighted syndromes (SWS)	85,50%	76,50%
Support vector machine (SVM)	83,3 %,	80,00%
Multiplicative neural network (MNN)	81,00%	67,70%
Method based on voting by Logical regularities (LR)	89,90%	77,80%

Besides for evaluating of relationship uni-dimensional OVP model was used. It was shown that for many algorithms calculated severity estimates rather strongly correlate with indicators of DD variant of ACE coding gene. At that very good correlation exists for ensemble of 4 algorithms: LDF, k-NN, SVM,SWS. Results of studies are given in Table 2.

Table 2. Statistical validity of relationship between calculated severity estimates and genetic factors

Recognition model	WMW		X^2		OVP	
	Full set	Without AS	Full set	Without AS	Full set	Without AS
LDF	0.007	0,01	0.07	0,05	0.08	0,04
SVM	0.024	0,02	0.017	0,01	0.05	0,01
SWS	0.06	0,02	0.07	No	0.07	0,05
Collective solution	0,014	0,05	0.014	0.0038	0.007	0,01

Distributions of calculated severity stages in group with DD variant and in group with alternative variants are given in table 3. It is seen that that in group with DD variant number of cases with calculated third stage of severity slightly exceeds number of cases with the calculated first stage of severity. At the same time number of cases with the first stage more than two times exceeds number of cases with the third stage.

Tabl. 3. Distributions of calculated severity stages in group with DD variant and in group with DI or II variants

	ID or II variants of gene coding ACE number of cases	DD variant of gene coding ACE number of cases
First severity stage	82	22
Third severity stage	39	28

So it was shown that computer estimates of DE severity in group of patients with variant DD of ACE significantly differ from DE severity in group of patients with variants ID and II. It is interesting to unveil causes of existing effect.

Regularities systems comparing

Possible approach is comparing of regularities that tie *DE* severity and corresponding levels of clinical, biochemical or genetic indicators in two groups of patients. Let $\tilde{R}_{dd}$ is set of valid regularities that were found for group of patients $\tilde{S}_{dd}$ with variants *DD*. Then effect of variant *DD* may be compared with effect of another variant zz by evaluating accordance between $\tilde{R}_{dd}$ and regularities characterizing dependence of *DE* severity on the same explanatory variables in group of patients $\tilde{S}_{zz}$ with variant zz. This problem may be also referred to as problem evaluating discrepancy between regularities in two groups of objects. Different approaches may be used. For example regularities system $\tilde{R}_{zz}$ may be found for group of patients $\tilde{S}_{zz}$. Then we may compare systems $\tilde{R}_{dd}$ and $\tilde{R}_{zz}$. Effect of variant may be described by selecting following regularities from $\tilde{R}_{dd}$:

a) in case regularity $r_{dd}^{ij} \in \tilde{R}_{dd}$ exists for pair of explanatory variables (X_i, X_j) and there no regularities for (X_i, X_j) in $\tilde{R}_{zz}$;

b) in case regularities $r_{dd}^{ij} \in \tilde{R}_{dd}$ and $r_{zz}^{ij} \in \tilde{R}_{zz}$ exist for pair of explanatory variables (X_i, X_j) but difference between regularities is sufficiently great .

Difference between regularities may be evaluated as difference between forecasting functions associated with regularities. Union of regularities sets satisfying conditions (a) and (b) will be referred to as $\tilde{R}_{dd\setminus zz}$

There is important drawback in described approach. In case (a), absence of regularity in $\tilde{R}_{zz}$ may be related not to quite different effect of zz variant relatively *DD* variant. Regularity for zz variant may be less expressed. Thus it is not revealed at corresponding significance level. So another approach was developed that allows describing difference between two types of effects more exactly.

It is supposed that regularity $r_{dd}^{ij} \in \tilde{R}_{dd}$ was found with the help of OVP technique by and r_{dd}^{ij} consist of k subregions $q_1, \ldots, q_k$. Let $n_i^1(r_{dd}^{ij}, \tilde{S}_*)$, $n_i^3(r_{dd}^{ij}, \tilde{S}_*)$ are numbers of cases with calculated first and third severity

stages in dataset $\tilde{S}_* \cap q_i$, $fr_i^1(r_{dd}^{ij}, \tilde{S}_*) = n_i^1(r_{dd}^{ij}, \tilde{S}_*)/[n_i^1(r_{dd}^{ij}, \tilde{S}_*) + n_i^3(r_{dd}^{ij}, \tilde{S}_*)]$. Set of ratios $\{fr_1^1(r_{dd}^{ij}, \tilde{S}_*), \ldots, fr_k^1(r_{dd}^{ij}, \tilde{S}_*)\}$ describes empirical dependence of DE on (X_i, X_j) in dataset $\tilde{S}_* \cap q_i$

Deviation between dependences may be evaluated with the help of functional

$$F_Q^{\Delta}(r_{dd}^{ij}, \tilde{S}_{dd}, \tilde{S}_{zz}) = \sum_{i=1}^{k} \{[fr_i^1(r_{dd}^{ij}, \tilde{S}_{dd}) - fr_i^1(r_{dd}^{ij}, \tilde{S}_{zz})]^2 \sqrt{n_i^1(r_{dd}^{ij}, \tilde{S}_{dd}) * n_i^1(r_{dd}^{ij}, \tilde{S}_{zz})}\}$$

So optimal partitions are searched by $\tilde{S}_{dd}$ and functional $F_Q^{\Delta}(r_{dd}^{ij}, \tilde{S}_{dd}, \tilde{S}_{zz})$ is calculated by $\tilde{S}_{dd}$ and $\tilde{S}_{zz}$ Important issue is validation of differences evaluated with the help of $F_Q^{\Delta}(r_{dd}^{ij}, \tilde{S}_{dd}, \tilde{S}_{zz})$. The same variants of permutation tests that were used in previous main version of OVP technique may be also used for comparing of two set of regularities. Pairs of artificial datasets $(\tilde{S}_{dd}^r, \tilde{S}_{zz}^r)$ are generated from $\tilde{S}_{dd}$ and $\tilde{S}_{zz}$ by random permutations of Y values relatively fixed position of $\mathbf{x}$ descriptions. Then again optimal partitions are found by $\tilde{S}_{dd}^r$ and functional F_Q^{Δ} is calculated by $(\tilde{S}_{dd}^r, \tilde{S}_{zz}^r)$. The same types of permutation test that was used in previous OVP technique are used in new developed method for evaluating discrepancy between regularities in two groups. Values of functional F_Q^{Δ} calculated by $(\tilde{S}_{dd}^r, \tilde{S}_{zz}^r)$ are compared with functional F_Q^{Δ} value for initial pair $(\tilde{S}_{dd}, \tilde{S}_{zz})$ and p-values are evaluated as fractions of $(\tilde{S}_{dd}^r, \tilde{S}_{zz}^r)$ pairs, for which F_Q^{Δ} value exceeds F_Q^{Δ} value for pair $(\tilde{S}_{dd}, \tilde{S}_{zz})$. Functional F_{2Q} and p_1, F_{2Q}^1, p_2, F_{2Q}^2 values are calculated according the same scheme that was used in previous variant of OVP technique. System of regularities that is received using validation with the help of functional F_Q^{Δ} and characterizes deviation between dependences existing in groups $\tilde{S}_{dd}$ and $\tilde{S}_{zz}$ will be referred to as $\tilde{R}_{dd\backslash zz}^{\Delta}$. Table 4 represents discrepancy between effect of gender on calculated severity estimated in groups with DD and II variants of ACE.

Table 4. Discrepancy between effect of gender on calculated severity estimated in $\tilde{S}_{dd}$ and $\tilde{S}_{ii}$

	DD		*II*	
	First stage	*Third stage*	*First stage*	*Third stage*
Female	*23*	*13*	*18*	*5*
Male	*3*	*12*	*5*	*0*

It is seen from table that there are 5 cases with calculated first stage of severity and no cases with calculated third stage in males with II variant of ACE. In males with DD variant of ACE there are 3 cases with first calculated severity stage and 12 cases with third calculated severity stage. So, rate of calculated third stage increases dramatically for DD variant. Statistical validity of difference between dependences of calculated severity stage on gender in group with DD and II variant of ACE was evaluated at $p<0.01$ with the help of permutation test using functional F_Q^{Δ}.

Figure 1 represents discrepancy between effect of cholesterol and thyrocsyn containments on calculated severity in group $\tilde{S}_{dd}$ with DD variant of gene coding ACE and in group $\tilde{S}_{id}$ with ID variant of gene coding *ACE*. At the left side of figure regularity is represented that ties calculated *DE* severity and two abovementioned explanatory

variables in group $\tilde{S}_{dd}$ and at the right side $\tilde{S}_{id}$ group empirical distribution is represented for the same pair of explanatory variables. It is seen that quadrant II at left part of figure contains 4 cases with calculated third severity stage and the same quadrant II at right part of figure contains 10 cases with calculated first severity stage. Statistical validity of difference between distributions represented al left and right parts of figure was evaluated at $p<0.01$ with the help of permutation test using functional F_Q^{Δ} .

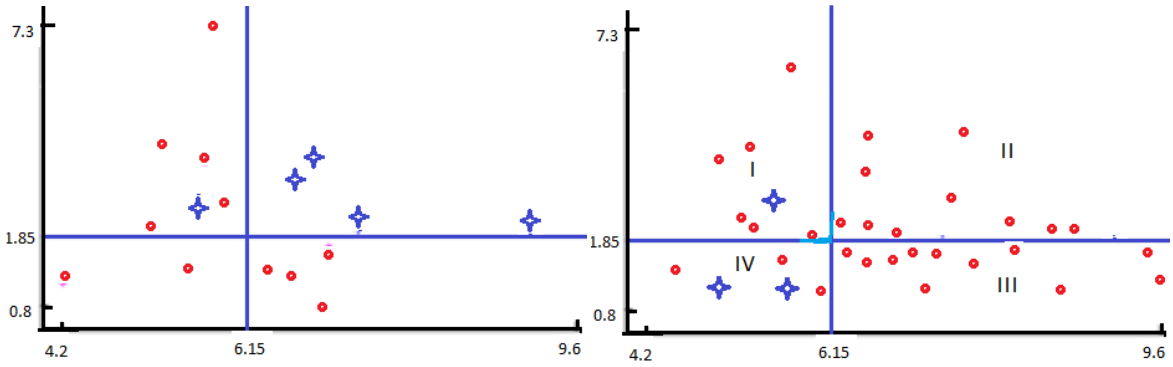


Fig 1. At the left part of figure regularity for Axis X corresponds containment of cholesterol in blood (mmmol/l), Y-containment of T3 (mmol/l)

- case with calculated third stage of severity,

- case with calculated first stage of severity.

Index γ characterizing importance of explanatory variables also may be used in method for evaluating discrepancy between regularities in two groups. It may be calculated as $\gamma(X_i,\ \tilde{R}^{*}_{dd\backslash zz}) = \sum_{r_*^{ij}\in\tilde{R}^{*}_{dd\backslash zz}} F^{1}_{2Q}(r_*^{ij}) + \sum_{r_*^{ji}\in\tilde{R}^{*}_{dd\backslash zz}} F^{2}_{2Q}(r_*^{ji})$, where $\tilde{R}^{*}_{dd\backslash zz}$ is equal $\tilde{R}^{\Delta}_{dd\backslash zz}$ or $\tilde{R}_{dd\backslash zz}$

Let by $\tilde{R}_{dd}$ and $\tilde{R}_{id}$ systems of two-dimensional regularities that were found in group $\tilde{S}_{dd}$ and in group $\tilde{S}_{id}$ with correspondingly. System $\tilde{R}^{\Delta}_{dd\backslash id}$ includes two-dimensional regularities characterizing deviation between dependences existing in groups $\tilde{S}_{dd}$ and $\tilde{S}_{id}$. Ten explanatory variables with greatest values of γ indices in discussed systems of regularities are represented in table 5.

It is seen that γ indices differently evaluate significance of explanatory variables for regularities from system $\tilde{R}_{dd}$ and $\tilde{R}_{id}$. In group $\tilde{S}_{id}$ the most informative variable is indicator of acute stroke. The γ index for acute stroke is more than 1.5 times than γ index for $\beta-$lipoproteins levels that is the most important variable in $\tilde{R}_{id}$ system besides indicator of acute stroke. In group $\tilde{S}_{dd}$ the most informative variable is indicator of thyroid ganglions that are detected by ultrasonic scanning. The γ index for this variable is more than 1.5 times greater than γ index for acute stroke. The only variables that belong to ten most important in $\tilde{R}_{dd}$ and $\tilde{R}_{id}$ are indicator of acute stroke and indicator heart conduction abnormality detected by ECG. It is seen that discrepancy between $\tilde{R}_{dd}$ and $\tilde{R}_{dd/id}$ is even greater. The only variable that is important in both sets of regularities is indicator of thyroid ganglions that are detected by ultrasonic scanning.

Table 5. Ten explanatory variables with greatest values of γ indices calculated are represented in table 5.

$\tilde{R}_{dd}$		$\tilde{R}_{id}$		$\tilde{R}^{\Delta}_{dd \setminus id}$	
Variable	γ indices	*Variable*	γ indices	*Variable*	γ indices
Thyroid ganglions by US scanning	13,175	Acute stroke	24,7728	Tortuous vessels by US diagnostics	15,1705
Alcohol	9,0115	β – lipoproteins	16,3555	Fibrosis gradations	7,9377
Acute stroke	8,2958	Abnormality of conduction by ECG	15,4514	Thyroid ganglions by US scanning	7,5158
Abnormality of conduction by ECG	7,5347	Hepatosis	15,2711	Alanin-amin-transferaze gradations	7,1757
Glucose	6,9559	Obesity stages	14,6469	Ultrasonic scanning-Sz	6,343
Hydrocephaly by MRT	6,8861	High density lipoproteins	13,2375	Br	5,7808
Nodular goiter	6,6777	H-H- variant of LPL gene	11,3931	Ultrasonic scanning-EC	5,6907
Red(blood)cells	6,4063	Type II diabetes	10,1628	Very low density lipoproteins	5,6695
Pancreas fibrosis	6,2065	Systolic pressure	10,1422	Free cholesterol gradations	4,9893
Prostate cancer	6,1434	Age gradations	9,8108	Ultrasonic scanning-atherosclerosis	4,6961

Conclusion

Results that are represented in paper may be summarized as follows. Method for computer evaluating of severity in patients with discirculatory encephalopathy was developed. It is appeared that severity estimates calculated with the help of computer recognition methods by variety of clinical, biochemical or instrumental indicators rather strongly correlates with indicator of DD variant of gene coding for angiotensin-converting enzyme. At that severity of DE is higher in patients with DD variant. To study the effect of ACE coding gene variants it was suggested to compare dependence of DE severity on variety of factors in groups with different variants of ACE gene. Method for evaluating of discrepancy between dependencies existing in two independent groups. At the first step optimal partitioning is searched by one of the groups. At the second step $F^{\Delta}_{Q}(r^{ij}_{dd}, \tilde{S}_{dd}, \tilde{S}_{zz})$ functional is calculated by regularity and two compared groups. Validity of discrepancy found at these two steps is evaluated with the help of special variant of permutation test using numerous recalculation of both steps for pairs of artificial random datasets. New method was developed that allows to calculate coefficients characterizing importance of single explanatory variables. So each system of regularities may be evaluated by vector of such coefficients. It was shown that there is statistically significant deviation between dependences of DE on variety of clinical and biochemical indicators in groups with DD and ID variant of gene coding ACE.

Acknowledgements

The paper is published with financial support by the project ITHEA XXI of the Institute of Information Theories and Applications FOI ITHEA (www.ithea.org) and the Association of Developers and Users of Intelligent Systems ADUIS Ukraine (www.aduis.com.ua).

Bibliography

[V.V.Ryazanov ,2003] Ryazanov V.V. About some approach for automatic knowledge extraction from precendent data // Proceedings of the 7th international conference "Pattern recognition and image processing", Minsk, May 21-23, 2003, vol. 2, pp.35-40.

[Gorman, 2001] T.W. O'Gorman An adaptive permutation test procedure for several common test of significance. Computational Statistics & Data Analysis. 35(2001) 265-281.

[Kuncheva, 2004] L.I. Kuncheva, Combining Pattern Classifiers. Methods and Algorithms. Wiley Interscience, New Jersey, 2004.

[Senko et al, 2003] Senko O.V., Kuznetsova A.V., Kropotov D.A. (2003). The Methods of Dependencies Description with the Help of Optimal Multistage Partitioning. Proceedings of the 18-th International Workshop on Statistical Modelling Leuven, Belgium, 2003, pp. 397-401.

