I T H E A

International Journal

INFORMATION THEORIES & APPLICATIONS

2004 Volume 11 Number 1

International Journal
INFORMATION THEORIES & APPLICATIONS
ISSN 1310-0513
Volume 11 / 2004, Number 1

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

IJ ITA welcomes scientific papers connected with any information theory or its application. Original and non-standard ideas will be published with preferences.

IJ ITA rules for preparing the manuscripts are compulsory.

The **rules for the papers** for IJ ITA as well as the **subscription fees** are given on *www.foibg.com/ijita*.

The camera-ready copy of the paper should be received by e-mail: foi@nlcv.net.

Responsibility for papers published in IJ ITA belongs to authors.

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

International Journal "INFORMATION THEORIES & APPLICATIONS" Vol.11, Number 1, 2004
Printed in Bulgaria
Edited by the **Institute of Information Theories and Applications FOI ITHEA**, Bulgaria, in collaboration with the V.M.Glushkov Institute of Cybernetics of NAS, Ukraine, and the Institute of Mathematics and Informatics, BAS, Bulgaria.
Publisher: FOI-COMMERCE - Sofia, 1000, P.O.B. 775, Bulgaria. www.foibg.com, e-mail: foi@nlcv.net

ISSN 1310-0513

Preface

Verba volant, scripta manent !

The "**International Journal on Information Theory and Applications**" (**IJ ITA**) has been established in 1993 as independent scientific printed and electronic media. IJ ITA is edited by the *Institute of Information Theories and Applications FOI ITHEA* in collaboration with the leading researchers from the Institute of Cybernetics "V.M.Glushkov", NASU (Ukraine) and Institute of Mathematics and Informatics, BAS (Bulgaria).

During the years, IJ ITA became as well-known international journal. Till now, including this volume, more than **350** papers from more than **600** authors have been published. IJ ITA authors are widespread in **38** countries all over the world: ***Armenia*, Belarus, Belgium, *Bulgaria*, *Canada*, *Czech Republic*, *Denmark*, Egypt, Estonia, Finland, France, *Germany*, *Greece*, *Hungary*, *Ireland*, *Israel*, *Italy*, *Japan*, *Kirghizia*, *Latvia*, Lithuania, *Malta*, Mexico, Moldavia, *Netherlands*, Poland, Portugal, *Romania*, *Russia*, Scotland, Senegal, *Serbia and Montenegro*, *Spain*, Sultanate of Oman, *Turkey*, *UK*, *Ukraine*, and *USA***.

Volume 11/2004 of the IJ ITA contains **59** papers written by **115** authors from **24** countries (*marked in italics above*), selected from several international conferences, seminars and workshops organized or supported by the Journal.

At the first place, the main source for selection were the **ITA 2004** Joint International Events on Information Theories and Applications, (June 14-24, 2004, Varna, Bulgaria):

- Second International Conference i.TECH 2004 "Information Research, Applications and Education",
- XXIX-th International Conference ICT&P 2004 "Information and Communication Technologies & Programming",
- III-th International Workshop on General Information Theory,
- International INTAS-FET Strategic Workshop "Data Flow Systems: Algorithms and Complexity",
- Second International Workshop on Multimedia Semantic.

A special issue (No.3) of this volume contains papers from the International Seminar "Digitization of Cultural and Scientific Heritage" (27.08 – 03.09.2004, Bansko, Bulgaria).

Several papers were selected from the pool of papers directly submitted to IJ ITA.

Congratulations to *Mr. Tibor Vámos* and *Mr. Boicho Kokinov* who were awarded by the International Prize "**ITHEA**" for the year 2004. The "ITHEA" Prize has been established in 1995. It is aimed to mark the achievements in the field of the information theories and applications.

More information about the IJ ITA rules for preparing and submitting the papers as well as how to take out a subscription to the Journal may be obtained from www.foibg.com/ijita.

Krassimir Markov
IJ ITA Founder and Editor in chief

International Prize "ITHEA"

Awarded Scientists till 2004:

1995	*Sandansky*	*K. Bankov, P. Barnev, G. Gargov, V. Gladun, R. Kirkova, S. Lazarov, S. Pironkov, V. Tomov*
1996	*Sofia*	*T. Hinova, K. Ivanova, I. Mitov, D. Shishkov, N. Vashchenko*
1997	*Yalta*	*Z. Rabinovich, V. Sgurev, A. Timofeev, A. Voloshin*
1998	*Sofia*	*V. Jotsov*
1999	*Sofia*	*L. Zainutdinova*
2000	*Varna*	*I. Arefiev, A. Palagin*
2001	*St.Peterburg*	*N. Ivanova, V. Koval*
2002	*Primorsko*	*A. Milani, M. Mintchev*
2003	*Varna*	*T. Gavrilova, A. Eskenazi, V. Lozovskiy, P. Stanchev*
2004	*Varna*	*B. Kokinov, T. Vamos*

IJ ITA major topics of interest include, but are not limited to:

INFORMATION THEORIES

Artificial Intelligence
- *Computer Intellectualisation*
- *Intelligent Networks and Agents*
- *Intelligent Technologies*
- *Knowledge Discovery and Engineering*
- *Knowledge Acquisition and Formation*
- *Distributed Artificial Intelligence*
- *Models of Plausible Reasoning*
- *AI Planning and Scheduling*

Bioinformatics
Business Informatics
Cognitive Science
Decision Making
Education Informatics
General Information Theory
Hyper Technologies
Information Models
Intellectualisation of Data Processing
Knowledge-based Society
Logical Inference
Natural language Processing
Neuroinformatics
Philosophy and Methodology of Informatics
Quality of the Programs
Software Engineering
Theory of Computation

APPLICATIONS

Business Information Systems
Communication Systems
Computer Art and Computer Music
Hyper Technologies
Intelligent Information Systems
Multimedia Systems
Programming Technologies
Program Systems with Artificial Intelligence
Pyramidal Information Systems
Very Large Information Spaces

COMPUTER DEMOCRACY – OUR NEXT STEP IN EUROPE

Tibor Vámos

Introduction

After about a quarter of a century of enlightened development and ongoing preparatory technological, scientific and political activities we are arrived at the realization period of the idea. The two major technological vehicles of progress are the World Wide Web, the most democratic international forum of information exchange and the advent of public key cryptography as a combined philosophical and practical device of individual integrity and collective responsibility.

The Two Major Technological Vehicles

A detailed explanation was given in detail in earlier papers and talks, several ones here in Bulgaria. A short summary to refresh your memory:

www:

- accessible all over the world, even if it is forbidden by (the authorities);
- difficult to trace the receiving user and even the dissemination origin;
- instant information regardless of distance;
- the creation of new groups within a global society;
- provides a stimulus for global standards of reasonable, acceptable, communication among different cultures, a real global human society.

Unfortunately as *Virtue* and *Evil* accompany all human related issues, we meet the *Evil* in:

- terrorism, crime, populist deceptive politicians, people spreading hate, misleading pseudo-science;
- an ocean of information without reliable and well oriented browsing facilities;
- the increasing orientation to business interests and not the original aim of free access to information.

To combat these problems, devoted professional people have developed new tools for elevating the *Virtues* and fighting against the Evil. We, the scientific brotherhood are active, too.

PKC: (public key cryptography)

- defends the individual against all kinds of mishandling of his/her personal data, ideas, views;
- elevates the responsibility and self-defence of the information issuers by preserving an unalterable document of the originally sent message;
- enables legally constituted public authorities to control malevolent information flow;
- enables all active constitutional, legal players in the information flow to control the legal conditions.

These two vehicles create not only technological tools for our global human efforts but a highly general instructive metaphor for future coexistence, the mutual responsibility of different people towards each other and their communities, i.e. a renewed and realistically establish-able *New Agora* in the Athenian, the first drafted but *never substantiated* democracy.

The Current State in Europe

I suppose that you have some more detailed surveys on the subject here, at this conference or at any other meeting devoted to the subject. The best references are collected at the special homepage of e-government. I strongly recommend subscribing to this and join the group of these benevolent people who try to digest an immense and almost impossible amount of information. I am not an agent of this group but for the third year a

happy consumer of their blessed activities and, organizational contacts. This helped us a lot in our successful joining process, which was celebrated on May 1 but prepared during the last quarter of a century.

Indicating the headlines only:

- The efforts are concentrated under the project name: E-2005

  that means a deadline –

  - for adopting standards for EU information; interoperation among the member countries. The IDA project of the EU is the main channel of communication, discussions and setting of standards based on these preparations;
  - for standards of realization on the principles reflected in the PKC ideology, i.e. protection of privacy balanced against the common interest of public democracy and defence of both the individual and communities against mistreatment by terrorism, fraud and political adventurers;
  - creating and distributing technology standards for all these purposes
  - helping people who are handicapped in education, social environment or some other condition to get appropriate support and equal opportunities.
- No EU country can serve as a general and completed standard due to the both highly different traditions of a democratic society and their technological status. In some countries, having a long positive experience in living in a people-serving and autonomous society, people see no problem with a more transparent system based primarily on a single natural identifier. This can be the normal data of domicile or birth register, too. In our kind of countries, people living a long time, i.e. centuries long, under foreign rule, consider the state authorities to be organizations against their civil interests, and traditionally regard underhand dealing as a national virtue. Therefore, they try to defend the individual at any cost against any imaginable governance intrusion.
- The interests of individual and community protection are different and their common view is relative to historical age and type of political system. We can state a positive viewpoint: in the last fifteen years these relations have changed a lot, sometimes also in the minds of the people, especially in that of the younger generations. Technology and experience of protection, and of course inimical actions, have changed even more.
- The EU boards have taken steps to reach a consensus, i.e. appropriate standards according to the technological and psychological possibilities, putting in effect our common constitutional principles. We can learn a lot using the American technology and legislation experience but Europe has its own nearly three millennia experience of its own to reach from, sometimes more tragic, sometimes more human.
- The EU Constitution has accepted currently sets out the general principles outlined in the PKC ideology. Relevant additions should be the separation of personal identification and other data, having a virtual envelop and opening operation. All actions should be registered in a no erasable and no alterable way, monitored by legally elected, independent, responsible bodies. All data unifications should be erased after the action, except the result and the record of the action. All kinds of these data procedures should be permitted by the individuals concerned and communicated to them but the actions of legal authorities (prosecutors and courts) should be carried out under well-controlled legal conditions (e.g. communication of the action only and not the result, time limits for action and secrecy, notification of people for whom the action and the data should be opened or closed).

The Hungarian Story and Experience

The present Hungarian practice is one of the most rigorous in Europe forbidding all kinds of data unifications except those based on prosecutor's or court's decisions. The previous system of a personal identifier (an 11 digit string composed of gender and birth data and a four digit zip code) was abolished though hidden in some way until now by certain authorities. A set of three different and not unified codes was legalized: one for domicile registry (2 characters, 6 digits), one for taxation (10 digits) and one for social security (9 digits). All these happened nearly fifteen years ago, immediately after the fall of the uncontrolled police control system and at a time of very low-level civilian computer usage. The international state of legislative and cryptographic practice

was lower by an order of magnitude and not only the US but the whole world lived before the drama of Sept. 11th and the massive experience of hacker and virus creating operations.

Electronic signature is generally not used though it is legislated. The reason, similar to the general European experience, is the exclusive financial condition: groups receiving the authorization power would like to receive high profits and for justification they started or demanded immense investments for explosion safe buildings, hardware and software systems, all separate for different purposes. The obsolete legislative situation and the particular interests of the different political groups and authorities supported these exaggerated demands.

We have now arrived at a point of almost general consensus for a revision of the early nineties' views and the introduction of current algorithmic software tools. Possessing an excellent school of algorithmic procedures and probability theory we are ready to create a highly safe system. I refer to the schools of Rényi (our academic institute of mathematics recently adopted the name of Rényi) and Erdős.

Politics and Science

Unfortunately, any kind of legislative action largely depends on mostly unintelligent, corrupt, malevolent, erratic politicians and their sycophants in dependent positions. In addition the situation in the daily press is submerging into a tabloid level, even the broadsheet newspapers are more and more interested in scandals and sensational news.

We proposed and partly realized a common effort of all sensible decision makers to unify our forces in a reasonable and given solution. Three branches of the government worked or shifted work and related financial responsibly to each other in the fuzzy channels of bureaucracy. The best educated and experienced, benevolent civil servants stood frustrated within the whirl of irresponsible politics.

The Academy of Sciences, being a partly independent and respectful body tries to convince the responsible decision making persons to consider national interest as a higher principle than their own financial and power involvement. The Committee, appointed by the President of the Academy includes outstanding personalities in the legal, social, computer related sciences, senior figures of our information history and the Ombudsman of data protection, who should be independent of political parties and elected by the Parliament. No active politician or government administrator is among the membership to maintain the Committee's political independence as a body. The Committee has no claim for any intervention but works with all state related people and organizations that are willing to do so.

I would like to mention that we found in a small minority of politicians several devoted and able people, who joined political groups in the hope of improving the regrettable situation. However, they all are subdued by the overpowering, negative influence of the more aggressive unscrupulous powers. These positive actors, sitting on both sides of the political divide welcomed the initiative of the Academy and are meaningfully cooperating with us.

We have had to experience the disastrous influence of political splits in relevant non-political problems and the dysfunctional organization of the political system, in its personal selection constraints and in overburdening practice, extending political and administrative activity far beyond the really necessary principal tasks. The operation stimulated thinking about the revision of state administration practice, returning to much older ideas of democratic and professional governance by adoption of both new tasks and technologies.

According to our observations, similar problems arose in every developed country and organization, even in multinationals and other international bodies. Thus the problem is less an issue of unrealistic ethical philosophical judgement but much more a social, cultural and organizational issue, i.e. an information science related question of our age and our intellectual communities.

A consensus of relevant thought in the legal profession has now been reached. Those who were pioneers of our present democratic constitutional order advocate the need for rational revision and that provides additional support of the need, as a priority, professional quality in all public affairs. We refer to the great Greek thinkers on city-governance (πολιτεια) especially *Aristotle* and the funeral speech of *Pericles*, reported later by *Thucydides* and to the *Founding Fathers* of the US through, their essays and papers in the *Federalist*. From the 19th century we have also had a wonderful tradition in Hungarian history, starting with the *Sage of the Country by,* Ferenc Deák.

The Proposal of the Hungarian Academic Committee

The proposal is clear:

- for the equal opportunity of citizens the right for electronic signature on an equitable basis, i.e.
  - it should be given free of charge for those whom it is a financial burden and not expensive for anybody. (.e.g. in relation to the taxation system)
  - electronic signature should be the only required authorization for any kind of public activity. If possible, this should be extended to banking operations, too;
  - all public authorities should participate in the popularisation, education and, training of different layers of society for usage and for being conscious of one's rights;
  - the state and, all accountable public authorities related to the electronic signature issue should be responsible for the preservation of the Civil Rights of individual citizens and any of their respective legal groups.

The measures are detailed above and should follow the agreements of the EU. EU conformity is the basis of interoperation and is a constitutional requirement. According to our legal experts this requires no fundamental change in our legal system, only some further updating and corrected interpretations, and the constitutional empowerments for participation in the EU.

- Technological means should not be included in the legal regulations, the system must be flexibly open for any kind of realization, i.e. currently traditional authorized handwriting, smart card, SIM-card used in mobile systems, biometrical (fingerprint, fundus, DNS, etc.) data.
- The Law should take care of independent and open operational authorities prescribing algorithms, the code length for citizens and prosecution and, other safety conditions related to data and their handling personal.

Going Together – Neumann and Athanasoff – Iliev

Bulgaria and Hungary have much common ties in our history, beginning with the Huns for those who believe the Hungarians are the successors of Attila and the ancient Bulgars who are really supposed to be the descendants, with lesser and greater Byzantine influence, with the tragedy of a certain city called Varna in 1444, with Turkish, German and Russian domination but most important of all should be the future, based on another lesson: of Neumann and Athanasoff.

Both were pioneers of the computer age, Neumann in mathematical and logic theory, Athanasoff more in technology. Neumann had to leave his country to avoid being a victim of the Holocaust, Athanasoff's family left for a better life, both, subsequently, had more possibility to develop their genius.

Now we enter a new age, based on our common three millennia old European history and, hopefully, our talent find a home within a more peaceful, less hatred-contaminated world, preparing a common home for our descendants. I remember here my friend Lubomir Iliev who passed away not too long ago and was not only a great mathematician and teacher of computer science but, at the same time, a representative of European cultural tradition and values. We always considered the two be inseparable by regarding these subjects as both metaphors and parallel realities.

These are the main lessons of that progress: preserving individual values within a cooperative, empathy driven human community. Let us hope that this comes true!

Author Information

Tibor Vámos – Computer and Automation Research Institute, Hungarian Academy of Sciences, H-1111 Budapest, Lágymányosi u. 11, Hungary; e-mail: vamos@sztaki.hu

TOWARDS ACTIVE VISION IN THE DUAL COGNITIVE ARCHITECTURE

Adrian Nestor and Boicho Kokinov

Abstract: The paper describes an extension of the cognitive architecture DUAL with a model of visual attention and perception. The goal of this attempt is to account for the construction and the categorization of object and scene representations derived from visual stimuli in the TextWorld microdomain. Low-level parallel computations are combined with an active serial deployment of visual attention enabling the construction of abstract symbolic representations. A limited-capacity short-term visual store holding information across attention shifts forms the core of the model interfacing between the low-level representation of the stimulus and DUAL's semantic memory. The model is validated by comparing the results of a simulation with real data from an eye movement experiment with human subjects.

Keywords: active vision, cognitive architecture, eye movements, structural descriptions, visual attention.

Introduction

Traditional cognitive models typically isolate a small piece of a seemingly cognitive process and simulate it. However, in all domains of cognitive modeling as well as in the field of applied cognitive systems research there is a new trend towards integrating various pieces of cognitive processes and even whole processes. The current paper presents an attempt to integrate visual perception with higher-level cognitive processes like thinking and memory. The traditional goal of computer vision was to infer the structure of a three-dimensional world out of two-dimensional images with an emphasis on the lower levels of visual processing [Marr, 1982]. Automating object and scene recognition would be the next step towards an integrated visual system and, more generally, towards an integrated cognitive system.

If one gives proper credit to the claim that high-level perception, delivers the representations which form the raw material for thinking, reasoning or decision-making, then it seems critical to understand the process of constructing and making available such representations. The need to justify the format and the availability of input representations was one of the most powerful criticisms directed against traditional AI [Chalmers, French and Hofstadter, 1992]. On the other hand, integrating high-level vision with cognition in total disregard of lower-level visual processing is not necessarily a step forward. Going all the way from the visual input to high-level cognition may be the right approach but one which may be still hardly within the researchers' grasp today.

One of the most obvious difficulties with this approach may be the need to use expertise from very different fields of research. The design of cognitive architectures drawing on the joint efforts of many experts from different fields is an enterprise particularly suitable for dealing with such a difficulty. Recent attempts in this direction may be noticed. The cognitive architecture ACT-R [Anderson and Lebiere, 1998] has been lately enriched with a model of eye movements [Salvucci, 2001]. Another cognitive architecture, EPIC [Meyer and Kieras, 1997], was designed and centered on the perceptual processing of stimuli in different modalities. A similar attempt is presented below towards integrating the cognitive architecture DUAL [Kokinov, 1994, 1997; Petrov and Kokinov, 1999] with a model of visual attention and perception. Beyond DUAL's need to justify its input and to construct its own perceptual representations, this attempt is also motivated by the search for a more principled approach to modeling visual perception, an approach more faithful to the sequencing of stages and the organization of the human visual system.

Another difficulty familiar to any vision researcher is the complexity of the visual information available at the front end of the system. One way to cope with this complexity is to confine the range of possible stimuli to a predefined type. The use of microdomains with simple predefined stimuli and rules is the modeling counterpart of this experimental practice. The TextWorld microdomain, a microworld made up of blocks of text, is our candidate for this role.

However, the use of microdomains might not be enough to deal with this complexity. *Active vision,* a concept proposed by computer vision theorists in order to surpass the limitations of the image-based approach advocated

by Marr [Marr, 1982] could be a realistic answer to this problem. Allowing vision to selectively attend and process parts or aspects of the available information instead of massively storing and processing all information present on the retina is not only a way to ignore irrelevant stimuli but also a necessity for a limited-capacity processing and memory system. Thus, rather than making and working on an internal copy of the outer world, an active vision system will tend to use the 'world as its own memory' [O'Regan, 1992] accessed according to the needs and the goals of the system. The serial deployment of attention and its visible counterpart, eye movements, is the way humans instantiate this principle. The model described below embodies the idea of active vision by modeling attention shifts and conditioning high-level processing of a stimulus by the availability of attention. Additionally, this offers the possibility to compare directly the performance of the model against eye movement data obtained from human subjects.

DUAL and Visual Processing

DUAL is a general cognitive architecture designed to provide a basis for modeling high-level context-sensitive cognitive processes. Although accounting for perceptual processes did not form a part of the initial motivation for its construction, a series of features characteristic to this architecture make it suitable and challenging as a framework in which to cast a model of visual perception and attention.

Hybridity in the form of a mixture of symbolic and neural network mechanisms and computations is perhaps DUAL's most significant feature. At the lowest level DUAL may be described as a large collection of units, called *DUAL agents,* reminiscent of Minsky's [Minsky, 1986] 'society of mind'. Each of these agents can be described by the symbol it stands for and by its level of activation. Agents communicate with each other both by sending symbolic messages and by spreading activation via weighted links. Coalitions of agents representing events, situations or objects tend to form themselves dynamically based on the level of activation and the links connecting a given set of agents. Finally, the set of all active units at a time may be described as the working memory of the system while the set of all units forms DUAL's long-term memory.

While hybridity has been successfully used in modeling high level-cognitive processes like analogical reasoning [Kokinov and Petrov, 1997] one can hardly imagine a domain which is more in need of such hybrid resources than visual processing. On the one hand, most of the image-based processes, e.g. the computation of a salience map [Itti and Koch, 2000], seem to be most appropriately described as the result of massively parallel numerical computations. On the other hand, perceptual primitives [Marr, 1982] and the classical structural description approach to object recognition [Biederman, 1987; Marr and Nishihara, 1978] encourage the appeal to symbolic representations and computations. In an attempt to connect a raw visual input with DUAL's semantic memory we take advantage of this hybridity by combining massively parallel activation-based computations with a serial attention-based symbolic processing mechanism instantiating the principle of active vision.

Another related point regards the way the relationship between symbolic and numerical processing is conceived of. Embracing the idea of a symbolic processor as an engine running on connectionist energy, DUAL conditions the possibility and the speed of symbolic computations on the activation level of its agents [Petrov and Kokinov, 1999]. Thus, a critical aspect in the functioning of the system is the identification of an initial source of energy or activation which enables the system to start and keep running. The current model elaborates on this topic by generalizing the concept of visual attention as a resource to be allocated [6]. Visual attention in our account is a limited and carefully managed source of energy selectively allocated to some part of the available information enabling its detailed symbolic processing. Moreover, attention spreads activation in the entire system playing the role of DUAL's energy source.

A feature which distinguishes DUAL from other architectures, typically production-rule architectures like ACT-R [Anderson and Lebiere, 1998] or EPIC [Meyer and Kieras, 1997], is the lack of a central mechanism controlling the functioning of the entire system. In DUAL each agent runs independently and in parallel with other agents using only local information obtained from its immediate neighbors. Elaborating a visual processing mechanism in the frame of a decentralized system like DUAL is surely a challenging task and a new manner of approaching vision in a cognitive architecture.

Finally, one of the most important principles underlying DUAL's development is the search for a less modular account of cognitive processes typically studied independently as part of different fields of research. Integrating memory and analogical reasoning in DUAL [Kokinov and Petrov, 1997] is one notable achievement in this

direction. In the current model of visual processing we take a step forward in the same direction by describing perception as an interactive process, driven both by a low-level raw visual stimulus and by the current state and contents of DUAL's memory. Moreover, we explain and implement the categorization-based stage in object and scene recognition as a form of automatic analogy-making adding further grounds for the claim that high-level vision and analogy are at their core one and the same process [Chalmers et al., 1992].

The Model

The structure of the model is sketchily depicted in Fig 1. The visual input corresponding to a TextWorld stimulus is presented on a two-dimensional visual array representing the front end of the system. Perceptual primitives like blobs and terminations are immediately generated by cheap parallel computations. Attention is con trolled at a time by an object which allocates it selectively to some area of the stimulus. A detailed symbolic representation is constructed for this area which tends to fade away as attention is withdrawn from it and allocated to another one. Categorization takes place for the visual memory contents by retrieving and mapping object and scene categories from DUAL's semantic memory onto current visual memory representations. Each of these processes will be briefly described below.

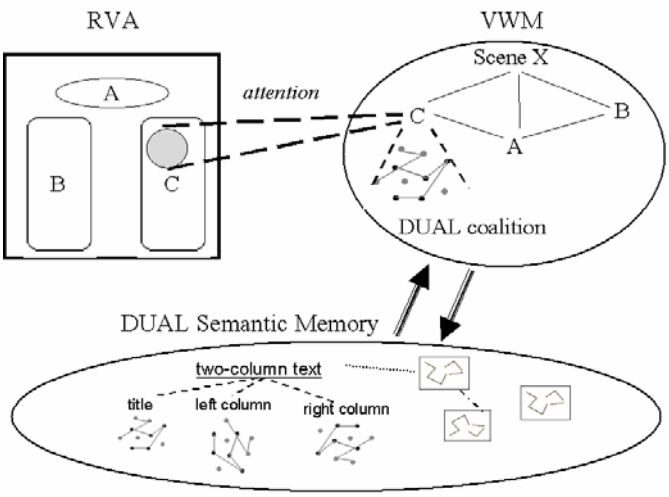


Fig. 1. The three main components of the model: the retinotopic visual array (RVA), the visual working memory (VWM) and DUAL's semantic memory. Attention is allocated to an area of the visual array by the object in VWM controlling attention while scene and object categories corresponding to the contents of VWM are retrieved from the semantic memory

TextWorld and the Retinotopic Visual Array. Stimuli are presented to the model as matrices containing filled or empty cells. Filled cells tend to group together in blocks giving the stimulus the outer appearance of a familiar configuration of typewritten text typical of TextWorld stimuli (see Fig. 2). In tests with human subjects the matrix is invisible and filled cells contain unreadable scrambled letters encouraging the subjects even more to think of the stimulus as a text format. TextWorld objects in a stimulus may also be manipulated according to specific rules giving the subjects the possibility to perform formatting tasks.

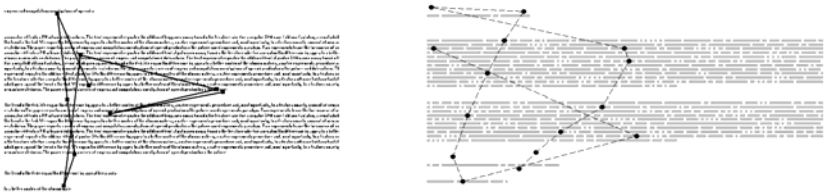


Fig. 2. A typical TextWorld stimulus with experimental (left) and simulated (right) eye movements data. Fixations are represented by points and saccades by lines connecting consecutive fixations

The type of input the model may receive for the time being is severely restricted. TextWorld stimuli are static, two-dimensional, black-and-white blocks of a roughly regular shape. The simplicity of our stimuli is the price for attempting to model vision and attention as a whole rather than focusing on a particular stage or subprocess as most models in the field do. Consequently, the types of representations and processes the models appeals to are designed to fit the specificity of our microdomain.

The first component of the model is the Retinotopic Visual Array (RVA) which in our account is a set of DUAL agents "sitting" in the cells of an imaginary matrix. Presenting a stimulus to the RVA comes down to clamping the state of each agent in RVA to filled or empty given the state of the corresponding cell in the stimulus. Each agent communicates in the array only with its immediate neighbors – at most 8 other agents – and it is unaware of its absolute coordinates in the array. By not deploying absolute coordinates the model faces a more challenging task and at the same time a more psychologically plausible one. Indeed, object representations held in VWM do not seem to be encoded in terms of absolute position but rather as configurations of stimuli in the world outside [Jiang, Olson and Chun, 2000]. Thus, we attempt to capture at least some aspects of the transition from the low-level retinotopically encoded information to the high-level spatiotopically encoded one, a transition without which visual information would be of very limited use.

Cheap parallel image-based processes take place at this level running on residual activation originating from previous or concurrent attentional processes. One such process determines whether an agent is a termination (end of row or column). For this purpose each active agent runs a local procedure checking the status of its neighbors: filled or empty. The procedure is complicated by the fact that empty cells are interspersed among filled ones just as blanks separate words in a written text. For this reason, termination detection is not always perfect although it should be reliable enough for approximating contours and edges.

A second process in the same category is blob detection. As each agent contains only local information, the perception of a cloud of filled cells as a single entity is not a trivial task. The model implements blob detection as a stepwise procedure in which neighboring cells gather initially into small groups, then intersecting groups fuse into larger ones and the process repeats itself until a whole block of cells is recognized as a single unit leading to its storing in VWM as a preattentive object [Wolfe and Bennett, 1997].

However, although cheap, the detection of perceptual primitives requires agents running symbolic procedures and is, therefore, dependent on the activation level. For this purpose, the activation spread in the system has to be a very efficient and reliable mechanism. Each DUAL agent in the system computes its activation level based primarily on the net input from its neighbors. A continuous version of the Grossberg activation function [Grossberg, 1978] governs the activation level of the agents as described below:

$$a(t_0) = a_0$$
$$da/dt = -d.a(t) + E.n(t).[M-a(t)] \quad (1)$$

where a is the activation level as a function of time, n is the net input to the node, M is the maximal activation value, and d and E are parameters that control the rate of decay and excitation. Activation values range between 0 and M and at the beginning of each run the activation of an agent is set to a small random value. The net input is a weighted sum of the activation of neighboring agents with normalized incoming link weights. In addition, all activation levels falling below a threshold θ are set to 0. Thus, some agents in the system will be unable to run their symbolic procedures at a time at least until they recover from their shortage of energy with the aid of their neighbors.

Visual Working Memory. An impressive amount of recent data document the existence and the properties of a specialized short-term memory store holding visual information. It has been argued, for instance, that this store is different from conceptual working memory and its functioning does not interfere with the functioning of the latter [Luck and Vogel, 1997], that its organization is based on the configuration of the visual stimuli and not on their absolute locations [Jiang et al., 2000], that detailed object representation it hosts tend to fade away as attention is withdrawn from them [O'Regan, 1992] and that its capacity is of about 3-4 items [Luck and Vogel, 1997]. More importantly, it seems to encode information not in a sensorial form, i.e. complete, metrical and noncategorical, but rather in a sparse abstract code [Carlson-Radvansky and Irwin, 1995] which may recommend the use of structural descriptions as a visual representational format.

Our proposal for a Visual Working Memory (VWM) system is in line with such evidence although adapted to the peculiarities of our TextWorld stimuli. VWM is the key component in the model by its ability to construct and to hold symbolic representations of scenes and objects. A scene layout is represented in VWM as a set of objects together with their spatial relations. Similarly, an object is represented as a set of lines or bars and their spatial relations. However, the mechanisms responsible for constructing scene layouts and object representations are very different.

Object agents are generated in VWM as a result of the blob detection process. They connect with all cells which enter their composition at the RVA level as well as with other object agents via links encoding their spatial relations. Thus, a scene layout will be represented in VWM as a coalition of DUAL agents each representing an object or a relation. The detection of object relations is a preattentional process in the sense that two objects do not need to be attended in order to get connected. However, layout detection is generated at the attended area and it uses energy originating in the deployment of attention like any other symbolic process in DUAL. Markers are launched from the attention spot in different directions and each time a marker reaches RVA agents belonging to different objects these objects are connected by a link carrying the signature of the marker. For instance, if a right-propagating marker is launched by an RVA agent belonging to object A and this marker reaches an agent belonging to object B by propagating on a straight line from A to B, then the two corresponding objects are connected in VWM by a "right-of" relationship indicating that B is to the right of A. However, this mechanism is not bound to find all relations in the input as markers may get stuck on their way in inactive agents. Thus, an object in VWM may be left out of the scene coalition if it is not very active at that moment. This could account for the fact that without giving attention to an object, in addition to missing a detailed representation of that object, one may not even notice the presence of the object, a phenomenon called *inattentional blindness* or *inattentional amnesia* [Wolfe, 1999].

A mixed object and location-based account of attention control is proposed for simulating attentional shifts and for generating energy in the system. At the level of VWM, objects compete for the control of attention on the basis of their activation level – more active agents are more likely to seize attention. Once an object wins, its activation level is set to M (see equation 1) and the object starts functioning as an energy generator in the system. A second type of selection takes place at the RVA level. An object in control of attention allocates it to some fragment of itself on the RVA in the form of a limited fixed-size circular area – the 'spotlight of attention' [Posner, 1980]. The selection here is also based on the activation level of RVA agents competing for the role of spotlight center. As termination agents are given higher E excitation rates they tend to be more powerful competitors and to capture attention more often than other agents. RVA agents in the spotlight receive temporarily high E excitation values so that activation builds up shortly in the area. This activation is spread in turn to the RVA agents around leading to a halo of activation around the spotlight. A series of shifts on the surface of the same object is sooner or later terminated as the VWM object loses attention and another one seizes it.

Most importantly, activation enables RVA agents to support the cost of building symbolic representations. Thus, agents sufficiently active tend to group themselves in lines or bars and to generate line agents in VWM which subordinate them. Line agents form a coalition by connecting in turn the object they belong to and by connecting with each other via links encoding their spatial relations similar to the way objects connect with each other. However, unlike scene coalitions, object coalitions have only an ephemeral existence and the agents they are made up of disappear as attention is withdrawn from the object fragment they represent and their activation drops below the threshold. In this manner the model instantiates the so-called principle of *visual transience* [8]: detailed object representations fade away as attention shifts away from them.

DUAL's semantic memory and categorization as analogy-making. The representational format utilized by DUAL's memory and the process of analogy –making in the AMBR model have been explained and detailed elsewhere [Kokinov and Petrov, 1997] so they will not be detailed here. Our view of categorization as a form of analogy-making is based on the fact that both types of processes require the retrieval of a base from memory and its mapping with a given target. In the case of categorization the base should be a category retrieved from the semantic memory of the system and the target an instance of this category constructed on the spot.