[Sen'ko et al, 1998] Sen'ko O.V., Kuznetsova A.V. (1998). The use of partitions constructions for stochastic dependencies approximation. Proceedings of the International conference on systems and signals in intelligent technologies. Minsk (Belarus), pp. 291-297.

[Kuznetsova et al, 2000] Kuznetsova A.V., Sen'ko O.V., Matchak G.N., Vakhotsky V.V., Zabotina T.N., Korotkova O.V. The Prognosis of Survivance in Solid Tumor Patients Based on Optimal Partitions of Immunological Parameters Ranges //J. Theor. Med., 2000, Vol. 2, pp.317-327.

[Sen'ko et al, 2006] Oleg V.Senko and Anna V. Kuznetsova, The Optimal Valid Partitioning Procedures . Statistics on the Internet ***http://statjournals.net/***, April, 2006

[Senko et al, 2010] Senko O, Kuznetsova A., Malygina N, Kostomarova I.,2010 Methods for evaluating of regularities systems structure. In book «New Trends in Classification and Data Mining.» ITHEA. Sofia. Bulgaria. P. 40-46, 2010.

[Vodolagina et al,2010] Vodolagina N.N., Kostomarova I.V., Kuznetsova A.V., Malygina N.A. Senko O.V. Study of relationship between severity discirculatory encephalopathy and polymorphism of some genetic factors with the help of pattern recognition methods. Reports of 3-rd international conference «Mathematical biology and bioinformatics», Puschino,Russia, October 2010, p.238-239.

[Maksudov, 1975] Maksudov G.A. Discirculatory encephalopathy. In book «Vascular diseases of nervous system» under editorship of acad. AMS USSR E.V. Shmidt.

Authors' Information

Senko Oleg Valentinovich *– Leading researcher in Dorodnicyn Computer Center of Russian Academy of Sciences, Russia, 119991, Moscow, Vavilova, 40,* senkoov@mail.ru

Kuznetsova Anna *– senior researcher in Institute of Biochemical Physics of Russian Academy of Sciences, Russia, 117997, Moscow, Kosygina, 4,* azfor@narod.ru

Malygina Galina *–chief of laboratory in Russian Research Institute of Gerontology, Moscow, Russia*

Kostomarova Irina *- Russian Research Institute of Gerontology, Moscow, Russia*

EVALUATION OF GREEDY ALGORITHM OF CONSTRUCTING (0,1)-MATRICES WITH DIFFERENT ROWS[1]

Hasmik Sahakyan, Levon Aslanyan

Abstract: *An approximation greedy algorithm is considered for reconstruction of (0,1)-matrices with different rows. Numbers of pairs of different rows is taken up as a quantitative characteristic, maximization of which, when appropriate, leads to matrices with different rows. Properties of the algorithm are studied and the performance is evaluated based on series of experiments.*

Keywords: *(0,1)-matrices, greedy algorithms.*

ACM Classification Keywords: *F.2.2 Nonnumerical Algorithms and Problems: Computations on discrete structures.*

Introduction

$(0,1)$-matrices with prescribed row and column sums is a classical matter which appears in many branches of applied mathematics. There is a known result by Ryser for a pair of vectors being the row and column sums of a (0,1)-matrix ([R, 1957]). (0,1) matrices with given row and column sums and with special geometric properties/constraints are addressed for example in [DCh, 1999], [BDLNP, 1996], [W, 2001].

Consider the n dimensional unit cube. Vertices of the cube are coded by n-tuples of 0,1 values, and in this way any vertex subset has been presented as a $(0,1)$-matrix, where rows correspond to vertices and thus all rows are different. Row sums indicate the layers of the cube containing the corresponding vertices. i-th column sum identifies the number of vertices in the vertex subset with 1 value in i-th position. Let $R=(r_1,\cdots,r_m)$ and $S=(s_1,\cdots,s_n)$ denote the row and column sum vectors of a $(0,1)$-matrix of size $m\times n$. Now existence of such matrix is equivalent to the existence of m vertices situated in r_1-th, r_2-th, etc. r_m-th layers such that s_1 vertices/tuples contain 1 in the first position, s_2 vertices contain 1 in the second position, etc. and s_n vertices contain 1 in the n-th position. In other words s_i and $m-s_i$ are partition sizes of the vertex subset on i-th direction. In case when the particular positions of vertices are not important and we are just interested in existence of vertex subsets that have given partition sizes – we search out a subclass of matrices with column sums $S=(s_1,\cdots,s_n)$ and with m rows which are different. Both cases (with or without $R=(r_1,\cdots,r_m)$) are known as algorithmically open problems (no polynomial algorithm is known).

For the problem with $S=(s_1,\cdots,s_n)$ and m a greedy algorithm was proposed in [S, 2010 G], which is proven to be optimal in local steps. The current research intends to evaluate performance of this algorithm. Series of experiments made for this purpose, taking into account special properties of the algorithm. We propose that the given greedy algorithm performs a rather good partitioning of matrix columns. For $S=(s_1,\cdots,s_n)$ and m, supposing the existence of row different matrices, greedy algorithm leaves at most 2 length intervals.

$(0,1)$ matrices with different rows

Consider a $(0,1)$-matrix of size $m \times n$. Let $R = (r_1, \cdots, r_m)$ and $S = (s_1, \cdots, s_n)$ denote the row and column sum vectors of the matrix respectively, and let $U(R,S)$ be the class of all $(0,1)$-matrices with row sum R and column sum S. Clearly $0 \le s_i \le m$ for $1 \le i \le n$ and $0 \le r_i \le n$ for $1 \le i \le m$. A necessary and sufficient condition for existence of $(0,1)$ matrices in the class $U(R,S)$ was found by H. J. Ryser. In difference to this case we consider a specific subclass of $U(R,S)$ where all the rows of matrices are different, and in particular, we will consider the class $U(S)$ of all $(0,1)$-matrices with column sum $S = (s_1, \cdots, s_n)$ and with m rows that are all different.

Now we formulate two basic postulations related to the problem posted:

(P1) ***Clarify the Existence issue of*** $(0,1)$ ***matrices in*** $U(S)$,

(P2) ***Construct*** $(0,1)$ ***matrix or matrices in class*** $U(S)$.

We will address, mainly, the construction issues, **(P2)**. Consideration includes also the approximate case and in this context it is reasonable to introduce quantitative characteristics, so that optimized values of such characteristics lead to matrices with different rows in case when the latter exist. As such measure we consider "number of pairs of different rows" and its maximum – which was first considered in [S, 1995]. If $\Im(S)$ denotes the class of $(0,1)$-matrices of size $m \times n$ having column sum vector $S = (s_1, \cdots, s_n)$ (in this way $U(S) \subseteq \Im(S)$), then C_m^2 is the maximum possible number of the pairs of different rows for matrices of $\Im(S)$. This maximum is achieved for matrices of $U(S)$.

Therefore, if $U(S)$ is not empty, then a matrix $M \in \Im(S)$ with maximum number of pairs of different rows will solve the existence problem **(P1)**. In general, when the maximum value is less than C_m^2, then this is an indication that $U(S)$ is empty. Thus, further we will consider the modified postulation:

(P2') **Construct** $(0,1)$ **matrices with** m **rows and the given column sum, that have maximum number of pairs of different rows.**

The requirement of different rows implies a simple restriction on number of possible rows: $m \le 2^n$.

Below we bring some statements/properties from [S, 2009, S, 2010] that we take into account during the experiments.

Lemma 1 [S, 2010]. Suppose that $S = (s_1, \cdots, s_n)$ is the column sum vector for some $m \times n$- matrix with different rows. Then there exists a $(2^n - m) \times n$-matrix with different rows such that $(2^{n-1} - m + s_1, \cdots, 2^{n-1} - m + s_n)$ serves as its column sum vector.

This is a simple but useful property. It says: regardless the issue (existence, construction or characterization of column sum vectors), it is sufficient to consider the case $m \le 2^{n-1}$.

Lemma 2 [S, 2009]. Assume that $U(S)$ is not empty for a given $S=(s_1,\cdots,s_n)$ and suppose that $s_k > m/2$ for some k. Then $U(S')$ is not empty as well for $S'=(s_1,\cdots,s_k-1,\cdots,s_n)$.

Greedy algorithm for solving (P2')

The greedy approach is a recursive execution of a procedure that minimizes/maximizes the increase of the objective function.

Now we describe an algorithm G that constructs a matrix column by column: starting from the first one and adding a column in each step. The objective function is D: $A \in \Im(S) \to$ "number of pairs of differing rows of A"; and the goal is to construct a matrix with the greatest possible value of D, that is $A_{opt} \in \Im(S)$ such that $D(A_{opt}) = \max\limits_{A\in\Im(S)} D(A)$. Let A_G denote a matrix constructed by G and let $A_{G,k}$ is the matrix at the k-th step of G (in this way $A_{G,n} = A_G$). $\Delta D(A_{G,k})$ denote the increase of objective function during the k-th step of G, that is: $\Delta D(A_{G,k}) = D(A_{G,k}) - D(A_{G,k-1})$.

Algorithm *G*

Step1. First column consists of s_1 ones placed in the first s_1 rows-positions followed by $m-s_1$ zeros in others. Two intervals are the result: – the s_1-length interval of ones, and the $(m-s_1)$-length interval of zeros. We denote these intervals by $d_{1,1}^G$ and $d_{1,2}^G$. Hereafter the first sub-index will indicate the number of column and the second – the number of interval within the column. So construction of the first column is in unique way:

$$\begin{cases} d_{1,1}^G + d_{1,2}^G = m \\ d_{1,1}^G = s_1 \end{cases}$$

A pair of rows consists of identical when they belong to the same interval; otherwise it consists of differing rows. Therefore the first column will produce $d_{1,1}^G \cdot d_{1,2}^G$ pairs of different rows. So the increase of objective function is: $\Delta D(A_{G,1}) = d_{1,1}^G \cdot d_{1,2}^G$.

Let we have constructed the first $k-1$ columns. In general, $(k-1)$-th column consists of 2^{k-1} intervals filled by ones and zeros accordingly. Since among them 0-length intervals are possible and they cannot be used anymore, let us assume that the $(k-1)$-th column consists of p non-zero length intervals denoted by $d_{k-1,1}^G, d_{k-1,2}^G, \cdots, d_{k-1,p}^G$. Recall that the rows coincide within the intervals and differ otherwise. If in some column j we get all one length intervals, then at this moment non repetition of all rows, and hence the maximum number of pairs of different rows is already provided. Further constructions can be arbitrary.

Step k. During this step each $d_{k-1,i}^G$ length interval will be partitioned into $d_{k-1,i,0}^G$ and $d_{k-1,i,1}^G$ length intervals filled by zeros and ones respectively: $d_{k-1,i}^G = d_{k-1,i,0}^G + d_{k-1,i,1}^G$ such that

$\sum\limits_{i=1}^{p} d_{k-1,i,0}^G = m - s_k$ and $\sum\limits_{i=1}^{p} d_{k-1,i,1}^G = s_k$.

The increase of objective function during the k-th step is:

$$\Delta D(A_{G,k}) = \sum_{i=1}^{p} d^G_{k-1,i,1} \cdot d^G_{k-1,i,0} .$$

We will release partitions to minimize the length differences of intervals.

The idea is in following: if $s_k = m - s_k, k = 1,\cdots,n$, then in each step we would split every interval into 2 equal (± 1 in case of odd) parts and fill by zeros and ones respectively which will lead to all one length intervals in logarithmic number ([K, 1973]) steps. Furthermore, among all integer partitions of $d^G_{k-1,i}$: $d^G_{k-1,i} = d^G_{k-1,i,1} + d^G_{k-1,i,0}$, the largest product $d^G_{k-1,i,1} \cdot d^G_{k-1,i,0}$ is achieved when $d^G_{k-1,i,1} = d^G_{k-1,i,0}$. Thus following this strategy would bring to the goal, but in general at each step k we have $s_k - (m - s_k)$ extra ones (or $(m - s_k) - s_k$ extra zeros). Trying to be closer to equal lengths of intervals we 1) distribute the extra $s_k - (m - s_k)$ ones (or $(m - s_k) - s_k$ extra zeros) among intervals keeping a "homogeneous" distribution; and then 2) split the remaining intervals into 2 equal parts – putting equal number of zeros and ones.

The lengths of intervals/partitions and their parity will be the same independently of occurrence of extra ones or extra zeros. Thus without loss of generality we may assume that $s_i \geq m - s_i,\ i = 1,\cdots n$.

Now describe the process in detail.

Let $r_k = s_k - (m - s_k)$ and assume that there are l odd length intervals among the intervals of $(k-1)$-th column. It is easy to check that r_k and l have the same parity and hence $r_k - l$ is an even number. Construction of the k-th column is in 2 phases: distribution of r_k "extra" ones during the first, and distribution of remaining ones during the second phases.

1) phase

a) $r_k \leq l$

Chose arbitrary r_k intervals among the l odd intervals and put a 1 in each.

b) $r_k > l$

All l odd intervals get a 1. After this we put ones two by two in intervals starting from the intervals of even length then altering from odd to even, and continuing the process until all r_k ones have been exhausted. If during the process some short intervals have been filled, they do not participate any longer. Let T denote the maximum length of those intervals filled during this process. It is worth to mention that after putting l ones on odd intervals, we get all even lengths, and $r_k - l$ is even as well, so in this way the process of distribution is correct. After this phase there remain equal numbers of zeros and ones.

2) phase

a) $r_k \leq l$

Half of the remaining $l - r_k$ odd intervals get one 0, others – one 1, after that all intervals have been split into equal parts and receive equal number of zeros and ones.

b) $r_k > l$

all intervals have been split into equal parts and receive equal number of zeros and ones.

Let c_i indicates difference between the distributed ones and zeros on i-th interval: $c_i = d^{G_1}_{k-1,i,1} - d^{G_1}_{k-1,i,0},\ i = 1,\cdots,p$.

Thus on k-th column lengths of partitions are the following:

$d^{G_1}_{k-1,i,1} = \frac{d^{G_1}_{k-1,i} + c_i}{2},\quad d^{G_1}_{k-1,i,0} = \frac{d^{G_1}_{k-1,i} - c_i}{2}$ filled by 1 and 0 respectively for $i = 1,\cdots,p$. Each of the intervals may be of 0-length. It is worth to mention that: $\max\limits_{i,j} | c_i - c_j | \le 2$ for all i, j pairs of even $(\ge T)$-length and odd $(\ge T+1)$-length intervals.

Note. The need to choose some quantity of even (odd) intervals among all even (odd) intervals will cause branching during the first phase in case of $r_k > l$. This may bring not uniqueness in constructions.

Nevertheless all branches maximize the increase of the pairs of different rows in each local step.

If the n-th column consists of only 1-length intervals then algorithm G produces a matrix with different rows. Otherwise within intervals of greater lengths – rows are coinciding.

Theorem 1

(1) Each step of the algorithm G is optimal: it provides the maximum increase of the objective function – pairs of differing rows;

(2) All optimal constructions in each column are those according to G.

However algorithm G does not provide the global optimum.

Properties of G

In this section we study properties of G.

For a given vector $S = (s_1,\cdots,s_n)$ consider the class $U(S)$ of all $(0,1)$-matrices with column sum $S = (s_1,\cdots,s_n)$ and with m rows which are all different. Suppose that $U(S)$ is not empty and $A \in U(S)$. Interchanging ones with zeros in some column i of A will not produce repeating rows and hence it will lead to a matrix A^i that belongs to the class $U(S^i)$ where $S^i = (s_1,\cdots,m - s_i,\cdots,s_n)$. So A is easily transformable to A^i and vice versa. Thus having a goal the existence/construction issues in $U(S)$ we assume without loss of generality that all $s_i \ge m/2$. Concerning the order of components of S, - it is not important while considering the existence problem. As regards the construction process, it may have an influence. However we will study

properties of the algorithm assuming that $s_1 \geq \cdots \geq s_n$. Thus hereafter in this paper we will assume that $s_1 \geq \cdots \geq s_n$ and $s_i \geq m/2$ for $i = 1,\cdots,n$.

Below we will prove a property of algorithm G consisting in following:

Theorem 2. If for a vector $S = (s_1,\cdots,s_n)$ algorithm G constructs a matrix from $U(S)$, and $s_k > m/2$ for some k, then G constructs also a matrix from $U(S')$, where $S' = (s_1,\cdots,s_k - 1,\cdots,s_n)$ [1].

While Lemma 2 states that "good" vectors supporting existence of matrices in $U(S)$ are those having coordinate values close to $m/2$, the theorem 2 confirms that these are "good" also from the point of view of construction by G.

Before proving the theorem we bring some notes/properties concerning the constructions of algorithm G taking into account the order $s_1 \geq \cdots \geq s_n$:

Quantities of intervals increase from column to column - it is obvious.