Scene layouts in VWM are mapped onto scene categories in DUAL's semantic memory and object representations in VWM are mapped onto object categories in the semantic memory. As category representations

similarly to scene layouts and object representations in VWM are represented by DUAL coalitions, categorization is modeled by the process of mapping of different symbolic structures. Scene-context effects on object recognition can easily be explained in this framework as scene categories activate object categories which enter their composition and, therefore, facilitate the recognition of objects belonging to these categories. The top-down control of attention also falls naturally out of this schema: categories for objects which are deemed important for the current goals of the system will be represented by active agents in the semantic memory; these categories will activate their instances in VWM; finally, more active objects in VWM will seize more easily attention and will use it longer.

However, unlike analogy-making studied as a reasoning process, object and scene recognition are fast, reliable and automatic. Therefore, we view categorization as an automatic form of analogy-making or, complementarily, we explain analogy-making as an extension of categorization in the area of thinking and reasoning.

Eye movements simulation. The model has been tested by comparing its performance with the performance of human subjects in order to check its adequacy as a cognitive model. One such test is the simulation of eye movement data in an observation task with TextWorld stimuli.

Eye movements data were collected from a group of 12 subjects asked to look at a series of TextWorld stimuli including the one in Fig 2. Each stimulus was presented for 5 seconds and the resulting sequences of fixations were used for computing a transition frequency matrix for each stimulus apart. Such a matrix records the frequency of consecutive fixations in two areas of a stimulus. For instance, a cell in row A and column B represents the frequency of transitions from A to B.

The experimental data were simulated by the model as stimulus-driven overt attention shifts. For this purpose, only the performance of RVA and VWM coupled together was considered. The stimulus in Fig 2 was presented on RVA and the model was run 12 times on this stimulus while initializing RVA agents to small random values at the beginning of each run. Different sequences of attention shifts were recorded on each run mainly because of the stochastic capture of attention by objects in VWM and by object areas in RVA. A simulation run ended when the number of shifts equaled the average number of saccades for a stimulus in the experiment.

The experimental and the simulation transition frequency matrices for the stimulus in Fig 2 were not different from each other as estimated by a chi-square test ($\chi^2(24) = 23.6$, $p< 0.5$). In order to ensure this is the result of meaningful shared structure rather than the result of a lack of structure, the simulation and the experimental matrices were compared with transition frequency matrices generated from vectors recording independent fixation frequencies for each object in the stimulus [Stark and Ellis, 1981]. Both the experimental and the simulation data proved significantly different from the latter ones ($\chi^2(24) = 81.2$, $p< 0.01$; $\chi^2(24) = 76.3$, $p< 0.01$) certifying the presence of structure in the fixation sequences. However, a larger range of stimuli should certainly be tested before being able to claim that model performance and human performance are indistinguishable as far as fixation sequences are concerned.

Conclusions

The construction of a model of visual perception and attention in the framework of the cognitive architecture DUAL is advantageous both for the existing architecture and for the new model we presented above. DUAL becomes capable of processing its visual input – at least as far as TextWorld stimuli are concerned – instead of running on ready-made symbolic representations. The model, on the other hand, draws on DUAL's knowledge representation and processing mechanisms and, furthermore, it earns the ability to interface higher-level cognition. No doubt, the model should be evaluated in its own right as an artificial system whose construction is inspired by the structure and the functioning of the human visual system. For this purpose, eye movements experimental data have been compared with simulation data produced by the model alone without plugging it into DUAL's memory and resources. However, the main thrust of the model is to allow the exploration of processes emerging out of the interaction of perception with high-level cognition, e.g. the interaction of stimulus-driven and goal-directed attention control or scene context effects on object recognition. The integration of perception and cognition is a major goal for any cognitive architecture. The results above represent DUAL's first steps towards reaching this goal.

Bibliography

[Anderson and Lebiere, 1998] J. R. Anderson and C. Lebiere. The atomic components of thought. Lawrence Erlbaum Associates, Hillsdale NJ, 1998.

[Ballard, 1991] D. H. Ballard. Animate vision. Artificial Intelligence, 48 (1991) 57-86.

[Biederman, 1987] I. Biederman. Recognition by components. Psychological Review, 94 (1987) 115-177.

[Carlson-Radvansky and Irwin, 1995] L. A., Carlson-Radvanski and D. E. Irwin. Memory for structural information across eye movements. Journal of Experimental Psychology: Learning, Memory and Cognition, 21 (1995) 1441-1458.

[Chalmers, French and Hofstadter, 1992] D. Chalmers, R. French and D. Hofstadter. High-level perception, representation, and analogy: Journal of Experimental and Theoretical Artificial Intelligence, 4 (1992) 185-211.

[Eriksen and Yeh, 1985] C. W. Eriksen and Y. Yeh. Allocation of attention in the visual field. Journal of Experimental Psychology: Human Perception and Performance, 11 (1985) 583-597.

[Grossberg, 1978] S. Grossberg: A theory of visual coding, memory, and development. In: Formal Theories of Visual Perception. Ed. L. Leeuwenberg and J. Buffart. Wiley, NY, 1978.

[Hollingworth and Henderson, 2002] A. Hollingworth and J. M. Henderson. Accurate visual memory for previously attended objects in natural scenes. Journal of Experimental Psychology: Human Perception and Performance, 28 (2002) 113-136.

[Jiang, Olson and Chun, 2000] Y. Jiang, Y., I. R. Olson and M. M. Chun. Organization of visual-short term memory. Journal of Experimental Psychology: Learning, Memory and Cognition, 26 (2000) 683-702.

[Kokinov, 1994] B. Kokinov. The DUAL Cognitive Architecture: A Hybrid Multi-Agent Approach. In: Proceedings of ECAI'94. Ed. A. Cohn. John Wiley & Sons Ltd., London, 1994.

[Kokinov, 1997] B. Kokinov. Micro-Level Hybridization in the Cognitive Architecture DUAL.. In: Connectionist-Symbolic Integration: From Unified to Hybrid Architectures. Ed. R. Sun and F. Alexander. Lawrence Erlbaum Associates, Hilsdale NJ, 1997.

[Kokinov and Petrov, 1997] B. Kokinov and A. Petrov. Integration of Memory and Reasoning in Analogy-Making: The AMBR Model. In: The Analogical Mind: Perspectives from Cognitive Science. Ed. D. Gentner, K. Holyoak and B. Kokinov. MIT Press, Cambridge MA, 2001.

[Itti and Koch, 2000] L. Itti and C. Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. Vision Research, 40 (2000) 1489-1506.

[Luck and Vogel, 1997] S. J. Luck and E. K. Vogel. The capacity of visual working memory for features and conjunctions. Nature, 309 (1997) 279-281.

[Marr, 1982] D. Marr. Vision. Freeman, NY, 1982.

[Marr and Nishihara, 1978] D. Marr and H. K. Nishihara. Representation and recognition of spatial organization of three-dimensional shapes. Proceedings of the Royal Society, London, Series B 200 (1978) 269-294.

[Meyer and Kieras, 1997] D. E. Meyer and D. E. Kieras. A computational theory of executive cognitive processes and multiple-task performance: Part 1. Basic Mechanisms. Psychological Review, 104 (1997) 3-65.

[Minsky, 1986] M. Minsky, M. The society of mind. Simon and Schuster, NY, 1986.

[O'Regan, 1992] J. K. O'Regan. Solving the 'real' mysteries of visual perception: The world as an outside memory. Canadian Journal of Psychology, 46 (1992) 461-488.

[Petrov and Kokinov, 1999] A. Petrov and B. Kokinov: Processing Symbols at Variable Speed in DUAL: Connectionist Activation as Power Supply. In: Proceedings of the 17th IJCAI. Ed. T. Dean. AAAI Press, 1999.

[Posner, 1980] M. I. Posner. Orienting of attention. Quarterly Journal of Experimental Psychology, 33 (1980), 3-25.

[Salvucci, 2001] D. Salvucci. An integrated model of eye movements and visual encoding. Cognitive Systems Research, 1 (2001), 201-220.

[Stark and Ellis, 1981] L. W. Stark and S. R. Ellis. Scanpaths revisited: cognitive models direct active looking. In: Eye Movements: Cognition and Visual Perception. Ed. D. Fisher, R. A. Monty and J. W. Senders. Lawrence Erlbaum Associates, Hillsdale NJ, 1981.

[Wolfe, 1999] J. M. Wolfe. Innattentional amnesia. In: Fleeting Memories. Ed. V. Coltheart. MIT Press, Cambridge MA, 1999.

[Wolfe and Bennett, 1997] J.M. Wolfe and S. C. Bennett. Preattentive Object Files. Vision Research, 37 (1997), 25-44.

Authors' Information

Adrian Nestor – Central and East European Center for Cognitive Science, New Bulgarian University, 21 Montevideo Str., Sofia 1618, Bulgaria; e-mail: adriannestor@students.nbu.bg

Boicho Kokinov – Central and East European Center for Cognitive Science, New Bulgarian University, 21 Montevideo Str., Sofia 1618, Bulgaria; Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Bl.8, Acad. G. Bonchev Str., Sofia 1113, Bulgaria; e-mail: bkokinov@nbu.bg

FORMAL DEFINITION OF THE CONCEPT "INFOS"

Krassimir Markov, Krassimira Ivanova, Ilia Mitov, Evgeniya Velikova-Bandova

Abstract: The concept INFOS is very important for understanding the information phenomena. Because of this, it is basic for the General Information Theory. The more precise formal definition of this concept is given in the paper.

Introduction

The genesis of the concept of **Infos** started from the understanding that the concept **"Information Subject"** is perceived as human characteristic. It is clear that in the nature there exist many creatures which may be classified to this category. To exclude the misunderstandings we decide to introduce new word to denote all possessors of the characteristics of the Information Subject. This word is **"***INFOS***"**.

The concept **"Infos"** is basic for the General Information Theory [Markov et al. 2003]. Its definition is only the starting point for further investigations and building the *Infos Theory*.

The variety of types of Infoses in the real world needs to be investigated and classified in the future research. At the first step, we may propose that may be at least two main types of Infoses exist:

- *infogens* - the natural creatures;
- *infotrons* - the artificial creatures.

Also, the Infos Theory needs to give answers to many other very important questions, such as:

- What is the nature of the activity of the Infos?
- What is the difference between the living level of the Infos and the non-living one?
- Is it true that the boundary between non-living and living entities is self-reflection and internal activity for satisfying the secondary (information) possibilities for internal or external contact?

Etc.

It is impossible to answer to all questions in a single article. We may make only the next little step. This is the aim of the present paper.

The formal definition of the concept of Infos needs theoretical basis which may support further mathematical investigations in this area. The first attempt was made in [Markov et al., 2003]. This brochure is the first attempt to compound all definitions of the main concepts of the General Information Theory, arisen during the years.

In [Markov et al., 2003a] we have continued our work with some new results, which make the previous formal definitions more precise. The main concepts, defined in this article are: "Entity", "Impact", "Interaction", "Reflection", and "Information".

The present work extends the basic definitions of [Markov et al., 2003a] with formal definitions, which concern the concepts "Information Witness", "Activity", and "Infos".

The Information Witness

Let's remember the *definition* of "Entity" [Markov et al., 2003a].

The **entity** A is the couple $A=(E_A,R_A)$ where:

E_A is a collection of sub-sets of a set M_A;

$R_A=\{r_i|i\in I, I \text{ is a set}\}$ is a nonempty set of relations in E_A, i.e.

$r_i\subset E_A\times E_A=\{(X,Y)|X,Y\in E_A\}$ is a relation and $\check{r}_i=r_i\cup\{(X,Y)|(Y,X)\in r_i)\}$, $\forall i\in I$; and:

1. $\varnothing\in E_A$;
2. $M_A=\cup X,\ X\in E_A$;
3. $\forall r\in R_A$ and $\forall X,Y\in E_A \Rightarrow ((\exists(X,Y)\in\check{r})$ or
$(\exists\ Z_1,..,Z_p\in E_A,\ Z_k\neq\varnothing,k=1,..,p : (X,Z_1)\in\check{r},(Z_1,Z_2)\in\check{r},\ldots,(Z_p,Y)\in\check{r})$ □

Definition 1. Every relation $r_i \in R_A$ from the set of relations R_A is said to be a **state** of $A=(E_A,R_A)$.■

Let's mention that often the set M_A and collection E_A are compounded by entities.

In the entities with very large and complicated structure which are built by another entities from lower levels, the impact may be considered as a basis for relations (X,Y) which may be considered as a part of any state of the whole entity (relation $r_i \in R_A$).

Building the relationship between the entities is a result of the **contact** among them. During the contact, one entity **impacts** on the other entity and vice versa. In some cases the opposite impact may not exist, but, in general, the contact may be considered as two mutually opposite impacts which occur in the same time.

In [Markov et al., 2003a] we have formally defined the concepts "impact", "reflection", and "self-reflection".

Definition 2. Let $B=(E_B,R_B)$ is an entity**.** We say that B has **possibility for reflection**, if there exists any entity $A=(E_A,R_A)$ and $\psi_f=(A\rightarrow B)_\psi$ is an direct impact of A on B, and state $g \in R_B$ of B, which contain the reflection F_ψ.■

Definition 3. Let $A=(E_A,R_A)$ is an entity**.** We say that A has **possibility for self-reflection**, if there exists any entity $B=(E_B,R_B)$ and transitive self-impact $\xi(A\rightarrow B\rightarrow A)$ and state $h \in R_A$ of A which contains the reflection F_ξ.■

Let's remember the *definition* of Information [Markov et al., 2003a].

Let:

$A=(E_A,R_A)$ and $B=(E_B,R_B)$;

τ is an impact of A on B, i.e. $\tau=(A\rightarrow B)_\tau$, $\tau \in \Omega_{AB}$;

$\exists$ entity $C=(E_C,R_C)$: $C\neq A$, $C\neq B$;

$\exists$ $\psi=(B\rightarrow C)_\psi$ which can be composed with $\tau=(A\rightarrow B)_\tau$;

$\exists$ transitive impact $\xi=\{\tau,\psi\}=(A\rightarrow B\rightarrow C)_\xi$;

$\exists$ impact $\varphi=(A\rightarrow C)_\varphi$, $\varphi \in \Omega_{AC}$ and $\varphi\neq\xi$;

F_φ is a reflection of the impact φ and F_ξ is a reflection of the impact ξ.

F_τ is **information** for A in B if $\exists r \in R_C:(F_\varphi\rightarrow F_\xi)_r$. □

The entity A is called **source**, the entity B is called **recipient**. The relation $r \in R_C$ for which $(F_\varphi\rightarrow F_\xi)_r$ is called **reflection evidence** and the entity C is called **information evidence**.

Let denote the information F_τ for A in B with information evidence C by $\mathbf{F_\tau=inform(A,B:C)}$.

F_τ is the reflection of an impact and we consider it as sub-entity of B.

Let denote by $\mathbf{r=evidence(A,B:C)} \in R_C$ the state of entity $C=(E_C,R_C)$ in which there exist the evidence $F_\tau=inform(A,B:C)$.

In [Markov et al., 2003a] we have formally defined the concept "interactive reflection".

Definition 4. If V_{AB} is an interactive reflection of between entities A and B, and entity C contains reflection evidences for all reflections of V_{AB} than C is called **information witness**.

Activity

Every forming relationship as well as every relationship unites the entities and this way it satisfies some theirs possibilities for building the relationship by establishing the contact. In other words, for creating the forming relationship we need:

- entities, from which the new entity is able to built;
- possibilities of the entities for establishing the contact by satisfying of which the forming relationship may be originated.

The forming relationship is the aggregate of the satisfied possibilities for establishing the contact.

It is clear that after establishing the relationship we may have any of two cases:
- all possibilities of the entities for establishing the contact are satisfied by such possibilities of other entities;
- there are any free possibilities after finishing the establishment of the new relationship - on the low levels of the entity or, if it is a new entity, on the level of the whole entity. Disintegration of the entity may generate any possibilities too.

In the second case, the entity has *"free valency"* which needs to be satisfied by corresponded contacts with other entities. We may say the entity has *activity* generated by the free possibilities for establishing the contacts with the entities from the environment.

The process of interaction is satisfying the possibilities for contact of the entities. From point of view of the entity, the interaction may be external or internal.

During the interaction given entity may be destroyed partially or entirely and only several but not all parts of the destroyed entity may be integrated in the new entity. This means that there exist both constructive and destructive processes in the process of interaction between entities. The determination of the type of the interaction depends on the point of view of given entity. The interaction dialectically contains constructive and destructive sub-processes.

If the entity is a complex, it is possible for it to have an opportunity of self-reflection. In such case, it is able to reflect any reflection, which has been already reflected in it. In this case, because of the new internal changes (self-reflection) the entity may obtain any new *"secondary activity"*.

The secondary activity is closely connected to the structural level of the entity, which correspond to the level of the self-reflection. This way the secondary activity may be satisfied by internal or external entity from point of view of the given entity. In other words, *the resolving* of the secondary activity may be *internal or external*.

Definition 5. Let $A=(E_A,R_A)$ is an entity and $r\in R_A$ is a state of A.
- $(X,\varnothing)\in$ř where $X\in E_A$, $\varnothing\in E_A$, is called **free valency** of A in the state r;
- the set P_r of free valences for the state $r\in R_A$ is called **activity** or **expectation for contact** of A in the state r: $P_r = \{(X,\varnothing) \mid X\in E_A, \varnothing\in E_A, (X,\varnothing)\in$ř$\}$ ■

During the establishment of the information relationship it is possible to be generated any secondary free activity (possibilities on the low levels of the entity or on the level of the whole entity) which needs to be satisfied by corresponded contacts with other entities.

The secondary activity in the information witness generated by the information relationship is called *"information activity"*.

Definition 6. Let $A=(E_A,R_A)$, $B=(E_B,R_B)$ and $C=(E_C,R_C)$ are entities; $F_\tau=\text{inform}(A,B:C)$ is an information for A in B and $r=\text{evidence}(A,B:C)$, where $r\in R_C$ is a information evidence of $\text{inform}(A,B:C)$. In such case:
- $(X,\varnothing)\in$ř where $X\in E_C$, $\varnothing\in E_C$, is called **free information valency** of C based on the $\text{inform}(A,B:C)$;
- the set $P_r = \{(X,\varnothing) \mid X\in E_C, \varnothing\in E_C, (X,\varnothing)\in$ř$\}$ of free valences of the state $r\in R_C$ is called **information activity** or **information expectation** of C based on the $\text{inform}(A,B:C)$.■

INFOS

On given level of complexity of the entities a new quality becomes - the existing self-reflection and internal activity based on the main possibilities for contact of the sub-entities as well as on the new (secondary) possibilities created after internal self-reflection.

The internal activity may be resolved by:
- the internal changes which lead to partial internal disintegration of the sub-entities and theirs a posterior internal integration in the new structures;
- the external influence on the environment.

The internal changes may lead to removing of some sub-entities if they have no possibilities for integration with the others, i.e. if they have no free valences to be resolved in the process of integration.

The external influence is the most important. The impact on the entities around the entity is the way to resolve its activity. The destroying of the external entities and including the appropriate theirs parts in itself is the main means to exist and satisfy the free valences.

One special kind of activity is the information one. The secondary activity need to be resolved by relevant to the information valences corresponded (information) valences. So, not every entity may be used for resolving the secondary activity.

This way, the entity needs a special kind of (information) contacts and (information) interaction for resolving the information activity.

Definition 7. The Information Witness $C=(E_C,R_C)$, which has:

- possibility for reflection in a state $r_1 \in R_C$;
- possibility for self-reflection in a state $r_2 \in R_C$;
- (primary) activity in a state $r_3 \in R_C$;
- a state $r_4 \in R_C$ in which C has non-empty information expectation (information activity)

is called **Infos**. ■

Conclusion

What gives us the concept "INFOS"?

At the fist place, this is the common approach for investigating the natural and artificial information agents.

In other hand, this is the set of common characteristics which are basic for all entities, which we may classify to the category of the INFOS.

And, at the end, this is a common philosophical basis for understanding the information subjects.

Our main goal is to provoke the scientists to continue the research in this important area and to make the next step.

Acknowledgements

This work is partially financed by project ITHEA-XXI of the Institute of Information Theories and Applications FOI ITHEA.

Bibliography

[Markov et al, 1993] Kr. Markov, Kr. Ivanova, I. Mitov. Basic Concepts of a General Information Theory. IJ "Information Theories and Applications", FOI ITHEA, Sofia, 1993, Vol.1, No.10, pp.3-10.

[Markov et al, 2003] Kr.Markov, Kr.Ivanova, I.Mitov. General Information Theory. Basic Formulations. FOI-Commerce, Sofia, 2003.

[Markov et al, 2003a] K. Markov, K. Ivanova, I. Mitov, E. Velikova-Bandova. The Information. IJ "Information Theories and Applications", FOI ITHEA, Sofia, 2003, Vol.10, No.1, pp.5-9.

Authors' Information

Krassimir Markov – ITHEA - FOI Institute of Information Theories and Applications; Institute of Mathematics and Informatics, BAS; P.O. Box: 775, Sofia-1090, Bulgaria; e-mail: foi@nlcv.net

Krassimira Ivanova – Institute of Mathematics and Informatics, BAS, Acad.G.Bontchev St., bl.8, Sofia-1113, Bulgaria; e-mail: foi@nlcv.net

Ilia Mitov – ITHEA - FOI Institute of Information Theories and Applications, P.O.Box: 775, Sofia-1090, Bulgaria; e-mail: foi@nlcv.net

Evgeniya Velikova-Bandova – Sofia University, Faculty of Mathematics and Informatics, Bulgaria; e-mail: velikova@fmi.uni-sofia.bg

ONLINE GENETIC ALGORITHMS

Alfredo Milani

Abstract: This paper present a technique based on genetic algorithms for generating online adaptive services.

Online adaptive systems provide flexible services to a mass of clients/users for maximising some system goals, they dynamically adapt the form and the content of the issued services while the population of clients evolve over time.

The idea of online genetic algorithms (online GAs) is to use the online clients response behaviour as a fitness function in order to produce the next generation of services. The principle implemented in online GAs, "the application environment is the fitness", allow modelling highly evolutionary domains where both services providers and clients change and evolve over time.

The flexibility and the adaptive behaviour of this approach seems to be very relevant and promising for applications characterised by highly dynamical features such as in the web domain (online newspapers, e-markets, websites and advertising engines). Nevertheless the proposed technique has a more general aim for application environments characterised by a massive number of anonymous clients/users which require personalised services, such as in the case of many new IT applications.

Keywords: genetic algorithms, adaptive web, evolutionary computation

1. Introduction

The research on the topic of adaptive systems has mainly focused on architectures based on knowledge representation and reasoning [1], fuzzy reasoning [2][3] and probabilistic models [4]. These approaches are often able to give an adequate account of uncertainty and dynamical aspects of the domain, but they also require a great effort in building a detailed model of the problem. Despite of the good qualitative response, they often reflect too rigidly the domain constraints at modelling time. When the environment, i.e. the constraint of the domain, evolves, the system performance tend to decrease until the model needs to be modified or redesigned.

The increasing diffusion of mass services based on new information technologies (ITs) poses new requirements and goals on adaptive systems which are seemingly contradictory, such as the problem of providing adaptive personalised services to a mass of anonymous users [4]. Sometimes models of user behaviour [1] for the new services does not even exist, and, in addition, services and technologies appear and disappear very quickly thus vanishing the effort of building accurate models.

The growing interest in self adaptive and self modelling systems is partially motivated by these reasons.

The two leading approaches to self adaptation, i.e. genetic algorithms [5][6] and neural networks [7][8] are characterised by somewhat symmetrical features which are worth to be pointed out: *neural networks (NNs) tends to be online systems while GAs operate offline*. GA usually operates offlinein the sense that they can be seen as building a simulated application environment in which they evolve and select the best solution among all the generations, under the well known Darwinian principle of "survival of the fittest".

Some works [9][10][11] have introduced "real world" issues into the GA loop, in the interactive GAs approach [12] the user is inserted in the algorithm with the role of providing the fitness functions by interacting with the GA, in other works still following the offline approach [13][14] about machine learning by GA, historical real data are used as fitness function.

Despite of their offline nature GA are able of a highly dynamical behaviour. The main reason is that the knowledge about "reasoning" structure of GA is embedded in the population chromosomes: when the population evolves the structure evolves as well. GA concepts such as *cross over* and *mutation* have no counterpart in NNs

approach, but they are a powerful tools which can allow a GA to make fast hill climbing of local minimum and plateau in optimisation problems [6].

The idea of bringing these adaptive features in the *online system* scenario is made more challenging from the facts that the population of clients asking for services is evolving over time, then their response to services changes.

In this paper we propose a new approach, *online genetic algorithms* (online GAs) which tries to combine timely responses with the adaptive behaviour of GAs. The basic idea of online GAs is to evolve populations by using the application world as a fitness function, under the principle "the real world is the fitness".

The goal of systems based on online GAs is to give a timely response to a massive set of clients requesting services, and to be able to adapt services to clients, both changing over time in unknown and unpredictable way.

As noted in the beginning, it is not realistic to rely on the hypothesis of detailed user models [1][15]. The increasing consciousness of privacy issues, legal limitations on personal info [16] and the growth of mobile and pervasive interfaces accessible from casual users, often make the user model impossible to collect. The anonymity of users/clients is then a structural constraint in mass adaptive services.

In the next paragraphs we will motivate the online GAs approach by analysing the features of a sample dynamical scenario regarding an online newspaper management system.

The principles and the architectural scheme of the online genetic algorithms approach will be presented, an example application and experimental results will be discussed.

2. The Online Adaptive Scenario: Web Newspaper

Let us consider as a typical scenario for online genetic algorithms: the problem of managing the generation of an online newspaper with the goals of maximising customers, i.e. readers, contacts.

The problem, well known to journal editors, is to build a newspaper in order to publish news according to the newspaper politics and mission, and selling it at its best. Selling news in this contexts means the goal of capturing readers attention for reading the articles, and for, possibly, satisfying the newspaper advertisers. Online readers browse time by time the newspaper web site and read the news which interests them. It is assumed that a good journal will collect a great number of contacts and many users will spend time in reading it. Managing editors of online newspapers have a great advantage with respect to their hardpaper colleagues: while a conventional paper journal is limited and bounded to a single daily edition (except the cases of extraordinary events), an online editors, instead, can make timely adaptation of the newspaper to the latest news, thus maximising the impact of the newspaper on the readers.

Online media have the likely feature that can be produced and delivered instantaneously such that, in principle, each user can read his own single, personalised and different copy of the journal.

The main issues, and source of difficulties, in the newspaper scenario are the lack of information about the users and the unpredictable dynamical evolution of all the elements which characterised it, in particular:

- anonymity of clients
- dynamical evolution of potential services
- dynamical evolution of clients
- dynamical evolution of client goals

these evolutionary features are shared by a wide class of online problems.

2.1. Anonymity of clients

Anonymity of clients means that no hypotheses can be made about profiles of the users of online services. As discussed in the introduction the typical assumption for online newspaper is that the information available to the system comes from anonymous user sessions, where users cannot be identified, nor recognised from previous sessions [16][17].

2.2. Dynamical evolution of potential services

The purpose of online systems is to provide the best of their currently available services for maximising the client impact [18], the situation is made more complex since *the services that are issued by the providers can vary over time in unpredictable way*.

News, seen as services, are characterised by a lifetime cycle (i.e. they appear, disappear and are archived), and the news flow is by its nature unpredictable. Thus the news editor task is to select according to the editorial line, which news best interest and impress their readers, among the available ones.

2.3. Dynamical evolution of clients

The set clients connected with the online system evolves over time in unpredictable way. The set of connected clients are not always the same, since new clients come and previous sessions disconnect.

In the case of online newspaper there can be made some general assumption about the target users. Users are assumed to have somewhat homogeneous features like in the case of readers of newspaper specialised in economics, politics, sports etc. Nevertheless the instantaneous audience profile of online newspaper can vary over time. For instance students can connect mainly in the afternoon, while corporate workers can connect in different time range. In addition, external factors and unpredictable events, such as holidays or exceptional events, can make different classes of readers to connect in unexpected time/dates.

Even assuming that we have a way of determining the ideal journal for the current audience given the currently available news, the newspaper edition will be no more adequate after some time, since the audience will change unpredictably.

2.4. Dynamical evolution of client goals and attitudes

Goals and attitudes of the single clients can vary and depend on time.

As we as pointed out before, external events of general interest can make the journal audience vary, but can also make the interests of the audience to vary. Economical or political events can induce a shift in the typical interest of the readers. Moreover even assuming to have a fixed audience, with fixed goals, is not possible to produce a fixed "ideal newspapers", since people expect that newspaper vary: it would be unlikely to read every day the same identical news; typical users of online newspapers connect to the system many times a day, expecting to read more news on topics of their interest.

2.5. Model of Service Impact Factor

A model of the impact factor of service cannot be easily defined and require classification effort.

The goal of the newspaper editor is to catch the attention of most of its readers by selecting the appropriate news and preparing a suitable edition according to the newspaper editorial line, i.e. mission, policy and cultural goals.

The typical tools available to an editor to maximise the impact of the service he provides (i.e. the news) are: *selections* of the news among the continuous flow (deciding which news are currently published and which news go to archive); *location* of the news in the grid of the newspaper layout (the position of the news usually reflect is evidence or priority in editor's intention); *presentation form* of the news, which regards aspects such are selecting a *title* for the news, and or selecting a possible *picture* accompanying it, and sometime also long or short versions of the article. These tasks are usually regarded to as an "art" which the newspaper editor performs by the help of his/her experience.

It is worth noticing that some factors, such as the news position in the layout, are not necessarily determining the readers' priority. A well prepared journal, for example, usually offers a mix of different news (i.e. not many news on the same topic). The visibility strictly depend not only in the position but also in the context in which news are presented. Sometimes hot emerging topics require breaking these rules and, when it happens most part of the journal news are devoted to a single topic.

The next paragraph will describe a framework based on genetic algorithms for providing adaptive services in highly evolutionary environment to a massive audience of anonymous users, such as in the newspapers scenario.

3. Online GA Schema

GA have been classically proposed for use in an *offline schema*. In the offline approach populations of solutions are evolved offline for a given number of generations in order to produce the best evolved solution (usually determined in the last generation) which given as system output. For instance in classical optimisation problems [6] GA are used for exploring a search space of solutions and the best minimum/maximum value found over all generation is produced. In GA applied to learning problems, such as discovering stock market rules [14], real data about stock market are used to evolve the population, but again, the best solution is computed offline, and it is used in the *real market* afterwards. A different approach is that of Interactive Genetic Algorithms [9] [12] [19] where the real world is included into the GAs loop under the principle that "the user is the fitness", i.e. the user participates to a cooperative optimisation process. In some interactive GAs applications to robot learning [13], the real world is used to evolve the solution, but GAs uses real world in an offline phase of training.

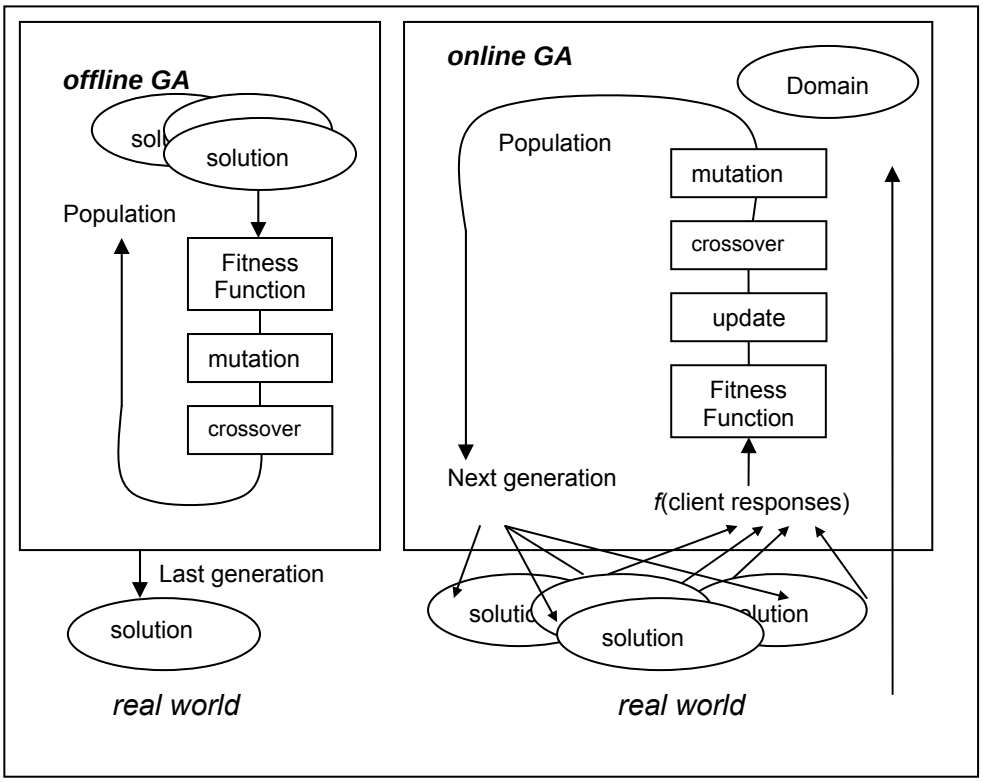


Fig.1 Online GA

In the online GAs approach we propose to literally implement the evolutionary metaphor which originally motivates GAs. In our proposal the basic elements and concepts of GAs such as population, chromosomes, crossover, mutation and selection mechanism still exists but, they are extended with the innovative but simple assumption that "*the real world is the fitness*", i.e. the application world, representing the environment, is used to selects the best surviving fittest solutions, moreover all generated solutions are output to the system and no interactive cooperation is required to users/clients.

The basic scheme of an online GAs versus offline GAs is depicted in the fig.1.

In *online GAs* a population of solutions is evolved with usual genetic mechanisms, with the difference that the solutions in the population are actually "executed", and the client/users behaviours/responses upon solutions are used as a fitness to evolve the next generation of online solutions. In other words the fitness function resides in the real world and it is expected to give timely response to the evolution of clients' population and domain modification.

Updating Phase

Since in online GAs the problem domain also evolves unpredictably (for instance the provided services changes, i.e. the flow of incoming/outcoming news), then there is the need for a novel phase of *updating*, not present in classical GA. The purpose of the *updating* phase is to establish and implementing a policy about how to adapt the current population of solutions to domain changes (such as how to replace an article which is disappeared from the journal, because expired or delete after explicit editor's decision).

Online GAs can be used when it is required to dynamically adapt to evolution and changes in the problem domain, moreover application domains best suitable for the adoption of online GAs are characterised by:

- a solution space with "many" valid solutions to explore, i.e. the solution space with not unique or few valid solutions
- a set of clients which require solutions to be used immediately
- an optimisation function which measures the efficacy of a solution given in output, which can be "sensed" by the system in the external world as a response/result produced by clients

It is worth noticing that online genetic algorithms would not be a real possibility without the new ITs. A massive diffusion of the internet, mobile phones services, on demand phone services has made possible application servers where a huge number of anonymous clients (with no distinction among final users or software agents) are concurrently requiring services in an automated framework which directly connects consumers to service providers. The services providers are usually optimising very simple functions which are completely inside their "sensing" scope such as *time-spent*, *services bought*, *money charged*, *advertising clicked*.

4. An Online GA

The pervasive dynamical and unpredictable evolution of all the key elements in the newspaper scenario represents a difficult challenge for adaptive systems, which should provide adequate services in answers to clients' requests.

News to be offered in the newspaper is continuously flowing in from news agency and journalists. Different classes of anonymous individual readers continuously connect and disconnect in order to read interesting news The goals and interests of the individuals vary in an unpredictable way (people get bored of old news). The impact of news upon users depends of the form, the position and the context in which the news is given, and it is hard to be deterministically modelled [9][20]. Finally the editor policy represents a pervasive constraint to be respected throughout the journal editing.

In the following we present the architecture of a sample online GAs applied to the newspaper evolutionary scenario.

4.1. Domain and Constraints

A newspaper has a typical layout and structure in term of sections of topics, which are part of the recognizable corporate image. No editors are available to modify it, moreover the editor usually want to have control over the proposed news in order to implement the editorial policy.

In order to reflect these constraints the structure and layout of the journal do not evolve, and the editor decides which news include/exclude in/from the newspaper and how to assign (or remove) them to sections, let the sections, for instance: *TopStories, National, International, Sports, Health and Technology*. A limited set of headlines (for example 4 headlines) is reported in the front page for each section; the sections occupy fixed layout positions; the section headlines are chosen among the articles available inside the sections.

For each single article we will assume that the newspaper editor provide a set of possible alternative formats for each article, i.e. alternative titles, texts and pictures to be used for presentation.

The task of the editor is to decide how to update the set of news and formats, while the online GA actually build the newspapers deciding which articles will be inserted into the sections headlines and which alternative formats will be used in the articles presentation.

4.2. Population and Individuals

The individuals which compose the population of the current generation consist of the different versions of the newspaper which have been issued to the currently requesting online readers.

4.3. Time Intervals

In order to make the online GA having a sufficient number of individuals in the population, and a sufficient time to evaluate user response, i.e. fitness of the individual, it is needed to fix a time interval value, i.e. the duration of the minimal interval of time from one generation to the next one. If, for instance, a newspaper has 6000 contacts per hour, a time granularity of 1 minute guarantee, guarantee an average population of 100 individuals, but doe not allows evaluating responses whose duration is greater than one minute.

4.4. Fitness Function

The fitness function measures the adequacy of the solution in term of client response.

According to the anonymity hypothesis the system is able to "sense" user sessions, but not to recognise user from previously started session. Sensing data are easily collected from the web server log files [Etzioni2000]. In the newspaper problem the fitness of a given solution k (i.e. the individual version of the newspaper) is defined as

$$F(k) = w_s t_s + w_{ca}\, n_{ca} + w_{cn}\, n_{cn} + w_{int}\, (\Sigma_{(i=1..n_{cn})}\, t_{si}/t_r) \,/\, n_{cn} - w_{nohl}\, c_{nohl}$$

The listed parameters reflect the general criteria that reward as positive, in particular: t_s is the *total time spen*t on the newspaper (measured as the time between the first and the last browser request); n_{ca} the *number of clicks on* newspaper *advertising*s; n_{cn} *number of clicks on news* (i.e. how many news have been red; the Σ term computes the *average interest of news*, where the interest is measured as the time spent t_{si} on a single news with respect to the time t_r needed to read the news (skipping rapidly a news means little interest versus carefully reading it); the minus terms c_{nohl} in $F(v)$ penalises the situations in which the readers find *no interest in headlines* and go straight to sections to read particular news, i.e. in other words it penalise at a certain extent the journal versions in which the content is interesting while presentation is not. Weights w_s, w_{ca}, w_{cn}, w_{int} and w_{nohl} are used to tune the contributions of the respective terms to the global fitness.

4.5. Chromosomes

The individuals, i.e., the single newspaper versions, are encoded by a set of *sections vectors* each one encoding a section of the newspaper.

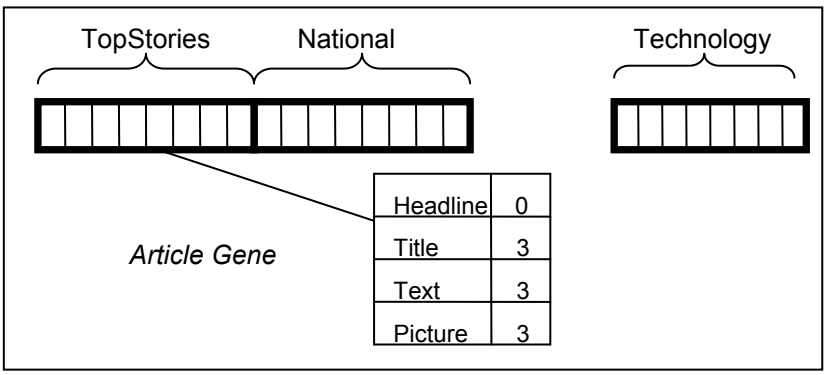


Fig.2 Individual chromosomes

Each element in the journal chromosome specifies a single news in term of its position in the section headlines (0 means not in headlines), and its presentation i.e. values indicating which title, text, and picture, the newspaper edition will contain for the given article among the different available versions.

4.6. Selection

A standard *proportional to fitness* selection method is used in order to determine the intermediate population used for crossover and mutation. The more the fitness is high more chances are given to individuals to survive. On the intermediate population thus determined crossover and mutation are applied.

4.7. Crossover

The purpose of crossover is to generate a new journal version from two individual chromosomes. The two offspring replace the parents. Again a proportional to fitness reproduction criteria is used.

The crossover is operated *section by section* on the whole chromosome. For each section a linear crossover point is determined (see dashed line in the figure 3) for splitting the section subvector. The respective subsection of the two parents is then combined.

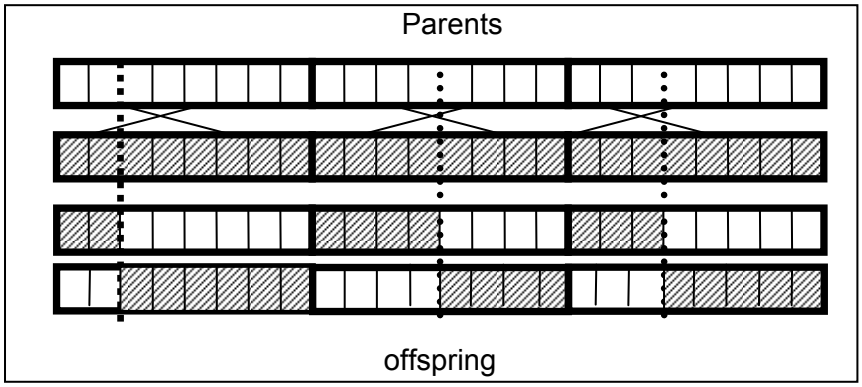


Fig.3 Segmented crossover

Restoring valid solutions can be necessary after crossover reproduction. Suppose that a given section is allowed h headlines; the split point position can divide the section segments such that one offspring segment contains more than h headlines while the other has less than h, i.e. the solution is not valid. In this case in order to restore a valid solution we move headlines from the longest to the shortest one selecting them randomly. Another case of invalid solution is when two headlines in a section points to the same position (another headline position must be empty), in this case the tie is broken randomly. Note this criteria guarantee that all headlines in the parents will be again headlines in the offspring.

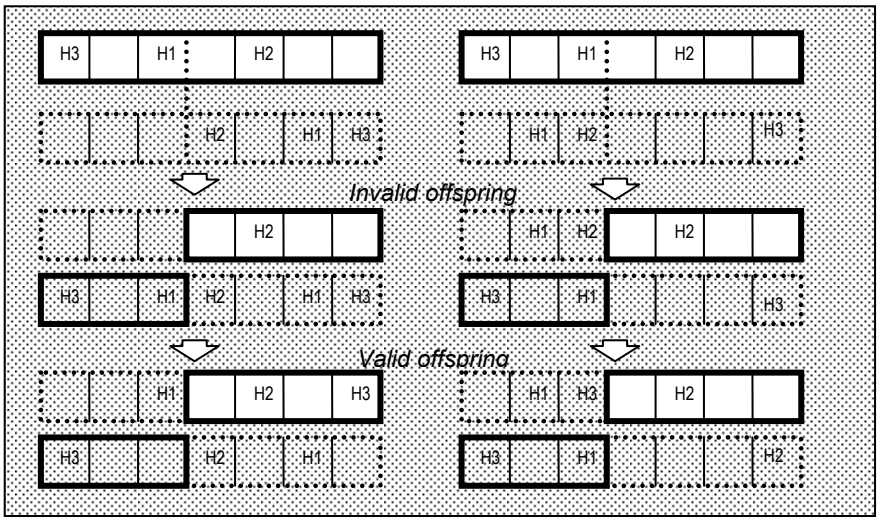


Fig. 4 Restoring Valid Offspring after crossover

4.8. Mutation

Mutations are operated at different levels with different priorities.

- *headline mutation*, is the operation which moves a news from sections into headlines and vice versa, since

an headline mutation is a dramatic change in a newspaper version, the probability P_h of headline mutation is kept relatively low, on the other hand and additional factor P_{new} is considered, P_{new}, is giving more probability to become section headline to new articles versus old ones;

- *format mutation*, this mutation tends to adapt the form in which the single news are given, i.e. order, titles, alternative texts and accompanying pictures, the probability P_f of this mutation is slightly high than the previous one.

A *format mutation* is realised by choosing randomly a format component (*order*, *title*, *text* and *picture*) and a feasible random value in the domain of the format component (e.g. one title over three available candidates). Format mutation of *ordering* is only applied to headline news and consists in swapping an article with another one randomly selected among the headlines.

Headline mutation is realised by randomly selecting the incoming article (taking into account of P_{new} to give priority to new articles), selecting the outcoming article and swapping them among headline and section.

4.9. Update adaptation phase

The adaptation phase concerns the problem of adapting the population of solution which were made invalid by external modifications. For example when the news editor decides that an old article has to be archived and/or a new one has to be inserted into the journal, some individual in the current population could be no more valid. In restoring the validity of the solution we use the following criteria:

- incoming news are added to the respective section with maximum P_{new}
- outcoming news not appearing in section hot headlines are simply deleted from the section
- outcoming news which are on the section headlines are replaced by shifting up the section headlines, and operating a format mutation on the last position where simple insertion replace swapping

The array representation is updated accordingly.

5. Experiments

Experiments to test online GA for newspaper management described in the previous paragraph have been designed and are under implementation. The first stage will take place using the "What's new?" list box in our University home page. The "What's new?" is an approximation of a single section of a newspaper. The "What's new?" area is located right in the center of the home page and it contains a set of headlines which link to announcements of University activities and events. The WebMaster policy limits the number of headlines to a maximum of 8, but much more departments, administrative offices and other institutions are competing for having their announcements on the home page box. There are averages of 30 candidate announcements per day, some 22 of which are forced to reside in the internal "What's new?" section. The system will be implemented in php in connection with an Apache server on a Linux platform. The idea is to provide to about an half of the thousands contacts a day, the GA managed version of "What's new?" and give the rest the usual static version for comparisons. Two further versions are under implementation: a newspaper manager, based on the content management system Nuke [21], and a simulation environment which aims at compare the online GA with respect to simulated user response.

6. Conclusions

Online GAs represent a new approach to systems which provide adaptive services to a large number of anonymous clients/users which evolves over time in unpredictable way.

The basic idea of online genetic algorithms is that “the world is the fitness”, i.e. the fitness function resides in the application environment and it can be evaluated by sensing the environment i.e. by evaluating clients/users response to the current solutions. A phase of adaptation is added to usual GA schema for restoring validity to solutions made invalid by evolution in the problem domain.

Online GAs are related with interactive GAs methods [12][9][10], in which the real world appear in GA in the form of the user cooperation to the selection process, or in the form of environment guided training [13]. The main difference between online GAs and interactive GAs is that in interactive GAs, GAs are used in a sort of offline simulation in order to select a *final* optimised solution or behaviour, used by the application. Instead online GAs based applications made immediate use of the solutions population.

The main issues which motivate the adoption of online GAs have been discussed in the framework of the newspaper scenario. Online GAs represents an answer in all those situation in which adaptation is required, while few o no data are available about users' profiles and attitudes [17].

The increasing diffusion of massive distributed services based on the new ITs, the increasing consciousness and laws about the privacy issues, motivates the apparently contradictory request of providing adaptive services to unknown users in dynamical domains.

Preliminary experimental results on a simplified version of the newspaper application confirm the validity of the online GAs approach.

Open theoretical and practical issues need to be further investigated in the framework of online GAs such as the problem of time granularity with respect to the time needed for fitness evaluation; defining effective methods for tuning GA parameters and weights, discussing typical GA issues such as co-evolution [17] in the context of online Gas, and providing extensive experimental evaluation of the performance.

Moreover the integration between online GA and other non evolutionary techniques such as fuzzy and probabilistic analysis are worth to be investigated.

References

[1] A. Kobsa and W. Wahlster, editors. User Models in Dialog Systems. Springer Verlag, London, 1989.

[2] L. A. Zadeh: Fuzzy Sets Information and Control 8(3): 338-353 (1965)

[3] M.A.S.; Monfared, S.J. Steiner Fuzzy adaptive scheduling and control systems in Fuzzy Sets and Systems Vol.115, n. 2 pp. 231-246, 2000.

[4] J.Binder, D.Koller, S.Russell, K.Kanazawa, Adaptive probabilistic networks with hidden variables. Machine Learning, 1997.

[5] J. Holland, Adaptation in Natural and Artificial Systems, University of Michigan Press, 1975.

[6] Whitley, D. An overview of evolutionary algorithms. Information and Software Technology, (2001).

[7] J. A. Anderson.An Introduction to Neural Networks,MIT Press Boston 1995

[8] M. T Hagen, H. B. Demuth, M. Beale, Neural Network Design, PWS Publishing Co. Boston 1996.

[9] T. Masui. Graphic object layout with interactive genetic algorithms. Proc. IEEE Visual Languages '92, 1992.

[10] J. G. Peñalver and J. J. Merelo. Optimizing web page layout using an annealed genetic algorithm as client-side script. In Proceedings PPSN, Parallel Problem Solving from Nature V, Lecture Notes in Computer Science. Springer-Verlag, 1998.

[12] H. Takagi, Interactive evolutionary computation: fusion of the capabilities of EC optimization and human evaluation in Proceedings of the IEEE01, IEEE Press, 2001.

[13] M. Dorigo and U. Schnepf, Genetics-based Machine Learning and Behaviour Based Robotics: A New Synthesis, IEEE Transactions on Systems, Man and Cybernetics, vol.23,n.1,pp. 141-154,1993.

[14] L. A. Becker, M. Seshadri, GP-evolved Technical Trading Rules Can Outperform Buy and Hold, in 3rd International Workshop on Computational Intelligence in Economics and Finance. Sept 2003.

[15] J. Kay, B. Kummerfeld, P. Lauder, Managing private user models and shared personas in Proceedings of Workshop on User Modelling for Ubiquitous Computing, User Modelling 2003.

[16] M. K. Reiter and A. D. Rubin, Crowds: anonymity for Web transactions, ACM Transactions on Information and System Security,vol.1, n.1,pp.66-92, 1998.

[17] N. Kushmerick, J. McKee, F. Toolan, Towards zero-input personalization: Referrer-based page prediction, Lecture Notes in Computer Science, vol.1892, Springer-Verlag, 2000.

[18] M. Koutri, S. Daskalaki, N. Avouris, Adaptive Interaction with Web Sites: an Overview of Methods and Techniques, in Proc. of the 4th Int. Workshop on Computer Science and Information technologies CSIT02, Patras Greece,(2002).

[19] A. Oliver, N. Monmarché, G. Venturini, Interactive design of web sites with a genetic algorithm. Proceedings of the IADIS International Conference WWW/Internet, pages 355-362, Lisbon, 2002.

[20] J.González; J.J.Merelo; P.A.Castillo; V.Rivas; G.Romero; A.Prieto. Optimized web newspaper layout using simulated annealing. In Sánchez-Andrés, Mira, editor. IWANN99, LNCS. Springer-Verlag, June 1999.

Author Information

Alfredo Milani – Department of Mathematics and Computer Science, University of Perugia

SIMULATING MEMBRANE SYSTEMS IN DIGITAL COMPUTERS[1]

**Fernando Arroyo, Carmen Luengo, Luis Fernandez,
Luis F. de Mingo, and Juan Castellanos**

Abstract: *Membrane Computing started with the analogy between some processes produced inside the complex structure of living cells and computational processes. In the same way that in other branches of Natural Computing, the model is extracted from nature but it is not clear whether or not the model must come back to nature to be implemented. As in other cases in Natural Computing: Artificial Neural Networks, Genetic Algorithms, etc; the models have been implemented in digital computers. Hence, some papers have been published considering implementation of Membrane Computing in digital computers. This paper introduces an overview in the field of simulation in Membrane Computing.*

Keywords: Simulation, Membrane Computing, Multiset, Evolution Rules, Membrane Structure.

Introduction

Membrane Computing is inspired in the structure and functioning of living cells. It was considered in October 2003 by Thomson Institute for Scientific Information as a fast emerging research area. This research area was initiated at the end of 1998 by Gh. Paun (in a paper which was published in 2000 in the Journal of Computer and System Sciences) [2]. Nowadays, Membrane computing is one of the most popular research topics in the European community of cellular computing.

The membrane systems, also named P systems, are a class of distributed parallel computing devices of a biochemical type, which can be seen as a general computing architecture where different types of objects can be processed by different operations. The basic model consists of a membrane structure (several membranes hierarchically embedded in a main membrane *skin*). Membranes define regions where different objects are placed. These objects are processed according to given evolution rules, which are associated with the regions. When a rule is applied in a region, the objects present in the region are modified, some of them are sent out or in it. The evolution rules can also dissolve the membrane. In that case, all the objects present in the membrane remain free in the membrane that includes the dissolved one; however, rules associated to the dissolved membrane are removed. The *skin* membrane never is dissolved because then the system can not be considered a system anymore. As can be seen, the system is governed by evolution rules, membranes are considered as separators and communication channels. The application of evolution rules is made in a nondeterministic and maximally parallel manner; at each step, all objects which can evolve must evolve in every region of the system.

This kind of systems compute by passing from a configuration to another configuration by applying evolution rules in the way described above. A computation is considered complete when it halts, e.g. when no further rules can be applied to the objects present in the last configuration. The result of a halting computation can be made in two different ways: by considering the multiplicity of objects presents in the halting configuration inside a determined region, or by concatenating the symbols which are sent out of the system considering the order in which they were sent out. In the first case, vectors of natural numbers are computed while in the second case, languages are generated.

Many variants of P systems have been considered [1]. In some of them, the number of membranes can only decrease by dissolving membranes as result of applying evolution rules. However, many of them the number of membranes can be increased using some biological features of living cells, for example: by division. Some other variants consider membranes not only passive objects of the system, these kind of systems are based in biological processes performed by membranes when chemical compounds pass through the membrane (*protein gates* or *protein channels*).

[1] Work partially supported by contribution of EU commission Under The Fifth Framework Programme, project “MolCoNet” IST-2001-32008.

Membrane Systems with Symbol-Objects: The Simplest Class

P systems with symbol-objects have been presented as a class of non-deterministic parallel computing devices whose their main ingredients are: a membrane structure, multisets of objects, and evolution rules. In this section, they will be formally defined and so, the dynamic of such systems.

The very important ingredient of P systems is the membrane structure. A membrane structure is a three-dimensional arrangement of vesicles in which the only important thing is the mutual relationship between membranes related to be in or out a determined membrane. This relation is defined as an adjacency relation. Figure 1, illustrates several notions about membranes and the adjacency relation. It is important to note the one-to-one correspondence between membranes and regions they enclose. A membrane structure can be pictorially represented in two-dimensional way in form of a Venn diagram and in term of an undirected tree, like in figure 1. The tree representation looses the biochemical intuition, but it makes clear the fact that the position of membranes do not matter

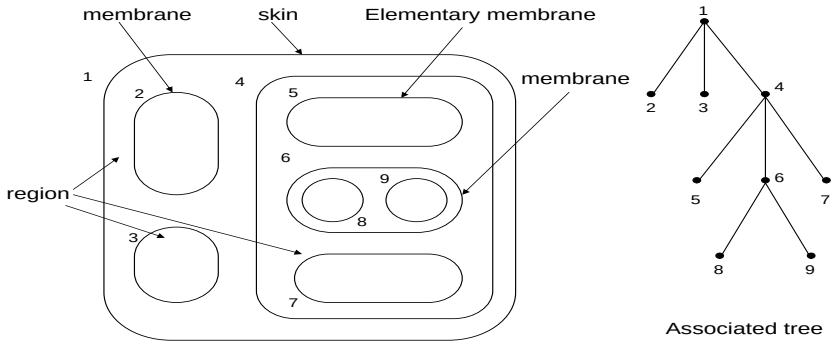


Figure 1: Membrane Structure and its associated tree.

There is one more representation for membrane structure in term of strings of matching parentheses. For example, the membrane structure of figure 1 is represented by the following parenthetic expression:

$$[_1[_2]_2[_3]_3[_4[_5]_5[_6[_8]_8[_9]_9]_6[_7]_7]_4]_1$$

Because the membranes have labels, the pairs of corresponding parentheses also have labels. Of course, the same membrane structure may be represented by different parenthetic expressions.

The second important ingredients are the multisets of objects, which are represented by strings of symbols associated with the objects. In turn, the evolution rules for objects will be given as multiset rewriting rules, of the form $u \rightarrow v$, where u and v are strings (understood as representations of multisets). It is important to note that, working with multisets, means to work with multiple copies of objects; when saying, for instance, that "and object a is processed" (for example, passed through a membrane), we understand "a copy of the object a is processed".

With these considerations, a membrane system with symbol-objects is a construct

$$\Pi = (V, T, \mu, \omega_1, \ldots, \omega_m, (R_1, \rho_1), \ldots, (R_m, \rho_m), i_0)$$

Where:

i. V is an alphabet, and its elements are called objects;

ii. $T \subseteq V$ is the output alphabet;

iii. μ is the membrane structure and membranes are labelled in a one-to-one manner from 1 to m, being m the number of membranes in the membrane structure,

iv. $\omega_1, \ldots, \omega_m$ are multisets of objects over V and they are associated to the regions defined by the membrane structure;

v. $R_1,\ldots,R_m$ are finite sets of evolution rules and they are associated to the regions defined by the membrane structure;

- An evolution rule is a pair (u,v) that is usually represented by $u \rightarrow v$ and
  a. u is a multiset of objects over U,
  b. $v=v'$ or $v=v'\delta$, where v' is a multiset of objects over $Ux(\{here, out\} \cup \{in_j \mid j = 1,\ldots,m\})$ and δ is a symbol which is not in U.

vi. $\rho_1,\ldots,\delta_m$ are partial order relations defined over the set of rules of the regions defined by the membrane structure and they specify a priority relation in $R_1,\ldots,R_m$ respectively,

vii. i_0 is a label which determines the output membrane of the system.

The symbols: *here*, *out*, in_j, are named targeting commands. They specify the membrane target, e.g. the membrane that have to pass objects produced by the rule. In order to simplify the notation, *here* is omitted when the rules are written.

P systems compute starting from an initial configuration and passing from one configuration to another by the application of evolution rules to objects inside regions until reaching a halting configuration. This configuration is defined as a configuration in which there is no rule applicable to the objects present in the last configuration. Because these systems compute by transition, they are named *Transition P systems.*

A ***configuration*** is defined as a (k+1)-tuple $(\mu,\omega_1,\ldots,\omega_m)$ where μ is the membrane structure having m membranes and $\omega_1,\ldots,\omega_m$ are the objects multiset associated to the regions defined by the membrane structure [3].

The application of evolution rules in every region of the system is made in a non-deterministic and massively parallel way; e.g. every object present in a region that can evolve by the application of an evolution rule must evolve. Moreover, every region in the system evolves in parallel at the same time with the regions of the system.

A transition in a P system is got by passing from a configuration to another consecutive one. More precisely, given two configurations $C_i=(\mu^i_1,\omega^i_1,\ldots,\omega^i_m)$ $C_j=(\mu^j_1,\omega^j_1,\ldots,\omega^j_n)$ of a P system, Π, it is said that there is a transition from C_1 to C_2 and it is represented by $C_i \rightarrow C_j$, if we can pass from C_i to C_j by using the evolution rules from $R^i_1,\ldots,R^i_m\$$ in the regions $i_1,\ldots,i_k$.

Transition P systems are devices very powerful. They are computationally complete. There are many other variants of P systems, but the essential set of specification described above act as framework for many of them.

Some Previous Formalizations

There are two main components in the simulation of membrane systems: the static and the dynamic. The first one is related to how to describe the main ingredients of the systems in a static way, the second one is related to give semantic to the static component; e.g. to make evolve the system from a static configuration to the next one by the application of rules associated to the regions of the systems.

The static structure of Transition P systems is related to everything concerning to membranes, and objects to be contained in them [4]. So, it is related to multisets, evolution rules and regions. There are some relations among these elements: multisets will be necessary for defining regions and evolution rules; evolution rules and again multisets will define a region, and finally regions will give the static structure definition of a Transition P system. Therefore, it is very important to determine a data structure for defining multisets, evolution rules and regions and the operations that can be done over such data structures [3] and [7]. Let us to provide some definitions for these elements.

Multisets provide a compact representation for words generated by a given finite alphabet when the order of symbols occurrences does not matter. In fact, multisets are sets of objects in which the number of occurrences of its different elements are taken into account. A multiset over a given finite object set U can be formally defined as follows:

Definition: Let U be an arbitrary finite set and let N the natural number set. A multiset M over U is a mapping from U to N.

$$M : U \to N$$
$$a \to Ma$$

There is another common and very useful representation for multisets of objects in membrane computing, it is the polynomial representation:

$a^{Ma} b^{Mb} \ldots z^{Mz}$ Where $U = \{a, b, \ldots, z\}$

It is very easy to define some algebraic operation in the set multiset over a given set of object *U*. In particular, addition and subtraction of multiset and multiset inclusion. Moreover, let *M(U)* the set of multiset over the set *U*, then it is easy to prove that *(M(U),+)* is a monoid with identity element (the empty multiset).

Evolutions rules are the active elements of P systems, they make evolve the system until reaching a halting configuration. In this section, evolution rules are described in terms of algebraic structures. They are defined without considering the dynamic implications they have in the evolution of P systems. First, they are defined in terms of multisets of objects over different finite sets of objects. Moreover, evolution rules are here defined independently of P systems and they are considered as static components. After that, some operations are defined and characterized; these operations will be considered in a different section of this paper for described the dynamic of evolution of P systems.

Definition: Let *L* be a label set, let *U* be a set of object, and let *D =*$\{out, here\} \cup \{in_j \mid j \in L\}$ the subset of labels to which rules can sends objects.

An *evolution rule* with label in *L* and objects in *U* is a tern *(u, v,* δ*)*, where *u* is a multiset over *U*, *v* is a multiset over *U x D* and $\delta \in$*{dissolve, not dissolve } = {true, false}*. Let *R(U,L)* be the set of evolution rules with label in *L* and objects in *U*. Then, several operations over evolution rules can be defined:

Definition: Addition of evolution rules:

$$+ : M(U,L) \times M(U,L) \to M(U,L)$$
$$((u_1, v_1, \delta_1), (u_2, v_2, \delta_2)) \to (u_1 + u_2, v_1 + v_2, \delta_1 \vee \delta_2)$$

Definition: Product of a natural number by an evolution rules:

$$\bullet : N \times M(U,L) \to M(U,L)$$
$$(n, (u, v, \delta)) \to (nu, nv, \delta)$$

It can be also easily demonstrate that *(M(U,L),+)* is a monoid with identity element. Moreover, linear combination of evolution rules can be defined and they accomplish some interesting properties related to linear dependency.

Regions act as bags for containing multisets of objects and evolution rules. They are defined by the membrane structure of the P systems. Hence, it could be said that a determined membrane structure define a region structure for the system. Moreover, it is very interesting, for developer of simulations of P system, to consider regions as the container elements of P systems and to define an algebraic structure for representing the information of P systems. In this way, regions can be defined by:

Definition: A region with labels in *L* and objects in *U* is a tern *(l,* ω*, (R,* ρ*))* where $l \in L$, ω is a multiset over *U*, *R* is a set of evolution rules with labels in *L* and objects in *U* and ρ is a partial order relation over *R*.

With the previous definitions, a Transition P system can be defined as a tree of regions in the following manner:

Definition: A *Transition P system* with labels in *L* and objects in *U* is a pair of elements whose first one is a region with labels in *L* and objects in *U* and the second one is a set of Transition P systems with labels in *L* and objects in *U*, and regions are uniquely labelled.

$$\Pi = ((l, \omega, (R, \rho)), \Pi\Pi)$$

Where *reg=(l,*ω*,(R,*ρ*))* is a region with labels in *L* and objects in *U* and $\Pi\Pi$ is a set of Transition P systems with labels in *L* and objects in *U*.

This static structure of Transition P systems provides a formal and very practical representation for describing P systems configurations. In fact, as it will be show below, it can be used for defining a language for describing P systems in a static manner.

The formalization of the dynamic component of Transition P systems involves the use of algebraic operations defined over evolution rules [5] and [6]. In these papers is described how to obtain linear combination of evolution rules, which are able to make evolve the different regions of the P systems by the application of only such linear combinations –complete multisets of evolution rules- and how the systems evolve using them.

On the Simulation of P Systems

P systems have been demonstrated to be a very powerful computational devices in the theoretical framework (*in info*) but they are today restricted only to this theoretic framework. It is not so clear if they will be implemented in *vivo*, in *vitro* or only in simulations running in digital computer. The straightforward way to demonstrate the computational possibilities for P systems is to developed software simulations to be run in digital computers of general purpose or in specific hardware specifically developed for a specific class of membrane systems. Some published papers deal with software simulations of Transition P systems with different programming paradigms like functional programming, object oriented programming, distributed programming, etc. [8], [9], [10], [11], [12], and [13].

To develop a friendly software application for simulating P systems executions in digital computers it is needed to consider at least three main components, like in Model-View-Controller (MVC) architecture. The Model is responsible for type abstract data, the View is responsible for showing the results to the user through a graphical interface and the Controller is in charge of the requests made by the user. Following with these considerations, three physical components must be implemented in different subsystems: The Graphical User Interface (GUI), which permits the interactions between the user and the application and represents the static structure of the system in a friendly way. The static component describes the P system structure including membrane structure, multisets of objects, evolution rules and their associated priority relations. Finally, the Simulator admits as input the static component of the P system, implements the evolution of the system starting from an initial configuration, and obtains every new possible configuration to which transit.

In order to describe the static structure of P systems, it is needed to represent them very precisely. An specific language have been developed to describe Transition P systems. It was named *Bio-language* for Transition P systems [7] and it is described in term of rewriting rules as follows:

- *Syntax of the bio-language*
  - *V a set of objects*
  - *L a set of labels*
  - *Transition P system (TPS):* $\Pi \Rightarrow$ *[l Region; {*Π*}]l*
  - *Region* $\Rightarrow$ *Objects, Rules, Priorities*
  - *Objects* $\Rightarrow$ *{on} where o*$\in$*V n*$\in$*N*
  - *Rules* $\Rightarrow \lambda$ *| rule {, rule}*
  - *Priorities* $\Rightarrow \lambda$ *| ri < rj {, ri' < rj'} where i, i', j ,j'*$\in$*N*
  - *Addresses* $\Rightarrow$ *here | out | I where I* $\in$ *Labels*
  - *ObjTarg* $\Rightarrow$ *{(o, Addresses)n} where o* $\in$ *V and n* $\Rightarrow$ *1 | 2 | …*
  - *Rule* $\Rightarrow$ *rm: Objects* $\rightarrow$ *ObjTarg* δ *where* $\delta \in$*{dissolve, not dissolve}*

As it can be seen, the bio-language uses the representation of Transition P systems as tree of regions. It describes every component of the system including the membrane structure, the multiset of objects, the evolution rules and the partial order relation defined among evolution rules.

This is an example of bio-language for a specific variant of P system. What is important here is to demonstrate the feasibility of formally describing membrane systems and to use this formalization as inputs to the appropriate simulator, which is in charge of implementing the semantic for the membrane system.

The simulator will implement the evolution of the membrane system. It is in charge of parsing the static structure of the P system into abstract data structures and then to transform them in the appropriate way implementing the

evolution of the systems. It will provide a set of possible configurations to which the systems could transit starting from the initial configuration. Every new configuration of the system will be described in terms of the static structure using the bio-language for the membrane system. Hence, once has been chosen one configuration to transit the new configuration can be used as input again to the simulator and continue the software execution. Hence, for each variant of P systems, a determined bio-language can describe the static structure of the corresponding system and the specific simulator must implement the dynamic behaviour of the system. An exhaustive list of published papers in Membrane Computing can be found in [14].