Quantities of odd intervals do not decrease from column to column - indeed, each odd interval being split produces one odd interval and each even interval produces 0 or 2 odd intervals.

If during the construction of some column an interval (say length d) receives some number of extra 1s (say c) then all partitions of d in the consequent columns receive $\leq c$ extra 1s (independent of size of partition) - this follows from the constructions.

Now let us prove Theorem 2.

G constructs a matrix from $U(S)$ means that on the n-th column it receives all 1-length intervals. Now we prove that starting with $S' = (s_1,\cdots,s_k - 1,\cdots,s_n)$, G will construct all 1-length intervals as well.

The first $k-1$ columns constructed by G are the same for S and S'. Now consider constructions of the k-th columns by s_k and $s_k - 1$ respectively (denoted them by $G(s_k)$ and $G(s_k - 1)$) and convince that in the latter case it will bring to a matrix from $U(S')$. It is worth to mention that $s_k > s_{k+1}$.

Assume that there are l odd length intervals in $(k-1)$-th column and r_k is the difference between ones and zeros for the k-th column: $r_k = s_k - (m - s_k)$. r_k and l have the same parity and hence $r_k - l$ is even number. k-th column had been constructed in 2 phases: distribution of r_k "extra" ones during the first, and distribution of remaining ones during the second phases.

Now let r'_k denote the new difference: $r'_k = (s_k - 1) - (m - (s_k - 1)) = s_k - (m - s_k) - 2$ so $r'_k = r_k - 2$.

---

[1] *If equal components exist in S then choosing $s_k > m/2$ to decrease it by 1, we take the last such coordinate keeping the vector $S' = (s_1,\cdots,s_k - 1,\cdots,s_n)$ ordered decreasingly.*

We will follow constructions of k-th column for both r_k and $r_k - 2$ extra ones, and compare the lengths of intervals.

Consider possible cases:

a) $r_k \le l$ hence $r'_k \le l$

Constructions and reasoning are the same as for s_k:

Phase 1

Chose arbitrary r'_k intervals among the l odd intervals and put a 1 in each.

Phase 2

Half of the remaining $l - r'_k$ odd intervals get one 0, others get one 1, after that all intervals have been split into equal parts and receive equal number of zeros and ones.

Lengths had not been changed and hence constructions of remaining columns will be the same which leads to a matrix from $U(S')$.

b) $r_k > l$

Since $r_k - l$ is even number, then $r_k - l \ge 2$. Consider cases:

1. $r'_k \le l$

It follows that $r'_k = l$. This means that allocating 1 to each odd interval, all r'_k extra ones are exhausted in $G(s_k - 1)$:

Phase 1

All l odd intervals get a 1 value.

Phase 2

All intervals have been split into equal parts and receive equal number of zeros and ones.

Meanwhile in $G(s_k)$ two extra 1s remained for distributing among even intervals.

We denote by d the length of an even interval which received two extra 1s in $G(s_k)$. Since other lengths are the same in both $G(s_k - 1)$ and $G(s_k)$, we will follow only partitions of d.

$$G(s_k): \quad d \;\begin{cases} (d+2)/2 = d/2 + 1 \\ (d-2)/2 = d/2 - 1 \end{cases}$$

$$G(s_k - 1): d \;\begin{cases} d/2 \\ d/2 \end{cases} \qquad (1)$$

Consider cases:

- $d/2$ **is odd**

Then $d/2+1$ and $d/2-1$ are even.

Since the order of coordinates in S is decreasing and $s_k > s_{k+1}$, then in both cases these intervals do not receive extra ones. In $(k+1)$-th column we get the following partitions coming from d :

$G(s_k)$: $(d/2+1)/2$, $(d/2+1)/2$, $(d/2-1)/2$, $(d/2-1)/2$, and

$G(s_k-1)$: $(d/2+1)/2$, $(d/2-1)/2$, $(d/2+1)/2$, $(d/2-1)/2$.

So – they bring to the same lengths.

- $d/2$ **is even**

$d/2+1$ and $d/2-1$ are odd.

In $(k+1)$-th column we will get:

$G(s_k)$: $(d/2+1+1)/2$, $(d/2+1-1)/2$, $(d/2-1+1)/2$, $(d/2-1-1)/2$, and

$G(s_k-1)$: $d/4$, $d/4$, $d/4$, $d/4$. Thus:

$$G(s_k): \; d/2 \;\begin{cases} \nearrow\, d/4+1 \\ \searrow\, d/4-1 \end{cases} \qquad G(s_k-1): \; d/2 \;\begin{cases} \nearrow\, d/4 \\ \searrow\, d/4 \end{cases} \tag{2}$$

Depending on parity of $d/4$, the process will be continued altering within the above cases. It stops in some column $(k+t)$ where $d/(2^t)$ is the first odd length and in this way the interval lengths for $G(s_k)$ and $G(s_k-1)$ are the same. Otherwise $d=2^q$ for some q and process stops in $(k+q)$ column where all lengths are 1. It will happen before the column n, since by assumption G constructs a matrix from $U(S)$.

2. $r'_k > l$

Phase 1

All l odd intervals get a 1. After this we put ones two by two in intervals starting from the intervals of even length then altering from odd to even, and continuing cyclically until all r_k ones have been exhausted.

Phase 2

all intervals have been split into equal parts and receive equal number of zeros and ones.

Suppose that phase 1 stopped on odd intervals:

$G(s_k)$: t of odd intervals received c extra ones (c is odd), others received $c-2$ or filled completely, even lengths either received $c-1$ or are filled;

$G(s_k-1)$: $t-1$ odd intervals received c extra ones, others received $c-2$ extra ones or filled completely, even lengths either received $c-1$ or filled.

We denote by d the length of an odd interval which received extra two 1s in $G(s_k)$. Since other lengths are the same in both $G(s_k-1)$ and $G(s_k)$, we will follow only partitions of d.

$$G(s_k): \quad d \;\begin{cases} (d+c)/2 \\ (d-c)/2 \end{cases}$$

$$G(s_k-1): d \;\begin{cases} (d+c-2)/2 = (d+c)/2-1 \\ (d-(c-2))/2 = (d-c)/2+1 \end{cases}$$

Phase 1 stopped on even intervals:

$G(s_k)$: q of even intervals received c (c is even) extra ones, others received $c-2$ or filled completely, odd lengths either received $c-1$ or filled;

$G(s_k-1)$: $q-1$ even intervals received c extra ones, others received $c-2$ extra ones or filled completely and odd lengths received $c-1$ or filled.

We denote by d the length of an odd interval which received the extra two 1s in $G(s_k)$. Other lengths are the same. The picture is as in previous case.

For easy construction we bring the details of the following particular case. In general, the reasoning is in an analogous way.

Suppose that r'_k is just the number of odd intervals + 2 times the number of even intervals.

Phase 1

All odd intervals get a 1 and all even intervals get two 1s.

Phase 2

All intervals have been split into equal parts and receive equal number of zeros and ones.

Then one interval receives 3 extra 1s in $G(s_k)$. We denote by d the length of this interval. Again other lengths are the same and we will follow only partitions of d.

$$G(s_k): \quad d \;\begin{cases} (d+3)/2 = (d+1)/2+1 \\ (d-3)/2 = (d-1)/2-1 \end{cases}$$

$$G(s_k-1): d \;\begin{cases} (d+1)/2 \\ (d-1)/2 \end{cases}$$

Consider cases:

$(d+1)/2$ is odd and $(d-1)/2$ is even.

Then $(d+1)/2+1$ is even and $(d-1)/2-1$ is odd.

Since $s_k > s_{k+1}$, then the number of extra ones decreased by two and taking into account that the number of odd intervals did not decrease, we make the following conclusion: in both cases all odd intervals get at most one extra 1 in both cases. Even intervals will get either 2 or 0 extra ones (we will assume that the same intervals in both cases get the same amount of extra ones).

First consider case of 0 extra ones on even intervals. So in $(k+1)$-th column we get the following partitions coming from d :

$G(s_k)$: $((d+1)/2+1)/2$, $((d+1)/2+1)/2$, $((d-1)/2-1+1)/2$, $((d-1)/2-1-1)/2$

$G(s_k-1)$: $((d+1)/2+1)/2$, $((d+1)/2-1)/2$, $(d-1)/4$, $(d-1)/4$.

So the lengths are:

$G(s_k)$: $(d-1)/4+1$, $(d-1)/4+1$, $(d-1)/4$, $(d-1)/4-1$

$G(s_k-1)$: $(d-1)/4+1$, $(d-1)/4$, $(d-1)/4$, $(d-1)/4$.

Thus all intervals of $(k+1)$-th column are the same in both $G(s_k)$ and $G(s_k-1)$ except the following lengths: $(d-1)/4+1$ and $(d-1)/4-1$ in $G(s_k)$; $(d-1)/4$ and $(d-1)/4$ in $G(s_k-1)$.

The continuations are as in (2).

Now consider the case of 2 extra ones on even intervals. Partitions in $(k+1)$-th column coming from d are the following:

$G(s_k)$: $((d+1)/2+1+2)/2$, $((d+1)/2+1-2)/2$, $((d-1)/2-1+1)/2$, $((d-1)/2-1-1)/2$

$G(s_k-1)$: $((d+1)/2+1)/2$, $((d+1)/2-1)/2$, $((d-1)/2+2)/2$, $((d-1)/2-2)/2$

So all intervals in $(k+1)$-th column are of the same length except the following:

$G(s_k)$: $(d-1)/4+2$ and $(d-1)/4$

$G(s_k-1)$: $(d-1)/4+1$ and $(d-1)/4+1$.

Continuations are in analogous way.

$(d+1)/2$ is even and $(d-1)/2$ is odd.

Then $(d+1)/2+1$ is odd and $(d-1)/2-1$ is even.

This is similar to previous cases.

Evaluation of Results – Experiments

Before experimental treatment of algorithm G we refer to a special subset of $U(S)$, - class of matrices with column sums obeying the property of being a "boundary vector".

Definition [S, 2009]. Let ψ_m denote the set of all vectors $S = (s_1, \cdots, s_n)$ with $0 \le s_i \le m$, such that $U(S)$ is not empty. Vector $\widehat{S} \in \psi_m$ is called an (upper) boundary vector for ψ_m if no vector greater than $\widehat{S}$ belongs to ψ_m.

Further the boundary vectors are found among the vectors that correspond to monotone Boolean functions.

It follows from the Theorem 2 that more "difficult" vectors for the algorithm G are boundary ones. Thus for estimation of the performance of G it is reasonable to address this set of vectors.

All monotone Boolean functions are composed for $n = 6$ with $m = 1, \cdots, 2^{n-1}$ one values. For example for $m = 28$ there are 390050 functions. Corresponding column sum vectors are calculated. As different matrices may have the same column sum vector, there were repetitions among them. Besides, only vectors with decreasing order of coordinates are taken. Because algorithm G is sensitive to this order, it may depreciate the ability of algorithm. Further in case of branching in some column, among all possible partitions only one is selected (an arbitrary choice), that also may decrease abilities of the algorithm. However all these steps simplify the calculations.

Algorithm G succeeds if in the last column of constructions only one length intervals occur. Experiments show that for some cases in the last column there remain 2-length intervals, in most cases only one such interval is present. Table 1 shows the result.

Table 1.

Number of rows	Number of vectors	One 2-length interval	More 2 length intervals	Number of rows	Number of vectors	One 2-length interval	More 2 length intervals
<12		0	0	22	172	59	1
12	23	5	0	23	193	63	3
13	30	5	0	24	214	90	2
14	41	11	1	25	232	99	3
15	51	11	0	26	255	119	2
16	69	18	0	27	265	108	5
17	83	26	1	28	290	134	5
18	97	33	1	29	287	119	11
19	115	34	0	30	287	159	3
20	134	40	1	31	284	121	15
21	149	44	1	32	253	0	0

Besides the case $n = 6$, special "bad" vectors are composed from the point of view of the algorithm G where:

- decrease order is the worst for the given vector
- there is only one successful choice of partitioning in case of branching, and we didn't take it, etc.

On the other hand series of easy constructible monotone Boolean functions are considered in up to 11 dimensional unit cubes and column sums are calculated for corresponding $(0,1)$-matrices. The result of performance of algorithm G for all these vectors remained unchanged: there are no intervals of > 2 length in the last column.

At this moment it is only a supposition based on experiments that algorithm G will not produce matrices with > 2-length intervals in the last column. Further efforts are required to prove or reject this supposition. However assuming true, this property let us estimate the performance of G.

The worst case with this assumption is a matrix with all 2-length intervals in the last column. This implies $m/2$ pairs of coinciding/repeating rows in the matrix. Denote by m_G the number of different rows by G and taking into account that there are C_m^2 different pairs of rows in matrices with m different rows, we get $m_G = m(m-1)/2 - m/2 = m/2(m-2)$. Let A_{opt} be an optimal algorithm constructing target matrices and hence $m_{A_{opt}} = m(m-1)/2$. Compose the performance ratio: $\frac{m_{A_{opt}}}{m_G} = 1 + \frac{1}{m-2}$.

Bibliography

[DCh, 1999] Durr Ch., Chrobak M., Reconstructing hv-convex polyominoes from orthogonal projections, Information Processing Letters 69 (1999) pp. 283-291.

[BDLNP, 1996] E. Barcucci, A. Del Lungo, M. Nivat, and R. Pinzani, Reconstructing convex polyominoes from horizontal and vertical projections, Theoret. Comput. Sci., 155:321{347, 1996.

[W, 2001] G.J. Woeginger, The reconstruction of polyominoes from their orthogonal projections, Inform. Process. Lett., 77:225{229, 2001.

[R, 1966] H. J. Ryser, Combinatorial Mathematics, 1966.

[K, 1973] D. Knuth, The Art of Computer Programming, vol.3. Sorting and Searching, Addison-Wesley Publishing Company, 1973.

[S, 1995] H. Sahakyan, Hierarchical Procedures with the Additional Constraints, II Russia, with participation of NIS Conference, Pattern Recognition and Image Analysis: New Information Technologies, Ulianovsk, 1995, pp.76-78.

[S, 2009] H. Sahakyan, "Numerical characterization of n-cube subset partitioning", Discrete Applied Mathematics, vol. 157, pp. 2191-2197, 2009.

[S, 2010] H. Sahakyan, "Boundary elements of the n-cube subset partitioning characterization", submitted to Discrete Applied Mathematics.

[S, 2010 G] H. Sahakyan, "Approximation greedy algorithm for reconstruction of (0,1)-matrices with different rows"

Authors' Information

Hasmik Sahakyan – Leading Researcher, Institute for Informatics and Automation Problems, NAS RA, P.Sevak St. 1, Yerevan 14, Armenia, e-mail: hasmik@ipia.sci.am

Levon Aslanyan – Head of Department, Institute for Informatics and Automation Problems, NAS RA, P.Sevak St. 1, Yerevan 14, Armenia, e-mail: lasl@.sci.am

HISTOLOGY IMAGE SEGMENTATION

Francisco J. Cisneros, Paula Cordero, Alejandro Figueroa, Juan Castellanos

Abstract: *In this article, a technique for the segmentation of the components of histological images will be explained. To be able to do a study about the various microscopic components of the animal tissues and to reach to the histological images; first it is tried to be obtained some sections of the tissues and then; dye them in accordance with the different components which wanted to be studied. The image-analysis is a statistical work and most of the time, the data which is reached in the end, depends on the observer who is carrying out the study.*

Keywords: *Image Processing, Image Segmentation*

ACM Classification Keywords: *I.4.3 Image Processing and computer vision – Enhancement, J.3 Biology and Genetics*

Introduction

Histology is a branch of science which studies the microscopic structures and the functions of the organic tissues. The actual means of technology; such as the microscopes, monitorized stages, slide loaders, cameras and PC; have improved the use of histology due to the incremental number of daily analyzed images. This increase allows having more precise data about the case-studies as we can have more reliable results. However, as the analysis of all sections of the tissues require a long time, only some sections of the sample are examined and the data from this sample is generalized to be used for statistical techniques (stereology). It has two faces; on the one hand, the analysis requires a long time while on the other - and also in the expanse of the other- it contains the problems of precision and reliability of the data. The possibility of the new automatic techniques in the image analysis addresses these problems and provides new advantages.

The automatic analysis of the images allows us not only to have reliable data but also it accelerates the acquisition process of the data. In addition to that, to be able by managing them digitally, it can also be reached to some other evaluation methods and data which are normally, expensive or not feasible; like the descriptors of textures or other descriptors which enables better analysis of the images. Thanks to those techniques, now we can carry out studies and reach conclusions which were not possible before.

The classic technique which is used to analyze the histological images is the stereology. The stereology is a method which enables the observers, to split three dimensional structures into two whenever they are parallel and equidistant. With the help of mathematical formulas which depend on the geometric probability, the results can be statistically significant.

Image analysis process requires a set of procedures; experimentation, processing and analysis procedures. First of all obtaining organic tissues to be analyzed, in order to get information, during the experimentation process these tissues are cut into thin slices to be observed under a microscopy. Before observation, slices are subjected to staining techniques to obtain the prepared pieces to be inspected and enhance contrast in the microscopic image. In observation process a set of images is taken through microcopy and then these images are processed using different techniques. In this context, it is necessary the segmentation technique's calibration in order to make plausible the distinction of the different areas to be identified in the image. Once the calibration is reached, images are segmented and descriptors are obtained from them (number of segmented regions, proportions, perimeter, etc). Image analysis and classification processes are carried out thanks to the information collected by these kinds of descriptors.

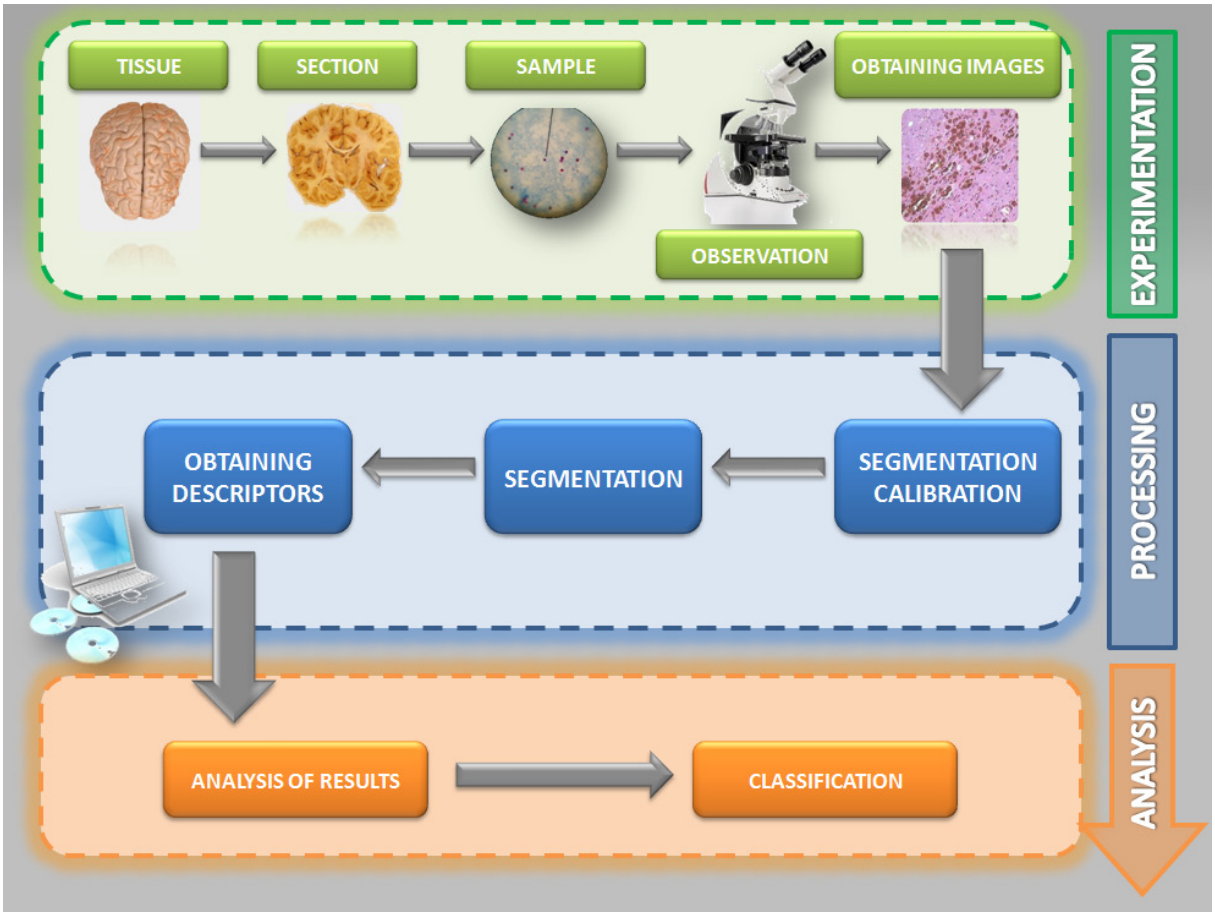


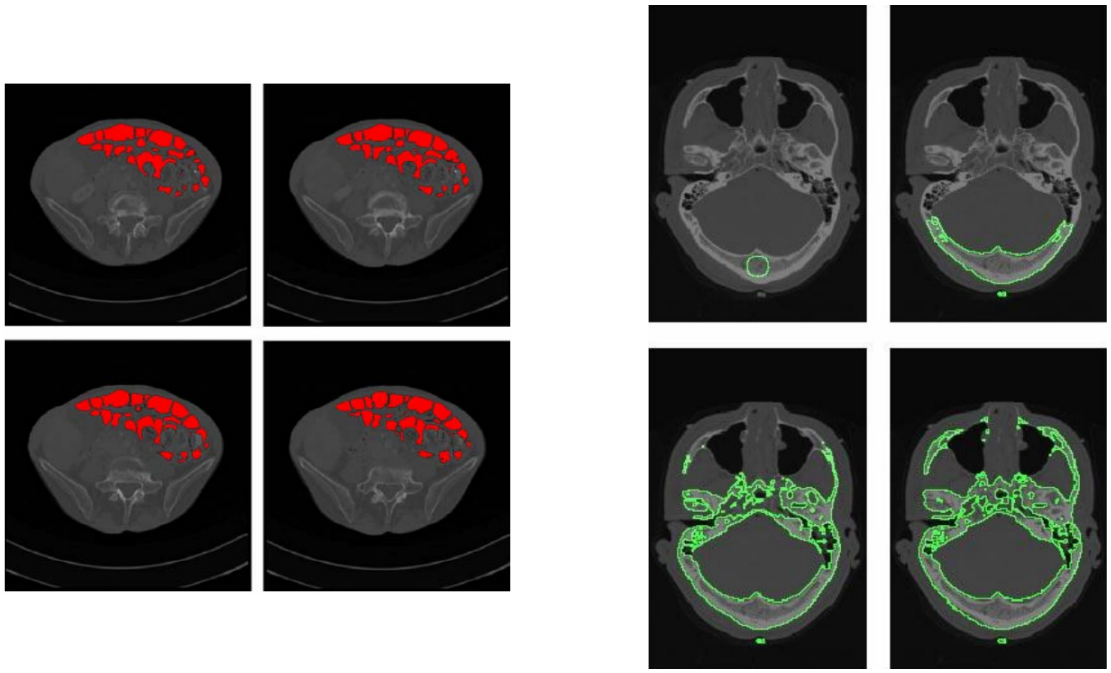
Figure 1.Experimentation, processing and analysis procedures of histology images.

Segmentation

Image segmentation is defined to be the process that subdivides an image into its constituent parts or objects [R.C. Gonzalez and P.Wintz, 1987]. One of main challenges in image analysis is to identify and examine the tissue that is often not present in the whole image of the study; the isolation of the desired tissue is carried out using segmentation processing. In addition, histology aims to analyze certain components of the same biological tissue and it is necessary to distinguish it within the image, this is the reason for using staining techniques for the different components to highlight structures for viewing, often with the aid of different microscopes. Identification of components in an image is easy for humans, but in order to get it using computational processes is necessary to apply segmentation techniques. After segmentation process you get an image divided into different compounds wanted to be analyzed, and it would be possible to be counted and analyzed with specific descriptors.

Nowadays there are many segmentation techniques and not all of them are valid for all environments. Due to a universal segmentation technique does not exist, it is necessary to use a concrete segmentation for each environment. Some main problems found during segmentation processes for the analysis of images are related to highly irregular image structures, inconsistent staining, non-uniform illumination, out-of-focus image components, and variability in the objects of interest. Segmentation techniques are divided primarily into these different topics*: Pixel-based-methods* as Threshold Segmentation, based on the categorization of pixels of an image according to a certain threshold (one pixel corresponds to a point on the digital image); Edge-based-methods as Edge Detection [Pal NR, 1993][Marr D, 1980], which consists of finding edges in the image in order to extract the closed polygons found; and finally *Region-based-methods* as Region Growing [MacQueen, 1967] based on group pixels with similar characteristics.

Segmentation techniques based on Edge Detection using minimum cost functions and information from the limits of the structures to determine the borders of homogeneous regions. Some of these kinds of techniques use different filters to detect perimeters and they are usually based on the concept of gradient. When the result does not provide closed elements it is advisable to apply additional techniques to connect them like morphological operators, for example. The main problem of these techniques is the threshold used to determine the existence of an edge which depends heavily on image contrast and details. The contrast, brightness and detail of the structures may be different from a tissue to another similar depending on the stain used and the quality of it. This technique can be interesting for segmenting images whose components are not clearly separated and parts of them overlap [J. Díaz, 2007].



2a) 2b)

Figure 2. 2a) Region-growing segmentation from CT images. 2b) Automatic segmentation using a shape-based approach. Images from [J. Díaz, 2007].

Region-based-methods attempt to reconstruct different anatomical structures using a certain similarity criteria from the pixels of each selected area. Some of these segmentation techniques are known as Region-growing (figure 2a) approaches, starting from one or more initial regions called seeds which can be initialized manually or automatically using and heuristic method the initial regions are expanded to neighboring volumetric pixels following certain homogeneity criteria [D.L. Pham et al, 2000]. Region-growing techniques like Clustering have problems due to the great irregularity of the images and the variability of components to be segmented which sometimes cause an over-segmentation by the segmentation of not interesting components or interesting objects in too many regions.

Other methods of this family are Shape-based-approaches (figure 2b), which continuously keep a representation of the target element's shape. These shapes, also known as active shape models and active contour models, can be modified manually getting a better adaptation to the element or can be adjusted automatically according to the value of a certain energy function. Related to the application of shaped-based approaches it is found the well-known Level set-method, the main idea of this method is to evolve a curve towards the lowest potential of a cost function which must be defined imposing certain smoothness constraints based on Lagrange's method in order to parameterize the objective contour according to some sampling strategy and then evolve each element according to the image and internal terms [A. Tsai et al, 2003].

Histology image processing

Thanks to the experience in other histological images processing works, we observed that using thresholding techniques, we obtained very precise results in a really short computing time. However, with other methods like region growing techniques, we couldn't segment properly all the components we wanted to analyze. And also with edge detection techniques, in many times, we obtained an excessive segmentation plus additional processing to solve this problem.

Recall that histological images normally have the feature to tint the components to study in a specific color by being able to differentiate to the rest of the components by the color they present. Because of that, thresholding has turned out to be the most precise and efficient technique. This method examines the properties of each pixel in order to evaluate his color. The calibration of this technique is very simple because it just requires selecting a color sample of the component we want to analyze. This color sample turns out to be the same for every histological image accomplished with the same preparation (same dye, same tissue and same exposition time) .

Thresholding requires a measure to determine the type of component of the evaluated pixel. We propose the Mahalanobis distance to determine the kind of component being evaluated. In short, we measure the distance between the color of the pixel and the average color each one of the component we want to segment.

First the image is segmented using pixel-based techniques to classify pixels into component 1 and component 2 to be distinguished. Pixel-based techniques do not consider the spatial context but only decide on the basis of the color features at individual pixels. Pixel determination, component 1 vs. component 2, was based on the mahalanobis distance [Mahalanobis, 1936] to the chromatic gamma of the component 1 and setting an a priori threshold for the classification of first component pixels. Chromatic range was determined from seven component representative sampling. Measure is subjective and must be calibrated for each type of staining.

1. Select the color sample of the component we want to segment
2. Obtain the average color of the sample
3. For each pixel:
    1.1. Measure the Mahalanobis distance between the color of the pixel and the average color of the sample
    1.2. If this distance is lower than 0.15 the pixel is checked as a part of the component we want to segment
4. It subdivides the image in each component, which keeps isolated

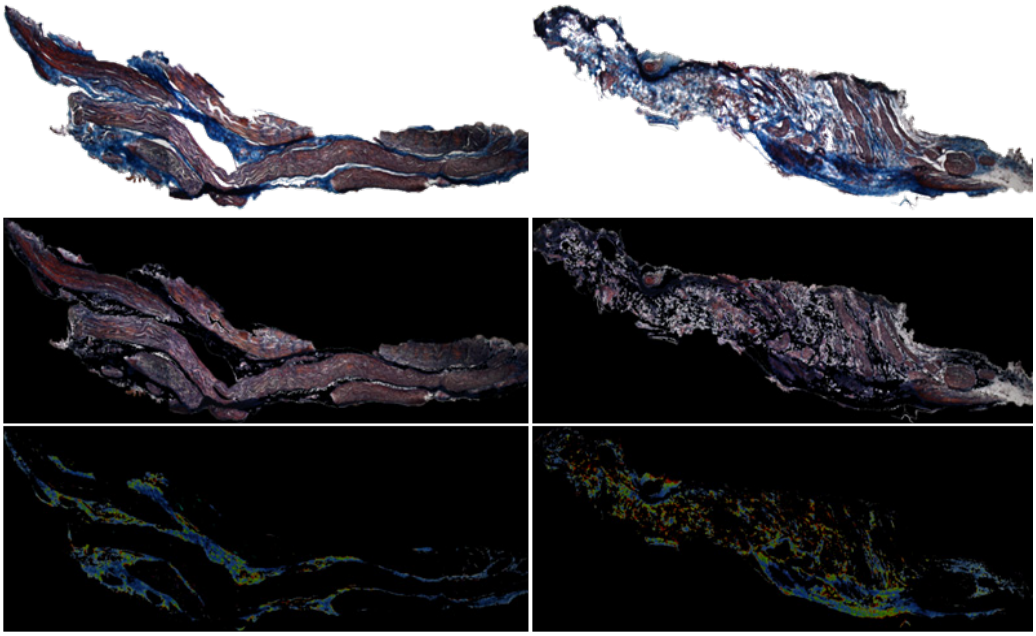


Figure 3. Histological neuroma image example [Herrera-Rincon, 2011]. Top: images of two neuromas (Prosthetic left and Amputated right). Middle: images after component 1 segmentation. Bottom: images after component 2 segmentation.

Figure 3 shows an example of the type of segmentation used in the article "Effects of Stimulation on the Peripheral Neuroprosthetic Nerves of Amputees: a Study Digital Image Processing" [Herrera-Rincon et al., 2011]. It was necessary a high level precision during the segmentation in order to make a correct count of component 1 percentage against component 2.

Experimentation

The following images were segmented using various techniques. In each example the result of applying the proposed segmentation is shown to be compared with others. After this results are discussed.

Figure 4 shows images obtained from tissues where it is intended to count cancer cells [Loukas et. al, 2003]. The technique used for segmentation is Laplacian of Gaussian (LoG) edge detector [Marr D, 1980]. For this technique we segmented only components (cells) in brown. As can be observed, for brown cells the results have been similar, though more problems that arises due to the proximity of all components to target.

The Figure 3 [Sertel et. al, 2008] show histological images of five major cytological components in the FL tissue: nuclei, cytoplasm, extra-cellular material, red blood cells (RBC) and background regions. Having nuclei and cytoplasm regions dyed with hues of blue and purple, extra-cellular material dyed with hues of pink and RBCs dyed with hues of red, H&E-stained FL images provides useful visual clues for segmentation. In addition to these components, there are also background regions that do not correspond to any tissue component. With this a priori knowledge on FL images, we performed the segmentation using K-means clustering algorithm [MacQueen, 1967] to identify these cytological components. With the proposed segmentation we obtain very similar results.

For this case we have selected the purple components (the nuclei). For each component is necessary to complete the complete process.