Conclusions

Membrane computing is a very active research field, many different variant of P systems are actively modified providing new variant of the model. Simulations of membrane systems are very powerful tools for researchers in order to check their working thesis. Many of the programming paradigms can be use to develop software simulations of membrane system (object oriented programming, functional programming, distributed programming, etc.) exploding different attributes of the membrane system. What is very important is to build different programming approaches for developing simulations and to explore which ones are useful to bring to membrane computing the necessary tools for making it useful for researchers in computer science and biology and to explore new un-conventional computing models.

References

[1] Gh. Paun: *Membrane Computing. An Introduction*, Springer-Verlag, Berlin, 2002

[2] Paun G.: *Computing with Membranes*. In: Journal of Computer and Systems Sciences, 61, 1. (2000) 108--143

[3] Gh. Paun, G. Rozenberg*, A Guide to Membrane Computing*, Theoretical Computer Science, vol 287, nr 1 (2002), pages 73-100.

[4] A. V. Baranda, J. Castellanos, R. Gonzalo, F. Arroyo, L. F. Mingo, *Data Structures for Implementing Transition P Systems in Silico*, Romanian J. of Information Science and Technology , 4, 1-2 (2001), 21-32

[5] A. V. Baranda, J. Castellanos, F. Arroyo, R. Gonzalo, *Towards an Electronic Implementation of Membrane Computing: A Formal Description of Nondeterministic Evolution in Transition P Systems*, Proc. 7th Intern. Meeting on DNA Based Computers, Tampa, Florida, USA, (N. Jonoska, N.C. Seeman, eds.) Lecture Notes in Computer Science 2340, Springer Verlag, 2002, 350-359.

[6] F. Arroyo, A.V. Baranda, J. Castellanos, C. Luengo, L.F. Mingo, *A Recursive Algorithm for Describing Evolution in Transition P Systems*, Pre-Proceedings of Workshop on Membrane Computing , Curtea de Arges, Romania, August 2001, Technical Report 17/01 of Research Group on Mathematical Linguistics, Rovira i Virgili University, Tarragona, Spain, 2001, 19-30.

[7] F. Arroyo, A.V. Baranda, J. Castellanos, C. Luengo, L.F. Mingo, *Structures and Bio-Language to Simulate Transition P Systems on Digital Computers*, in Multiset Processing. Mathematical, Computer Science, and Molecular Computing Points of View (C.S. Calude, Gh. Paun, G. Rozenberg, A. Salomaa, eds.) Lecture Notes in Computer Science 2235, Springer-Verlag, 2001, 1-16.

[8] F. Arroyo, C. Luengo, A.V. Baranda, L.F. de Mingo*, A software simulation of transition P systems in Haskell*, Pre-Proceedings of Second Workshop on Membrane Computing, Curtea de Arges, Romania, August 2002 and Proc. of WMC02, Curtea de Arges, Romania, (Gheorghe Paun, Grzegorz Rozenberg, Arto Saloma, Claudio Zandron Eds.) Lecture Notes in Computer Science 2597, Springer-Verlag, 2003, 19-32.

[9] G. Ciobanu, D. Paraschiv: Membrane Software. *A P System Simulator*, Pre-Proceedings of Workshop on Membrane Computing, Curtea de Arges, Romania, August 2001, Technical Report 17/01 of Research Group on Mathematical Linguistics, Rovira i Virgili University, Tarragona, Spain, 2001, 45-50 and Fundamenta Informaticae, 49 , 1-3 (2002), 61-66.

[10] G. Ciobanu, G. Wenyuan: *A parallel implementation of transition P systems*, Artiom Alhazov, Carlos Martín-Vide and Gheorghe Paun (eds.): Preproceedings of the Workshop on Membrane Computing; Tarragona, July 17-22 2003, 169-184

[11] M. Malita: *Membrane Computing in Prolog*, Pre-proc. Workshop on Multiset Processing , Curtea de Arges, Romania, CDMTCS TR 140, Univ. of Auckland, 2000, 159-175.

[12] Y. Suzuki, H. Tanaka: *On a LISP Implementation of a Class of P Systems*, Romanian Journal of Information Science and Technology, 3, 2 (2000), 173-186.

[13] A. Syropoulos, E.G. Mamatas, P.C. Allilomes, K.T. Sotiriades, *A distributed simulation of P systems* , Artiom Alhazov, Carlos Martín-Vide and Gheorghe Paun (eds): Preproceedings of the Workshop on Membrane Computing; Tarragona, July 17-22 2003, 455-460

[14] http://psystems.disco.unimib.it/

Authors' Information

Fernando Arroyo – Dpto. Lenguajes, Proyectos y Sistemas Informáticos- Escuela Universitaria de Informática; Universidad Politécnica de Madrid - Crta. de Valencia km. 7 - 28031 Madrid SPAIN - farroyo@eui.upm.es

Carmen Luengo – Dpto. Lenguajes, Proyectos y Sistemas Informáticos- Escuela Universitaria de Informática; Universidad Politécnica de Madrid - Crta. de Valencia km. 7 - 28031 Madrid SPAIN - cluengo@eui.upm.es

Luis Fernandez – Dpto. Lenguajes, Proyectos y Sistemas Informáticos- Escuela Universitaria de Informática; Universidad Politécnica de Madrid - Crta. de Valencia km. 7 - 28031 Madrid SPAIN - setillo@eui.upm.es

Luis F. de Mingo – Dpto. Organización y Estructura de la Información - Escuela Universitaria de Informática; Universidad Politécnica de Madrid - Crta. de Valencia km. 7 - 28031 Madrid SPAIN - lfmingo@eui.upm.es

Juan Castellanos – Dpto. Inteligencia Artificial - Facultad de Informática - Universidad Politécnica de Madrid; 28660 Boadilla del Monte - Madrid SPAIN - jcastellanos@fi.upm.es

TEACHING STRATEGIES AND ONTOLOGIES FOR E-LEARNING [1]

Tatiana Gavrilova, Michael Kurochkin, and Victor Veremiev

Abstract. The paper presents one approach aimed at developing teaching strategies based on the principles of ontological engineering. The research framework is targeted on development of methodology and technology that will scaffold the process of knowledge structuring for e-learning. The structuring procedure is the kernel of ontology development. Ontologies that describe the main concepts of the domains are used both for teaching and assessment techniques. Special stress is put on visual design as a powerful learning mindtool. The examples are taken from the courses on the foundations of artificial intelligence and intelligent systems development. These courses are delivered by the authors in St.Petersburg State Polytechnical University at School of Computer Science and in Poland in the First Independent University.

Keywords: E-learning, Ontologies, Visual Knowledge Engineering, Expert Systems Building Tools, Knowledge Acquisition, Knowledge Sharing and Reuse.

1. Introduction

The drawback of e-learning is lack of feedback from the teacher or tutor. That is why the courseware should be more precisely structured that in face-to-face teaching.

The idea of using visual structuring of teaching information for better understanding is not new. Concept mapping [Sowa, 1994; Jonassen, 1998, Conlon, 1997] is scaffolding the process of teaching and learning for more than 20 years. Visual representation of the general domain concepts is facilitative and helps both learning and teaching. A teacher now has to work as knowledge analyst or knowledge engineer making the skeleton of the studied discipline visible and showing the domain's conceptual structure. This structure is now called "ontology". However, ontology-based approach is rather young. It was born in knowledge engineering [Boose, 1990; Wielinga, Schreiber, Breuker, 1992], then it was transferred to knowledge management [Fensel, 2001].

[1] The work is partially suported by Russian Foundation for Basic Studies (grant 02-01-00466)

The short prehistory of knowledge engineering (KE) techniques and tools (including knowledge acquisition, conceptual structuring and representation models), the overall overview of which is presented in [Adeli, 1994; Scott, Clayton, Gibson, 1994], is an ascending way to the development of the methodology that can bridge a gap between the remarkable capacity of human brain as a knowledge store and the efforts of knowledge engineers to materialise this compiled experience of specialists in their domain of skill.

Beginning from the first steps to nowadays knowledge analysts have been slightly guided by cognitive science. So major part of KE methodology suffer of fragmentation, incoherence and shallowness.

The last years the main interest of the researchers in this field is concerned with the special tools that help knowledge capture and structuring. This generation of tools is concerned with visual knowledge mapping to facilitate knowledge sharing and reuse [Eisenstadt, Domingue, Rajan, Motta, 1990; Tu, Eriksson,. Gennari, Shahar,. Musen, 1995; Johnassen, 1998]. The problem has been partially solved by developing of knowledge repositories called ontology servers where reusable static domain knowledge is stored (e.g. projects as Ontolingva, Ontobroker, KA2, etc.)

In tutoring systems teachers are supposed to reuse the domain ontologies in order to support the description of the discipline they taught and the problem-solving methods of their domain. The idea is to allow teachers to model both domain and problem-solving knowledge using the same visual language. Ontology design also may be used as an assessment procedure. Students show their knowledge and understanding while creating ontologies.

Knowledge entities that represent static knowledge of the domain are stored in the hierarchical order in the knowledge repository and can be reused by other teachers. At the same time those knowledge entities can be also reused in description of the properties or arguments of methods of another knowledge entity. Concept maps modelling language that is designed in the framework of the described project is based on a class-based object-oriented language which is aimed to support typing and parameterisation of knowledge entities. Due to the class subsumption and polymorphism of classes the reasoning process becomes extremely flexible. This non-formal system allows to reason on a large set of knowledge and to apply problem-solving rules described for the higher level knowledge entities to the lower level knowledge entities based on the class inheritance. In contradistinction to ontology server approach where static knowledge described is very specific to the domain, the approach which is taken in the paper simplifies reusability of the dynamic knowledge and as a consequence building of large-scale knowledge bases with a flexible reasoning capability.

The proposed ideas and methods may be applied to those tutoring systems where general understanding is more important that factual details. We used such approach in teaching Artificial Intelligence, Neuroscience and Computer Graphics.

2. Ontological Engineering

An ontology is a set of distinctions we make in understanding and viewing the world. There are a lot of definitions of this milestone term [Neches et al,1991; Gruber, 1993; Guarino et al, 1995; Gomez-Peres, 1999]:

1. Ontology defines the basic terms and relations comprising the vocabulary of a topic area, as well as the rules for combining terms and relations to define extensions to the vocabulary.
2. Ontology is an explicit specification of a conceptualization.
3. Ontology as a specification of a conceptualization.
4. Ontology as an informal conceptual system.
5. Ontology as a formal semantic account.
6. Ontology as the structured vocabulary.
7. Ontology is a hierarchically structured set of terms for describing a domain that can be used as a skeletal foundation for a knowledge base.

All these definitions together clarify the ontological approach to knowledge structuring on one hand, on the other hand give enough freedom to the open thinking. So ontological engineering gives the intuitively clear representation of company structure, staff, products and relationship among them.

Many researchers and practitioners argue about distinctions between ontology and user's conceptual model. We supposed that ontology corresponds to the analyst's view of the conceptual model, but is not the model itself.

Ontologies are useful structuring tools, in that they provide an organising axis along which every student can mentally mark his vision in the information hyper-space of domain knowledge. Rather often we can't express all the information in one ontology, so subject knowledge storage includes a set of ontologies. Some problem may occur when jumping from one ontological space to another. But constructing of meta-ontologies may help.

Ontology development also faces the knowledge acquisition bottleneck problem. The ontology developer comes up against the additional problem of not having any sufficiently tested and generalised methodologies recommending what activities to perform and at what stage of the ontology development process these activities should be performed. That is, each development team usually follows their own set of principles, design criteria and steps in the ontology development process. The absence of structured guidelines and methods hinders the development of shared and consensual ontologies within and between teams, the extension of a given ontology by others and its reuse in other ontologies and final applications [Guarino, Giaretta, 1998].

Till now, few domain-independent methodological approaches have been reported for building ontologies [Swartout, Patil, Knight, Russ, 1997; Fensel, 2000]. These methodologies have in common that they start from the identification of the purpose of the ontology and the need for domain knowledge acquisition. However, having acquired a significant amount of knowledge, major researchers propose a formal language expressing the idea as a set of intermediate representations and then generating the ontology using translators. These representations bridge the gap between how people see a domain and the languages in which ontologies are formalised. The conceptual models are implicit in the implementation codes. A reengineering process is usually required to make the conceptual models explicit. Ontological commitments and design criteria are implicit in the ontology code.

This paper proposes the most explicit way for ontology design - to use the visual representation in a form of a tree or set of trees.

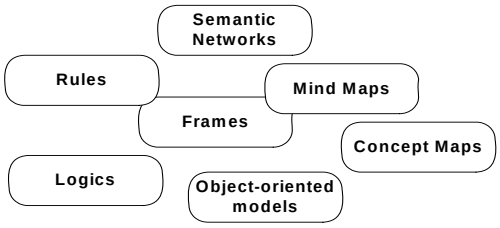


Figure 1. Unstructured set of Knowledge Representation Models

Figures 1 and 2 illustrate the idea how ontology bridge the gap between chaos of unstructured data (names of different models and techniques for knowledge representation) and clear knowledge of modern classification. Our approach shows that ontology development process needs some creative efforts of meta-concepts definition that helps to name the groups and stucture the chaos.

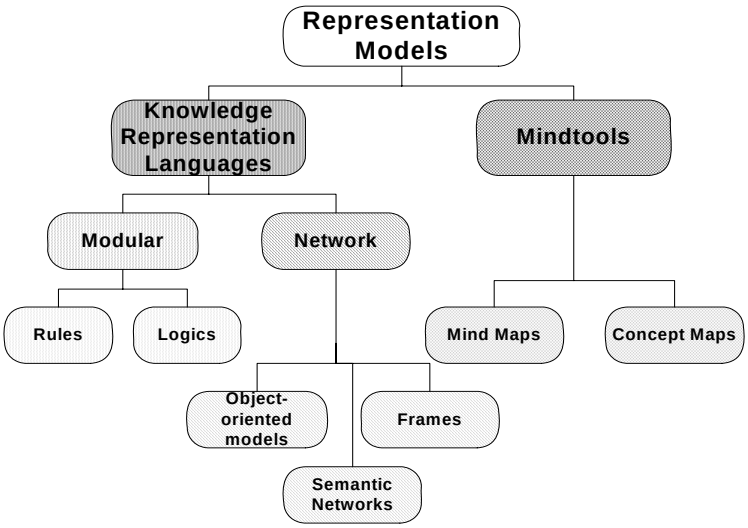


Figure 2. Ontology "Knowledge Representation Models"

Ontology developers (who are unfamiliar with or simply inexperienced in the languages in which ontologies are coded, e.g. DAML, OIL, RDF) may find it difficult to understand implemented ontologies or even to build a new ontology.

It is easier for any educationalist simply to draw the ontology using well-known to everybody "pen and pencil" technique then to study these languages.

3. Object-Structured Approach

Although there are some methods that are rather powerful and versatile [Kremer, 1998], the teacher as knowledge analyst is still weakly supported while working for a set of ontologies to describe the main subject knowledge. This process is the most important and critical stage in the courseware preparation life cycle - transition from elicitation to conceptualisation by understanding and realisation of the subject information structure and main reasoning way. The teacher may do this sophisticated g procedure alone or to ask for help of professional analysts.

In this way, a special methodology Object-Structered Analysis (OSA) has been developed [Gavrilova and Voinov, 1992-2000], which is intended to help knowledge analyst to perform the most informal step of knowledge acquisition, concluding in prior conceptual structuring of the subject domain. The approach presents the enhancement of classical structured analysis methodology [Sowa, 1994; Yourdon, 1990] to knowledge engineering.

OSA is based on decomposition of subject domain into several (3-8) strata (Tab.1). The number of strata is considered by the analyst. This multi-step and time-consuming procedure is methodological base for effective constructing of subject ontologies.

s1	*WHAT FOR* Knowledge	*Strategic* Analysis: Targets, Aims, Requirements, Constraints.
s2	*WHO* Knowledge	*Organisational or Historical* Analysis: Main Researchers, Human Resources, Actors.
s3	*WHAT* Knowledge	*Conceptual* Analysis: Main Concepts, Processes, Entities and Relationships between them.
s4	*HOW TO* Knowledge	*Functional Analysis*: Main Algorithms, Decision Procedures, Business Processes Modelling, Decision Making Models.
s5	*WHERE* Knowledge	*Spatial* Analysis: Geography, Environment, Communications, etc.
s6	*WHEN* Knowledge	*Temporal* Analysis: Historical Dates, Schedules, Time Constraints, etc.
s7	*WHY* Knowledge	*Causal* Analysis: Explanations to Decision Making Models.
s8	*HOW MUCH* Knowledge	*Economical* Analysis: Resources, Losses, Incomes, Revenues, SWAT, etc.

Filling that matrix is performed into two steps:

Step 1. Global (vertical) analysis, i.e. decomposition of the heterogeneous domain information into the groups related to mentioned above methodological strata.

Step 2. Local analysis of each individual stratum (horizontal), concluding in maintenance of gradually detailed structures. The number of levels depends on peculiarities of the subject domain and could vary dramatically for different strata. From the point of view of methodology the number of levels $n<3$ indicates ill-structured domain knowledge.

The first level (or column 2 in the table) corresponds to the discipline information as a whole. The second one corresponds to the problem that is studied now. The others may correspond to particular sub-problems, depending on the required reasonable deepness of detailing. The procedure of the described analysis may be performed both in top-down and bottom-up strategies, including their possible mixture.

The formation of strata with more of less definite meaning as described in Tab.1 allows to avoid many traditional didactic mistakes in teaching and learning. The minimal obligatory set of strata for the course structuring development is:

s3: Conceptual Structure or subject ontology.

s4: Functional Structure or main problem solving procedures.

Other strata are designed and developed if needed by subject peculiarities, e.g. spatial and temporal analysis strata (s5 and s6) may be formed in those disciplines which study construction or management where the issues of scheduling, real-time operations, real object manipulation are substantial.

Step 1 algorithms may be sketched in such form:

1.1: Gather all the data and knowledge of discipline identification

1.2: Select a set of N strata to be formed ($N \geq 3$).

1.3: For each i-th stratum select a subset of all available information, relevant to that stratum and represent it in way appropriate to that stratum (see below).

1.4: If there remains unused bulk of information, increase number of strata and repeat step 1.3. Otherwise, begin the horizontal analysis of each declared stratum.

Step 2 is horizontal analysis of strata that depends on the number of columns in OSA matrix and may be performed in two ways: deductive (top-down) and/or inductive(bottom-up). As the most essential stratum is s3 (WHAT-analysis), the horizontal analysis for it is concluded by resulting conceptual structure or a set of the domain ontologies.

Analogous algorithms were developed and practically tested and evaluated by the authors during developing of distance learning courses for different branches of computational science and for artificial intelligence (AI).

Table 1. Matrix for OSA

Level → **Stratum↓**	**Domain Level in general (u_1)**	**Problem Level (u_2)**	**Sub-Problem Level (u_3)**	**.....**	**(u_n)**
Strategic Analysis s_1	E_{11}	E_{21}	E_{31}	E_{i1}	E_{n1}
Organisational Analysis s_2	E_{21}				
Conceptual Analysis s_3	E_{31}				
Functional Analysis s_4	E_{41}				
Spatial Analysis s_5	E_{51}				
Temporal Analysis s_6	E_{61}				
Causal Analysis s_7	E_{71}				
Economical Analysis s_8	E_{81}				
.....				E_{ij}	
s_m	E_{m1}				E_{mn}

4. Teaching Ontologies in Artificial Intelligence

We have developed more than 20 teaching ontologies (What-knowledge conceptual structures s3) helping to understand and to remember main concepts of AI. Fig.3 shows one of them (it includes a part of Fig.2).

We worked out several tips to add expressiveness to the ontology on the design stage.

1. Use different font sizes for different strata
2. Use different colours to distinct the subset or branch
3. Use vertical layout of the tree
4. If needed use different shapes of nodes

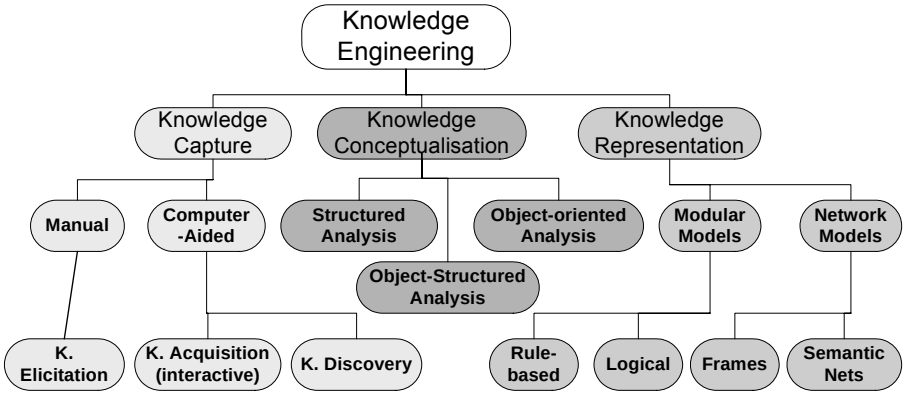


Figure 3. Ontology "Knowledge Engineering"

It is possible to use any of graphical editors to design ontology, e.g. PaintBrush, Visio, Inspiration. But computer program, which could be really useful for a knowledge engineer on the described stages of structuring of the subject domain, should necessarily follow the phenomenological nature of the knowledge elicitation and described above algorithms. This program must not frustrate the knowledge engineer with any "game rules" which were not evident for him/her. Ideally, it should adjust itself for particular cognitive features of the knowledge engineer. Moreover, each of the stages of analysis described above may be represented visually in its proper terms, as is already approached in some commercial expert system shells.

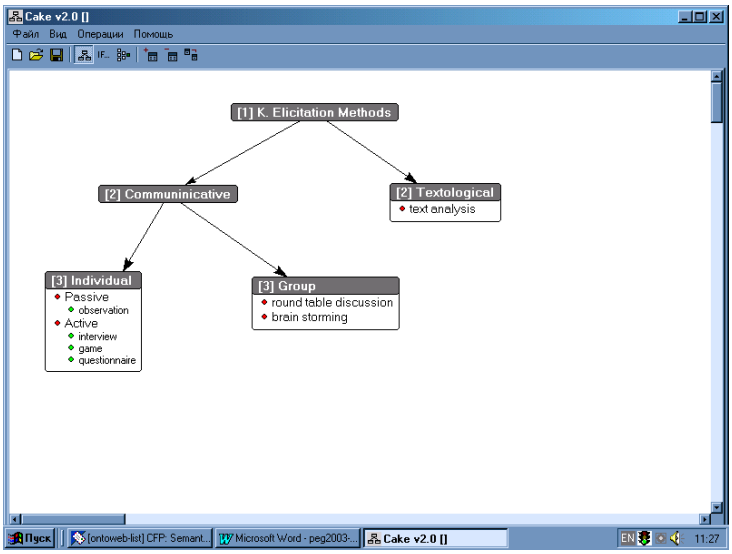


Figure 4. Screenshot with a part of ontology "Knowledge elicitation techniques".

A special visual tool was developed and named CAKE-2 (Computer Aided Knowledge Engineering by leading programmer Tim Geleverya, previous release by Alex Voinov). CAKE illustrates the idea of knowledge mappability, that find another application in the data mining and structuring for heterogeneous data base design. Its first prototype is described briefly in [Gavrilova, Voinov, 1996]. CAKE-2 proposes a kind of a visual knowledge representation language, which analogues may be found in a wide range of visual software construction tools – from large CASE's to Visual Basic. In particular, it supports the principle of a bi-directional mutually unambiguous correspondence between the two-dimensional visual object description syntax with the traditional one-dimensional one.

CAKE-2 is based on classical structured analysis methodology [Yourdon, 1989] enriched by new results that gives a teacher the opportunity to use special graphical interface to create ontology, to save it and to compile into the knowledge base (if needed).

Fig.4 presents CAKE's screenshot with fragment of the ontology of knowledge elicitation methods.

5. Discussion

Our approach puts stress on the knowledge structuring for better understanding of main course ideas in e-learning. The use of visual paradigm to represent and support the teaching process not only helps a professional tutor to concentrate on the problem rather than on details, but also enables pupils and students to process and understand great volume of information.

A better apprehension of teaching information might be achieved by imposing a knowledge structure on it. This may improve later usage of this information, comparing, generalisation, and so on. Therefore, a visual knowledge structure editor plays here a role of a two-dimensional, pictorial conspectus of the regarded piece of information.

Acknowledgements

The authors are very grateful to Tim Geleverya who assisted in the described research.

References

Adeli H. (1994) *Knowledge Engineering.* McGraw-Hill. New-York.

Boose, J.H.(1990) Knowledge Acquisition Tools, Methods and Mediating Representations. In: Motoda, H., Mizogochi, R., Boose, J., Gaines, B. (Eds.): *Knowledge Acquisition for Knowledge-Based Systems,* IOS Press, Ohinsha Ltd., Tokyo,123-168.

Conlon T. (1997) Towards Diversity: Advancing Knowledge-based Modelling with Knowledge Acquisition. In *Proceedings of 8th International PEG Conference.* Sozopol, Bulgaria.,379-386.

Eisenstadt M., Domingue J., Rajan T., Motta E. (1990) Visual Knowledge Engineering. *In IEEE Transactions on Software Engineering*, Vol.16, No.10, 1164-1177.

Fensel D. (2001) *Ontologies: A Silver Bullet foe Knowledge Management and Electronic Commerce*. Springer.

Gavrilova T.A., Voinov A. (1996) Visualized Conceptual Structuring for Heterogeneous Knowledge Acquisition // Proceedings of *Int. Conf. on Educational Multimedia and Hypermedia EDMEDIA'96, MIT, Boston, USA*, 258-264.

Gomez-Perez A. Tutorial on Ontological Engineering. *Int. Conference on artificial Intelligence IJCAI'99*.

Gruber, T. (1993) A translation Approach to portable ontology specifications. *Knowledge Acquisition*. Vol. 5, 199- 220.

Guarino, N., Giaretta, P. (1998) Ontologies and Knowledge Bases: Towards a Terminological Clarification. In *Towards Very Large Knowledge Bases: Knowledge Building & Knowledge Sharing*. IOS Press, 25- 32.

Jonassen, D.H. (1998). Designing constructivist learning environments. In C.M. Reigeluth (Ed.), *Instructional design models and strategies*, 2nd ed. Mahwah, NJ: Lawrence Eribaum.

Kremer R. (1998) Visual Languages for Knowledge Representation. In Proceedings of KAW'98 *Eleventh Workshop on Knowledge Acquisition, Modeling and Management*, Banff, Alberta, Canada, 124-129.

Neches et al, (1991) Enabling Technology for Knowledge Sharing. *AI Magazin,*. Winter, 36- 56.

Sowa, J. F. (1984) *Conceptual Structures: Information Processing in Mind and Machine*. Reading, Massachusetts, Addison-Wesley.

Scott A., Clayton J.E., Gibson E.L. (1994) *A practical Guide to Knowledge Acquisition*, Addison-Wesley Publ.

Swartout B., Patil R., Knight K., Russ T. (1997) Toward Distributed Use of Large- Scale Ontologies Ontological Engineering. *AAAI- 97 Spring Symposium Series,* 138- 148.

Tu S., Eriksson H., Gennari J., Shahar Y., Musen M. (1995). Ontology-Based Configuration of Problem-Solving Methods and Generation of Knowledge-Acquisition Tools: Application of PROTEGE-II to Protocol-Based Decision Support. *Artificial Intelligence in Medicine,* N7, 257-289.

Wielinga B., Schreiber G., Breuker J. (1992) A Modelling Approach to Knowledge Engineering. *Knowledge Acquisition*, 4 (1). Special Issue, 23-39.

Yourdon, E. (1989) Modern Structured Analysis, Prentice-Hall Bristol, Computer Science Department.

Authors' Information

Tatiana Gavrilova, Michael Kurochkin, Victor Veremiev –

Saint-Petersburg State Polytechnical University; Intelligent Computer Technologies Dpt.

Politechnicheskaya 29/9, St. Petersburg - 195251, Russia, Tel: + 07 812 371 82 72, Fax: + 07 812 373 06 89; gavr@limtu.spb.su

HANDWRITING IN FORENSIC INVESTIGATIONS

Georgi Gluhchev

Abstract: The process of automatic handwriting investigation in forensic science is described. The general scheme of a computer-based handwriting analysis system is used to point out at the basic problems of image enhancement and segmentation, feature extraction and decision-making. Factors that may compromise the accuracy of expert's conclusion are underlined and directions for future investigations are marked.

Keywords: Handwriting, Identification, Verification, Screening, Forensic Investigation, Image Enhancement, Feature Extraction, Decision-making.

1. Introduction

The individual's authentication becomes a serious problem concerning different areas of the social and economic relations in the world. Its importance increases as a factor in the prevention of terrorist actions and illegal access to important information. The problem attracts the attention of researchers all over the world and during the last years a few international projects were launched aimed at the development of reliable systems for authentication using biometric information.

Writer authentication is one of the broadly investigated modality in this aspect. Until now it was mainly used in forensic investigations dealing with handwritten document analysis or signature verification. Despite of the long history in that respect handwriting investigation still remains a difficult, time-consuming and subjective process, where qualified experts evaluate the similarity between letters, strokes and writing styles on the basis of their experience. In all cases of writer identification the objectiveness of the analysis and reliability of conclusion are of great importance. However, the inevitable variation in writing under different conditions and psychological state or when a significant time gap exists between the incriminated and reference documents may mislead the expert. Also, extremely difficult are cases where the handwriting is deliberately changed. In such situations different experts may disagree as to who is the writer of a particular document and a wrong conclusion may be drawn.

Other problems in writer identification concern the expert's workload during the analysis and difficulties stemming from sometimes poor quality of handwritten materials.

To overcome the problems quantitative methods for objective handwriting analysis and adequate decision-making have to be developed and implemented.

To achieve this, serious scientific investigation is required in order to develop appropriate methods for feature measurement, selection of reliable sets of features, evaluation of the minimal number of handwriting elements which is necessary for the reliable decision-making, suggestion of robust classification algorithms dealing with mixtures of continuous and categorical variables. Major difficulties in this direction stem from the qualitative character of most of the handwriting parameters used by the experts.

Despite that the problem of writer identification is of great practical importance in forensic investigations a relatively small number of papers have been published until now [5,11,14,22]. The existing computer systems are aimed especially at the screening of similar handwritings from a large data base of handwritings. After that the identification problem is carried out manually by an expert on the basis of his own experience and subjective evaluation of the similarity between the handwritings under investigation. For this he compares visually strokes, letters or combinations of letters performing sometimes simple measurements.

The goal of the paper is to sketch the frame of a computer-based handwriting investigation system and discuss the problems of its major components (Fig. 1). It is organized in the following way: Section 2 concerns the improvement of image quality; Section 3 deals with feature extraction techniques; Section 4 describes decision-making and Section 5 points out at the unsolved problems.

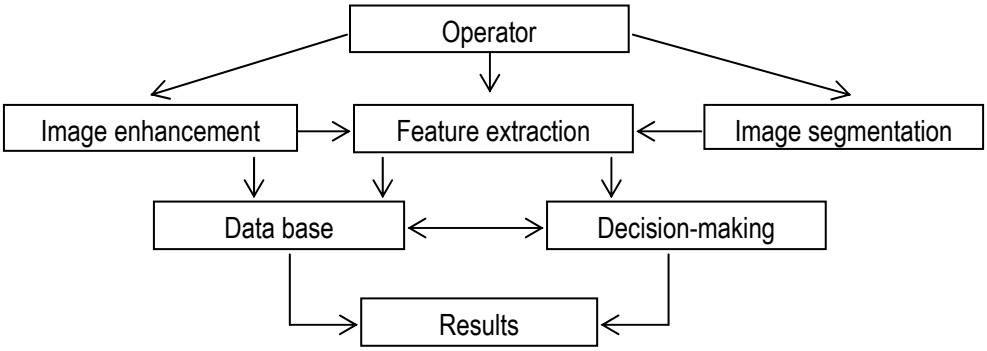


Figure 1. Block-diagram of a handwriting investigation system

2. Image Enhancement

The block-diagram in Fig. 1 shows that the development of a computerized handwriting investigation system follows the general methodology for the development of image processing and pattern recognition systems.

Since very often the handwritten materials are of poor quality, it is necessary to achieve some pre-processing. Its goal is two-fold: a) to improve image quality including contrast enhancement, random and structured noise reduction, and edge sharpening [2,3,9,17,19,21]. In such a way the image will offer better possibilities for the automatic analysis; b) to correct strokes and complex lines using morphological operations. This is especially important for the analysis of specific handwriting features, where the skeleton of the characters is used. Morphology allows to automatically connecting disrupted lines or cutting-off wrongly connected strokes.

3. Image Segmentation

Image segmentation is an essential step in the automatic document analysis. In handwriting analysis additional difficulties may arise due to the possibility of significant variation in rows', words' and letters' position.

The automatic segmentation includes background elimination as a first stage. Approaches based on histogram analysis (uniform background is supposed), locally adaptive binary trees and heuristic approaches (non-uniform background or presence of structured noise is assumed) are used.

The second stage concerns separation of rows. This operation is based on the analysis of histogram-like graphs obtained via horizontal projections of pixels, Hough transform or analysis of envelopes of continuous lines. While this operation could be easily achieved, special techniques are required for a proper detection of under-row and over-row elements of some characters. Also, the medial axis of the row may not be presented as a single straight line which requires piece-wise presentation.

The segmentation continues with the separation of the words. Different cases of concise writing or writing where letters are not connected between them are a challenge. For the word separation vertical projections of pixels from the corresponding row are analyzed, distances between envelopes of continuous lines are used, separation lines parallel to the predominant slope of the vertically oriented strokes could be applied, as well.

However, the most difficult problem concerns segmentation of letters and strokes. Except some special cases, e.g. child's writing, their delineation may be quite difficult even for a human being. A proper solution of this problem could be achieved via a man-machine dialog. The operator has to identify some specific points like end-points or vertexes of a polygon that encompasses a handwriting element. After that lines could be automatically investigated for the detection of points of intersection or bifurcation, local extremums and like [12,13,20].