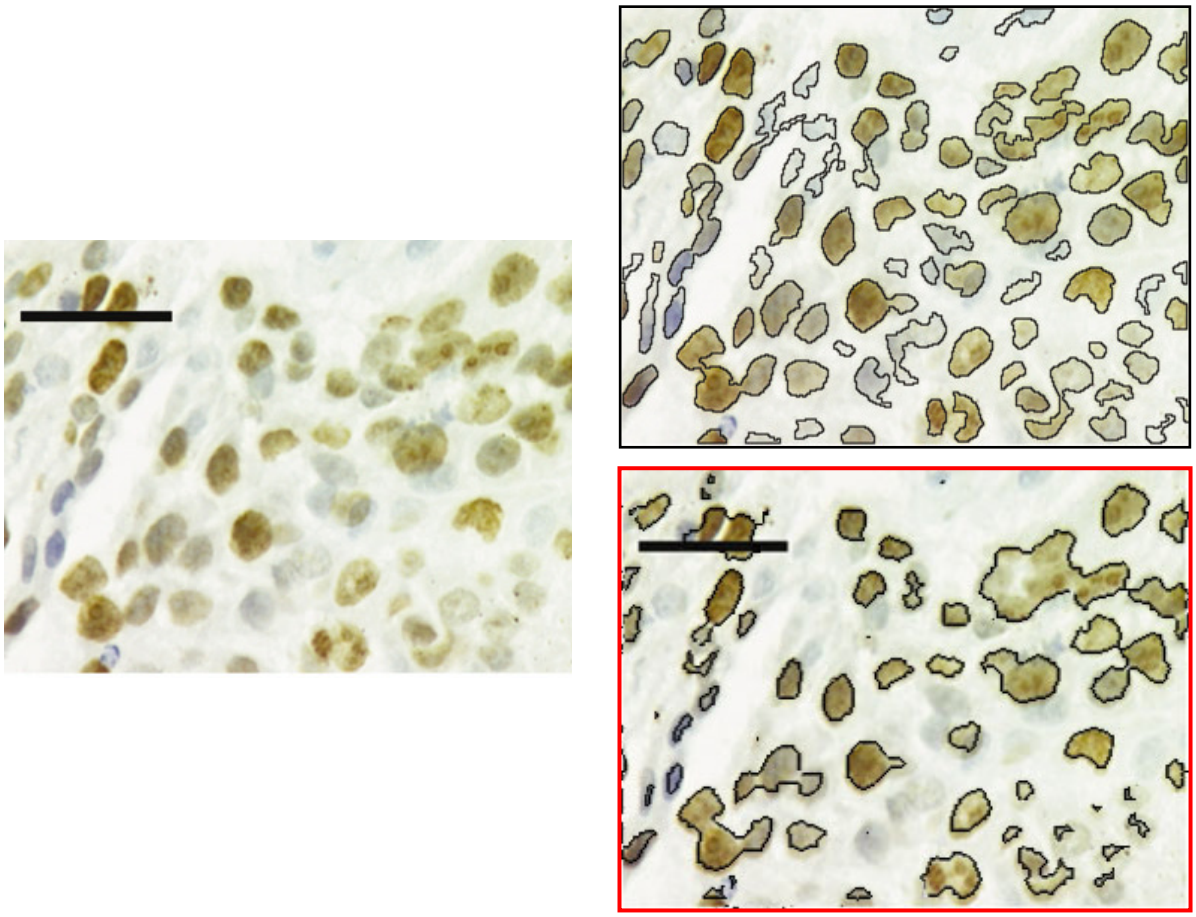


Figure 4. Example using Edge Detection [Loukas et. al, 2003]. Left, original image. Right, segmentation image (top, with LoG, bottom with our technique)

Local and Global Gaussian Mixture Models for Hematoxylin and Eosin Stained Histology Image Segmentation (LG-GMM) algorithm employs the unique characteristics of the H&E staining protocol: hematoxylin stains nuclei blue; eosin stains the cytoplasm pink, red blood cells red; air spaces are white. Therefore, we can obtain that nuclei are always represented by dark points. As it can be observed, in this case the results are practically the same, even more precise. This is due to good staining and spatial distance of the components to be detected.

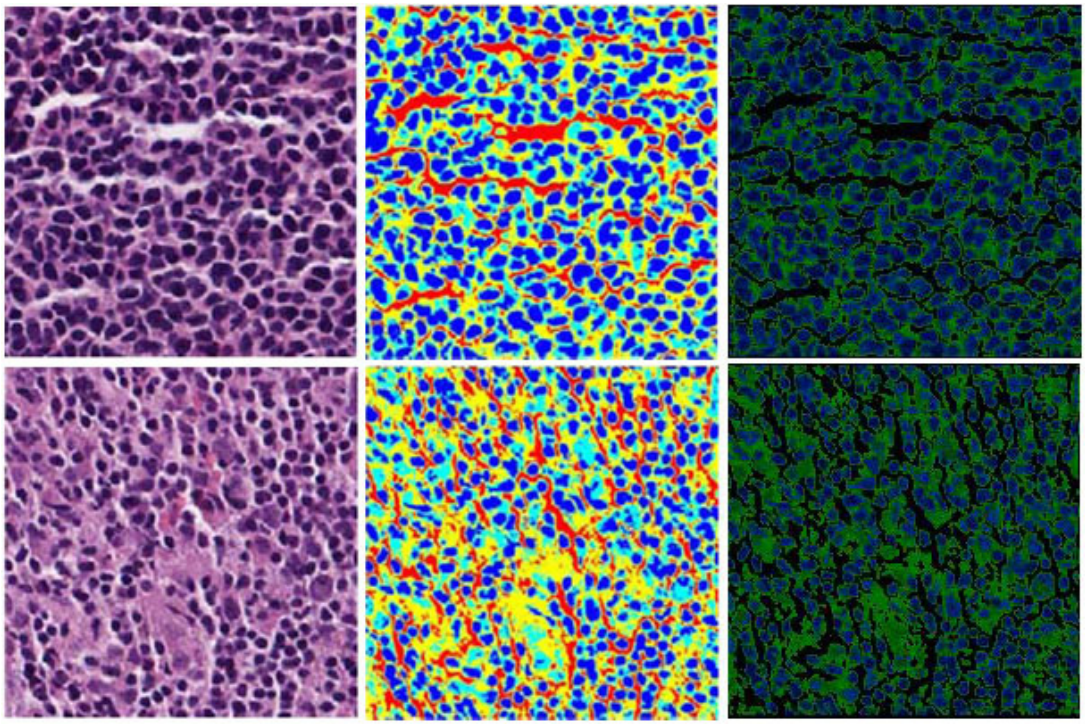


Figure 5. Example using K-means clustering [Sertel et. al, 2008]. Left column shows sample H&E-stained FL images. The corresponding segmentation with K-means results are shown in the center column. In these color labeled images, blue corresponds to nuclei, cyan to cytoplasm material and red and grey to background, and RBCs, respectively. The corresponding segmentation with us technique results is shown in the right column. In these, dark green to cytoplasm material, black to background, and green to RBCs.

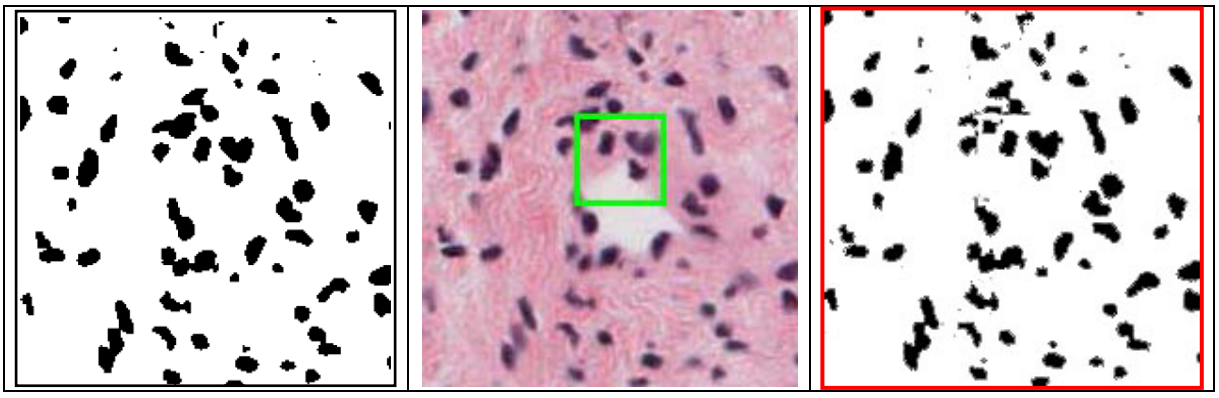


Figure 6. Example using LG-GMM segmentation [He et al, 2010] for nuclei detection example. First column shows the real image. The corresponding segmentation results are shown in the second column, the top correspond to LG-GMM segmentation, bottom correspond to the technique proposal in this article.

Conclusions

As shown in the results, the segmentation technique we propose has a very similar segmentation to that obtained by other techniques. The calibration of segmentation proposed is very simple, and the code's implementation is trivial in languages like MATLAB.

There are many image segmentation techniques, some of them very complex. Many of the papers published in recent years on the analysis of histological images have very sophisticated techniques for segmentation. This means spending more time in understanding, implementation and processing. The proposed technique provides results comparable and even better than those obtained with other techniques shown. Moreover, given the small number of parameters presented and easy understanding, it can be adapted quickly to any histological image and its processing speed is difficult to improve because of the small number of operations performed.

In this way we can conclude that the proposed segmentation technique has significant advantages for analysis of histological images. Especially in preparations with different dyes and whose elements have a considerable spatial separation.

References

[Herrera-Rincon et al., 2011] Celia Herrera-Rincon, Francisco Jose Cisneros, Juan Castellanos, and Fivos Panetsos. Effects of Neuroprosthetic Stimulation on the Peripheral Nerves of Amputees: a Digital Image Processing Study. 2011 5th International IEEE/EMBS Conference on Neural Engineering (NER). 2011

[He et al, 2010] Lei He, L. Rodney Long, Sameer Antani, George R. Thoma. Local and Global Gaussian Mixture Models for Hematoxylin and Eosin Stained Histology Image Segmentation. 10th International Conference on Hybrid Intelligent Systems. 978-1-4244-7364-9/10/$26.00 ©2010 IEEE pg 223-228. 2010

[Sertel et. al, 2008] Olcay Sertel, Jun Kong, Umit V. Catalyurek, Gerard Lozanski, Joel H. Saltz, Metin N. Gurcan. Histopathological Image Analysis Using Model-Based Intermediate Representations and Color Texture: Follicular Lymphoma Grading. J Sign Process Syst (2009) 55:169–183. DOI 10.1007/s11265-008-0201-y. 2008

[J. Díaz, 2007] Jose Díaz Iberri. Inspección Interactiva de Estructuras Anatómicas Tubulares. Master Thesis in Computer Sicinces. Universidad Politécnica de Cataluña. 2007

[A. Tsai et al, 2003] Andy Tsai et al. A Shape-Based Approach to the Segmentation of Medical Imagery Using Level Sets. IEEE Transactions on medical imaging. Vol. 22 No. 2. February, 2003

[Loukas et. al, 2003] Constantinos G. Loukas, George D. Wilson, Borivoj Vojnovic, and Alf Linney. An Image Analysis-Based Approach for Automated Counting of Cancer Cell Nuclei in Tissue Sections. Cytometry Part A 55A:30–42. 2003

[D.L. Pham et al, 2000] Dzung L. Pham et al. Current Methods in Medical Image Segmentation. Annual Review of Biomedical Engineering, volume 2, pp 315-337. 2000

[Pal NR, 1993] Pal NR, Pal SK. A review on image segmentation techniques. Pattern Recognition 1993;26:1277–1294. 1993

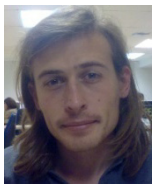
[R. C. Gonzalez and P. Wintz, 1987] R. C. Gonzalez and P. Wintz, Digital Image Processing. Addison-Wesley, Reading, Massachusetts. 1987

[Marr D, 1980] Marr D, Hildreth EC. Theory of edge detection. Proc R Soc London 1980;B207:187–217. 1980

[MacQueen, 1967] MacQueen, J. B. Some Methods for classification and Analysis of Multivariate Observations. In Proc. of 5th. Berkeley symp. on mathematical statistics and probability. (vol. 1, pp. 281–297) Berkeley, Univ. of California Press. 1967

[Mahalanobis, 1936] P. Mahalanobis, "On the generalized distance in statistics" Proc. Nat. Inst. Sci. India (Calcutta) , 2 (1936) pp. 49–55. 1936

Authors' Information



Francisco J. Cisneros *– Natural Computing Group. Universidad Politécnica de Madrid, Boadilla del Monte, 28660 Madrid, Spain. email:* kikocisneros@gmail.com

Paula Cordero *– Natural Computing Group. Universidad Politécnica de Madrid, Boadilla del Monte, 28660 Madrid, Spain. email: pcordero@gcn.upm.es*

Alejandro Figueroa *– Natural Computing Group. Universidad Politécnica de Madrid, Boadilla del Monte, 28660 Madrid, Spain. email: afmeana@gmail.com*

Juan Castellanos *– Natural Computing Group. Universidad Politécnica de Madrid, Boadilla del Monte, 28660 Madrid, Spain. e-mail: jcastellanos@fi.upm.es*

DIFFERENTIAL EVOLUTION – PARTICLE SWARM OPTIMIZATION

Nuria Gómez Blas, Alberto Arteta, Luis F. de Mingo

***Abstract**: This paper shows the Particle Swarm Optimization algorithm with a Differential Evolution. Each candidate solution is sampled uniformly in [−5,5] ^{D}, whereDdenotes the search space dimension, and the evolution is performed with a classical PSO algorithm and a classical DE/x/1 algorithm according to a random threshold.*

***Keywords**: Benchmarking, Black-box optimization, Direct search, Evolutionary computation, Particle Swarm Optimizacin, Differential Evolution*

***Categories:**G.1.6 [Numerical Analysis]: Optimization-global optimization, unconstrained optimization; F.2.1 [Analysis of Algorithms and Problem Complexity]: Numerical Algorithms and Problems.*

Introduction

A variety of general search techniques can be employed to locate a solution in a feasible solution space. Most techniques fit into one of the three broad classes. The first major class involves calculus-based techniques. These techniques tend to work quite efficiently on solution spaces with “friendly” landscapes. The second major class involves enumerative techniques, which search (implicitly or explicitly) every point in the solution space. Due to their computational intensity, their usefulness is limited when solving large problems. The third major class of search techniques is the guided random search. Guided searches are similar to enumerative techniques, but they employ heuristics to enhance the search [14].

Evolutionary algorithms (EAs) are one of the most interesting types of guided random search techniques. EAs are a mathematical modeling paradigm inspired by Darwin's theory of evolution. An EA adapts during the search process, using the information it discovers to break the curse of dimensionality that makes non-random and exhaustive search methods computationally intractable. In exchange for their efficiency, most EAs sacrifice the guarantee of locating the global optimum.

Differential evolution (DE) and Particle Swarm Optimization are both stochastic optimization techniques. They produce good results on both real life problems and optimization problems. A simple mixture between those two algorithms, called Differential Evolution – Particle Swarm Optimization (DE-PSO), is explained in the following sections. The explanation will no longer use the sine function, but the more frequently used sphere function. Also note that the explanation for this algorithm will not use a single value, but arrays (vectors) to represent particles and velocities. Therefore it is compatible with more dimensions [2, 3, 4].

Differential evolution

Differential evolution (DE) is a simple evolutionary algorithm for numerical optimization whose most novel feature is that it mutates vectors by adding weighted, random vector differentials to them [1].Differential evolution is, like PSO, a stochastic and population-based optimization technique. It was first introduced in 1996 by Price and Storn

[13]. The algorithm is a modification of genetic annealing, which is beyond the scope of this paper and will not be discussed. Differential evolution is capable of handling non-differentiable, nonlinear and multimodal objective functions and is fairly fast in doing so. DE has participated in the First International IEEE Competition on Evolutionary Optimization (ICEO) and was proven to be one of the fastest evolutionary algorithms [10]. The DE algorithm also works with a population of potential solutions. The principle is the same as PSO: a particle can gain by using information from other particles as well as the results of their own search. However, in the case of differential evolution, that information is sampled randomly. The classical DE algorithm works as shown in the pseudo code of figure 1.

```
1  For each particle
2     Initialize particle and fitness
3  End
4
5  Do
6  For each particle
7     Create mutant
8     Perform crossover
9     If an offspring is better than the parent
10       Replace parent in the next generation
11    Calculate fitness
12 End
13 While maximum iterations is not reached
```

Figure 1: *Pseudo code for DE written in Netbeans.*

Step 1: define a population

For the purpose of this example, a simple 2-dimensional (having 2 variables x1 and x2) sphere function is suitable to search for a minimum. The function rule is as follows:

$$\text{Rand}(0,1) * (\mathbf{b}u - \mathbf{b}l) + \mathbf{b}$$

By drawing the function on a graph it is easily perceived that the global minimum is at coordinates [0,0]. To prove this general truth, a population of 5 particles can be randomly initialized in a search space bounded by [-10,10]. Each particle is presented by an array, or formally speaking by a vector, of D values where D represents the dimension of the search space. Each component of the array can be randomly initialized using equation (a) to initialize a random particle between bounds where rand(0,1) indicates a random number between 0 and 1, bu is

the value of the upper bound and bl is the lower bound. Because the problem requires the search for a global minimum, the particle with the smallest fitness value is closest to the optimum. The table below shows 5 particles and their fitness values upon initialization. The population size does not change during the process.

#	Particle	Particle fitness value
1	[5,-9]	$5^2+(-9)^2=106$
2	[6,1]	$6^2+1^2=37$
3	[-3,5]	$-3^2+5^2=34$
4	[-7,4]	$-7^2+4^2=65$
5	[6,7]	$6^2+7^2=85$

***Table 1**: Each particle is an array of D values where D is the dimension.*

Step 2: mutation

The next step consists of making a mutated vector through differential mutation. It is a process whereby for each particle 3 different randomly chosen particles (r1,r2,r3) create a mutated particle mi (i=1,2,…,population size). This mutated vector is obtained by applying the formula below for each dimension:

$$mij = \mathbf{r}1j + \mathbf{F} * (\mathbf{r}2j - \mathbf{r}3j)$$

With j being the current dimension and F being the positive differential weight value usually between 0 and 1. This value is initialized before the loop and thus the same for every particle in the population. The formula adds the weighted difference between 2 population members to a third member. The table below shows an example of a mutation with the differential weight F set to 1.

Particles for mutation	Mutant vector
2, 4 and 5	[-7,-2]

***Table 2:** Mutant vector for target particle 1 with F=1.*

Step 3: crossover

Crossover is a term used when parameters of the mutant are mixed with parameters of the target vector to form a trial vector z. Which parameters should be crossed is depending on probability.Figure2 shows that 2 out of 5 parameters of the mutant vector were chosen to cross over.

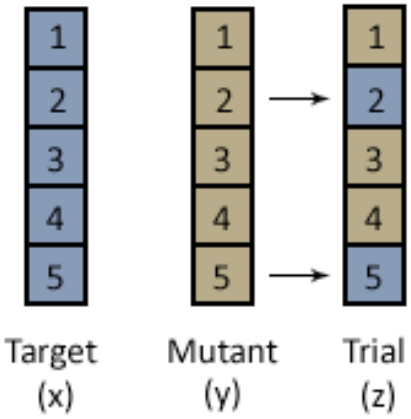


Figure2: Crossover principle.

In Differential Evolution there are 2 main forms of crossover: binomial and exponential. Both ways use a parameter called the Crossover Probability (CR) which is a value between 0 and 1. To understand DE, it is enough to only explain binomial crossover. Many publications explain other ways and should be looked into when more information is necessary.

```
1  For each value i in vector
2    If rand(0, 1) < CR
3      z_i = y_i
4    Else z_i = x_i
5  End
```

Figure3: Pseudo code for binomial crossover (for a minimization)

By applying binomial crossover, one should assign a random value between 0 and 1 and compare it to the crossover probability CR. If CR is bigger, the parameter zi becomes the mutant parameter yi. If it is smaller, zi is equal to the parameter of the current target element, xi, and thus stays the same.

Target particle x	Mutant y for x	Trial vector z
[5,-9]	[-7,-2]	[5,-9]

Table 5: *Binomial crossover for a component of particle 1 with CR=0.5 and twice rand=0.9135 for convenience.*

Step 4: selection

Once the new trial vector z is created, the algorithm has to decide whether or not it should become a member of the new generation (or otherwise known as iteration t+1). This is done by comparing fitness values with the original element of the population. If the fitness of the trial vector is better (i.e. lower for a minimization), itshould be the new member of the population. If it is not better, the current member should be kept unchanged. This process continues, for example, until a maximum number of iterations is exceeded.

Variants of DE

The above mentioned is only the basic variant of DE, but there are many others. In order to classify variants, the following notation is used: DE/x/y/z, with the following components:

X: specifies the vector to be mutated.This can be 'rand' for a randomly chosen population vector or 'best' for the vector with the best fitness value.

Y: The number of difference vectors used.

Z: specifies the crossover scheme. This can be 'bin' for a binomial scheme or 'exp' for an exponential crossover.

Based upon this notation, the version explained in this paper is written as DE/rand/1/bin. DE/best/2/bin is an example of a variant proposed by Price and is proven to be effective. 2 difference vectors are used and seem to improve the diversity of the population if the population size is high enough. For this method, formula (b) can be rewritten:

$$(c)\ vij = \mathbf{x}best + \mathbf{F} * (\mathbf{r}1j + \mathbf{r}2j - \mathbf{r}3j - \mathbf{r}4j)$$

Hybrid DE- PSO

DE-PSO is basically a Differential Evolution algorithm mixed with ideas of Particle Swarm Optimization. It was proposed by Millie Pant, Radha Thangaraj, Crina Grosan and Ajith Abraham in their paper called "Hybrid Differential Evolution – Particle Swarm Optimization Algorithm for Solving Global Optimization Problems". This section explains how the algorithm works according to the very detailed pseudo code presented in figure4, written as a minimizer.

```
For each particle
   Initialize particle and velocity
End
Set Global best and personal best
Do
For each particle
   Choose 3 random particles (r1≠ r2 ≠ r3)
   For each dimension j
      Perform DE part to create mutated U.
   End
   If evaluate (U) < evaluate (X) then
      X = U
   Else
      For each dimension j
         Perform PSO part to create Z
      End
   End
   If evaluate (Z) < evaluate (X) then
      X = Z
   Else
      X = X
   Update personal best
End
Update global best
While stop condition is not reached
```

Figure4: *Pseudo code for DE-PSO*

Step 1: defining population

This step is completed in the same way as PSO and DE. Please note that you should define your population and their velocities as two 2-D arrays when programming. The higher the dimension, the more values each particle will have. A particle in a 2-D search space can be defined as 2 points on a graphs, namely x and y. When you have a particle in a 3-D space, it needs 3 points to be represented on x,y,z axes. To initialize the velocities, one can either choose random values between the predefined bounds, or zero.

Step 2: performing DE

First, 3 random particles (r1,r2 and r3) should be chosen in such a fashion that they are different from each other. The same principle is applied in DE mutation. The second task is to create a mutated value for eachdimension j of the particle. This is done by utilizing equation (b). When the mutation is done for every dimension, one must evaluate the mutated particle with the fitness function, and compare it to the evaluation of the non-mutated,

current particle. If it is smaller, the mutated particle should replace the old one in the population. Though if it is bigger, the particle swarm optimizer should be triggered.

Step 3: performing PSO

If the DE-part of the algorithm did not find a better solution, the PSO is activated. A new particle should be created according to formulas (d) for the velocity and (e) for the new position. A basic velocity and position clamping should be performed here as well. It is enough to just check if the new velocity or position exceeds the bounds in which the algorithm is performed. If the newly created particle is proven to be better, it should be replaced by the old one for the next generation. Both personal best and global best particle vectors should be updated as well. This process is again repeated until the stopping condition is met.

$$(d)\ \boldsymbol{vi}(t+1) = \mathbf{w}^{*}\boldsymbol{vi}\ (t) + \boldsymbol{c1}^{*}\boldsymbol{r1}^{*}(\ \boldsymbol{p}Best - \boldsymbol{xi}) + \mathbf{c}2^{*}\ \boldsymbol{r2}\ ^{*}(\ \boldsymbol{g}Best - \boldsymbol{xi})$$

$$(e)\ \boldsymbol{xi}(t+1) = \boldsymbol{xi}\ (t) + \boldsymbol{vi}\ (t+1)$$

Expansion

More hybrid versions between Particle Swarm Optimization and Differential Evolution have been proposed. One of those is the version proposed by José García, Enrique Alba and Javier Apolloni in their work: " Noiseless Functions Black-Box Optimization: Evaluation of a Hybrid Particle Swarm with Differential Operators". Their model is also simple and is proven to obtain an accurate level of coverage range. The algorithm also contains 2 main parts. The first is the differential variation, where new velocities and positions are calculated according to:

$$(f)\ \mathbf{v'}ij = \mathbf{w} * \mathbf{v}ij + \boldsymbol{\mu} + \boldsymbol{\varphi} * (\mathbf{g}bestj - \mathbf{x}ij)$$

$$(g)\ \mathbf{x'}ij = \mathbf{x}ij + \mathbf{v'}ij$$

Where j is the dimension and i=1,2,…, population size. μ is a scaling factor (μ = UN(0,1)) and φ is the social coefficient (φ = UN(0,1)). The second part is the mutation, which is calculated according to formula (a). It is basically a new particle position between the specified bounds.

Bibliography

[1] S. H. Brooks. A discussion of random methods for seeking maxima.Operations Research, 6:244– 251, 1958.

[2] S. Finck, N. Hansen, R. Ros, and A. Auger. Real-parameter black-box optimization benchmarking 2009: Presentation of the noiseless functions. Technical Report 2009/20, Research Center PPE, 2009.

[3] N. Hansen, A. Auger, S. Finck, and R. Ros. Real-parameter black-box optimization benchmarking 2009: Experimental setup. Technical Report RR-6828, INRIA, 2009.

[4] N. Hansen, S. Finck, R. Ros, and A. Auger. Real-parameter black-box optimization benchmarking 2009: Noiseless functions definitions. Technical Report RR-6829, INRIA, 2009.

[5] M. J. D. Powell. The NEWUOA software for unconstrained optimization without derivatives.LargeScaleNonlinearOptimization,pages255–297, 2006.

[6] J. Nelder and R. Mead. The downhill simplex method. Computer Journal, 7:308–313, 1965.

[7] T Jayabarathi, Sandeep Chalasani, Zameer Ahmed Shaik, Nishchal Deep Kodali; "Hybrid Differential Evolution and Particle Swarm Optimization Based Solutions to Short Term Hydro Thermal Scheduling", WSEAS Transactions on Power Systems Issue 11, Volume 2, pp. , ISSN: 1790-5060, 2007.

[8] Piao Haiguo, Wang Zhixin, Zhang Huaqiang, "Cooperative-PSO-Based PID Neural Network Integral Control Strategy and Simulation Research with Asynchronous Motor Controller Design", WSEAS Transactions on Circuits and Systems Volume 8, pp. 136-141, ISSN: 1109-2734, 2009.

[9] Lijia Ren, Xiuchen Jiang, Gehao Sheng, Wu B;"A New Study in Maintenance for Transmission Lines", WSEAS Transactions on Circuits and Systems Volume 7, pp. 53-37, ISSN: 1109-2734, 2008.

[10] Kenneth Price. Differential evolution vs. the functions of the second ICEO. In Proceedings of the IEEE International Congress on Evolutionary Computation, pages 153–157, 1997.

[11] Kenneth Price, Rainer M. Storn, and Jouni A. Lampinen. Differential Evolution: A Practical Approach to Global Optimization (Natural Computing Series). Springer- Verlag New York, Inc., 2005. ISBN 3540209506. URL http://portal.acm.org/citation.cfm?id=1121631.

[12] K.V. Price. Differential evolution: a fast and simple numerical optimizer. In Fuzzy Information Processing Society, 1996. NAFIPS. 1996 Biennial Conference of the North American, pages 524–527, 1996. doi: {10.1109/NAFIPS.1996.534790}.

[13] Storn, R., Price, K., "Differential evolution—A simple and efficient heuristic for global optimization over continuous spaces", Technical Report TR-95-012, International Computer Science Institute, Berkeley, CA, USA (1995).

[14] Paul K. Bergey, and Cliff Ragsdale; Modified differential evolution: a greedy random strategy for genetic recombination. The International Journal of Management Science. Volume 33, Issue 3, June 2005, Pages 255-265. doi:10.1016/j.omega.2004.04.009

Acknowledgment

This work has been partially supported by the Spanish Research Projects:

TRA2010-15645. "COMUNICACIONES EN MALLA PARA VEHICULOS E INFRAESTRUCTURAS INTELIGENTES" (Mesh communication with intelligent vehicles). (2010)

TEC2010-21303-C04-02. "ESTRUCTURAS RESONANTES PARA APLICACIONES DE SEÑAL FOTONICA DE BANDA ANCHA ". (2010).

Authors' Information

Nuria Gómez Blas *– Associate professor U.P.M Crtra Valencia km 7, Madrid-28031, Spain; e-mail: ngomez@eui.upm.es Research: DNA computing, Membrane computing, Education on Applied Mathematics and Informatics*

Albeto Arteta *– Associate professor U.P.M Crtra Valencia km 7, Madrid-28031, Spain; e-mail: aarteta@eui.upm.esResearch: DNA computing, Membrane computing, Education on Applied Mathematics and Informatics*

Luis F. de Mingo *– Associate professor U.P.M Crtra Valencia km 7, Madrid-28031, Spain; e-mail: lfmingo@eui.upm.es Research:,Artificial Intelligence, Social Intelligence, Education on Applied Mathematics and Informatics*

COMPUTATIONAL MODEL FOR SERENDIPITY

A. Anguera, M.A. Diaz, A. Gutierrez

Abstract: *In recent years there have been various attempts and studies that are eager to serendipity in Computer Science. Authors such as Campos and Dias have tried to model the serendipity in order to get serendipitous behaviors in Information Retrieval (IR). There have been attempts to introduce serendipity in Recommender Systems (RS), although the latter proposals have led efforts using metrics for measure Serendipity in those RS, rather than to emulate Serendipitous behaviors in recommender systems. However, so far there haven´t been succeeded in designing a model which can be applied to different Web browsing. The main problem we have detected analyzing the proposals in this field is that the solutions provided do not take into account the two aspects of the concept of serendipity. Do not consider that in addition to the accidental discovery of information that is not sought for, is also required some characteristics in the user like sagacity, perception, flexible thinking or intensive preparation. If we could develop a model that support any search engine or search tool, we would facilitate an incredible advantage to the user by offering specially information that the user is not focused upon. The aim of this paper is to propose a computational model that supports Serendipity and induce and facilitate serendipity through the use of a special-purpose designed system.*

Keywords*:* Supporting serendipity, : Designing Serendipity Serendipia, Data Minig, Artificial intelligence, SOM, Models of Computation

ACM Classification Keywords: H.2.8 Data Mining, F.1.1 Models of Computation, I.2 Artificial intelligence

Introduction

Making a journey through the history of discoveries, scientists or not, appears again and again, a phenomenon that is sometimes overlooked and sometimes is neglected, component by chance, coincidence, luck or even a "fluke" that entails the seemingly mysterious aspects it contains. This phenomenon, called Serendipity, refers to the accidental discovery of something valuable, find by, perhaps, luck, chance, fluke and sagacity of something that was not expected. It is also defined as the ability to make important discoveries by accident, or the ability to extract knowledge of random events. There are many examples of the value of serendipitous discoveries.

Although the importance of the phenomenon called serendipity, in the field of computer science there are not many jobs about it. What´s more, the few we founded are focused in the field of information retrieval and how to provide users of such systems serendipitous discoveries. Computer science, assessing the role of serendipity in discoveries, has tried to design systems that procure the serendipity. In any case, this research had its focus mainly on only one aspect of serendipity: "that of chance encounters". A step in the study of this subject is to consider not only the user's role in the perception of serendipity, but also how user perception should be assumed by the system itself to increase the chance of serendipity.

Applying agents to web browsers, some systems have designed trying to facilitate or induce serendipity. The first, Letizia [Lieberman 95], although no mention of the word serendipity, is described as an attempt to anticipate

what items may be of interest to the user, through inferences from his owns navigation behaviors. It doesn´t have predefined objectives or goals, and Letizia guides a search for linked documents , making recommendations to anticipate future needs.

Another relevant work in this field is the one done by Campos and Dias [Campos, 2001]. The authors, on the one hand, they develop what they call "Serendipity Equations", which would be better to call pseudoformal expressions, which help to discriminate whether a discovery is serendipitous or not. Campos and Dias identified four different types of serendipity, namely pseudoserendipity, serendipity, serendipity without "inspirational idea" and serendipity from incorrect knowledge. The authors provided a succinct description of the phenomenon in the following equations:

P1⊂(KP1) $\Rightarrow$ S1⊂(KP1, KM,KN) M⊂(KM)	Pseudoserendipity
P1⊂(KP1) P2⊂(KP2) $\Rightarrow$ M⊂(KM) S2⊂(KP2,KM,KN)	Serendipity
P2⊂(KP2) P1⊂(KP1) $\Rightarrow$ S2⊂(KP2, KN)	Serendipity wihtout an inspiring metaphor
P2⊂(KP2) P1⊂(KP1, EP1) $\Rightarrow$ S2⊂(KP2, KN)	Serendipity wiht incorrect knowledge

Table 1: Serendipity ecuations

P1	Original Problem.
KP1	Knowledge domain problem *P1*.
S1	Solution *P1*.
M	Metaphor.
KM	Knowledge domain *M*.
P2	Problem 2 inspirated by *M*.
KP2	Knowledge domain *P2*.
S2	Solution *P2*.
KN	New knowledge.
EP1	Incorrect knowledge for *P1*.