4. Feature Extraction

Feature extraction is the crucial problem that should be solved. While during the last decades a common methodology for the handwriting analysis has been set up, many of the suggested features are of qualitative

character and are prone to different evaluation from different experts. Also, there are no strong recommendations as to what number of features is to be used for a reliable decision-making.

Different types of features may be investigated including graphometric, densitometric, categorical, model-based and topological invariants [2,5,6,7,8]. From the expert's point of view they are classified as general and specific features.

The general features are of categorical type and describe qualitative characteristics as: degree of connection between letters (usually three degrees are accepted: low, moderate and high), slope (right, left, upright), motion (rectilinear, curvilinear, angular or arched, loop-like, oval, wavy or spiral), elaboration (presence of ornaments), direction of movement (clock-wise or counter-clock-wise), quantity of movement (average number of strokes used to draw separate letters) and like [2,7,8,12,16]. They are difficult for automatic evaluation and are specified by the expert.

Specific features admit quantitative evaluation. They are known as graphometric and are aimed at the automatic or semi-automatic measurement of the following characteristics: distances between rows, height and width of letters, distances between letters, size of the above-row and under-row elements, distances between words, predominant slope, geometric parameters of handwriting elements like strokes, fragments and/or combination of characters [2,11,22].

Figure 2. a) Original image, b) Areas of different pressure

While many of the above-mentioned features can be easily imitated, there are features that are not clearly seen and, therefore, difficult to falsify. In that respect a special attention is paid to the distribution of pressure alongside the strokes (Fig. 2). It could be analyzed in different ways looking for a reliable description, e.g. evaluation of the geometric parameters of areas of different pressure and their mutual disposition at different writing elements, or pressure change alongside the skeleton of the elements. In that respect promising results have been reported in [15].

Except the described features, which are reasonable and intuitively clear for the experts, other features that do not express a particular property of the handwriting may be used as well. These include topologically invariant points associated with a particular character. According to this approach characters are divided into specific segments that can be transformed and compared piece-wise.

Another approach is based on the presentation of characters as elastic changes of an ideal model. Thus, a transformation between the model and the real character can be evaluated and its parameters used for classification.

Very important is the problem concerning the reliability evaluation of different set of features. A number of approaches may be used for this based on the information theory, statistics and classification power.

5. Decision-Making

The overall estimation of the similarity between two handwritings must be obtained as a combination of decision-making classifiers.

The decision-making for the specific features is based on the evaluation of the similarity between particular elements from the handwritings under investigation. Since the overall estimation will be based first on the estimations of separate elements and second on groups of elements, multi-level classifiers are to be used. The first level will concern the comparison of basic elements like strokes, letters and signs of punctuation. At the

output of these classifiers every element will be assigned a number that reflects the degree of similarity between the handwritings under investigation. Since a particular element may be detected in a few places in the text, an average similarity relative to this element will be calculated at the second stage. After the similarity is evaluated for all different elements, an overall evaluation is being obtained at the third level. One of the basic problems that has to be solved here concerns the weight factors of the elements, i.e. their classification power. Different types of decision rules could be used, including statistical, linear, heuristic, and NN-based [1,4,6,18].

The authenticity, where a forgery is expected, would be predominantly verified using stable features like pressure distribution. This is usually applied to small pieces of written text like particular words or signatures. Depending on how the pressure will be measured (areas of different pressure or a function alongside a skeleton line) different comparison techniques could be applied.

The categorical features are mainly used for the search of similar handwritings in a large database. Also, for the sake of one-to-one comparison mixed variables discriminant techniques could be used. A simple approach for the analysis of mixture of categorical and continuous data requires arbitrary scoring of all the categorical variables followed by the use of standard methods for multivariate continuous data, which in the case of classification means the use of techniques such as linear or quadratic discriminant analysis.

6. Discussion and Future Work

Different aspects and major problems that are to be solved for the development of a computer-based handwriting investigation system are described.

While the pre-processing stage of digital images is thoroughly investigated during the last decades, the well-known approaches may not work properly due to possible damages, background variation and/or poor image contrast. This requires locally adaptive methods to be developed, reflecting the specificity of the investigated images.

A big challenge is the selection of a reliable set of features. A computerized system must include as much as possible features that experts are accustomed to, but at the same time, special attention must be paid to the measurement of some parameters that are difficult for expert's evaluation, and therefore difficult for imitation, e.g. curvature at characteristic points, line smoothness or pressure distribution. Also, the expert has to have the possibility to interfere and suggest his one selection of features. This requires a friendly man-machine interface to be available.

The decision-making seems to be the most expert-independent part of the problem, since various objective measures of similarity (parametric, non-parametric, clustering) have been developed in pattern recognition theory. The different levels of similarity estimation will require the development of multi-level hierarchical classifiers.

Age-due variations in handwriting or changes due to different diseases must be investigated as well. This will increase the possibility for reliable writer identification when a significant time-gap between the handwritten materials exists or in case of a psychiatric disease [10].

A successful solution to all of the discussed problems will allow developing of a reliable and user-friendly computer system for handwriting analysis that could be implemented in police departments, bank and notary offices. It must be noted that such a system will help the expert to do an objective analysis, not to replace him.

The obtained solutions to specific handwriting analysis problems could be easily incorporated in a more complicated access-permit systems or person authentication systems at check points.

Acknowledgements

This work is supported by the Ministry of Education and Sciences under contract # *И* 1302/2003.

References

[1] Alexandre, Luís A., A. C. Campilho, M. Kamel. "Combining Unbiased and Independent Classifiers Using Geometric Mean", 11th Conf. of Portuguese Association on Pattern Recognition, 2000

[2] Angelov, A., D. Nestorov, S. Benchev, G. Gluhchev, D. Kamenov, P. Veleva. "A system for the automated handwriting investigation", In: Information Bulletine of RIFSC, Sofia, 1997, pp. 127-133

[3] Dimov, D. T., G. Y. Gluhchev. “A Locally Adaptive Binary-Tree Method for Binarization of Text Images”, In: Progress in Handwriting Recognition, (Edts A.C. Downton, S. Impedovo), World Scientific, London, 1996, pp. 575-580

[4] Gluhchev, G., V. Shapiro, D. Lalchev. "Probabilistic estimate of handwriting similarity", In: Theory and application of cybernetic systems, Bulgarian Academy of Sciences, Sofia,1991, 3,pp. 90-93

[5] Gluhchev, G., V. Shapiro, S. Ogorelkov. "Automatic investigation of handwritten characters", In: Forensic science and criminology, Sofia, 1989, 5,pp. 135-142

[6] Heutte, L., T. Paquet, J.V. Moreau, Y. Lecourtier and C. Olivier. "A Structural/Statistical Feature Based Vector for Handwritten Character Recognition", Pattern Recognition Letters, vol. 19, no. 7, pp. 629-641, 1998.

[7] Heutte, L., J.V. Moreau, B. Plessis, J.L. Plagnaud and Y. Lecourtier. "Handwritten numeral recognition based on multiple feature extractors", 2nd IAPR International Conference on Document Analysis and Recognition, IAPR-ICDAR'93, Tsukuba, Japan, IEEE Computer Society Press, pp. 167-170, 1993.

[8] Heutte, L., J.V. Moreau and Y. Lecourtier. "A new feature extraction strategy for handwritten pseudo-character recognition based on multiple feature extractors", 6th IGS International Conference on Handwriting and Drawing, IGS-ICOHD'93, Paris, France, pp. 186-188, 1993.

[9] Hummel R.A. “Image enhancement by histogram transformation”, Comput. Graphics and Image Processing, 6, 1976, pp. 184-197

[10] Kalcheva, E., G. Gluhchev. “Evaluation of handwriting changes in case of neurological diseases”, Proceedings of the CompSysTech ’2000, Sofia, pp.V.2-1–V.2-5

[11] Klement, V. “An Application System for the Computer Assisted Identification of Handwritings”, Proc. Int. Carnahan Conf. Security Technology, Zurich, 1982, pp. 75-79

[12] Lickforman-Sulem, L. “Extraction d’elements graphiques dans les images de manuscripts, CIFED’98, Quebec, Canada, 1998, pp. 198-207

[13] Lickforman-Sulem, L. and C. Faure. “A Hough Based Algorithm for Extracting Text Lines in Handwritten Documents”, ICDAR’95, Montreal, 1995, pp. 774-777

[14] Lima, Rui M., A. J. C. Campilho. "A System for Image Analysis of Documents", 7th Conf. of Portuguese Association on Pattern Recognition, pp. 1.5.1-1.5.6, 1995.

[15] Nestorov, D., V. Shapiro, P. Veleva, G. Gluhchev, A. Angelov, I. Stoyanov. "Towards objectivity of handwriting pressure analysis for static images", 6th Int. Conf. on Handwritting and Drawing ICOHD '93, 5-7 July, Paris, pp. 216-218

[16] Nosary, A., L. Heutte, T. Paquet and Y. Lecourtier. "Defining writer's invariants to adapt the recognition task", 5th International Conference on Document Analysis and Recognition, ICDAR'99, Bangalore, India, IEEE Computer Society Press, pp. 765-768, 1999.

[17] Pizer, S.M., E.P. Amburn, J.D. Austin et al. “Adaptive hiastgram equalization and its variations”,Comput. Vision, Graphics and Image Proc., 39, 1987, 355-368

[18] Plessis, B., A. Sicsu, L. Heutte, E. Menu, E. Lecolinet, O. Debon and J.V. Moreau. "A multi-classifier combination strategy for the recognition of handwritten cursive words", 2nd IAPR International Conference on Document Analysis and Recognition, IAPR-ICDAR'93, Tsukuba, Japan, IEEE Computer Society Press, pp. 642-645, 1993.

[19] Pratt W. K. Digital Image Processing, 2nd edn, John Wiley & Sons, 1991

[20] Shapiro, V., G. Gluhchev, V. Sgurev. "Handwritten document image segmentation and analysis", Pattern recognition letters, North Holland, 1993, 14,pp. 71-78

[21] Shapiro, V., G. Gluhchev, V. Sgurev. "Preprocessing for automatic examination of handwritten documents", 7th Scandinavian conf. on image analysis, Denmark, 1991,v.2, pp. 790-797

[22] Steinke, K. “Entwicklung von Mustererkennungsverfahren zur textunabhangigen Analyse von Handschriftenbildern”, Dissertation TH Aachen, 1981

Author Information

Georgi Gluhchev – Institute of Information Technologies, BAS, Acad. G. Bonchev Str., bl.2, Sofia - 1113, Bulgaria; e-mail: gluhchev@iinf.bas.bg

NEURAL NETWORKS: A DIAGNOSTIC TOOL FOR GASTRIC ELECTRICAL UNCOUPLING?

Catherine Gooi and Martin Mintchev

Abstract: Neural Networks have been successfully employed in different biomedical settings. They have been useful for feature extractions from images and biomedical data in a variety of diagnostic applications. In this paper, they are applied as a diagnostic tool for classifying different levels of gastric electrical uncoupling in controlled acute experiments on dogs. Data was collected from 16 dogs using six bipolar electrodes inserted into the serosa of the antral wall. Each dog underwent three recordings under different conditions: (1) basal state, (2) mild surgically-induced uncoupling, and (3) severe surgically-induced uncoupling. For each condition half-hour recordings were made. The neural network was implemented according to the Learning Vector Quantization model. This is a supervised learning model of the Kohonen Self-Organizing Maps. Majority of the recordings collected from the dogs were used for network training. Remaining recordings served as a testing tool to examine the validity of the training procedure. Approximately 90% of the dogs from the neural network training set were classified properly. However, only 31% of the dogs not included in the training process were accurately diagnosed. The poor neural-network based diagnosis of recordings that did not participate in the training process might have been caused by inappropriate representation of input data. Previous research has suggested characterizing signals according to certain features of the recorded data. This method, if employed, would reduce the noise and possibly improve the diagnostic abilities of the neural network..

Keywords: Neural Networks, Gastric Electrical Activity, Gastric Electrical Uncoupling

1. Introduction

Neural networks are useful tools in medical settings. They have been applied successfully in classifying various forms of Parkinson syndrome [1], in diagnostic electromyography [2], and in studying breast cancer disease [3]. These applications are usually implemented using Kohonen self-organizing neural networks [1-3]. Kohonen maps build clusters based on the similarities among input data. In the present study this technique will be applied in diagnosing gastric electrical uncoupling.

1.1. Medical Condition

Gastric motor function is an important part of the digestive process, and entails the storing, mixing and grinding of food, as well as its movement towards the intestines. This process requires the coordination of gastric smooth muscle contractions. Similarly to cardiac contractions, stomach contractions are preceded by electrical activity. These electrical events determine the frequency, velocity and direction of the contractions [4]. Accordingly, abnormal gastric function can occur when electrical signals are not synchronized, i.e. the electrical signals are uncoupled. To detect uncoupling, internal recordings of gastric electrical activities are made [5]. From these recordings it is important to be able to categorize the severity of uncoupling. This paper proposes the use of neural networks for such categorization.

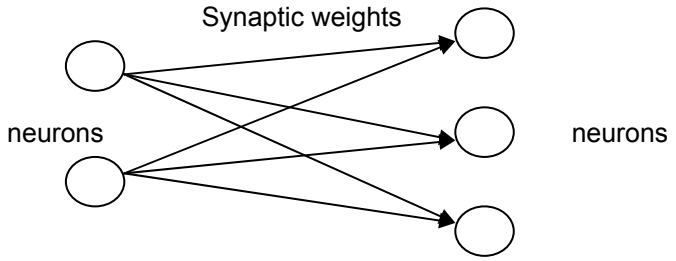


Figure 1: Simple Neural Network Architecture

1.2. What Is a Neural Network?

Neural networks consist of interconnected simple computing cells, referred to as “neurons” [6]. The strengths of the interconnections between these neurons are called synaptic weights (Figure 1).

Through modification of the synaptic weights, called learning or training, the neural network is able to store information.

1.3. Neural Network Training

In the first part of the training process the synaptic weights are initialized to small random values. Next, a training set of data is introduced to the network. There are two types of learning: (1) supervised learning, and (2) unsupervised learning. In supervised learning, a set of inputs along with target outputs is provided to the network. The network passes the inputs through the layers of neurons and modifies the synaptic weights according to a learning algorithm, which adjusts the outputs closer to those of the desired target outputs. In the case of unsupervised learning, no target outputs are provided. An example of unsupervised learning is the Kohonen map’s competitive learning [7-8]. Kohonen maps cluster input data according to their similarities.

1.4. Kohonen Maps

Kohonen self-organizing maps are a type of neural network. They consist of two layers of neurons, the input neurons and the output neurons. Each input neuron is connected to every output neuron [6]. An example of Kohonen map architecture is shown in Figure 2.

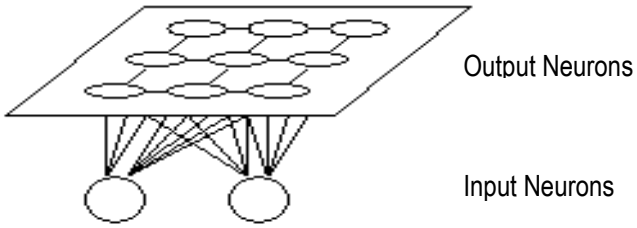


Figure 2: Kohonen Map

Kohonen maps learn in a competitive manner. First, the synaptic connections between the input and the output neurons are initialized, and each output neuron is characterized by a synaptic weight vector. Next, an input pattern (vector) is randomly selected. The Euclidean distance between the input vector and each synaptic weight vector is computed, and then the output neuron with the shortest Euclidean distance is declared the winning neuron. The synaptic weights of the winning neuron are adjusted, increasing the similarity between its synaptic weight vector and the input vector. Similarly, the weight vectors of the neurons in the proximity of the winning neuron are adjusted, increasing their similarity, but to a lesser degree than that for the winning neuron [8]. The algorithm for weight adjustment is:

$$\omega_{ij}(t+1) = \omega_{ij}(t) + \eta(t)\left(x_i(t) - \omega_{ij}(t)\right) \quad (1)$$

where $\omega_{ij}(t)$ is the synaptic weight value from input neuron i to output neuron j; $x_i(t)$ is the input to neuron i at time t; and $\eta(t)$ $(0 < \eta(t) < 1)$ is the learning rate coefficient.

Due to its ability to group similar data, competitive networks are particularly useful for diagnosis, allowing similarly characterized inputs to be clustered together. This learning method, however, does not allow the user to control the categories into which the input will be classified. Learning vector quantization (LVQ) networks, on the other hand, allow the user to classify the input vectors into predetermined categories.

1.5. Learning Vector Quantization (LVQ).

LVQ is a supervised learning technique employed in combination with Kohonen maps [8]. As illustrated in Figure 3, an LVQ network consists of an input layer, a competitive layer and a linear layer [9].

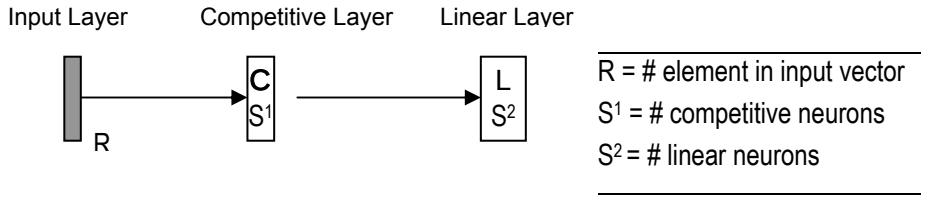


Figure 3: LVQ Network Architecture

The competitive layer classifies the inputs as described above, while the linear layer classifies the outputs from the competitive layer into target values. In other words, the outputs of the competitive layer are subclasses of the target layer. If the output of the input vector matches the target value, the weight vectors of the winning neuron, n_w, are modified with the following algorithm,

$$n_w(t+1) = n_w(t) + \eta(t)[x(t) - n_w(t)], \quad (2)$$

otherwise they are modified using:

$$n_w(t+1) = n_w(t) - \eta(t)[x(t) - n_w(t)], \quad (3)$$

Equation (2) moves the competitive neurons closer to vectors that belong in its same class, and Equation (3) moves the competitive neurons farther from vectors that do not belong in its same class. In Equations (2) and (3), $n_w(t)$ represents the winning neuron's present synaptic weight vector, i.e. at time *t*, $n_w(t+1)$ represents the winning neuron's modified synaptic weight vector, i.e. at time *t+1*, $\eta(t)$ $(0 < \eta(t) < 1)$ represents the learning rate coefficient, and $x_i(t)$ is the input to neuron *i* at time *t*.

2. Aim

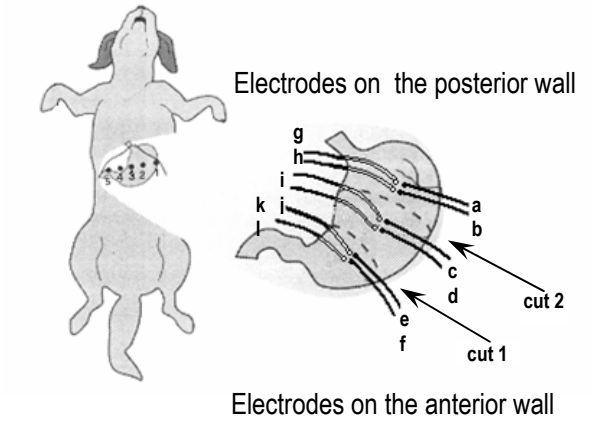
The aim of this paper is to apply Learning Vector Quantization neural networks in recognizing gastric electrical uncoupling from internal recordings of canine gastric electrical activity.

3. Experimental Design

3.1. Data Acquisition

In order to understand and recognize varying degrees of uncoupling, 16 anesthetized dogs underwent surgically induced gastric uncoupling [9]. Data were obtained from each dog in the three different states, basal, mild uncoupling and severe uncoupling. Six pairs of electrodes were placed on the gastric antral wall, three along the anterior and three along the posterior. These six pairs of electrodes provided 6 channels from which half-hour recordings of gastric electrical activity (GEA) were made for each state. During the first session the dogs were in the basal state, in the second session the stomach was divided by a single circumferential cut of the entire gastric muscle between the distal and the middle electrode sets, dividing the organ into two electrically active regions each oscillating at different electrical frequency, thus producing mild uncoupling. Finally, a second circumferential cut surgically divided the stomach between the middle and the proximal electrode sets, dividing the organ into three electrically active regions, simulating severe uncoupling (Figure 4).

Gastric electrical activity (GEA) signals were filtered in a frequency band of 0.02 - 0.2 Hz and digitized with a 10 Hz sampling frequency. In total, 18 000 samples were collected for each channel per session.



Internal Canine GEA

Channel	Electrode Combination
1	a-b
2	c-d
3	e-f
4	g-h
5	i-j
6	k-l

Figure 4: Data acquisition setup. The locations of the circumferential cuts are also denoted.

3.2. Neural Network Modelling

The aim of the neural network is to categorize the condition of the dogs into one of the 3 states:

1. Basal;
2. Mild Uncoupling: One circumferential cut;
3. Severe Uncoupling: Two circumferential cuts.

In view of the fact that the categories were predetermined, an LVQ network was chosen for the implementation. The network was created, trained and simulated using the Neural Network Toolbox from MatLab 6.0 (MathWorks, Natick, MA).

Since the neurons from the competitive layer form subclasses for the linear layer's target neurons, the number of neurons in the competitive layer should always be larger than the number of target neurons [9]. In addition, the number of neurons in the competitive layer should be smaller than the number of training examples, otherwise each training example would have a separate winning neuron in the competitive layer and the competitive layer would serve no purpose in the classification process. Given these limitations the number of neurons in the competitive layer was chosen to be six, two neurons belonging to each target class.

Next step in the implementation process was the training of the network. In order to train the network, data were to be selected and represented in a vector form acceptable to the input neurons.

3.3. Data Selection and Representation

To determine which data samples should be used, the data from dogs 1 through 16 were displayed using locally designed gastrointestinal signal acquisition and analysis software, GAS v. 3.0. Visual inspection indicated that channel 2 was not functioning for dog 1, thus it was not used for training the neural network. In addition, data from dog 2 and 3 were collected utilizing filters with smaller bandwidths. This was inconsistent with the data acquisition parameters utilized for the other records, so these two recordings were also disregarded when training the neural network. It was also noted that sometimes signals were not adequately registered within the first few seconds of recording, thus data for training and validation were extracted from the middle of the recording time. 5000 data samples from each channel were utilized. Each training session was, therefore, characterized by an input vector of 30 000 elements, 5000 from each of the six channels.

3.4. Training and Simulating

The LVQ network was trained by repeatedly feeding the network with data from dogs 4 through 13, and their corresponding target outputs. Each time the data was fed through the synaptic weights were modified according to the learning algorithm described earlier. The first output neuron was designated for the basal state, the second for mild uncoupling, and the third for severe uncoupling.

Data from dogs 4 through 16 were used for simulation and verification. The vectors for each case were input and the output was recorded and compared with the desired outputs. Outputs from dogs 4 to 13 provided verification

of the network ability to diagnose for cases it had seen before during the learning process and outputs from dogs 14 to 16 were used to demonstrate the network ability to generalize.

4. Results

4.1. Neural Network Training and Verification

Training was performed with thirty training vectors from Dog 4 to Dog 13 (three from each dog), and subsequently, seven simulations were executed with data from Dog 4 to Dog 16. The average result is shown in Table 1.

	Percentage of correct diagnosis
Dog 4 – Dog 13	89.5%
Dog 14 – Dog 16	31.0%

Table 1: Verification results

The performance of the network for diagnosing dogs within the training set was fairly high, at 89.5%. However, its generalization ability was poor and had an accuracy of only 31%.

5. Conclusion

The network diagnosed well for the training data, but was unable to provide accurate diagnosis for new cases. Performance of a neural network is directly related to the quality of its input data. Therefore, it is necessary for these data to contain sufficient information [10]. It is important to represent the input data in an appropriate fashion, eliminating noise where possible, and capturing characterizing features. Similar study [2] suggested that segments of the signal could be characterized by seven parameters, duration, spike duration, amplitude, area, spike area, phase and turns. This study involved diagnosing neuromuscular disorders based on the electromyography (EMG) recordings of muscle electrical activity. The study resulted in an accurate diagnosis in the order of 80%. As a proposal for further study, a similar approach might be applied for characterizing gastric electrical signals.

Acknowledgement

This study was sponsored in part by the Natural Sciences and Engineering Research Council of Canada

References

[1] Fritsch T., Kraus P.H., Pruntek H., Tran-Gia P.: "Classification of Parkinson Rating-Scale-Data Using a Self Organizing Neural Net", In: IEEE International Conference on Neural Networks, pp. 93-98, Mar28-April1 1993.

[2] Pattichis C.S., Schizas C.N., Middleton L.T.: "Neural Network Models in EMG Diagnosis", IEEE Transactions on Biomedical Engineering, v42, n5, pp. 486-496, Piscataway, NJ, May 1995.

[3] Allan R., Kinsner W: "A Study of Microscopic Images of Human Brest Disease Using Competitive Neural Networks", In: Canadian Conference on Electrical and Computer Engineering, v1, 2001.

[4] Sanmiguel C.P., Mintchev M.P., Bowes K.: "Electrogastrography: A noninvasive technique to evaluate gastric electrical activity", Canadian Journal of Gastroenterology, vol. 12, n. 6, September 1998.

[5] Mintchev M.P., Otto S.J., Bowes K.L.: "Electrogastrography Can Recognize Gastric Electrical Uncoupling in Dogs", Gastroenterology 112:2006-2011, 1997.

[6] Haykin S.: "Neural Networks a Comprehensive Foundation", New Jersey: Tom Robins, 1999.

[7] Kohonen T.: "Self-Organizing Maps", New York: Springer-Verlag Berlin Heidelberg New York, 1997.

[8] Beale R., Jackson T.: "Neural Computing an Introduction", Bristol and Philadelphia: Institute of Physics Publishing, 2001.

[9] Demuth H., Beale M.: "Neural Network Toolbox User's Guide", The MathWorks Inc., 1998.

[10] Noyes J.: "Training and Generalization", In: Handbook of Neural Computation, IOP Publishing Ltd and Oxford University Press, pp. B3: 5:4, 1997.

Authors' Information

Catherine P. Gooi – Department of Electrical and Computer Engineering, University of Calgary, Calgary, Alberta, Canada T2N 1N4

Professor **Martin P. Mintchev** – Ph.D. P.Eng. Department of Electrical and Computer Engineering, University of Calgary; 2500 University Drive NW, Calgary, Alberta, Canada T2N 1N4; Phone/Fax: (403)220-5309, e-mail: mintchev@enel.ucalgary.ca

FOURIER NEURAL NETWORKS: AN APPROACH WITH SINUSOIDAL ACTIVATION FUNCTIONS[1]

Luis Mingo, Levon Aslanyan, Juan Castellanos, Miguel Díaz, and Vladimir Riazanov

Abstract: This paper presents some ideas about a new neural network architecture that can be compared to a Fourier analysis when dealing periodic signals. Such architecture is based on sinusoidal activation functions with an axo-axonic architecture [1]. A biological axo-axonic connection between two neurons is defined as the weight in a connection in given by the output of another third neuron. This idea can be implemented in the so called Enhanced Neural Networks [2] in which two Multilayer Perceptrons are used; the first one will output the weights that the second MLP uses to computed the desired output. This kind of neural network has universal approximation properties [3] even with lineal activation functions.

Enhanced Neural Networks

The only free parameters in the learning algorithm are the weights of one *MLP* since the weights of the other *MLP* are outputs computed by a neural network. This way the backpropagation algorithm must be modified in order to propagate the *Mean Squared Error* through both *MLPs*.

When all activation functions in an axo-axonic architecture are lineal ones ($f(x)=ax+b$) the output of the neural network is a polynomial expression in which the degree n of the polynomial depends on the number m of hidden layers [2] ($n=m+2$). This lineal architecture behaves like *Taylor* series approximation but with a global schema instead of the local approximation obtained by Taylor series. All boolean functions $f(x_1,\dots,x_n)$ can be interpolated with a axo-axonic architecture with lineal activation functions with n hidden layers, where n is the number of variables involve in the boolean functions. Any pattern set can be approximated with a polynomial expression, degree $n+2$, using an axo-axonic architecture with n hidden layers. The number of hidden neurons does not affects the polynomial degree but can be increased/decreased in order to obtained a lower *MSE*.

This lineal approach increases *MLP* capabilities but only polynomial approximations can be made. If non lineal activation functions are implemented in an axo-axonic network then different approximation schema can be obtained. That is, a net with sinusoidal functions outputs *Fourier* expressions, a net with *ridge* functions outputs ridge approximation, and so on. The main advantage of using a net is the a global approximation is achieved instead of a local approximation such as in the Fourier analysis.

[1] Supported by INTAS 2000-626, INTAS YSF 03-55-1969, INTAS INNO 182, and TIC 2003-09319-c03-03.

Sinusoidal Activation Functions

In short, considering that only output neurons of the net computing weights have sinusoidal functions, then the output of this net is:

$$o_j = \sum_{i=0}^{N} w_{ij} o_i \text{ , where } f(x) = \sin(ax + b)$$

Taking into account that weights w_{JI} of the other net are computing by previous output o_j, we can say that $w_{JI}=o_j$, where $j=J*N+I$. Then, desired output follows equation:

$$o_k = \sum_{l=0}^{N_I} w_{kl} o_l \text{ , where } w_{kl}=w_{JI}=o_j=sin(.)$$

Therefore, considering p hidden layers in the net, output can be expressed as:

$$o_k = \sum_{n=0}^{n=l+2} A \sin^n(.) + B$$

Considering previous equation, equations involved in *Fourier* analysis are similar to those obtained by axo-axonic networks.

Results

This section uses three neural network architectures in order to evaluate their forcasting properties. First of all we will present obtained results of a Multilayer Perceptron and a Time-Delay Neural Network, and after that we will show the Time-Delay Enhanced Neural Network results.

The data set has been obtained from the IBEX 35 index. Each pattern represents a day in the stock market and it consists on 5 variables: open value, close value, min value, max value and volume. The pattern set is made up of 8 years, that is 2502 patterns. The desired output is the close value of the IBEX 35 index in time $t+1$, that is a forecasting of the future behavior of the signal, without iterative prediction.

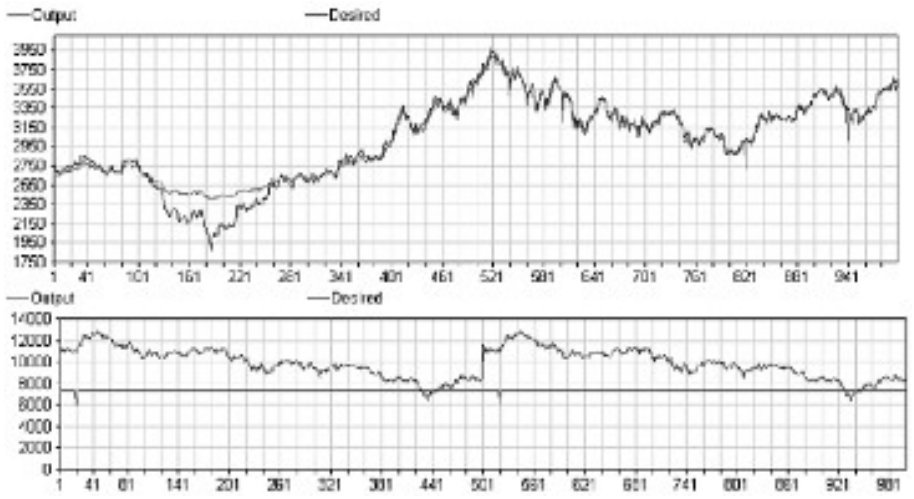


Figure 1. Desired and MLP net output of IBEX-35 stock market. First figure corresponds to the training set, and the second to the cross validation set. Training MSE=0.007585, Cross Valitation MSE=0.983621.

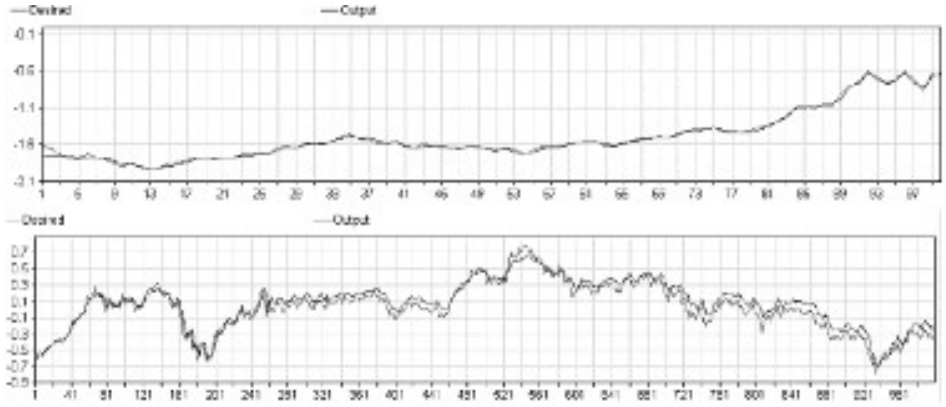


Figure 2. Desired and Time Delay Net output of IBEX-35 stock market. First figure corresponds to the training set, and the second to the cross validation set. Training MSE=0.000585, Cross Validation MSE=0.003621.

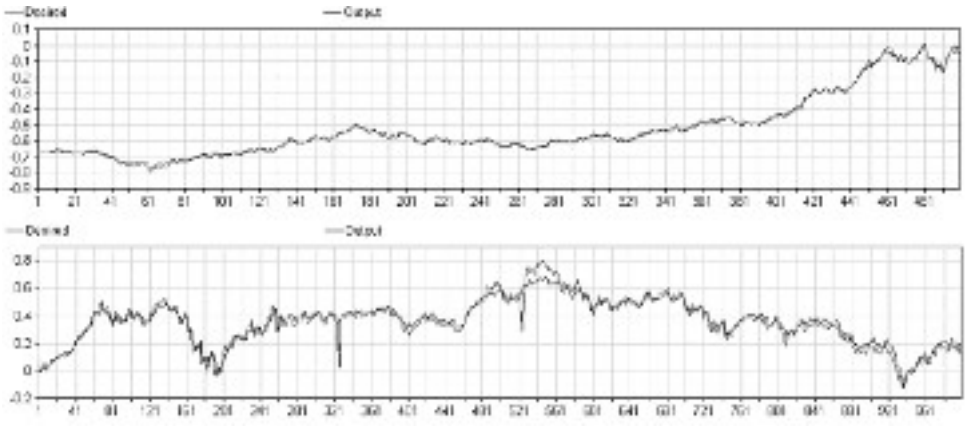


Figure 3. Desired and ENN output of IBEX-35 stock market. First figure corresponds to the training set, and the second to the cross validation set.Training MSE=0.001703, Cross Validation MSE=0.000803.