Table 2: Serendipity ecuations terms

From the formulation of the Serendipity equations, Dias and Campos developed Max, a software agent that uses information retrieval techniques and heuristic searches in information space to find information on the Internet that can motivate and promote serendipitous discoveries [Campos and Dias, 2001]. The mission of Max is to stimulate the user with unexpected information. They postulate six possible categories in the value of Max's suggestions:

- Category 1: the suggested pages were already known the suggestion has no value at all;
- Category 2: the suggested pages were not known and did not belong to any domain of interest, the suggestion has little value. Although not interesting at the present moment, the suggested pages may have some usefulness in the future (in the context of lateral thinking).
- Category 3: the suggested pages were not known, but belong to some domain of interest the suggestion has little value, since the user could have reached those pages otherwise (ex.: using a search engine);
- Category 4: the suggested pages were not expected, but they are slightly related to some domain of interest the suggestion is valuable, and it could hardly have been found by the user.
- Category 5: the suggested pages were not expected they did not belong to any current domain of interest of the user, but they sparked in him a new interest , the suggestion is extremely valuable since it is very improbable that the user would ever find the page on himself.
- Category 6: the suggested pages establish a new connection between two current domains of interest, the suggestion is extremely valuable (an insight occurred);

Another significant work about serendipity in the world of digital data and media is the one done by Beale [Beale, 2007], using artificial intelligence approaches to reduce the distance between the user interface and the system making it more natural for the user to be able to achieve their goals. Beale developed two differente systems which are able to, using appropriate technologies, keep the user at the centre of an interaction and still support them in making new discoveries in different ways, making for a more serendipitous environment.

For its part Ertzscheid and Gallezot (Ertzscheid and Gallezot, 2003) suggested serendipity is an increasingly valuable tool in helping in the search field to find information. Their starting point is the definition of serendipity as a leading model of access to information in the context of a search process or one of its iterations, considering that this model gives results that satisfy user´s needs in a casual way.

Serendipity can also occur during directed web search. Spink et al (Spink, 1998) suggest that results "partially relevant" may lead to the generation of new ideas and guidance for those seeking information. In this same field of Web search, Paul André and Teevan (André, 2009) conducted a study to determine if the results of web search results contain potentially serendipity. The participants were asked to categorize the results they obtained in their searches in two scales: Relevance (relevant, partially relevant and irrelevant), and degree of interest (Interesting, interesting part is not interesting.) The starting working hypothesis is that the results judged as not very relevant, but at least interesting, is an area where you can be serendipity. In addition to collecting relevant trials were also able to obtain additional information about each query and each result by examining search logs.

One of the areas where they have also developed different projects over the past years regarding the serendipity has been the Recommendation Systems (Recommender System, RS). Next we are going to expose some of these works of particular relevance.

In information recommendation, a critical aspect is to provide useful information to the user. These systems have, as a first objective, to provide personalized recommendations to improve user satisfaction.

In the last decade, several recommender systems have been developed and used in a variety of domains. As more and more are proposed recommendation techniques, researchers and practitioners face the problem of how to estimate or measure the value of the recommendations. From the perspective that good recommender systems recommend items that suit the preferences of the user, they are evaluated by using metrics related to accuracy. Two popular metrics for measuring accuracy are precision and recall, which were developed in the research area of information retrieval. In addition to precision and recall, various predictive accuracy metrics are used in recommender systems.

However, some recent papers argue that accuracy is not the only value that should be considered for the evaluation of recommender systems and that there are other important aspects to consider, and that assessments should focus on the future. The evaluation metrics for measuring user satisfaction move beyond accuracy to include novelty and serendipity

Computational Model

Although in recent years there have been efforts for implementing computer systems that directly support serendipity in search engines and information retrieval and also in recommendation systems, there has not been developed any computational model system for serendipity. The main cause of the lack of this model is one that, as pointed out Van Andel (Van Andel, 1994) the very fact of "program" serendipity invalidate the concept itself.

Another limitation detected by analyzing the scientific literature on this subject, is that in general, there has not been considered the dual nature of the concept of serendipity in the designing of these systems. The model we propose consider the concept from its duality, namely, first an unexpected event should occur, by chance or fluke, and also the necessary observer attitude that is necessary to make possible a discover of a new opportunity for generating knowledge.

Advances in information technology, especially in terms of the potential of current hardware and the new opportunities offered by the current in the field of human-machine interface, can face the design of this model from a new perspective that includes both the ability of the system itself to generate serendipity as the willingness of the user before it. It is from this perspective that arises from the computational model for serendipity.

As a first approximation to Serendipitous computational model, we describes the steps that should be reflected in to accommodate both the possibility of generation of serendipity as to the attitude that every time the person concerned to show that time that the system will offers. Therefore, the model should provide the opportunities that the computer system is implemented on he has the ability to present serendipitous results to requests for information made by the user, and the different answers you may give these results, the model must understand first the possibility that during the search for information arising actions not directly related to the exact terms and search domain user and, secondly, what action does the user to this new fact (not expected by him).

Before formulating the model, it should indicate that it should be collected in different situations about serendipity can occur. First, it may not occur either because the user serendipity is not interested in investigating the unexpected or because, even so, does not reach any valid conclusions (IF POSSIBLE, IF WORKABLE, undesirable outcome). In this frame of mind, you can find situations where the unexpected event

pseudoserendipia help solve the problem initially in the search (IF POSSIBLE, IF WORKABLE, undesirable outcome). Second and more interesting for what we are concerned, pure serendipity can occur if the act emerged by chance helps users find a solution to a problem that already existed but was not the target of their search or investigation of the discover so much a problem as a solution for the same had not been raised so far (IF POSSIBLE, IF WORKABLE, IF DESIRED).

1. Identifying a research problem.
2. Investigation of this problem: finding a solution.
3. In the research phase of a problem, the appearance of an unexpected event.
4. With the emergence of an unexpected event can take the following actions:
    A. Not investigate the surprising fact (IF POSSIBLE, IF WORKABLE, undesirable outcome). The user may not realize the importance that presents the surprising fact, not investigate. In this case there will be no serendipity. However, there are two situations:
        - Discarded, which implies that the user does not forget and we can be of any use in subsequent investigations (IF POSSIBLE, IF WORKABLE, undesirable outcome).
        - Record it, which that user or another we can be of any use in subsequent work. (IF POSSIBLE, IFWORKABLE, undesirable outcome).
    B. Investigate the surprising fact. In this case the user is aware of the importance of the event and therefore the need to investigate (IF POSSIBLE, IF WORKABLE, undesirable outcome). In this case there are alternatives:
        - You can investigate and do not come to any conclusion. In this case, there is no serendipity. These findings can be recorded and may be used for future research. In any case, will become part of the "background" of the user to possible future serendipitous discoveries (IF POSSIBLE, IF WORKABLE, undesirable outcome.
        - You can investigate and come to the conclusion that it is the solution.
            i. the problem he was studying, in which case it serendipity, but pseudoserendipity (IF POSSIBLE, IF WORKABLE, undesirable outcome)
            ii. a problem existed but that he was not studying at that time. In this case, it is Serendipity (IF POSSIBLE, WHICH CAN BE MADE IF, IF DESIRED)
            iii. that is an important discovery. It has just discovered a problem and its solution. Again, this is a case of serendipity (IF POSSIBLE, IF WORKABLE, IF DESIRED)
5. Knowledge generation and use.

Figure 1 shows the overall system architecture in which you can see the different modules that compose it. As shown in the figure, the user is logged by a search request on a topic of interest to him in a particular domain. The system collects the search significant tokens and sends them to a module that is responsible for transforming these tokens in serendipitous tokens enriched. This transformation is done using a neural network responsible for obtaining new words based on the concept of semantic distance. To develop the architecture of this network model we start from the SOM (Self-Organizing Maps) [Kohonen, 2002], [Somervuo, 2004] in which instead of syntactic distances between neurons arise semantic distances.

The tokens obtained in the neural network, process manager searches in launching these tokens in different search engines. Documents obtained from these searches are filtered and categorized by whether they belong to the initial search domain or a different one. These documents are presented to the user making an assessment based on the interest and use that report, both in relation to your original search with other search domains that had not initially plant.

This user interaction the system will learn to guide the user in subsequent searches.

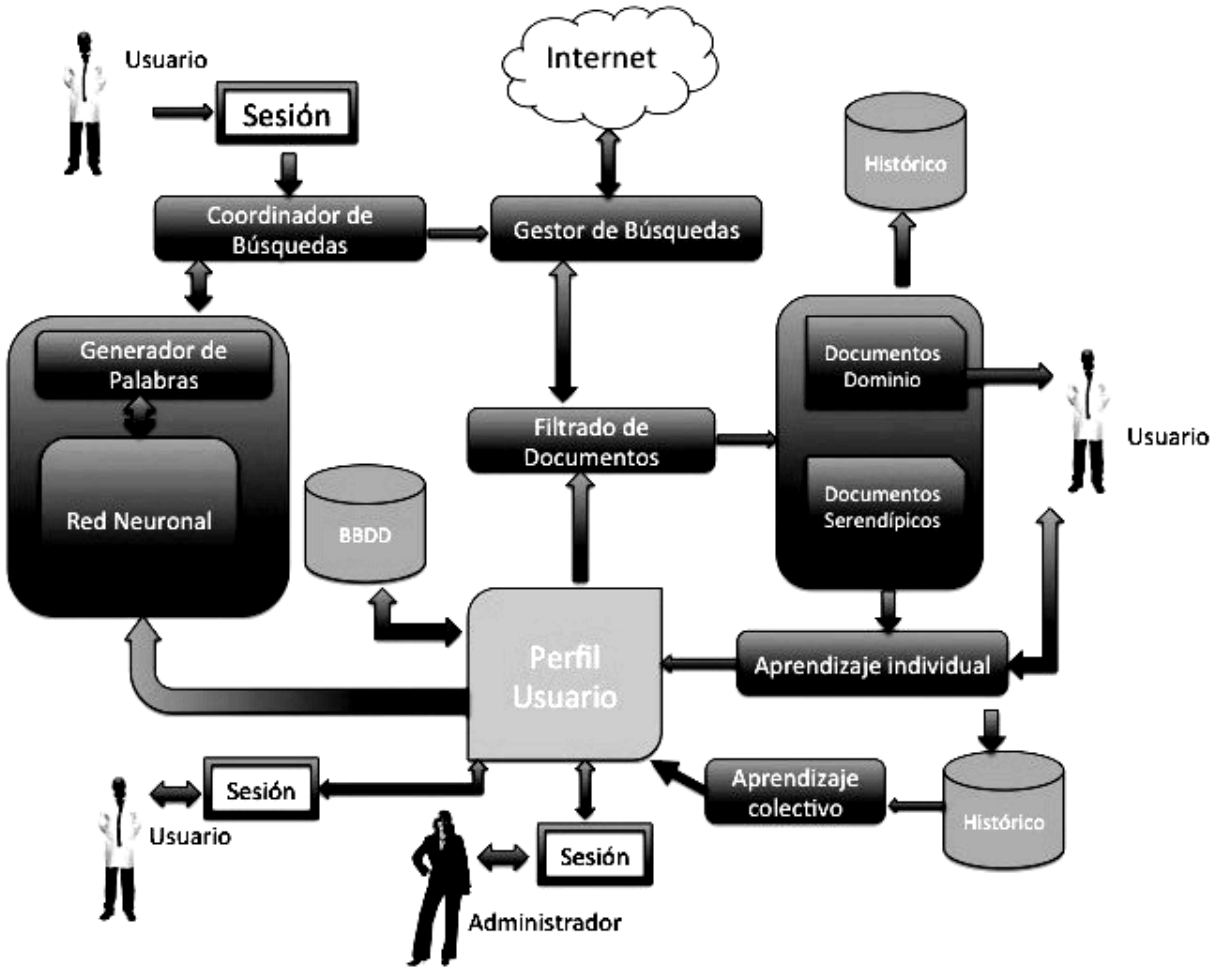


Figura 1: Modelo arquitectonico del sistema

Conclusions

The proposed model allows us to consider new lines of research in relation to the proposed neural network to generate the tokens Serendipic as regards the integration of serendipity in the various current systems in Internet search.

The system allows different forms of serendipity collected in the field equations and Dias, as the user's attitude to what the system offers and the final result is obtained.

The model provides a new way to treat the phenomenon of serendipity computationally integrating the two parts of the same: the emergence of a surprising fact (which provides serendipity generation module) and facilitates the user's predisposition to this fact

Bibliography

[André I, 2009] Paul André, Teevan J, Dumais Susan , Discovery Is Never by Chance: Designing for (Un)Serendipity. C&C09 Octubre 26-30, 2009, Berkeley, California 305-314

[Beale, 2007] Beale, Russell: Supporting serendipity: Using ambient intelligence to augment user exploration for data mining and web browsing. International Journal of Human-Computer Studies. 5-mayo-2007, pages 421-433

[Campos I, 2001] Campos, J. y Dias de Figuereido, A. 2001. The Serendipity Equations. Proceedings of the Workshop Program at the Fourth International Conference on Case-Based Reasoning., ICCBR 2001, Technical Note AIC-01-003. Washinton, DC: Naval Research Laboratory, Navy Center for Applied Research in Artificial Intelligence.

[Campos II, 2001] Campos y Dias. Searching the Unsearchable: Inducing Serendipitous Insights. 2001

[Ertzscheid, 2003] Ertzscheid, O. y Gallezot, G. Buscar falsedades y encontrar exactitudes, serendipia y búsqueda de información. CIFSIC Bucarest 2003. Comunicación y complejidad. 2003

[Kohonen, 2002] Kohonen, T., Somervuo, P.J.: How to make large self-organizing maps for nonvectorial data. Neural Networks. Vol. 15. Pp: 945-952.

[Somervuo, 2004] Somervuo, P.J.: Online algorithm for the self-organizing map of symbol strings. Neural Networks. Vol. 17, issues: 8-9. Pp: 1231-1239.

[Spink 1998] Spink, A., Greisdorf, H. & Bateman, J. From highly relevant to no relevant: Examining different regions of relevance. IP&M 34,5 (1998), 599-621

[Van Andel, 1994] Van Andel, Pek. "Anatomy of the unsought finding: serendipity: origin, history, domains, traditions, appearances, patterns and programmability". British Journal for the Philosophy of Science 45 (2): 631–648. 1994

Authors' Information

Aurea Anguera de Sojo Hernández- *Associate professor U.P.M. Crtra Valencia km 7, Madrid 28031, Spain;*

e-mail: aanguera@eui.upm.es

Miguel Angel Díaz Martínez- *Associate professor U.P.M. Crtra Valencia km 7, Madrid 28031, Spain;*

e-mail:mdiaz@eui.upm.es

Abraham Gutiérrez Rodríguez- *Asocciate professor U.P.M. Crtra Valencia km 7, Madrid 28031, Spain;*

e-mail:abraham@eui.upm.es

TWO APPROACHES TO ONE OPTIMIZATION PROBLEM IN RECOGNITION THEORY

Nataliya Katerinochkina

Abstract*: In recognition theory a number of optimization problems appear. The search for the maximum solvable subsystem of the system of linear inequalities is one of these tasks. In this paper two approximation solution methods for given problem are proposed. The first method is based on the theory of nodal solutions of systems of linear inequalities. But the second algorithm principally differs from the first: it is based on some reasoning of geometric nature. The comparison of the represented methods is produced in the number of problems.*

Keywords: *recognition, optimization, system of linear inequalities, nodal solution*

ACM Classification Keywords: *G.1.6 Optimization – Constrained Optimization, G.1.3 Numerical Linear Algebra – Linear Systems, I.5 Pattern Recognition*

Introduction

The solution of optimization problems in recognition theory is one of the most important stages in the synthesis of the high-precision algorithms of classification, recognition and forecast. The following optimization problems are here central: search for the maximum solvable subsystems of systems of inequalities, search for the minimum coverings of matrices, synthesis of the minimum formulas for realization of weakly determined Boolean and finite-valued functions, constructing the algebraic correctors of minimum degree, and others.

During the optimization of some models of recognition algorithms, e.g., the estimate-calculating algorithms (ECAs), a specific system of conditions is constructed. This system is described by the large number of linear inequalities and, as a whole, can be contradictory. It is necessary satisfy these conditions as precisely as possible. In the general case, the solution of this problem is reduced to the search for the maximum (according to the number of inequalities) solvable subsystem of the assigned system of linear inequalities.

In some cases, additional requirements are imposed on the desired solvable subsystem. For example, the given system of linear inequalities may be broken into blocks (subsystems) of identical power. It is necessary to find a solvable subsystem that contains the maximum number of complete blocks.