According to previous figures we can see that ENN with tap delays improves obtained results using TDNN (table 1), this fact is due to the universal approximation capabilities of ENN.

	ENN (fig. 3)	TDNN (fig. 2)
Training set	0.001703	0.000585
CV set	0.000803	0.003621

Table 1. MSE in training and cross validation sets.

Conclusions

Multilayer perceptrons are a subset of non lineal axo-axonic networks, since axo-axonic architectures with a given weights configuration behaves same way as a multilayer net. That is the reason why axo-axonic nets have more approximation capabilities than *MLP*. But the degree of output equation in a *MLP* can not be computed a priori such as in the axo-axonic architectures. Proposed architecture outputs a sinusoidal basis and a non lineal combination of it in order to obtained desired output and it can be compared to a *Fourier* analysis, moreover,

Fourier coefficients can be used to initialize weights of neural networks in order to start the learning process with a low error.

Axo-axonic architectures can be used to forecast signals since they behave like *Fourier* analysis. This kind of connections can also be implemented on *Time-Delay* networks to improve results when dealing periodic signals. Some applications have been developed in order to forecast stock markets, weather and load demand.

Bibliography

[1] Mingo L.F., Diaz M.A., Gisbert F., Carrillo J.D.: Enhanced Neural Networks: An Overview. Knowledge-based Intelligent Information Engineering Systems and Allied Technologies. Frontiers in Artificial Intelligence and Applications. IOS Press Ohmsha. ISSN: 0922-6389. ISSN: 1535-6698. Vol.: 69. Part II. Pp.: 1170-1173. 2001.

[2] Luis F. Mingo, Miguel A. Diaz, Perez M., Palencia V.: Enhanced Neural Networks as Taylor Series Approximation. Knowledge-Based Intelligent Information Engineering Systems \& Allied Technologies. Frontiers in Artificial Intelligence and Applications. Vol. 82. IOS Press. ISSN: 0922-6389. Pp.: 532-533. 2002.

[3] de Mingo L.F., Gisbert P., Carrillo J., Hernandez C.: Regression Hypersurfaces with a Modified Neural Network Architecture. Advances in Soft Computing. ISSN 1615-3871. Pp: 241-249. 2003.

Authors' Information

Luis Mingo – Dpto. Organización y Estructura de la Información - Escuela Universitaria de Informática; Universidad Politécnica de Madrid - Crta. de Valencia km. 7 - 28031 Madrid SPAIN - lfmingo@eui.upm.es

Miguel Díaz – Dpto. Organización y Estructura de la Información - Escuela Universitaria de Informática; Universidad Politécnica de Madrid - Crta. de Valencia km. 7 - 28031 Madrid SPAIN - mdiaz@eui.upm.es

Levon Aslanyan – Institute for Informatics and Automation Problems - National Academy of Sciences of Armenia; P. Sevak 1 Str., Yerevan 14 Armenia - lasl@sci.am

Juan Castellanos – Dpto. Inteligencia Artificial - Facultad de Informática - Universidad Politécnica de Madrid 28660 Boadilla del Monte - Madrid SPAIN - jcastellanos@fi.upm.es

Vladimir Riazanov – Computer Center of Russian Academy of Sciences, Vavilov, 40, Moscow - rvv@ccas.ru

MUSIC AS THE SOURCE OF INFORMATION INFLUENCE AND SOUL EDUCATION

Larissa Kuzemina

Extended Abstract

Unlike other works of art (painting, sculpture, etc.) a musical composition should be performed, it should sound to become accessible. Therefore, the role of the musical masterly performance is extremely important. But presently it has increased in importance when music through mass communication media i.e. radio, television, sound recording becomes in the full sense of the word the property of millions.

Art in all its genres as a means of information helps to recreate a picture of one or other epoch as a whole. Moreover, art has a profound impact on education: it can be positive or negative, creative or destructive. Let us dwell on such aspect of music as means of information and the value of musical mastery activity for brining information to hearers of the alternating generations.

Unlike other works of art (painting, sculpture etc.) a musical composition should be performed, it should sound to become intelligible. Therefore, the role of the musical masterly performance is extremely important. But presently

it becomes particularly great in the XXI century when music becomes a true property of the masses due to mass media – radio, television, sound recording.

The educative importance of music as a means of forming the spiritual culture of a person grows as well. Clearly, a great deal depends on performers. They can breath new life into the musical composition, but they can necrotize it. They can educate with music, they can awaken love in it, and, vice versa, discourage a person from interest, love for music etc.

At present the research on the theoretical and practical aspects of the musical masterly performance is paid a great attention. But far from being all is developed in the field of the mastery art theory. In particularly, the masterly performance concept developed by B.L.Yavorsky, one of the greatest native musicians-researchers, known, first of all, for his own theory of fret rhythm afterwards called a theory of musical thinking, was kept in the background. The views of the scientist were repeatedly the subject of speculations, debates, discussions. He repeatedly wrote himself for the press outlining the fundamental tenets of his theoretical system. Despite controversial points the system developed by Yavorsky appealed to researches for its singularity, novelty, breadth of phenomenon scope in art, stimulated the creative search.

The basis for this system is the association of the theory with the history of the musical practice development and, first of all, an objective scientific investigation into particularities of musical thinking as a reflection of the social reality.

Long 22 years the name of Yavorsky was not mentioned despite the fact that all the works on the musical theory with all their dissimilarity from the teaching of B.L. Yavorsky were linked in one way or another with it, borrowing from it at times rather significant and at other times side elements and appropriate terminological innovations (the case in the point are the works of 40-ies.

Interest in studying the heritage left by B.L. Yavorsky increased noticeably after publication of the first volume of the collection “B.L.Yavorsky. Articles, reminiscence, correspondence”. In the 70-ies the works on different aspects of theoretical heritage of Yavorsky appeared. Among them there are the papers by A.Farbshtain, V.Bajevsky, M.Aranovsky, Yu.Kohn, I.Naboque, L.Maslionkova and others. All the authors underline the historical importance of the theory created by Yavorsky in explaining continuous flow of a musical thought based on the principle of inclination to contradictory elements, in approaching music with a wide range of phenomena, interdisciplinary sciences and arts, this gives the possibility to reveal the features in common with different types of artistic creative work, types of thinking and perception, favors their mutual enrichment and further development. Theoretical tenets developed by Yavorsky are distinguished not only for a profound scientific substantiation, but also for interconnection. He is “the first Russian scientist who created the generalized system embracing all musical speech as a whole”.

Systematically the process of revealing the musical composition “biography” can be presented in the way, presented in figure 1.

As K.V. Zenkin asserts, the level of historical-theoretical generalizations reached by Yavorsky is the problem of the future of musical science and the history of culture. At the same time the author says about dialectics of historical concept created by Yavorsky, about filling it with a powerful pulse of Russian philosophical thought of XIX-XX centuries. Each epoch, according to Yavorsky, passes a number of development stages (cognitive, motive, emotional, volitional, contemplative) and at the same time it is the integrity. Style is the constant of the epoch. That is why speaking about the musical-historical concept of Yavorsky it is possible to speak about the concept of musical styles. The ideas of Yavorsky about music are the ideas about integral organism in the whole completeness of untangling the spiritual-psycho-physiological image of a person in the sequence of the age phases.

Yavorsky introduced the concept of a symbol into the musical theory and practice and made it in a conclusive and evident way. (K.Zenkin. Symbolism in Yavorsky’s thinking). The concept of a symbol was universal for Yavorsky, it was universal helping to recreate the integral philosophy of culture and art. Condition of searching for analogies between different arts was a thought about the unity of the universe, it was filled with a vivid sensual representation. Yavorsky’s world outlook was a dynamic one.

Unity of theory and practice was the basic one in the activity of Yavorsky. He paid a great attention to training musicians-performers, creation of theoretical foundations of performance mastery. In his new methods of educating musicians he was striving for realization of the universal concept of the mastery based on the synthesis of inventing and performance. Yavorsky introduces the idea of a symbol into mastery offering the **idea of the gesture-symbol,** which **must correspond to the symbol of movement** – volitional, flying, emotional, motive – depending on the style of the work being performed. Diversity of touches to a key, orchestration of the piano widened the boundaries of the performance individualization. Striving for a profound understanding of musical symbols Yavorsky turns to other arts having created the universal concept of mastery based on the styles.

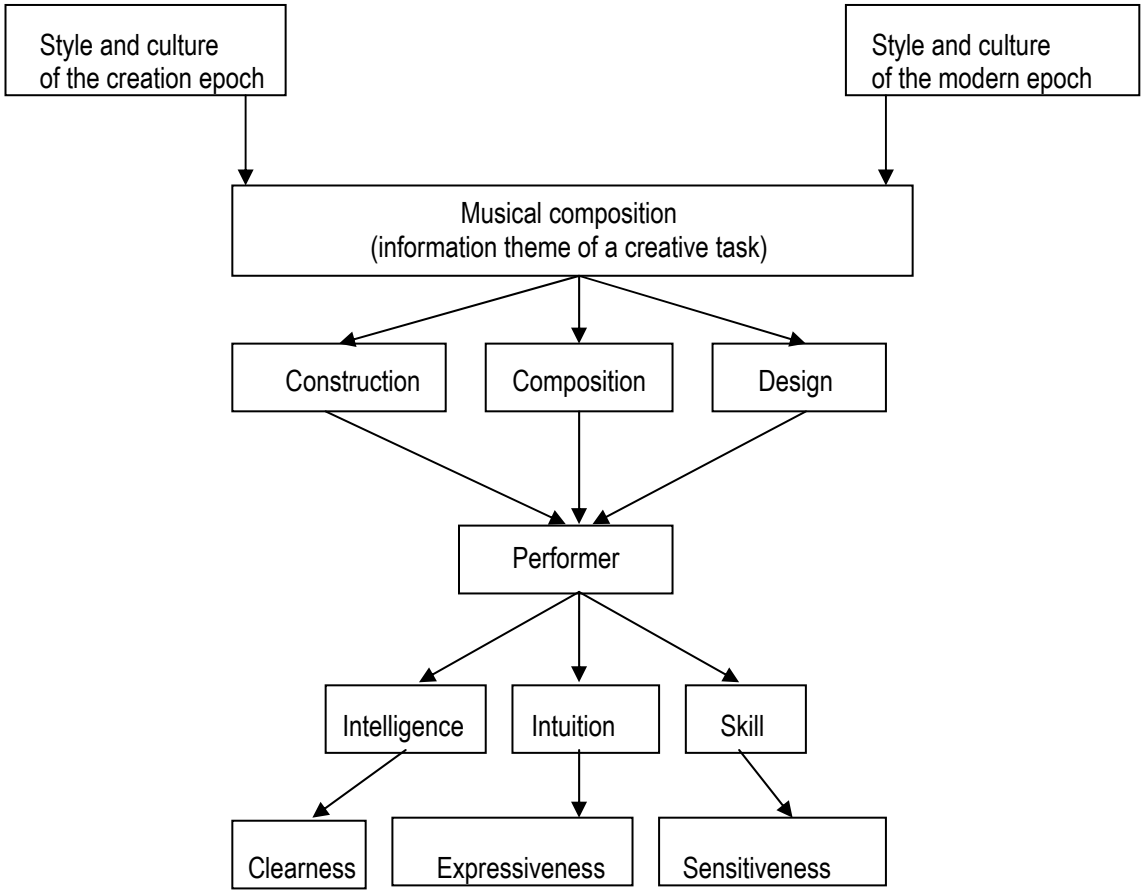


Figure 1

There sources of brining information and sources of its perception in the musical art.

According to this Yavorsky recognizes “five types of musical performance:

1. No mention could be made if poor performance is met so often.
2. Performance reproduced not by a performer himself, but representing totality of the impact of education, teacher or tradition on a performer; such a type is assumed to be named the unskilled performance […].
3. Performance of the temporal architectonic sound construction deprived of a vivid rhythm inherent to it, contributed into the composition by the author’s spirit and presenting on this basis the whole process of the author’s sensations. With such a type of performance the composition is perceived only by eyes: such a performance filled with good intentions can be called a bureaucratic one: it satisfies the most the modern middle-level cognition.
4. Performance when artistic composition is considered by a performer as a possibility to reveal his own personality without attending to the question whether this rhythmic of the composition fits to that rhythmic with

which it was created by the author. Here one is forced to come into collision with the problem of pithiness, interest, significance of the performer himself […]. With such a performance the composition is also perceived only with eyes and then the performer provides it with his vivid rhythm expressing his emotional experience and the hearer attention is taken up perceiving it, he pays little attention the composition or even doesn't notice it. It is needless to say that performance of the academic type cannot exist absolutely without the influence of the performer's personality, but this influence is not so strong to divert the hearer's attention from the sound architectonics and make him to concentrate himself on revealing the performer personality.

5. Performance, the author dreams about, where the author's personality (to what extent it has manifested itself) and the performer's personality harmonically merge; the vivid rhythm, resulting when performing the composition, owes to its origin to the presentation of the author's fillings revived in the fillings of the performer adding his own individuality not contradicting to the possibilities contributed by the author. With such a performance the performer perceived the composition not only with his eyes but he perceived its whole form with his essence and according to this form he reproduced fillings and transmitted them…" (the model is presented in the Diagram 1.)

According to this a specific performer's capability for perception generates a reciprocal capability in the hearers. It is no coincidence that the performers' typification continues here the hearers' typification:

"Equally there are five types of audience:

1. Inadequate audience, which is not able to listen: the reasons for this can be different.
2. The audience paying attention only to the scholastic accuracy of performance: absence of stoppage, errors, presence of highly specialized methods (scales, octaves, mediants, sixth, leaps, glissando, flageolets, etc.) what is the hand, finger, voice training, whether words are clear, how vowels and consonants are pronounced, whether the performance answers the prescribed tradition.
3. The audience requiring to perform the composition exactly as it was written, as it is seen with eyes, just with the eyes of the listener, but the listener's eyes know sometimes the author's remarks so poorly that attribute precision of the author's remarks to the performer's self-will. Such an audience likes to hear the compositions well-known to it; […] it particularly rises against manifestation of the performer's personality, […] having heard the performance presenting the same composition in the clearly expressed new presentation trying to elucidate the contents of the composition in its own way, maybe even at the expense of appearance, the audience does not become aware of the performance and, at best, gives no ear to it indifferently.
4. The audience searching not for a good performance of the composition but manifestation of the performer's personality; that is why deterioration of the composition has little or no effect on this audience and it makes difference between compositions only because they give more or less space for the performer's self-will, but those compositions that do not allow such a self-will for some reason or other are referred by this audience to the rank of scientific, serious, dry music […].
5. The audience […] capable to estimate the performance, where harmonically united the author's and performer's personalities, to estimate both the author and the performer taking advantage of their joint manifestation, to get the same fillings, imparting at the same time individual particularities inherent to every hearer".

For Yavorsky the work on the composition was the process of studying its "biography" i.e. getting a comprehensive and complete information about it.

In this respect memories of L.A. Averbuh about the work of Yavorsky on the march by Vagner –List from the opera "Tanguiezer": This march should draw in the people imagination the medieval splendid festival, the whole environment of the castle, grandeur of the feudal lord, diversity of the ranges and dresses, difference of steps, perspective of the long procession". Yavorsky told about a secluded life of the inhabitants of the medieval castles waiting impatiently for the festival held once a year, a herald was sent to a tower (here Yavorsky showed the pitch of the trumpet so that an image-bearing representation favors perspective space piano sounding). Representation of the order of the procession, a step which could be when foot-wear had non-bending soles, accompaniment of very picturesque, figurative detailed account using performance made it possible to obtain the desired result – the march was performed vividly, with necessary perspective of sounding, having preserved all the wealth of colors of the original text for the orchestra.

One of the examples in the system for performer's analysis developed by Yavorsky is the analysis of the of the play by List "Sposalizio" from the cycle "Years of wandering". The story of creation of the play from this cycle is known from the musical history. Yavorsky strove to find more vivid visual image which would help to perceive more profound the musical image (Yavorsky was far from artificial "attaching" of musical compositions to the works of art literature etc. It was one of the possible methods directed to mobilization of creative fantasy, creative imagination of the musician which facilitated in the given case interpretation of the program music and widened parallels with the compositions of adjoining arts, horizons in the field of music through the search of analogies). Raphael's picture of the same name was such a striking illustration.

The myth about the betrothal of Maria and Joseph overcame the double temporal distance: Raphael brought it in the conditions of the XVI century in Italy and then the impression produced by the picture was found in List's composition. Yavorsky told about the plot of the picture and paid attention to the peculiarities of the architecture of Italian cathedrals known to him from his visits to Italy, where a cathedral and a church bell tower were built together, about a peculiar sonority emerging due to connection of the sounding of the big bell and the bells of the small dome with silver hand-bells (held usually by boys). If the action takes place before the cathedral in the Raphael's then List's music is perceived as if the action takes place in the cathedral itself. That is why the scientist joins the plot of the picture with the description of the Italian cathedrals' architecture. Monographic data were accompanied with a number of requirements:

1. compiling of the vocabulary of terms met in the composition;
2. graphical presentation of the composition dynamic scheme;
3. definition of the instruments creating the timbre diversity.

All the requirements were directed to the revealing the constructive-composition traits of the composition and design corresponding to it. Yavorsky connected the composition fret instability with two moments: the first, with the instability of the composition's' plot itself, the second, with the visual perception of the architecture style of Gothic creating the feeling of instability. Development of the fret foundation of the composition was regarded as comparison of principles-symbols in the unity of constructive-composition traits: invocatory chime, blows of the bell, sounding of the organ and choir, intonation of the question-answer connected with the ceremony, a light chime. The constructive- composition scheme was created in correspondence with the sequence of the composition development.

Introduction (beginning) includes 29 bars and symbolizes the first sounds of the bell in the empty cathedral signifying the call. Gradually people arrive. The next 38 bars (8-30) are the plot of the composition – the choir sounds during the wedding, its melody is interlaced with the ringing of the hand-bells, medium-size bells, organ. Before the intonation of the question-answer appeared (bars 60 – 67) silence set in the cathedral. Then the sounding increases – the action is developing. The bar 67 is a distinctive division of two-part symmetrical structure of the composition: the first part consists of 67 bars, the second one consists of 66 bars + fermata. So, the events (peripetia) signify animation in the cathedral (45 bars) concerning the ceremony taking place on the background of the inviting ringing naturally brining to the culmination-triumph (bars 109 –112 + fermata). And in conclusion of the composition (bars 113-120) the intonations of the introduction sound mournfully, events move away from us, everything come to a standstill. The last 14 bars are culs-de-lampe of the composition.

Yavorsky presented his ideas of the figurative contents of the composition exactly and impressively in the criticism concerning its performance by Ferruccio Buzoni "giving neither the general sentiments in the cathedral and action, nor all-filling, all-penetrating bell sounding from the vague distant bell's call till the ecstatic symphony of bells ending with voices of hand-bells in the empty disappearing in darkness of the church".

Information underlies any activity (possession of information, its quantity and methods of its transmission). All this influences psychological self-programming and the choice of the corresponding valuable orientation.

Author Information

Kuzemina L.A. – Prof. of Kharkiv State University of Arts, Ukraine. kuzy@kture.kharkov.ua

DIGITAL CREATIVITY: ADVANTAGES, PROBLEMS, RESPONSIBILITIES

Velimir Velev

Abstract: The paper focuses on the rapid development of the digital culture and the challenges it imposes to human creativity. It analyses e-learning, digital entertainment, digital art and the issues of creativity and improvisation. It also presents a classification of the levels in the creative structure including hardware and software tools; product developers; creators and end users. Special attention is paid to the advantages of the new digital culture and the responsibilities of all people who create it or use it. We conclude that more attention should be paid to the threats and to ways of boosting positive creativity in the various fields of application of information and communication technologies.

Keywords: digital art, creativity, improvisation

Introduction

The rapid development of information and communication technologies (ICT) implied important changes in the human lives. The earliest uses of computers were very narrow-profiled and applied in specific sectors and professions. Nowadays, ICT accompany people from early childhood in the educational system, at work, in the societal organisation, in everyday life activities, and at leisure. This trend seems so natural that one important shift was left somewhat behind the focus: the function of computers changed. Starting with support for humans in routine tasks, nowadays computers became in the centre of wide range of tasks, which involve human creativity. This is obvious in areas related to computer art, but is equally important in entertainment and educational applications, which form the fabrics of the human personality, including the creativity.

The basic aim of this paper is to increase the awareness of creativity related issues in the modern computer world and its influence on the human personality.

We provide an analysis of the typical areas influenced by the new digital culture: e-learning, digital entertainment, digital art and discuss the issues of creativity and improvisation in the digital environment.

We continue with a presentation of our view on the digital culture ingredients and their inner dependencies. The basic ingredients of the digital culture include hardware and software tools, developers and users. The users basically form two groups – consumers and creators. One would expect that one consequence from the use of the ICT would be increased creativity and improvisation. This matter is not studied yet and we hope that this paper will raise the interest to it.

Finally, we present a SWOT (Strengths-Weaknesses Opportunities-Threats) analysis of the digital culture. Our basic conclusion is that more attention should be paid to raise the awareness in the threats and to boost positive creativity in order to support the development of the unique distinctive creative nature and potential of the human personality and to save the human values in the society as a whole through the various fields of application of ICT.

Forms of Life of Digital Culture

Nicholas Negroponte, the founder and the director of the Media Laboratory at the Massachusetts Institute of Technology, wrote in the introduction to his book *Being Digital*: “Computing is not about computers anymore. It is about living" (p. 6, [Negroponte 1996]).

The digital technologies led to the creation of a new world, a ‘different reality’ that we often call ‘virtual space’ or ‘virtual environment’ (VE) The technological developments, which make possible human beings to immerse into it, are becoming more and more mature. Almost everything which one could ever imagine could be modelled inside of the virtual environment: one can ‘live’ and ‘do’ there, i.e. one can simulate all types of human activity.

There are several application fields, which are particularly strongly influenced by the idea of virtual environment and human immersion:

- o E-learning
- o Digital entertainment
- o Digital art and its genres

E-learning

E-learning is defined as "The use of new multimedia technologies and the Internet to improve the quality of learning by facilitating access to resources and services as well as remote exchanges and collaboration."[1]This definition puts the emphasis on two basic issues: specific technologies (multimedia, Internet) and on the process of learning. In this field, there is enormous space for creativity. We are moving to the development of a virtual community, which shares a common educational content.

However, the e-learning should not be overestimated. Even if fields like linguistics or pure information, there is a small share, which can be filled in only by the real practice. Without it nowadays and in the future the process of learning could not be completed adequately. This is not only due to the fact that devices simulating physical sensation (haptic devices, olfactory interfaces, immersive visual interfaces) are still not perfect, but because of the impossibility to simulate and substitute the human presence. The contact with the other person, no matter whether he is a supervisor, partner or patient, and most over the vivid human feedback could not be replaced even in the most advanced digital technologies. To many people this statement could seem old-fashioned, but we should not neglect that in each human sphere there is a certain amount of knowledge, which can be transmitted only via personal contact.

Michael Polanyi, the creator of the theory of the personal knowledge, concludes that there are two types of human knowledge – the first one is obvious and logical, and the other one is hidden, peripheral [Polanyi 1962]. The second one is acquired through studying of the internal structure of the object of study and this type of knowledge can be transmitted only through personal contact between the teacher and student. The obvious knowledge is transferred logically and verbally. The hidden knowledge is transmitted only directly in the contact of teacher and student.

The informal personal knowledge could be transmitted only within the process of personal contact. Thomas Kuhn investigates the phenomenon why good students most usually are not becoming successful researchers [Kuhn 1962]. He suggested that this is due to their maturing basically through reading textbooks and the lack of personal contact and real practice.

Of course, we are not denying the role, place and the meaning of the e-learning. It is a new kind of enriching the forms of education nowadays. Being accessible, financially profitable and safe in the cases of simulating of critical situations, it is unique and can give us a lot, and shorten the periods of training. In the future, it could become obligatory way of knowledge acquisition. But we should understand that it could be necessary, but not sufficient form of training and it is better to use it as a step to the last phase of training: the real practice in our "material" world.

Digital Entertainment

In the USA the number of sales of games already overcomes book sales. In the UK the games are exceeding by 80% video rentals [Ross et al, 2003]. This illustrates the great power of games industry and opens the question how it is actually utilised by the society.

The possibilities for building edutainment (educational entertainment) products involving computer games in the field of cultural heritage are studied in [Sotirova, 2004]. This paper shows clearly that developing games with cultural and historical content is a difficult process and needs special measures from governments or other organizations which would support it – otherwise the companies would normally go for the development of more cost-effective products, which on the long-run do not contribute to the balanced development of the young generation.

In the digital entertainment we find the combination of spectacular scenery and personal involvement, 'real' participation in the virtual life. The attractive power of games is in immersing in worlds, which exist only in the imagination. While in the daydreams one can model the reaction of the opposite side and the outcome of the

[1] http://www.elearningeuropa.info, last visited on 22 April 2004.

events, here one meets other people or artificial intelligence, which are beyond his power. Here the opposite side is independent, acts on its own will, has its own logic and the outcome is not predictable. This builds the adventure spirit and makes the emotional experience especially strong.

Digital art and its genres

The search for the term "Digital art" returned 1,580,000 results on 22 April 2004. There are websites, which provides tens of thousands of pages of galleries and advice for the digital artist. A typical one is 'The Digital Art Community"[1] which offers tutorials for the people willing to become digital artists, e.g. 3D modelling/Animation, anatomy and Figure Drawing, Concept and Design, Digital Illustration and Painting, Photoshop Effects and Tips. On the one hand, we witness enormous amount of resources on this topics and of works called by their authors 'digital art'. On the other hand, the advice, which is given, most often is targeted to acquiring technical skills. Is this what makes the artist? Where one could learn creativity or improvisation?

Now in the digital environment we have plethora of new art directors, new scriptwriters, new artists – this is the new art. However, the aesthetic, moral and spiritual laws which are valid for the traditional art are valid here too; the same codex and the same principles because the essence of the art is the same, only the tools are different. But we should take into account that the new art compared to the traditional one is different in the sense of power of effect on the human consciousness. The power of art, the large media scale, the magic of the drugs, all of these are collected here, but strengthened several times. Its power of effect is tenfold bigger than that of the cinema and visual arts; hundredfold than that of the theater and thousandfold more than that of the literature, because it collects all of them together.

Digital creativity and improvisation

A substantial part of the software tools – these for graphical design, manipulation of 3D objects, space and animation, as well as the specialised applications for creative modelling, e.g. in fashion, agent and avatar technologies, stimulate the development of the creative nature of a growing number of people. Everyone could find something, which suits his personal taste; his individuality and interests.

For example, People Putty is a program from Haptek[2] that allows one to make his own interactive 3D characters changing skins, shapes, emotions and accessories. The game developers are also creators. Every digital maniac could create a game and settle it with creatures without being trained to draw, to write or to compose music. People can make different things: their own greeting cards, electronic journals, web pages and thus have numerous opportunities to create.

The most incredible is that everyone can try whatever he likes – to draw, to design, to fight, to test crazy ideas, to make impossible dreams become true. The most substantial advantage is that everyone can **improvise**, i.e. to entertain himself, to reveal **creativity**. Everyone could be an artist, architect, musician, designer, builder, warrior, politician, king, and even God! This is so easy, safe and without any responsibility! And that is the most dangerous side of it. False creative self-confidence is being formed and the need for preceding professional training disappears. This leads to invasion of unprofessionalism in the art. Creation of clichés suffocates originality and creative thinking. Every work despite its artistic value becomes public and natural censorship regarding artwork and artists is lacking. All this leads to decay of the esthetical criteria and expectations of the user.

The improvisation in the digital environment is basically taking the form of a game and creative entertainment, while the professional creative improvisation has a preliminary defined topic, clear idea and a structural skeleton around which the improvisation is built. The topic and the idea are leading in the choice of means of expression and the connections within the skeleton. The creative improvisation follows one general idea and a message. The difference in the digital improvisation is in the level of consciousness and preliminary planning. Digital improvisation can cover the range of irresponsible entertainment to the conscious creative act, but this depends on the user. Thus, the users can range from users consumers to users creators.

---

[1] http://www.gfxartist.com, last visited on 22 April 2004.

[2] http://www.haptek.com/, last visited on 16 May 2004.

The Digital Culture Ingredients

After presenting the fields, which are most strongly influenced by the modern digital culture, we would like to discuss what are its current ingredients and bearers. We start with the traditional grouping to *hardware* and *software* and add two core groups of people connected with the digital culture: *developers* and *users.*

Hardware

The computer hardware is the presupposition for the functioning of a specific application. It could facilitate or make more difficult the execution of a specific task. The current struggle of companies in this field is for building devices, which are *easy to use, do not require special education and effort*, and are *portable* (see the section on Human-computer Interaction in [Ross et al., 2003]). Additional trend in the recent years is that the specialisation of devices blurs and we use more and more *multifunctional devices*, e.g. mobile phones, which include game stations, digital cameras, audio recorders and players.

One trend of special importance in the current hardware developments is the increased role of *virtual environments (VE)* opposed to *augmented reality.* One important concept related to VEs is the human *immersion* (see, e.g., the section on Virtual reality in [Ross et al., 2003])

Virtual Environments

The basic role of VE technology is to provide tools for the development and use of computer-generated artificial environments. It is believed that thus users will interact with the environment in natural and easy ways. In this respect the interface issues play crucial role. Probably the basic reason for interest of professionals in VE systems is that they contribute to work on new application areas, which were too expensive or dangerous before (like underwater and space work, hazardous environments, simulation of natural calamities). In addition it is considered that well-designed VEs would provide more intuitive metaphors for human-computer interaction. VEs quickly become one of the basic ingredients of computer games.

VEs can be divided into three groups differing on the sense of immersion they provide: *non-immersive, semi immersive* and *fully immersive* [Ross et al, 2004]. Immersion is measured by the power of attention, which the user focuses on the activity and his/her sense of presence in the environment. The sense of immersion is directly connected with the number of senses involved in the user work. This goes beyond the visual and audio communication channels, which is traditionally used in human-computer interaction and involves devices, which use the sense of touch, sense of smell and movement.

For the visual channel, despite the display characteristics, other factors contributing to the sense of immersion include image quality (including number of dimensions) and the speed and level of interactivity of the system.

Non-immersive systems are the weakest ones. The VEs are simulated on a desktop computer with a standard high-resolution display. Keyboard, mouse, joysticks are employed for interaction with the system. Data gloves often are used in such systems to involve the sense of touch in the user work.

Semi-immersive systems provide better sense of immersion. This is achieved through the use of large screens of multiple projection systems such as CAVEs (Cave Automatic Virtual Environment), and shutter glasses.

Fully immersive systems currently are based on the use of head-mounted display (HMD) technology. They give highest sense of immersion, although the image quality compared to previous technologies is worse.

Health issues are often raised in connection with immersive systems. Dizziness and lost of sense of orientation can be caused by the use of VE devices. Most commonly these problems are connected with the use of HMDs which are reported to physical, physiological and psychological problems. HMDs may also lead to posture problems because of the additional load on the body.

Augmented reality is a concept taken in contrast to virtual reality. Unlike virtual environments and their imaginary experiences, augmented reality uses the modern computer technologies to supplement the perceptions of the user of his actual surrounding. The basic factor for building augmented reality is not in different devices, but in the content of information provided to the user and its relation to the real world surroundings.

Care for Impaired Users

Another important trend in current technological developments is the greatest care for users with visual and other impairments. Thus people who were not able to use computers before join the computer society.

In the last years, we witness technological changes in the hardware every six months, and the appearance of novel devices. The history of computing never witnessed such plethora of devices with different technical capabilities and underlying concepts. The challenge to the users in this respect is to be open to new developments; form clear views on the necessity to master a new device and quickly learn to use it.

Software

The next level of interest in the computer culture is related to the employment of new hardware possibilities in the development of software.

In software development nowadays, several trends are influencing the development of systems, which are more attractive to the users:

- Personalisation of users in the virtual environments through use of avatars.
- Tele-presence applications, which contribute to form the sense of presence at another location.
- Building virtual communities.

Personalisation of Users

The virtual presentations of human users in the computer environment are called avatars. Human presentations of computer processes or programs are called agents. Both avatars and agents assist in forming the sense of social dimension in the human-computer interaction. Their use does not involve heavy bandwidth necessary for use of other technologies presenting the real user in the virtual environment, such as video conferencing and this explains their growing popularity.

Telepresence

Telepresence is defined as "the use of technology to establish a sense of shared *presence* or shared *space* among separated members of a group" [Buxton 1992]. Telepresence is a step ahead of video conferencing. It is of special importance due to the fact that it enriches communication with non-verbal aspects: gestures, eye contact, and spatial perception. Telepresence blurs the sense of reality of the user and moves him/her to another place.

Virtual Communities

Modern ICT offer a broad range of tools supporting communication and help people to be in contact with other people from all over the world. In the beginning, the communication tools were text-based and asynchronous, such as the e-mail and mailing lists. Later, instant messaging, multi user dungeons and video conferencing added other means and synchronised the communication. Peer-to-peer technology allows people in a community to exchange resources [Ross et al., 2004]. Virtual community is defined as a network of individuals/organisations using digital technology to create and use shared experience and knowledge.

Knowledge and *emotions* are the two typical centres around which communities are being built. In the first case, people collect and exchange opinions, learning materials, information how to do something or where to find particular content, etc. In the second case, communities gather to fulfil emotional or communication needs.

The Humans' Basic Roles: Developers and Users

People immerse in the digital environment in two basic roles: developers and users. The users can be themselves creators, or passive consumers.

Fig. 1 presents the relationships between the different components of the digital culture.

In the digital eye centre (see 1) is the field of new digital technologies (I). Around it, from the centre towards the periphery, are differentiated its adjacent inheriting areas (II, III, IV). However, the links between them are neither linear nor hierarchical and these areas are overlapping and intersecting each other (see 2). Thus the image of the Digital Shell (see 3) is being formed. Surrounded by the amorphous area of the consumers (V), this metaphor gives an idea about the digital culture which conquers our society slowly, gradually, but undoubtedly.

The process is bi-directional (see (4)). The digital shell (A) develops and perfects its products around the needs, interests and expectations of its users (B). They play the role of the small grits which boost the development of mother-of-pearl – in this case of the research quests and the development of innovative technologies, while finally

the pearl itself – i.e. the specific product, appears (see C). On this stage the new product has been delivered back to the user and influences his future interests. Thus, the developers of the digital products represented by the digital shell strongly influence the interests of the mass user. This two-fold process enhances the development of each tendency and also of the mistakes, if they had been done.

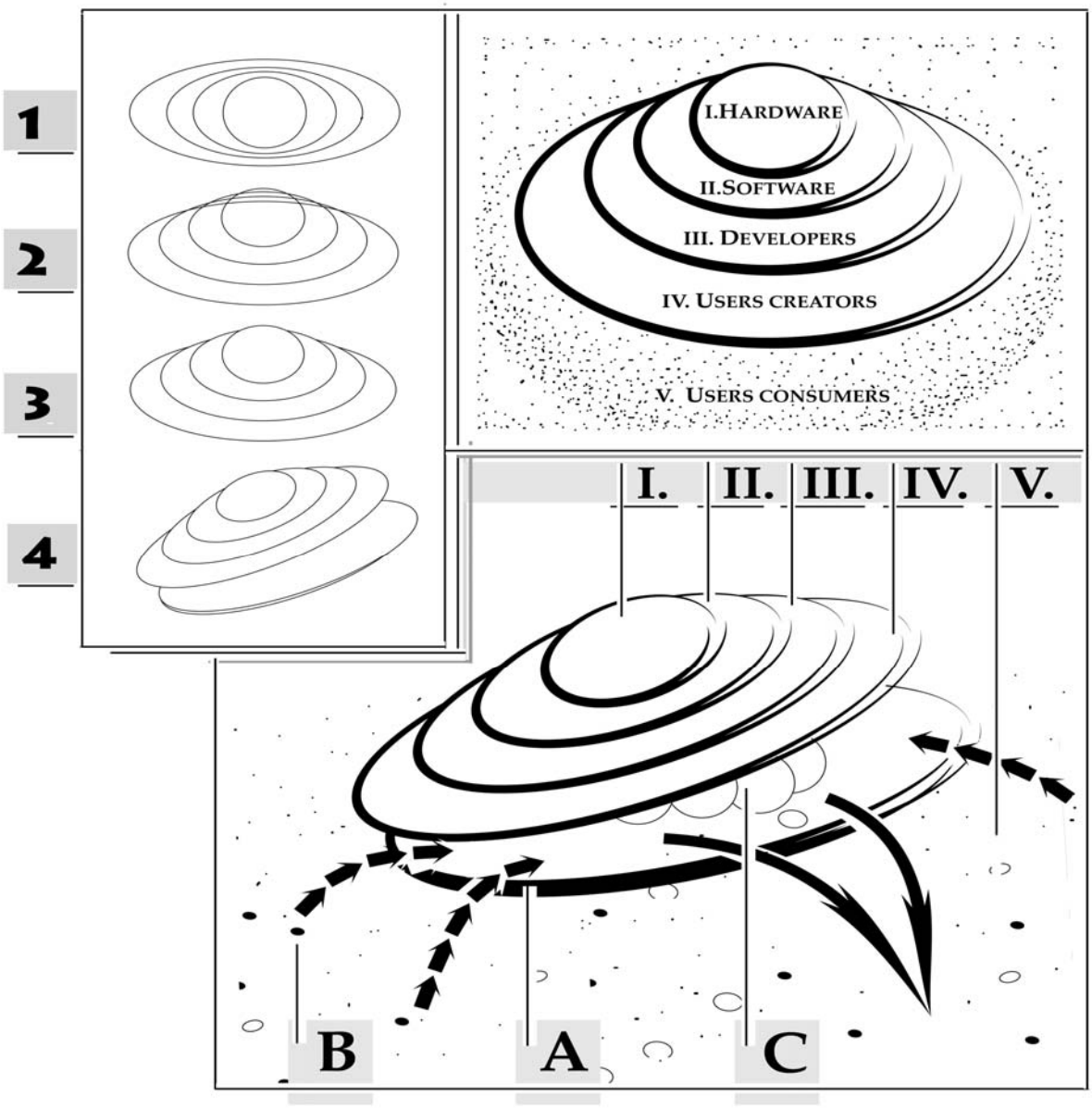


Fig. 1. The Digital Culture Ingredients

The digital shell should educate the users taking into account their interests, without allowing they to become leading ones. However, this is not easy to achieve in the market economy where the basic concern is the profit, which could lead to domination of the bad taste and low criteria due to the sporadic nature of the mass consumer.

Conclusion

Finally, we would like to summarise (see Table 1) the strengths-weaknesses opportunities and threats of the current virtual environments.

STRENGTHS	WEAKNESSES	OPPORTUNITIES	THREATS
Globality and dimension of the net: the new technologies can unify the world not only in the sense of information, but also in the sense of social issues. The problems being regional or private become global and shared.	The lack of order leads to chaos; many issues are not noticed which causes the 'voice in a desert' effect.	Dissemination of knowledge. Faster problem solving. Unifying the efforts of people from all over the world in problem solving. Building synergies in decisions making.	The effect of negative messages is multiplied.
Immense power of effect of the digital culture on the human consciousness.	Lack of order and control. Anonymity and lack of responsibility in the cases of misuse.	These technologies could contribute in the cases of global problems (ecological, terrorism, etc.) which solution requires behavioral change in humans. Raising of moral and ethical norms common to the mankind.	Manipulation of the consciousness, malicious use and propagation of false moral values. Deliberate or sporadic destroying of the moral ideals.
VE can be used to model and simulate almost everything one could imagine.	Immersion is still not 100% possible.	New possibilities are opened for training, education and entertainment.	The real life is shifted; asocial behavior can grow.
New educational opportunities: low price, mass coverage.	Knowledge obtained is incomplete (the practical experience and 'hidden' knowledge can be achieved only through personal face-to-face contact).	One can gain precious experience for situation, which are too dangerous to be learned in the real world (e.g., modelling of hazardous environments, calamities, space etc). Involvement of outstanding personalities in the learning process is easier in the VE than in the real world.	There is a danger to overestimate this type of education and to throw off the real practice, which could lead to abundance of low-level graduates.
Considerably greater possibilities for creative activities (professional, hobbies, and entertainment). Everyone could try himself and act as an artist, architect, musician, designer, animator, director, etc. without being formally trained to do this.	Diminishes the formal need for preceding professional training. Lack of natural censorship regarding works of art and artists, compared to classical arts. Every work despite its artistic value is made public.	Stimulation and development of the creative nature with appropriate creative games.	Invasion of unprofessionalism in the art. Building false creative self-confidence. Decay of the esthetical criteria and expectation of the user. Creation of clichés suffocates originality and creative thinking.

New possibilities for entertainment. Things impossible before become possible: immersion in activities and situations, wars, research missions, magic worlds, entry into impossible roles (warriors, magicians, emperors, Gods). Safety. Lack of fear of the consequences. Freedom and completeness of the experience.	Lack of connection between action and consequence in all personality aspects: physical, physiological, moral, social. Lack of risk, lack of fear of fatal errors.	The hypnotic power of the digital entertainment could be connected to the education in a way, which leads to improving its outcomes.	Weakening of the sensitivity and the instinct of self-preservation (making harm to self and others). Feeding the low human passions: aggression, power-loving, and cupidity. Lack of responsibility for one's actions in life. Lack of understanding of the importance and irreversibility of life events and developing of 'undo' behaviour.

Table 1. SWOT analysis: Digital Culture

This paper was written with the idea to attract the attention of IT professionals on the deep meaning of their work on the personal development and, respectively, on the society as a whole.

If we take as a starting point Nietzsche's thought "A person's maturity consists in having found again the seriousness one had as a child, at play."[1], we can say that current technologies are on the way to assist one to find the way back to his childhood by a number of external factors, such as hardware devices and interfaces. This cannot be claimed for seriousness, feeling, passion – and how often the developers or users themselves realise this?

It is unlikely that the ideological predecessors of the World Wide Web imagined what changes in human lives its appearance would actually cause. We should realize the responsibility. The mankind knows cases when its own creations turn against it.

We should not forget that digital technologies are not an end in themselves, but a tool for support and development in our lives. But they are also an element with an immense power, like the fire and nuclear energy. It depends only on us whether this element would serve us or will come out of control and how we will use it – for creation or for destruction.

Let we will be led in our digital creative acts by the humanity and the care for people; by the unique distinctive creative nature and potential of the human personality and the preservation of the human values in the society as a whole and not by the technology itself.

Bibliography

[Buxton, 1992] W.Buxton, *Telepresence: integrating shared task and person spaces*. Proceedings of Graphics Interface '92, 123-129., 1992. http://www.billbuxton.com/shared_space.html

[Kuhn 1962] T. Kuhn, *The Structure of Scientific Revolutions*, University of Chicago Press, 1962.

[Negroponte, 1996] N. Negroponte, *Being digital,* First Vintage Book Edition, 1996.

[Polanyi 1962] M. Polanyi, *Personal Knowledge: Towards a Post-critical Philosophy.* Chicago: University of Chicago Press. 1962.

[Ross et al., 2003] S. Ross, M. Donnely, M. Dobreva, *New Technologies for the Cultural and Scientific Heritage Sector (Technology Watch Report 1),* ISBN 92-894-5275-7, European Commission, 2003, 196 pp.

[Ross et al., 2004] S. Ross, M. Donnely, M. Dobreva, *Emerging Technologies for the Cultural and Scientific Heritage Sector (Technology Watch Report 2),* ISBN 92-894-5276-5, European Commission, 2004, 214 pp.

[Sotirova, 2004] K. Sotirova. *Edutainment games – Homo Culturalis VS Homo Ludens*, In: Review of the National Digitisation Center, Belgrade, Serbia and Montenegro, Vol. 4, 2004, pp. 84–98.

Author information

Velimir Velev – Assistant Professor, National Academy for Theatre and Film Arts, Studentski grad bl. 58-V-204, Sofia-1734, Bulgaria; e-mail: velimir_velev@abv.bg

[1] Friedrich Wilhelm Nietzsche, Beyond Good and Evil (1885-86)

PRACTICAL, COMPUTATION EFFICIENT HIGH-ORDER NEURAL NETWORK FOR ROTATION AND SHIFT INVARIANT PATTERN RECOGNITION

Evgeny Artyomov and Orly Yadid-Pecht

Abstract: In this paper, a modification for the high-order neural network (HONN) is presented. Third order networks are considered for achieving translation, rotation and scale invariant pattern recognition. They require however much storage and computation power for the task. The proposed modified HONN takes into account a priori knowledge of the binary patterns that have to be learned, achieving significant gain in computation time and memory requirements. This modification enables the efficient computation of HONNs for image fields of greater that 100 × 100 pixels without any loss of pattern information.

Keywords: HONN, higher-order networks, invariant pattern recognition.

1. Introduction

Invariant pattern recognition using neural networks was found to be attractive due to its similarity to biological systems. There are three different classes that use neural networks for invariant pattern recognition [1], that differ in the way invariance is achieved, i.e. Invariance by Training [2], Invariant Feature Spaces, or invariance by Structure, good examples are: the Neocognitron and HONN [3].

In third-order networks, which are a special case of the HONN, invariance is built into the network structure, which enables fast network learning with only one view of each pattern presented at the learning stage. However, an exponentially growing amount of interconnections in the network does not enable its usage for image fields larger than 18 x 18 pixels [3]. A few different solutions were proposed to minimize the number of the HONN interconnections. Weight sharing, by similar triangles [3]. Weight sharing by “approximately similar triangles” [4]-[5]. Coarse coding [6]. Non-fully interconnected HONN [7]. All these methods partially solve the problem of the HONN interconnections but do not help with larger images. Consequently, the research community in the field of invariant pattern recognition largely abandoned the HONN method.

In this paper, a modification for the third-order network is described. The proposed modification takes into account a priori knowledge of the binary patterns that must be learned. By eliminating idle loops, the network achieves significant reductions in computation time as well as in memory requirements for network configuration and weight storage. Better recognition rates (compared to conventionally constructed networks with the same input image field) are attained by the introduction of a new “approximately equal triangles” scheme for weight sharing. The modified configuration enables efficient computation of image fields larger than 100 × 100 pixels without any loss of image information — an impossible task with any previously proposed algorithm.

2. HONN Architecture

Following equation describes the output of a third-order network:

$$y_i = f\left(\sum_j \sum_k \sum_l w_{ijkl} x_j x_k x_l\right), \quad (1)$$

where *w* is the weight associated with a particular triangle, *y* is the output and *x* is a binary input, *j*, *k*, and *l* are the indices of the inputs.

A schematic description of this network is shown in Fig. 1.

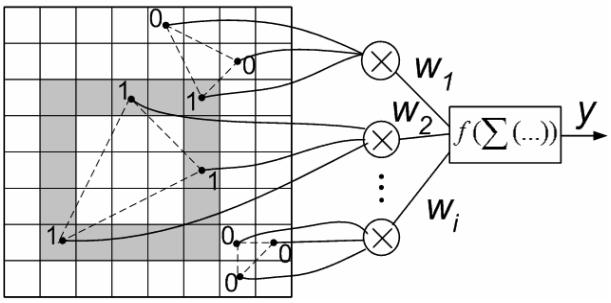


Fig.1. Schematic description of a third-order network

In the training phase, the perceptron-like rule is used:

$$\Delta w_{ijkl} = \eta(t_i - y_i)x_j x_k x_l \ , \tag{2}$$

where *t* is the expected training output, *y* is the actual output, η is the learning rate and *x* is a binary input.

The number of triangles (*NoT*) can be calculated with the following equation:

$$NoT = \frac{IN!}{(IN-3)!3!}, \tag{3}$$

where *IN* is the number of input nodes.

For image fields 100 x 100 and 256 x 256 the number of triangles will be 1.6662×10^{11} and 4.6910×10^{13} accordingly. As can be seen, the number of triangles grows very fast off the limits of any current hardware. A few techniques to reduce the number of weights have been proposed in the literature (as described in section 1), but they do not reduce computation time.

The problem of large computational demands arises since the network is constructed in the pre-processing stage before the learning phase. At this stage, all possible triangles are computed and pointers to the weights are saved [8]. In addition to the pointers, the weight array is also stored. At least two memory bytes are required for each pointer. If, for example, an input field of 100 x 100 pixels is given, the total number of bytes required to store the entire vector of pointers is 3.3324×10^{11} bytes. The memory and computation requirements are enormous. To work with large input patterns, significant network modifications are required.

3. The Proposed Modified HONN Method

As noted before, the input pattern is binary: edge or contour pixel has the value “1” and all other pixels have the value “0”. As can be seen from equation (1), each product with pixel value “0” will give “0” as a result. This means that only active triangles (in which all pixels belong to an object contour) will influence the result. In addition, the weights that belong to the inactive triangles will not be updated and will keep “zero” value during the learning process.

Following this observation, the network can be modified and all inactive triangles can be disregarded during the construction phase, which eliminates the idle loops from the computation. With this modification, the network configuration strictly depends on the input patterns that have to be learned.

In addition, to improve network performance regarding rotation, distortion and a number of learned classes we introduce an “approximately equal triangles” scheme for network construction. This scheme, in addition to the “approximately similar triangles” scheme presented in [4] for weights sharing, adds triangle area equality. This means that “approximately similar triangles” with “approximately equal” areas will share the same weight.

3.1 The Proposed Network Construction

The modified algorithm for network construction can be described as follows: 1. Load all patterns that must be learned. 2. Run through each image and save the coordinates of the contour (boundary) pattern pixels to the

different arrays. A set of such arrays is shown in Fig. 2. 3. Compute angles of all presented triangles and classify them in order to associate with a particular weight.

Indices X_{im} ,Y_{im} and n_{ij} correspond to pattern number (n), pixel number (m), weight index (j) and puttern number (i). The variable n_{ij} is the number of triangles from the particular pattern that correspond to the particular triangle weight index (class).

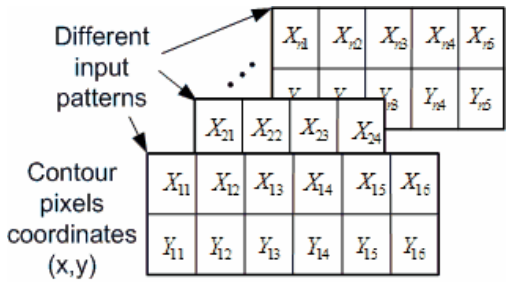


Fig.2. Arrays of the pixel coordinates of the object contours

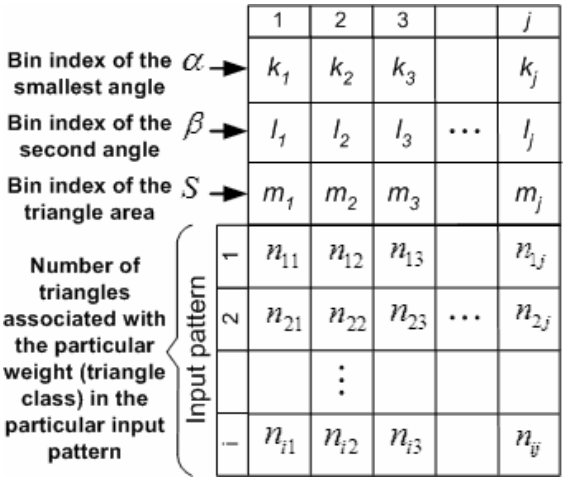


Fig.3. General array for classifying triangles by weight index.

The presented method of classification is based on “approximately equal triangles”. For the association of a triangle with a particular weight, the sets of possible values of the two smallest triangle angles (α, β) and the triangle area (S) are partitioned into bins defined by:

$$(k-1)\,w \le \alpha < kw,\ \ (l-1)\,w \le \beta < lw,\ \ (m-1)\,s \le S < ms, \tag{4}$$

where w is an angular tolerance, α and β are the smallest angles of the triangle so that $\alpha < \beta$, s is an triangle area tolerance, S is an area of the computed triangle, k, l and m are the bin indices associated with two angles and triangle area, respectively.

During the classification step, each triangle class is associated with a corresponding weight and is represented by three variables *k, l,* and *m.* The array of triangle classes is constructed as shown in Fig.3. After construction, only the array of triangle classes presented in Fig.3 must be stored in memory.

3.2 Network Training

The previously constructed array of triangle classes (Fig. 3) is used as the basis for learning in the training phase. In addition, a zero matrix of weights (**W**) with the size of *NoP x NoW* is constructed. Where *NoW* is the number of individual weights, *NoP* is the number of training patterns.

Output computation takes into account only information presented in the weights array (**W**) and in the triangle array (*N*) (Fig. 3). It follows the next equation for a particular input image:

$$y_i = f(\sum_j w_{ij} n_{kj}), \tag{5}$$

where i is the output index, j is the weight index, k is the pattern index, w is the weight, and n is the number of triangles that correspond to the particular triangle class (i.e., the particular weight).

All weights are updated only once after each iteration, according to the next equation:

$$\Delta w_{il} = \eta(t_i - y_i)\ \text{, if } N_{kl}>0\ ;\ \Delta w_{il} = 0\ \text{, if } N_{kl}=0, \tag{6}$$

where t is the expected training output, y is the actual output, η is the learning rate.

After the training phase is complete, only the array of learned weights and the corresponding coefficients *k, l,* and *m* that represent the equivalence class (from the upper three rows of the array in Fig. 3) must be saved.

3.3 Recognition

The algorithm for recognition can be described as follows: 1. Load pattern intended for recognition. 2. Construct an array of coordinates of contour pixels (as in the construction stage). 3. Construct a zero matrix (**N**) with the size equal to *1 x NoW.* This is a counter for triangles in the image, which correspond to the particular weight. 4. Run through the coordinate array and compute coefficients *k*, *l* and *m* for all possible triangles as was described in 3.1. After each computation, compare the newly found *k*, *l* and *m* with the ones previously saved (upper part of the array from Fig.3). If a matching class for the triangle is found, the counter corresponding to that triangle class position is increased by one ($n_{1j} = n_{1j} + 1$) Thus, during classification the nonzero one-dimensional matrix of counters (**N**) is built. 4. Compute outputs according to equation (5), using the weights array (**W**) built during construction phase and the triangle counters (**N**) built in the beginning of recognition phase.

4. Experimental Results

To study the performance of the modified network and compare computational resources with the conventional network, seven different object classes with 60 x 60 and 170 x 170 pixels were prepared. One object from each class was used in the training phase and 14 rotated patterns of each class were used in the recognition phase. Pattern examples are shown in Fig.4.

Fig.4. Pattern examples

The comparison for computational resource demands for 60 x 60 and 170 x 170 input fields are presented in Table 1.

As can be seen from the table, the gain achieved with the modified network in computational steps amount is four orders of magnitude for an input field 60 x 60 and five orders of magnitude for an input field of 170 x 170. This gain will be more significant with image size increase. In addition, the memory resources are minimized also.

Table 1: Comparison of the computational resources demands
("approximately similar triangles" scheme is used alone, the network was trained for first five pattern classes, w = π/180, m - not used).

Input field size	60 x 60		170 x 170	
Network type	Conventional	Modified	Conventional	Modified
Computational steps (number)	7.8×10^{9}	13.8×10^{5}	4.02×10^{12}	4.5×10^{7}
Total memory requirements (bytes)	15.5×10^{9}	81340	8.04×10^{12}	81340

Table 2: Recognition rate for a varying number of trained classes, angular and area similarities. Input pattern: 60 x 60 pixels.

Tolerance		Number of trained classes						Weight number
Angular (*w*)	Area (*S*)	2	3	4	5	6	7	
π/60	10	100	95	94	84	80	80	17286
π/60	20	100	95	95	85	80	80	8935
π/20	10	100	95	91	87	82	80	2502
π/20	20	100	95	88	80	-	-	1217

Table 3: Recognition rates of the net with the "approximately similar triangles" scheme alone.
Input pattern: 60 x 60 pixels.

Angular (w) tolerance	Number of trained classes				Weight number
	2	3	4	5	
π/225	100	80	75	60	8533

For comparison with the "approximately similar triangles" scheme, a few results are provided in Table 3. Results for the best configuration are shown only, but even this shows much worse recognition rates. The cause for this is that similar triangles with very large difference in size are associated with the same triangle class, as a result, some object classes will be associated with the same triangle class, preventing from the objects to have an individual triangle set associated with it.

From the experimental data provided, it can be seen that our method enables the possibility of large input field computation without significant resource demands. Translation invariance is built into the network, thus 100% translation invariance is achieved. All experimental data are provided for this particular data set. For other data sets, where object classes differ significantly in size and in form, much better recognition results can be achieved.

5. Conclusions

A modified High-Order Neural Network for efficient invariant object recognition has been presented. The proposed modification achieves significant gain in computation time and memory requirements. The gain in computation time is achieved by eliminating the idle loops, by taking a priori knowledge of training patterns. With the proposed modified HONN, large input patterns can be processed without large computation demands. Performance of the network is improved also significantly, by using the "approximately equal triangles" scheme.

References

[1] Barnard E., D. Casasent. 1991. Invariance and neural nets. IEEE Transactions on Neural Networks, vol. 2, no. 5, pp. 498-507.

[2] Wood J. 1996. Invariant pattern recognition: a review. Pattern Recognition, vol. 29, no.1, pp.1-17.

[3] Spirkovska L., M.B. Reid. 1992. Robust position, scale, and rotation invariant object recognition using higher-order neural networks. Pattern Recognition, Vol.25, No. 9, pp. 975-985.

[4] Perantonis S.J., P.J.G. Lisboa. 1992. Translation, Rotation, and Scale Invariant Pattern Recognition by Higher-Order Neural Networks and Moment Classifiers. IEEE Transactions on Neural Networks, Vol.3, No. 2, pp. 241-251.

[5] He Z., M.Y. Siyal. 1999. Improvement on Higher-Order Neural Networks for Invariant Object Recognition. Neural Processing Letters, Vol. 10, pp 49-55.

[6] Spirkovska L., M.B. Reid. 1993. Coarse-Coded Higher–Order Neural Networks for PSRI Object Recognition. IEEE Transactions on Neural Networks, Vol. 4, No. 2, pp. 276-283.

[7] Spirkovska L., M.B. Reid. 1990. Connectivity Strategies for Higher-order Neural Networks applied to Pattern Recognition. Proceedings of IJCNN, Vol. 1, San Diego, pp. 21 - 26.

[8] He Z. 1999. Invariant Pattern Recognition with Higher-Order Neural Networks. Master Thesis, School of Electrical and Computer Engineering, Nanyang Technological University, Singapore.

Authors' Information

Evgeny Artyomov – The VLSI Systems Center, Ben-Gurion University, Beer Sheva, Israel. e-mail: artemov@bgumail.bgu.ac.il

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

AN INTERACTIVE METHOD OF LINEAR MIXED INTEGER MULTICRITERIA OPTIMIZATION

Vassil Vassilev, Krasimira Genova, Mariyana Vassileva, and Subhash Narula

Abstract: The paper describes a learning-oriented interactive method for solving linear mixed integer problems of multicriteria optimization. The method increases the possibilities of the decision maker (DM) to describe his/her local preferences and at the same time it overcomes some computational difficulties, especially in problems of large dimension. The method is realized in an experimental decision support system for finding the solution of linear mixed integer multicriteria optimization problems.

Keywords: linear mixed integer multicriteria optimization, interactive methods, decision support systems.

1. Introduction

The reference point interactive methods and the classification-based interactive methods are the most widely spread methods [Gardineer and Vanderpooten, 1997], [Miettinen, 1999], [Vassilev et al., 2003] for solving linear problems of multicriteria optimization. This kind of interactive methods is also used [Climaco et al., 1997], [Vassileva, 2001] for solving linear mixed integer problems of multicriteria optimization.

Single-criterion problems of linear programming are used in the interactive methods solving linear problems of multicriteria optimization. They belong to the class of P-problems [Garey and Johnson, 1979] and therefore the time for solving a single-criterion problem does not play a significant role in the interactive methods of linear multicriteria optimization. In the development of the interactive methods main attention is paid to the possibilities offered to the DM to describe his/her preferences. Single-criterion problems of mixed integer programming are applied in the interactive methods solving linear mixed integer problems of multicriteria optimization. They are NP-hard problems and hence the time for single-criterion mixed integer problems solution has to be obligatory taken into account when designing interactive methods.

A learning-oriented interactive method is proposed in the present paper on the basis of new linear mixed integer classification-based scalarizing problems, intended to solve linear mixed integer problems of multicriteria optimization. The method increases DM's possibilities to describe his/her local preferences and also overcomes some computational difficulties, connected with single-criterion mixed integer problems solving.

2. Problem Formulation

The linear mixed integer problem of multicriteria optimization (denoted by **I**), can be formulated as follows:

(1) $$"\max"\{f_k(x)\}, \qquad k \in K$$

subject to:

(2) $$\sum_{j \in N} a_{ij} x_j \le b_i, \qquad i \in M$$

(3) $$0 \le x_j \le d_j, \qquad j \in N$$

(4) $$x_j - \text{integers}, \; j \in N'; \; N' \subseteq N,$$

where $f_k(x), k \in K$ are linear criteria (objective functions), $f_k(x) = \sum_{j=N} c_j^k x_j$, "max" denotes that all the objective functions should be simultaneously maximized; $x = (x_1, x_2 ..., x_j, ...x_n)^T$ is the variables vector; $K = \{1,2,..., p\}$, $M = \{1,2,..., m\}$, $N = \{1,2,..., n\}$ and $N' \subseteq N$ are sets of the indices of the linear criteria

(objective functions) of the linear constraints, of the variables and of the integer variables respectively. Constraints (2)-(4) define the feasible region X_1 for the variables of the mixed integer problem.

Several definitions will be introduced for greater precision.

Definition 1: The solution x is called an efficient solution of problem (**I**), if there does not exist any other solution $\overline{x}$, so that the following inequalities are satisfied:

$f_k(\overline{x}) \geq f_k(x),$ for every $k \in K$ and

$f_k(\overline{x}) > f_k(x),$ for at least one index $k \in K$.

Definition 2: The solution x is called a weak efficient solution of problem (**I**), if there does not exist another solution $\overline{x}$ such that the following inequalities hold:

$f_k(\overline{x}) > f_k(x),$ for every $k \in K$.

Definition 3: The solution x is called a (weak) efficient solution, if x is either an efficient or a weak efficient solution.

Definition 4: The vector $f(x) = (f_1(x),...,f_p(x))^T$ is called a (weak) non-dominated solution in the criteria space, if x is a (weak) efficient solution in the variables space.

Definition 5: A near (weak) non-dominated solution is a feasible solution in the criteria space located comparatively close to the (weak) non-dominated solutions.

Definition 6: A current preferred solution is a (weak) non-dominated solution or near (weak) non-dominated solution, if selected by the DM at the current iteration.

Definition 7: The most preferred solution is the current preferred solution, which satisfies the DM to the highest extent.

3. Scalarizing Problems

The linear mixed integer problems of multicriteria optimization do not possess a mathematically well-defined optimal solution. That is why it is necessary to choose one of the (weak) non-dominated solutions, which is most appropriate with reference to DM's global preferences. This choice is subjective and it depends entirely on the DM.

The interactive methods are the most widely used methods for solving linear mixed integer problems of multicriteria optimization [Climaco et al., 1997], [Vassileva, 2001]. Every iteration of such an interactive method consists of two phases: a computation and a decision one. One or more non-dominated solutions are generated with the help of a scalarizing problem at the computation phase. At the decision phase these non-dominated solutions are presented for evaluation to the DM. In case the DM does not approve any of these solutions as a final solution (the most preferred solution), he/she supplies information concerning his/her local preferences with the purpose to improve these solutions. This information is used to formulate a new scalarizing problem, which is solved at the next iteration.

The type of the scalarizing problem used lies in the basis of each interactive method. The scalarizing problem is a problem of single-criterion optimization and its optimal solution is a (weak) non-dominated solution of the multicriteria optimization problem. The clasification-based scalarizing problems are particularly appropriate in solving linear mixed integer multicriteria optimization problems, because they enable the decrease of the computational difficulties connected with their solution as well as the increase of DM's possibilities in describing his/her local preferences and also lead to reduction of the requirements towards the DM in the comparison and evaluation of the new solutions obtained. In the scalarizing problems suggested in this chapter, the DM can present his/her local preferences in terms of desired or acceptable levels, directions and intervals of alteration in the values of separate criteria. Depending on these preferences, the criteria set can be divided into seven or less criteria classes. The criterion $f_k(x), \quad k \in K$ may belong to one of these classes as follows:

$k \in K^{>}$, if the DM wishes the criterion $f_k(x)$ to be improved;

$k \in K^{\geq}$, if the DM wishes the criterion $f_k(x)$ to be improved by any desired (aspiration) value Δ_k;

$k \in K^{<}$, if the DM agrees the criterion $f_k(x)$ to be worsened;

$k \in K^{\leq}$, in case the DM agrees the value of the criterion $f_k(x)$ to be deteriorated by no more than δ_k;

$k \in K^{><}$, if the DM wishes the value of the criterion $f_k(x)$ to be within definite limits with respect to the current value f_k, $(f_k - t_k^- < f_k(x) \leq f_k + t_k^+)$;

$k \in K^{=}$, if the current value of the criterion $f_k(x)$ is acceptable for the DM;

$k \in K^{0}$, if at the moment the DM is not interested in the alteration of the criterion $f_k(x)$ and this criterion can be freely altered.

In order to obtain a solution, better than the current (weak) non-dominated solution of the linear mixed integer problem of multicriteria optimization, the following Chebyshev's scalarizing problems can be applied on the basis of the implicit criteria classification, done by the DM. The first mixed integer scalarizing problem [Vassileva, 2000] called *DAL* (desired and acceptable level) has the following type:

To minimize:

$$S(x) = \max\left[\max_{k \in K^{\geq}}\left(\bar{f}_k - f_k(x)\right)/\left|f_k^{'}\right|, \max_{k \in K^{\leq}}\left(f_k - f_k(x)\right)/\left|f_k^{'}\right|\right] \quad (5)$$

under the constraints:

$$f_k(x) \geq f_{k,}\ k \in K^{=}, \quad (6)$$

$$f_k(x) \geq f_k - \delta_k, k \in K^{\leq}, \quad (7)$$

$$x \in X_1, \quad (8)$$

where:

f_k, $k \in K$ is the value of the criterion $f_k(x)$ in the current preferred solution;

$\bar{f}_k = f_k + \Delta_k$, $k \in K^{\geq}$ is the desired level of the criterion $f_k(x)$;

$f_{\kappa}^{'}$, $k \in K$ is a scaling coefficient:

$$f_k^{'} = \begin{cases} \varepsilon, \text{if} \left|f_k^{'}\right| \leq \varepsilon \\ f_{k,} \text{ if} \left|f_k^{'}\right| > \varepsilon \end{cases},$$

where ε is a small positive number.