In other cases, we need to find the maximum solvable subsystem that contains the fixed inequalities of the assigned system. This task appears during the dynamic completion of the training sample.

Formulation of the problem

In our paper, we consider the general task of the search for the maximum solvable subsystem (MSS) of the system of linear inequalities. Assume that an inconsistent system of linear inequalities in the following form (1) is assigned:

$$\sum_{j=1}^{n} a_{ij} x_j \le b_i, \quad i = 1,\dots,m. \tag{1}$$

It is necessary to find the MSS of this system.

One approach to the solution of this problem is known (see [Chernikov, 1968], chapter 5). This method uses the convolution of the system of linear inequalities. However, it is unfit for the tasks of large dimensionality, since it requires the rapidly growing memory capacity.

Earlier the author constructed the precise solution algorithm for this problem (see [Katerinochkina, 2005]). This algorithm is effective for linear systems of small ranks with a large number of inequalities. Furthermore, the modifications of general method for some types of tasks with additional requirements on the desired solvable subsystem were developed. However, the precise algorithms of solution of such problems make large exhaustive search. Therefore, the rapid approximation methods should be developed for the tasks of large dimensionality.

In this paper, the author proposed two approaches to the solution of the formulated problem. The first approach is based on examining a special set of subsystems of the assigned system of linear inequalities. The second method uses some reasoning of geometric nature. Descriptions of both methods are given, and the comparison of their work in the number of tasks also is produced.

Description of the first method

Suppose that an unsolvable system S of form (1) is given.

DEFINITION 1. We say that a solvable subsystem P of system S is a *not extensible, solvable subsystem*, if the addition to P of any (not entering it) inequality of system S converts it into an inconsistent system.

The following assertion is valid.

LEMMA 1. Assume that the inconsistent system S has a rank $r, r > 0$. Then the rank of any not extensible, solvable subsystem of system S is equal to r.

DEFINITION 2. A subsystem whose power and rank are equal to r is called *an r-subsystem.*

Following [Chernikov, 1968], let us introduce a number of concepts necessary for further reasoning. Let us examine system P of form (1). We assume that system P is solvable and has a rank r different from zero.

DEFINITION 3. We call the solution of system P *a nodal solution*, if it turns into the equalities some r of its inequalities with linearly independent left sides.

DEFINITION 4. The subsystem of system P is called *a nodal subsystem*, if its rank is equal to the number of inequalities in it and all of its nodal solutions satisfy system P.

In [Chernikov, 1968] the following theorem is proven.

THEOREM 1.1. Each solvable system of the linear inequalities of form (1) with rank $r, r > 0$, has at least one nodal subsystem, and, hence, at least one nodal solution (each nodal subsystem is an r-subsystem).

Examine now an inconsistent system S in the form (1). Lemma 2 follows from Lemma 1 and Theorem 1.1.

LEMMA 2. Let the inconsistent system S have the rank r, $r > 0$. Then each not extensible, solvable subsystem of system S has at least one nodal r - subsystem.

Therefore, for the selection of all not extensible, solvable subsystems of system S, it is sufficient to examine all of its r -subsystems. For each of them, it is required to solve a simple task, i.e., to find one nodal solution. For this purpose, it is necessary to replace all inequality signs in a subsystem with the equalities signs and find one solution of the obtained system of linear equations. The obtained nodal solution must be substituted into all inequalities of system S, and the inequalities satisfied with this solution must be selected. The chosen inequalities form a solvable subsystem. After exhaustion of r -subsystems, we will obtain a certain set that contains all not extensible, solvable subsystems of system S. Among them, it is possible to select optimum subsystems with the required properties. Maximum subsystem (according to the power) among the chosen subsystems is the MSS of system S.

The described method combines finding the MSS with the simultaneous presence of one solution, which is the final goal of optimization problem.

However, the complete exhaustion of r -subsystems will be too long for tasks of large dimensionality. Therefore, a number of approximations methods have been developed.

In this paper, the approximate algorithm $\mathbf{A_w}$ of solution of the formulated problem is presented. This algorithm realizes a partial directed search of r -subsystems of the assigned system.

General work scheme of algorithm $\mathbf{A_w}$. First, we transform system S of form (1) as follows. We normalize the inequalities of the system so that the modulus of the first nonzero coefficient in each inequality would be equal to 1 (thus, we do not change the signs). Then, we place all inequalities in order of increasing of right sides and, afterwards, we examine inequalities in this order.

Algorithm $\mathbf{A_w}$ has several iterations. At any iteration, we first choose the so-called "base r -subsystem". Then, a set of associated subsystems is constructed by replacing some inequalities in the base r -subsystem. We apply the construction procedure of *the extended solvable subsystem* (ESS) to the base subsystem and associated r -subsystems; at the same time, we find one solution for each ESS. The maximum subsystem is selected from the obtained ESSs. The power of the latter subsystem is compared with value ρ, *i.e.,* the record power of the ESS obtained at previous iterations.

As a result, after all iterations, the record value of power ρ and the corresponding solution $\mathbf{x}$ are remembered. Maximum ESS, which is considered to be an approximation of the MSS of system S, is restored by this solution.

Let us now describe the procedures mentioned in the general work scheme of the algorithm.

Construction of base r –subsystems. Algorithm $\mathbf{A_w}$ has a parameter w that determines the maximum power of the intersection of base subsystems. If $r = 1$, we set $w = 0$; if $r = 2$, then $w = 1$. Generally, at $r > 2$, values of parameter w fall in following boundaries: $0 \le w \le r - 3$.

The first r (in the indicated order) inequalities with the independent left sides enter into the first base subsystem. For constructing the next base subsystem, we move the first $r - w$ inequalities, which entered into the previous

base subsystem, away from system S. We obtain a reduced system of linear inequalities, from which we again select the first r independent inequalities. We continue this process until it is possible to choose an r-subsystem from the reduced system. The total number b of base subsystems is estimated as follows:

$$b \le \left[\frac{m-r}{r-w} \right] + 1 .$$

Construction of the extended solvable subsystem (ESS). This procedure can be applied to any r-subsystem of system S. Let the r-subsystem B be given. First, we find the nodal solution of B. For this, we replace all inequality signs in the subsystem by the equality signs and find the solution of the obtained system of linear equations. We find only one solution $\mathbf{x}$; i.e., at $r < n$, we select one solution from the set. We then substitute the obtained solution $\mathbf{x}$ in all inequalities of system S. Also, we separate the inequalities satisfied by $\mathbf{x}$; these inequalities form the ESS. We remember the power ρ of the obtained ESS and the corresponding solution $\mathbf{x}$.

Construction of the set of subsystems associated with the base subsystem. The construction of associated subsystems for the base subsystem is tree-like process. Assume that, on some iteration, a base subsystem B is built. Let us apply the procedure for constructing an ESS. Thus, its nodal solution $\mathbf{x}$ is found and subsystem $R(B)$ (extension of B) with power ρ is built. Let us select the inequalities from system S that $\mathbf{x}$ does not satisfy. The set of these inequalities is denoted by $\overline{R}$.

STAGE 1. First we construct subsystems associated with the base r-subsystem B. These subsystems are subsystems of the first level. For this, we replace each of the r inequalities of system B with each of the $m-\rho$ inequalities from the set $\overline{R}$ in turn. We will obtain $r(m-\rho)$ subsystems associated with B; of these only r-subsystems are left. To each received r-subsystem B_i, we shall apply the procedure for construction the ESS. As a result, the extended solvable subsystem $R(B_i)$ with power ρ_i will be found.

DEFINITION 5. We call an r-subsystem B_i, which is associated with the base subsystem B, *the point of increase*, if $\rho_i > \rho$.

STAGE 2. Following cases are possible:

1) Among subsystems associated with B, there are no points of increase. Then, this branch of the process is broken and the transition to the next base subsystem is produced.

2) Points of increase exist and, among them, there is one maximum; i.e., for some r-subsystem B_t, we have $\rho_t > \rho$, and $\rho_t > \rho_i$ for all $i \neq t$. Then, for subsystem B_t, the associated subsystems are constructed exactly as was done for base subsystem B, which are subsystems of second level.

3) There are several points of the maximum increase, including subsystems $B_{j_1},...,B_{j_l}$, for which will be $\rho_{j_1},...,\rho_{j_l}$. In this case, we investigate these subsystems one by one until we find the point of increase for one of them.

Then we pass to the third stage, namely, we construct, as are above, the subsystems of third level for the maximum point of increase. We do not return to the remaining subsystems of the second level.

Let us put the following limitation on number u of the levels of this process:

$$u \le \left]\frac{r-w}{2}\right[- 1 .$$

This is the result of the requirement that the associated subsystems for the neighboring base subsystems not coincide.

Estimation of the algorithm complexity. After the end of all iterations of algorithm $\mathbf{A_w}$, we will obtain the record value ρ for the power of all constructed ESSs and the corresponding solution $\mathbf{x}$. The maximum ESS is restored based on this solution. This ESS is considered to be the approximation for MSS of system S.

Let us estimate the number q of all considered subsystems of power r. For one examined r-subsystem B, $q(B)$ subsystems are investigated, where

$$q(B) \le r^2(m-\rho)^2 u .$$

Further, we have:

$$q \le q(B)b \le r^2(m-r)^2 bu \le \frac{r^2(m-r)^3}{2},$$

where value b is the number of base subsystems.

This estimation is overstated; the number of the subsystems examined will be much lower in practice.

Description of the second method

This method is principally different from those mentioned above. It is based on some reasoning of geometric nature.

Description of the overall algorithm scheme. For the brevity we will sometimes write system S in the following form:

$$l_i(\mathbf{x}) \le b_i, \quad i = 1,...,m, \tag{2}$$

where $l_i(\mathbf{x}) = (\mathbf{a}_i, \mathbf{x}) \equiv \sum_{j=1}^{n} a_{ij} x_j$.

DEFINITION 5. The inequality $(\mathbf{a}, \mathbf{x}) \le b$ is called *improper*, if vector $\mathbf{a}$ is equal to zero.

Depending on the sign of value b, improper inequality is either contradictory or identical.

It is obvious that the improper inequalities can be not considered in the process of construction of the MSS of a system. Contradictory inequalities will enter not into one subsystem, but identical inequalities can be added to any solvable subsystem.

The following assertion is known (for example, see [Eremin, 2007]).

ASSERTION 1. Any inconsistent system of linear inequalities that does not contain improper inequalities can be subdivided into two solvable subsystems.

Actually, let us divide system (2) into two subsystems as follows.

Denote by M_+ the set of those indices $i \in \{1,\ldots,m\}$, for which the first of the nontrivial coefficients $a_{i1},\ldots,a_{in}$ is positive. Put also $M_- = \{1,\ldots,m\} \setminus M_+$. Then system (2) is decomposed into two subsystems:

$$l_i(\mathbf{x}) \le b_i, \quad i \in M_+; \tag{2.1}$$

$$l_i(\mathbf{x}) \le b_i, \quad i \in M_-. \tag{2.2}$$

It is easy to show that these subsystems are solvable and to find their particular solutions.

Let us describe the work scheme of the search algorithm for the MSS of system S in form (2).

STAGE 0. First we divide the assigned system into two solvable subsystems (2.1) and (2.2).

For each of these subsystems we find one particular solution. Let us denote them by $\mathbf{y}$ and $\mathbf{z}$ respectively.

Then cycle on i, $i = 1,\ldots,m$, is produced. In this cycle the i-th iteration consists of several stages.

STAGE i1. Let us note that the hyperplane $l_i(\mathbf{x}) = b_i$ corresponds to the inequality $l_i(\mathbf{x}) \le b_i$ of the given system. First we construct the points $p(\mathbf{y},i)$ and $p(\mathbf{z},i)$ that are the projections of points $\mathbf{y}$ and $\mathbf{z}$ on the hyperplane $l_i(\mathbf{x}) = b_i$.

STAGE i2. For the straight line passing through points $p(\mathbf{y},i)$ and $p(\mathbf{z},i)$, we find the points of its intersection with all remaining hyperplanes: $l_j(\mathbf{x}) = b_j$, $j = 1,\ldots,m$, $j \ne i$. Let these be the points $\mathbf{u}^1,\ldots,\mathbf{u}^l$, $(l \le m-1)$.

Stage i3. Then we substitute each obtained point $\mathbf{u}^j, 1 \le j \le l$, in all inequalities of system (2) and calculate the number of inequalities that this point satisfies. These inequalities obviously form the solvable subsystem. We compare the power of this subsystem with the record power obtained at the previous steps. In this case, we memorize the value of the record power ρ of solvable subsystem and the corresponding point $\mathbf{v}$.

As a result after all iterations, the record value of power ρ and the corresponding solution $\mathbf{v}$ are memorized. On this solution the solvable subsystem is restored. This subsystem is considered as the approximation for the MSS of system S.

Let us estimate the exponent of operations of the described algorithm. Algorithm contains m iterations, for each of them we find not more than m points. We substitute each point in m inequalities from n variables. In all, we obtain the order $m^3 n$ operations.

About the parameters of the represented algorithm. In the initial stage, system (2) is divided into two solvable subsystems determined above. For each of these subsystems it is necessary to find one particular solution. It is obvious that the particular solutions of subsystems (2.1) and (2.2) are built not uniquely.

Consider the coefficients matrix A_+ of system (2.1). By the transposition of its lines it is possible to lead it to the stepped form. In the obtained matrix the first nontrivial column of each step consists of the positive elements. Let the reduced matrix have k of steps. Let us denote by i_j, $j = 1,\ldots,k$, the number of the first nontrivial column of the j-th step and by l_j denote the number of the upper row of the j-th step. We will search for the particular solution of the subsystem, beginning from the inequalities that correspond to the k-th step of the given matrix.

Obviously, we can attach any values to variables $x_j, i_k < j \le n$. Then the subsystem, which corresponds to the k-th step, is led to the form:

$$a_{ji_k} x_{i_k} \le c_j,\ j = l_k, l_k + 1, \ldots, m, \quad a_{ji_k} > 0.$$

Then we can choose the value of variable x_{i_k} from the condition:

$$x_{i_k} \le \min_{l_k \le j \le m} \frac{c_j}{a_{ji_k}}.$$

Substituting the obtained values of variables into the previous inequalities, we find the values of remaining variables by analogy.

The particular solution of subsystem (2.2) is built according to the same scheme. Difference lies in the fact that the set of the values of the corresponding variable will be limited not on top but from below.

Thus, in the process of the search for the particular solutions we can attach any values to certain (free) variables and to others attach arbitrary values from the infinite set.

Hence such variables can be considered as the parameters. Varying these parameters, we will obtain different particular solutions.

It is possible to use this fact in the process of applying the represented method. Changing the relationships between the parameters, we can attempt to improve initial result. Thus, we can tune our algorithm to the task in the dialogue regime.

Comparison of two solution methods of the considered problem. We made a comparison of the represented methods (method 1 and method 2). About five ten tasks were solved by these methods. The identical results are obtained for 48% of these tasks. For 24% of tasks, method 1 gave the best result. For 28% of tasks, method 2 operated better.

Conclusion

In this paper, two search methods for the maximum solvable subsystem of the system of linear inequalities are proposed. The represented algorithms can be used together with others methods. Let us note that each of them can operate better than other on some tasks.

Acknowledgment

The paper is published with financial support by the project ITHEA XXI of the Institute of Information Theories and Applications FOI ITHEA (www.ithea.org) and the Association of Developers and Users of Intelligent Systems ADUIS Ukraine (www.aduis.com.ua).

Bibliography

[Chernikov, 1968] Chernikov S.N. Linear inequalities. - M.: Nauka. 1968. (In Russian)

[Katerinochkina, 2005] Katerinochkina N.N. The search methods for the optimum solvable subsystem of the system of linear inequalities // Mathematical methods of pattern recognition. Reports to the 12th All-Russian conference. 2005. Pp. 122-125. (In Russian)

[Katerinochkina, 2009] Katerinochkina N.N. Decision methods for some optimization problems in recognition theory // Pattern Recognition and Image Analysis, 2009, Vol. 19, No. 3, pp. 441-446.

[Eremin, 2007] Eremin I.I. Linear optimization and systems of linear inequalities. - M.: Publishing center “Academy”, 2007. – 249 p. (University textbook. Series “Applied mathematics and information theory”). (In Russian)

Authors' Information

Katerinochkina Nataliya *– senior researcher in Dorodnicyn Computing Centre of Russian Academy of Sciences, ul. Vavilova 40, Moscow, 119333, Russia; e-mail: nnkater@yandex.ru*

Major Fields of Scientific Research: discrete mathematics, mathematical cybernetics, pattern recognition, optimization.

TABLE OF CONTENTS