DAL scalarizing problem has three properties, which allow to a great extent the overcoming of the computational difficulties, connected with its solving as a problem of integer programming and also decrease DM's efforts in the comparison of new solutions. The first property is connected with this, that the current preferred integer solution of the multicriteria problem (found at the previous iteration), is a feasible integer solution of *DAL* problem. This facilitates the exact as well as the approximate algorithms for solving *DAL* problem, because they start with a known initial feasible integer solution. The second property is connected with the fact, that the feasible region of *DAL* problem is a part of the feasible region of the multicriteria problem (**I**). Depending on the values of the parameters $\Delta_k / k \in K^{\geq}$, $\delta_k / k \in K^{\leq}$ the feasible region of *DAL* problem can be comparatively narrow and

the feasible solutions in the criteria space, found with the help of approximate integer programming algorithms, may be located very close to the non-dominated surface of the multicriteria problem (**I**). The third property comprises DM's possibility to realize searching strategy of "not big profits – small losses" type. This is due to the fact, that such optimal solution is searched for with the help of *DAL* problem, which minimizes Chebyshev's distance between the feasible criteria set and the current reference point, the components of which are equal to the wished by the DM values of the criteria being improved and to the current values of the criteria being deteriorated. The (weak) non-dominated solution obtained and the current solution are comparatively close and the DM can easily make his/her choice.

The classification-oriented scalarizing problems are appropriate in solving integer problems of multicriteria optimization, because the computational difficulties, connected with their solving are decreased with their help and the requirements towards the DM in the comparison and evaluation of the new solutions obtained, are diminished. From a viewpoint of the information, required by the DM in new solutions seeking in this scalarizing problem, the DM has to define the desired or acceptable levels for a part or for all the criteria. With the help of the scalarizing problem below described, called *DALDI* (desired or acceptable level, direction and interval), the DM is able to present his/her local preferences not only by desired and acceptable levels, but also by desired and acceptable directions and intervals of alteration in the values of separate criteria. The mixed integer scalarizing problem *DALDI* has the following type:

Minimize:

$$(9)\qquad S(x)=\max\left[\max_{k\in K^{\geq}}\left(\bar{f}_k-f_k(x)\right)/\left|f_k'\right|,\ \max_{k\in K^{<}\cup K^{\leq}}\left(f_k-f_k(x)\right)/\left|f_k'\right|\right]+$$

$$+\max_{k\in K^{>}}\left(f_k-f_k(x)\right)/\left|f_k'\right|,$$

under the constraints:

$$(10)\qquad f_k(x)\geq f_{k,}\ k\in K^{>}\cup K^{=},$$

$$(11)\qquad f_k(x)\geq f_k-\delta_k,\ k\in K^{\leq},$$

$$(12)\qquad f_k(x)\geq f_k-t_k^{-},\ k\in K^{><},$$

$$(13)\qquad f_k(x)\leq f_k+t_k^{+},\ k\in K^{><},$$

$$(14)\qquad x\in X_1,$$

The scalarizing problem *DALDI* has characteristics similar to *DAL* scalarizing problem, but still there are two differences between them.

The first difference consists in this, that *DALDI* scalarizing problem gives greater freedom to the DM when expressing his/her local preferences in the search for a better (weak) non-dominated solution. Besides desired or acceptable values of a part or of all the criteria, the DM has the possibility to set also desired or acceptable directions and intervals of change in the values of some criteria The second difference between *DAL* and *DALDI* scalarizing problems concerns the possibility to alter their feasible sets (make them "narrower"), so that the feasible solutions are positioned close to the non-dominated (efficient) solutions of the multicriteria problem The more the criteria are, which the DM wishes to be freely improved or freely deteriorated, the smaller this possibility is. The narrow feasible regions of the scalarizing problems *DAL* and *DALDI* enable the successful application of approximate single-criterion algorithms, which is especially important when these problems are integer. It should be noted that scalarizing problem *DAL* is better than scalarizing problem *DALDI* in this aspect.

DAL and *DALDI* problems are nonlinear mixed integer programming problems [Wolsey, 1998]. Equivalent linear mixed integer programming problems can be constructed, [Vassileva, 2000], [Vassilev et al., 2003] with the help of additional variables and constraints

4. GAMMA-I1 Interactive Method

On the basis of scalarizing problems *DAL* and *DALDI*, a classification-oriented interactive method, called GAMMA-I1 is proposed for solving linear mixed integer programming problems of multicriteria optimization. The problems of mixed integer programming are NP-problems, i.e. the time for their exact solution is an exponential function of their dimension. That is why, in solving integer problems, particularly problems of larger dimension (above 100 variables and constraints), some approximate methods are used [Vassilev and Genova, 1991], [Pirlot, 1996]. Since finding a feasible solution is as difficult as finding an optimal solution, the approximate integer methods in the general case do not guarantee the finding of an optimal integer solution and of an initial feasible integer solution too. If the initial feasible integer solution is known and the feasible region is comparatively "narrow", then with the help of the approximate integer methods some satisfactory and in part of the cases optimal integer solutions could be found. The scalarizing problems *DAL* and *DALDI* have known feasible initial integer solutions.

DALDI scalarizing problem allows enlargement of the information with the help of which the DM can set his/her local preferences. This information expansion leads to the extension of the feasible set of criteria alteration in the criteria space and of the integer variables in the variables space. Hence, the approximate integer solutions of *DAL* problem (obtained with the help of an approximate integer method) are located closer to the non-dominated (efficient) set of the multicriteria problem, than the approximate solutions of *DALDI* problem. Therefore, if solving linear mixed integer problems of (*I*) type of large dimension, when the scalarizing problems have to be solved approximately in order to reduce the waiting time for new solutions evaluated by the DM, it is better to use *DAL* scalarizing problem than scalarizing problem *DALDI*.

Two different strategies are applied in the development of GAMMA-I1 interactive method in the process of searching for new solutions, that are evaluated. The first strategy, called integer strategy, consists in seeking a (weak) non-dominated integer solution at each iteration by exact solution of the corresponding linear mixed integer scalarizing problem. The second strategy, called approximate integer strategy, comprises searching for near (weak) non-dominated integer solutions at some iterations, approximately solving a respective linear mixed integer scalarizing problem. During the learning phase and in problems of large dimension up to the very end, only near (weak) non-dominated solutions can be looked for.

The interactive GAMMA-I1 method is designed to solve linear mixed integer problems of multicriteria optimization. The two strategies above described are realized in the method during the search for new solutions for evaluation in order to overcome the computational difficulties (particularly in solving problems of large dimension). The method is oriented towards learning and the DM has to determine when the most preferred solution is found. The algorithmic scheme of GAMMA-I1 interactive method includes the following basic steps:

Step 1. An initial near (weak) non-dominated solution is found, setting $f_k = 1,\ k \in K$ and $\bar{f}_k = 2,\ k \in K$ and solving *DAL* problem.

Step 2. The current (weak) non-dominated solution or near (weak) non-dominated solutions obtained are presented for evaluation to the DM. If the DM evaluates and chooses a solution that satisfies his/her global preferences, Step 6 is executed, otherwise – Step 3.

Step 3. A question is set to the DM what new integer solution he wishes to see – a (weak) non-dominated or a near (weak) non-dominated solution. Step 5 is executed in the first case, and Step 4 - in the second.

Step 4. The DM is asked to define the desired or feasible levels of the values of a part or of all the criteria. Scalarizing problem of *DAL* type is solved and then go to Step 2.

Step 5. The DM is requested to define the desired or feasible levels, directions and intervals of alteration in the values of a part or of all the criteria. Scalarizing problem of *DALDI* type is solved and then go to Step 2.

Step 6. Stop the process of the linear mixed integer multicriteria problem solving.

In GAMMA-I1 interactive method the DM controls the dialogue, the computing process and the conditions for canceling the process of linear mixed integer multicriteria problem solution..

5. Conclusion

The interactive GAMMA-I1 method is realized in the experimental software system MOLIP, developed at the Institute of Information Technologies of the Bulgarian Academy of Sciences. This system is designed for interactive solution of linear and linear mixed integer multicriteria optimization problems with different number and type of the criteria, with different number and type of the variables and constraints. MOLIP system functions in the environment of Windows 98 and higher versions and may serve for learning purposes, as well as for the solution of different applied problems. Our experimental results confirm that the computational effort and time are reduced considerably using heuristic integer algorithms in the learning phase and when solving large problems.

Bibliography

[Climaco, et al., 1997] J. Climaco, G. Ferreira, and M. E. Captivo. Multicriteria Integer Programming: An Overview of the Different Algorithmic Approaches In: Multicriteria Analysis (J. Climaco, Ed.). Springer, 248-258, 1997.

[Gardiner and Vanderpooten, 1997] L.R.Gardiner and D.Vanderpooten. Interactive Multiple Criteria Procedures: Some Reflections. In: Multicriteria Analysis (J. Climaco, Ed.). Springer, 290-301, 1997.

[Garey and Johnson, 1979] M. Garey and D. Johnson. Computers and Intractability: A Guide to the Theory of NP-completeness. W. H. Freeman, San Francisco, 1979.

[Miettinen, 1999] K. Miettinen. Nonlinear Multiobjective Optimization. Kluwer Academic Publishers, Boston, 1999.

[Pirlot, 1996] M. Pirlot. General Local Search Methods. European Journal of Operational Research, 92, 3, 493-511, 1996.

[Vassilev et al., 2003] V. Vassilev, K. Genova, M. Vassileva and S. Narula. Classification-Based Method of Linear Multicriteria Optimization. International Journal on Information Theories and Aplications, vol.10, 3, 266-270, 2003.

[Vassilev and Genova, 1991] V. Vassilev and K. Genova. An Algorithm of Internal Feasible Directions for Linear Integer Programming. European Journal of Operational Research, 52, 203-214, 1991.

[Vassileva, 2000] M. Vassileva. Scalarizing Problems of Multiobjective Linear Programming Problems, Problems of Engineering Cybernetics and Robotics, 50, 54-64, 2000.

[Vassileva, 2001] M. Vassileva. An Interactive Algorithm of the Multicriteria Linear Integer Programming. Problems of Engineering Cybernetics and Robotics, 52, 17-25, 2001.

[Wolsey, 1998] L. Wolsey. Integer Programming. Wiley – Interscience, 1998.

Authors' Information

Vassil Vassilev – Institute of Information Technologies, BAS, Acad. G. Bonchev St., bl. 29A, Sofia 1113, Bulgaria; e-mail: vvassilev@iinf.bas.bg

Krasimira Genova – Institute of Information Technologies, BAS, Acad. G. Bonchev St., bl. 29A, Sofia 1113, Bulgaria; e-mail: kgenova@iinf.bas.bg

Mariyana Vassileva – Institute of Information Technologies, BAS, Acad. G. Bonchev St., bl. 29A, Sofia 1113, Bulgaria; e-mail: mari@iinf.bas.bg

Subhash Narula – School of Business, Virginia Commonwealth University, Richmond, VA 23284 - 4000, USA; e-mail: snarula@vcu.edu

SITUATION CENTERS IN MODERN STATE

Alexander Kuzemin

Introduction

Development of information technologies and hardware-software means make it possible to integrate various engineering solutions of technical, technological and information nature within the framework of the single system of implementation.

The situation center (SC) represents the complex of hardware-software means for the personal and team work of the managers group. Its main task consists in supporting decision-making on technical and strategic management solutions based on visualization and analytical procession of information.

The SC ensures the support of preparation and decision making in particular:

- elaboration of the versions and recommendations taking into account various conditions and limitations;
- choice or creation of the situation analysis model;
- structurization of the problems and definition of the most informative parameters;
- monitoring of socio-economical and socio-political information.

Being based on the experience of creation and operation of the situation centers in the organs of the state management it is now possible to formulate the main types of solution concerning the contents of hard-ware-software complex of the situation center and operation regimes.

Regimes of holding actions in the SC can be realized in the form of the limited participation of experts or managers and in the form of television conferences (broadened content of participants), they assume joining with the use of telecommunication means of distant representatives from different organizations, enterprises or situation centers.

Thus, information and intelligent resources needed for consideration and elaboration of versions of solutions in a problem situation are consolidated in the situation center. Efficiency of the SC consists in the possibility to include the figurative associative thinking into the active decision-making work.

Designing and operation of the system of the interconnected federal and regional SC create premises for the transition to the fundamentally another technology of operation of the state management of all levels, namely, to the usage of the new “generation” of the decision-making methods directed to the work of experts and persons making decisions in fuzzy multivariant conditions.

Application of Information Technologies and Systems in the Practice of the Situation Centers Functioning

Information technologies and systems make the basis of the SC efficient functioning. In addition to that their practical application is based on the definition of the objects’ domains which in the state management are: the national security, economy, financial activity, social sphere, external and internal policy, branches of industry, fuel – power complex, transport, agriculture, development of regions, objective programs etc.

Means of information collection and storage is realized in the forms of based and data banks, knowledge bases as well as the systems of telecommunication information action.

Forms of information storage represent segments with databases of the problem and objective orientation as well as booths of data with the selective data in terms of the users’ orientation. The segments’ database and booths are universally connected with the operation and special databases in the federal and regional structures. They are united in the global network of data collection arriving from different sources.

In the context of development of network technologies it became urgent to ensure the users interface with the distant (external) sources of information. One of the efficient ways to solve the emerging problems is creation

of the unique mechanism of access to the internal and external resources in the form of the corporative information portal.

Information portal integrates such internal applications as: e-mail, access to databases, OLAP, Data Mining and DSS, systems of documents management with the external application (Internet news services and users' Web-nodes). Modern intelligent information technologies used in the SC can be represented by four groups of technologies: operative analysis, intelligent analysis, simulation and decision-making solutions support.

Technology of operative analysis of data makes it possible to estimate the state of the processes being observed to reveal and range the causes of significant variations to forecast development of the processes.

Recently a wide-spread occurrence was obtained by analytical systems based on OLAP technologies (OLAP – on-line analytical processing – operative analytical data processing), which make it possible to group the data into representation of information in the form of N-dimensional cube. Such a technology gives the possibility to an analyst to receive "multidimensional reports" from miscellaneous sources of information and form the necessary sampling in the form of different data cut.

Intelligent analysis of data ensures automated search of earlier unknown regularity in the databases of the information fund. The use of the acquired knowledge allows so simplify significantly the procedure of the informative analysis for the analysts and to increase efficiency of the models' design by experts. Methods of the artificial intelligence make the basis of the intelligent analysis.

Simulation of the decision-making process makes it possible to give a quantitative estimation and perform quantitative analysis of results of the made decisions.

In these technologies oriented to the users from the organs of the state power the models are used classified as the models of socio-economical processes, models of socio-political processes as well as the models of extraordinary situations.

Thus, in the socio-economical sphere the use of the models of regression analysis makes it possible to predict dynamics of macro economical indices, development of different branches of industry and agriculture, to perform comparative analysis of socio-economical situation of regions in Russian Federation, to estimate their investment attractiveness etc.

In the socio-political sphere, for example, the model calculation is widely used for procession and analysis of data of the sociological enquiries.

In the process of preparation of decisions in the extraordinary situations simulation models are widely used, they allow to analyze the development of the extraordinary situations, to estimate their consequences and calculate the required resources needed for elimination of resources damage.

Decision-making support systems are singled out separately; they are functionally oriented to the preparation of the analytical reports and documents, performance of the group expertises, development of recommendations and decision versions ranging. In these systems they used mainly methods of expert estimates, models of group estimates, method of analysis of hierarchy (method of pair comparisons), method of alternative decisions synthesizing.

Information-analytical systems input into software-hardware complex of the SC significantly extend its functional possibilities. In this framework of the highest organs of the state power the most required are the general mathematical information-analytical systems (IAS).

The means of the information presentation unite the following types of technologies:

- Cartography of the problem situations and objects of decisions;
- Structuring of fuzzy ideas and decision hypothesis;
- Multimedia imagery of the situation dynamics.

Problems of Information Security

Provision of the information security is one of the most important problems in application of the information technologies in the practice of the state management. Organization of the systems' functioning in the modern

communication means in conjunction with the necessity to perform the requirements on the user service significantly complicates this problem solution.

Two types of information struggle should be singled out, namely, information-psychological and information-technical ones. The main objects of impact and security in information technical struggle are information-engineering systems (communication systems, telecommunications system, radio electronic means etc).

Realization of information security hazard can result in a serious and, in a number of cases, catastrophic consequences; the main of them are as follows:

- violation of the state establishments, social organizations and institutions activity;
- variation of the individual or mass consciousness, moral and political, social and psychological climate in the society;
- infringement of the state national interests due to the drain of the most important information

Prospects on the Information Technologies Development and Proposals on the Situation Centers Development

Today the level of development of the information technologies makes it possible to envisage confidently the possibilities of creating the system of "electronic power" in the near future. The essence of the efficient control and the processes of decision-making management at the first stage and further in formation of the complex system of control of socio-economical and socio-political processes in the context of the state management.

Under conditions of the global information of the society the role and functions of management transform, this, respectively, is reflected on the realization of economic and social policy as well as on the support of democratic institutes of power. At present it becomes evident that in future the efficiency of the process of management will depend even greatly on the quality of information. In this case the main methodical directions of rising quality of the information support of the state management in the context of the considered problem are, at our point of view, as follows:

- development of methodological principles for introducing the situation centers as means of the system integration of the intelligent information technologies into practice of the state management of the federal and regions levels;
- development of the methodology for consolidation of the information resources and clear definition of their propagation limits, creation and development of the information funds on the basis of the distribution calculations methodology and data storage;
- creation of the decision-making system on the basis of the modern intelligent information means of data procession;
- development of a wide class of the socio-economical and socio-political processes models;
- orientation in development and application to the management of the intelligent information technologies aimed to creation of the systems of "electronic power";
- creation of the efficient technologies of the information security oriented to application to information-analytical systems of the state management.

Conclusion

Results of the imitative simulation systems' use in the situation centers are given in this work. Majority of the program complexes, used in the world for the economical, political, and financial simulation, are based on the methods of the so called system dynamics. The latter one, in its turn, uses the fuzzy cognitive maps' apparatus offered by Kosko at the beginning of the eighties and used for the first time in the field conditions during the political crisis in South Africa.

Author Information

Kuzemin A.Ya. – Prof. of Information Department, Kharkov National University of Radio Electronics, Head of IMD, (Ukraine), kuzy@kture.kharkov.ua

THE USE OF SITUATION REPRESENTATION WHEN SEARCHING FOR SOLUTIONS IN COMPUTER AIDED DESIGN SYSTEMS

Alexander Kuzemin, Mikhail Sorochan, Igor Yanchevskiy, and Asanbek Torojev

Abstract: Projects solutions reuse methodology is offered for software development. The main idea consists in connection of the system objective with the situation using the entities which describe the condition of the system in the process of the objective statement. Every situation is associated with one or several design solutions, which can be used at the development. Based on this connection the situation representing language has been created, it lets to express a problem situation using a natural language describe. The similarity measure has been built to compare situations, it is based on the similarity coefficients with adding the absent part weight.

Introduction

Business software development for modern organizations is a severe and difficult process. The latest researches in software companies show that more than a half of all projects exceed the time and resource budgets [1]. One of the commonly occurring tasks in computer aided design systems is an assembling of the object being designed using complete components or similar ones after some modification [2]. Any engineering project including a software one, which solves new problems by existing solutions reuse, has great possibilities to reduce resource and time costs to develop a new project [3].

But in component reuse the problem of finding the proper project solution matching requirements established to a desirable component arises. Industrial level design methodologies (RUP, UML, SADT etc.) have no embedded guidance for solutions reuse, their main goal is the design models visualization.

The Goal of the Work

The main goal of this research is to increase productivity in the computer aided software engineering systems by reuse of the solutions obtained before. Not only the completed program modules and components can be reused, but algorithms, diagrams and other project artifacts [4, 5]. Solutions reuse is effective only when the taken to search the existing solution than that spent for creating a new one. Therefore, when developing the reuse methodology the main attention should be paid to the algorithms of the project solutions search.

Statement of the Problem

To reuse design solutions effectively one has to solve two main tasks: the current solution classification in the knowledge base and search for the most suitable solution for a particular task. This can be done by developing a unified mechanism of describing solutions' parameters to compare them with the current task being solved by designer. Such descriptions are offered to shape in the form of situations, which consist of the natural language constructions describing the system state and its environment.

The Main Body

Modeling an organization problem domain a designer builds a system model which is a reflection in an observer mind of objects attributes and their relations in the research problem solution.

The Sm system as a reflection of the target organization for which the software is developed can be presented by a designer as [6]:

$$Sm = < A, R, Z, P, G >, \quad (1)$$

where:

A - elements (components) of the Sm system,

R - relations of the system elements,

Z - the system goals,

P - the system designer,

G - the designer language (the chosen modeling methodology)

The goal of the software being developed to meet the Z goals is satisfaction of system Sm. In the initial stages of the design process a designer P using different language means G carries out research on the target organization to clarify the relationship of A elements connected in the form of R to gain Z goal.

Environment together with the system concerned in it can be modeled in different ways. Traditional methods which describe the object as a set of mathematical formulas do not give the idea on the object being described by the given equation.

The main problem in data gathering by analysts from the problem domain experts is associated with different language entities operating with and the need in a common communication language [7].

In [7] the author puts forward two hypotheses concerning the presentation of information on the objects. The first one says that all information about the object under control can be presented by the common natural language means. The second one says that any text which describes the object under control may be translated into a formal language. There is no doubt that there may exist objects for which the mentioned hypotheses fail but experience has proved their consistency.

Considering the solution of some problem or achieving some goal one may say that there exist some initial conditions of the task. This information describes the condition of the modeled object and its environment in the moment when the need appears the need to solve the set appears. Such a state can be effectively presented by the situations presented in the [8].

The situation consists of entities sets. To find a design solution appropriate to the current situation one needs to compare the current situation with a set of etalon situations. Before the situation building it is necessary to identify entities composing the situation. Conventional classification methods such as the cluster analysis do not match here because it is unknown beforehand what is entity and how its attributes are correlated with etalon entity attributes. That's why before the situation description it is required to develop and guide some additional operations on the entities identification.

To identify correctly the entities belonging to the current situation description a categorical classification algorithm can be used. This algorithm lets to correlate effectively unknown entity with the known class. It is implied that there exist categories groups of the entity classes. These categories are built as hierarchies with subsumption relations.

When modeling the organization problem domain the designer finds himself in a certain situation which is defined by the group of the interrelated entities presenting different organization features and describing the current situation. An interrelation with some entity means that the entity exists in the current situation and in some way takes part in the action, it is a rule or deliverable, it acts as an observer or tool, it has some special feature etc. That is why the organization presentation may be considered as a system consisting of a set of entities which are interrelated between each other and form integrity and organic unity.

Having found himself in one of such situations a designer may have a need to create one or some design diagrams, require some guidance or need some other project artifact. The similarity of such situations is confirmed by many researches [9]. Therefore, the designer encountering meeting such situations may operate common design solutions too. Every time meeting the situation known before and for which there exists a group of design solutions the designer may use the solutions associated with the situation.

The situation contains sets of entities each of them reflects its qualities in any characteristic category. Such categories can present processes and activities, resources and goods, organization structure, deliverable structure and requirements, cost and so on. An entity set separated from entities' interrelations can hardly describe a situation because those entities are related to each other, they take part in variety of processes, appear and disappear, create new entities and cooperate between each other.

It is offered to divide categories into common main groups reflecting the main aspects of structural and dynamic organization presentation. Each of those groups will consist of categories sets presenting entities of the group.

Category C is presented as an hierarchy by the oriented graph of entities $C =< O, A >$ which is build by the subsumption indication where O is a set of graph nodes presenting entities $\{e_1, e_2, ..., e_n\}$ and $A = \{G_1, G_2, ..., G_{n-1}\}$ is a set of graph arcs representing generalization relation between two entities directed from a successor to an ancestor. Entity e_1 is a primary entity of category C which represent the most common entity in a hierarchy.

Every entity includes a pointer to the decision rule which defines a set of features distinguishing this entity from others.

The decision rule is presented as:

$$DR = \hat{}_i F_i \,, \tag{2}$$

where F_i is a separate feature.

To calculate the value of the decision rule DR each feature F_i gets the truth value if this entity has the feature otherwise it gets a false value.

The decision rule defines a group of features making it possible to distinguish the given entity from other ones. A feature is something that characterizes the entity defining its likeness or difference in a variety of entities. Feature may present both presence values (positive feature) and absence values (negative feature).

In a variety of features one selects the most typical and distinctive for this kind of entities. Therefore in entities formation abstracting and idealization are required. As a rule it is called an entity essential features allocation. It should be remembered that allocation of exactly essential features is not always performed in practice and even in some cases is unrealizable. The essence of objects, processes, relations and events is defined by science. In this case it is enough to point to the features which distinguish the discussed entities from others to avoid confusion. Along with essential features it is possible to use some features sufficient for distinction; they let to define an entity sufficiently and unambiguously not to confuse the entity with others though these features may be inessential for this entity. In practice such features are used even widely. In this work attention is paid to the essential features use. It is considered that it is sufficient to point on features unambiguously distinguishing some entity from others.

Each entity must contain only one semantic value which concerns only one essence. This makes it possible to avoid ambiguity using this entity in situations.

An unknown entity is classified by the entity selection strategy, for the category is $C =< O, A >$ it is presented as :

$$S_C =< P, D, p_0, D_0 >, \tag{3}$$

where $P = \{p_1, p_2, ..., p_n\}$ is a points set, $D = \{D_i\}, i = \overline{1, n}$ is a set of transitions to every point $p_i \in P$, references to primary point $p_0 \in P$ from which the identification begins, reference to the primary transition $D_0 \in D$ which assists to make decision whether the entity belongs to the category or not.

In transition $D_k = \{d_{j,k}\}$ parameters $j = \overline{1, n}, k = \overline{1, n}$ mean that the transition is directed from the point p_j to the point p_k or in other words from the entity class e_j to the entity class e_k .

The transition D_k from some entity class to the entity class e_k corresponds the decision rule DR_k .

The entity identification strategy in a category is not a separate set but only a semantic union over entities hierarchy in the category and their decision rules. Such separate representation allocation makes it possible to consider the strategy in algorithms as a single element. The strategy is not a static set. It is build depending on the current structure of the category and entities it consist of.

Thus, it is obvious that the possibility of transition to some point is defined only by decision rule connected with corresponding entity. $D_k = \{d_{i,k}\}$ transitions direction is backwards to arcs directions from the set

$A = \{G_1, G_2, ..., G_n\}$. Every $G_j(e_i, e_k)$ arc corresponds to the set of D_i transitions for the pair of points p_k and p_i.

Every strategy begins with the primary transition D_0. Passing through the transition means that the unknown entity may belong to this category. The primary transition leads to the primary point p_0 corresponding to the entity e_0 of the category C.

The situation is formed as a set of triples – micrositutations. The situation S is presented as an aggregate of n microsituations s_i [8]:

$$S = \{s_i\}, i = \overline{i,n}, s_i =< e_i, R_{e_i} >, \quad (4)$$

The part of the situation described by the pair $s =< e, R_e >$ is called a microsituation of this entity e.

The entity e is called the central entity of a microsituation $< e, R_e >$, set R_e is a context of the central entity e. An entity is called the central one when it is a central semantic unit where this microsituation is based. Set $R_e = \{r_i\}, i = \overline{1,m}$ consists of entities expressing the relation of the central entity e with other secondary entities participating in the microsituation.

The microsituation corresponds to the triple *subject – control action – object* a human mind is operating with [10]. The subject is the central entity, the control action is the context and object is the secondary entity. In a general case the sequence *subject – control action – subject – control action … object* [10] corresponds to the case when the secondary entity from one microsituation is the central entity in the other microsituation.

To compare microsituation let us define the similarity measure expressing the distance between situations. Its calculation is divided into calculation of the distances between microsituations and deeper between entities.

The general principle of calculation of the distance between two entities, microsituations and situations consists in separation of the common part of the objects being compared and calculation of how each of them differ from the common part. The idea of distance calculation is depicted on figure 1.

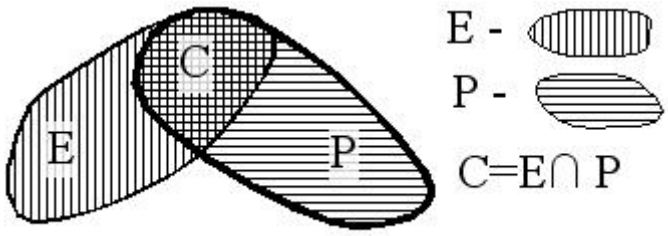


Fig. 1 – Correspondence between two sets

Comparison between two sets etalon E and problematic P is exemplified. It is also known that this sets have a common part $C = E \cap P$. Let us consider that etalon set E is linked with some solution D and it is necessary to calculate the similarity between the sets E and P to decide whether the solution D can be used with the set P. The similarity degree Sy is defined by a part of set C in every of comparable sets. But here it is necessary to take into account that solution D is connected with elements of the set E and in the case of $C \subset E$ that solution is leaned on the absent part E / C that is missing from the set P. This decreases the applicability degree of the solution D for problematic set. At the same time the set $C \subset P$ defines the part of problematic set P for which this solution D is applicable. Therefore, the similarity degree of the etalon and problematic sets

$Sy(E,P)$ in the common case is a part of the set C in the union $E \cap P$. But here the additional weight of absent part E/C weakening Sy must be also considered.

Let define the set A cardinal number as $\mu(A)$. Starting from Tversky's contrast model [11] and relying on the association coefficients (Zhakkar) and similarity (Roberts and Tanimoto), taking in account the remark on the absent part the similarity degree can be presented as:

$$Sy(E,P) = \frac{\mu(C)}{\mu(C) + \mu(P/C) + \frac{\mu(E)}{\mu(C)} \cdot \mu(E/C)}, \quad (5)$$

where $\mu(E)/\mu(C)$ is the absent part strengthening coefficient which is directly proportional to the part of the set E/C in the set E.

Thus, calculated in such a way the similarity Sy is dissymetric $\forall A \neq B, Sy(A,B) \neq Sy(B,A)$ and therefore cannot be a metric, but in the context of the considered problem it is an appropriate one.

The absent part strengthening coefficient is present in all further equations for distance calculation for situations, microsituations and entities. Similarity measure is expressed by the interval $[0,1]$. The same situations have the similarity equal to unity and for absolutely different situations the similarity is equal to zero.

Considering the situation S as a set of n microsituations $\{s_1, s_2, ..., s_n\}$, the distance between etalon S_1 and problematic situation S_2 will be presented as:

$$L(S_1, S_2) = \frac{1}{n} \begin{cases} \max\limits_{a_j} \sum\limits_{i=1}^{n} L(s_i, s_{d_i}), d_i \in a_j, a_j \subset T_m^n, npu\ n \le m \\ \max\limits_{a_j} \sum\limits_{i=1}^{m} L(s_{d_i}, s_i), d_i \in a_j, a_j \subset T_n^m, npu\ m < n \end{cases}, \quad (6)$$

where n and m is a quantity of microsituations for the situations S_1 and S_2, respectively; $L(s_1, s_2)$ is a distance between microsituations $s_1 \in S_1$ and $s_2 \in S_2$; T_n^m is a distribution of A_n^m groups of microsituations' numbers of n elements by m, $T_n^m = \{a_j\}, j = \overline{1, A_n^m}$; every element a_j consists of the set of microsituations numbers $a_j = \{d_1, d_2, ..., d_m\}, d_k \in N$.

The distance between microsituations is presented in the form of the following product:

$$L(s_1, s_2) = L(e_1, e_2) \cdot L(r_1, r_2) \cdot \frac{1}{n} \cdot \begin{cases} \max\limits_{a_j} \sum\limits_{i=1}^{n} L(e_i, e_{d_i}), d_i \in a_j, a_j \subset T_m^n, npu\ n \le m \\ \max\limits_{a_j} \sum\limits_{i=1}^{m} L(e_{d_i}, e_i), d_i \in a_j, a_j \subset T_n^m, npu\ m < n \end{cases}, \quad (7)$$

where e_1 and e_2 are the central entities of microsituations s_1 and s_2, respectively; r_1 and r_2 are the entities of microsituations s_1 and s_2 representing also features and relations; e_i and e_j are the secondary entities; n and m are amounts of entities taking part in relations r_1 and r_2; T_n^m is a distribution of A_n^m groups of entities numbers of n elements by m, every element a_j consists of the set of microsituations numbers $a_j = \{d_1, d_2, ..., d_m\}, d_k \in N$; $L(a,b)$ is a distance between entities a and b.

The etalon entity is an entity situated in the concerned category. The problematic entity is an entity taking part in the current situation description. The distance between the comparable entities is defined as a degree of feature

correspondence with the nearest paternal entity (NPE) and its remoteness from NPE. The distance between etalon e and problematic p entities is presented as follows:

$$L(e,p)=\frac{D_a}{D_a+(P_a-D_a)+\frac{E_a}{D_a}\cdot(E_a-D_a)}, \quad (8)$$

where D_a is a quantity of NPE features, E_a is a quantity of etalon entity features, P_a is a quantity of problematic entity features.

Design solutions search consists of the following stages:

Current situation description as a narrative text.

Revealing of the entities in the available categories from the obtained description.

Construction of the description in the situation representing language.

Search for the etalon situation close in its structure to problematic one within the defined similarity interval

Narrowing of the search by including entity objects attributes

Situation description looks as follows:

situation : microsituation-group | entity-declaration microsituation-group |
entity-declaration entity-attributes microsituation-group

Components of *entity-declaration* and *entity-attributes* are not compulsory. They may absent if there is no need to give names to same kind entities and (or) indicate values of their attributes.

Situations' description presented in the situation description language obey definite rules which let to understand the situations structure with some fixed meaning.

Axiom 1. The following statement holds for two microsituations s_1 and s_2 of some situation S. If the microsituation s_1 is described before the microsituation s_2 in the situation S description, then it can be said that a time interval $\Delta t \geq 0$ passed between the microsituations s_1 and s_2.

Theorem 1. If the entity e_2 inherits the entity e_1 - $G(e_2,e_1)$ in some entitiess C category then the entity e_2 can substitute the entity e_1 in the microsituations without loss of meaning and self-descriptiveness of the given microsituation.

Proof. According to the category structure and entity identification strategy to identify the entity e_2 it is necessary to give an affirmative answer to the decision rules $p_{1,}p_{k_1},...,p_{k_n},p_2$ corresponding to the enities $e_{1,}e_{k_1},...,e_{k_n},e_2$. This means that having identified the problematic entity as e_2 the above mentioned decision rules were passed including p_1, which is related to e_1. Therefore, the entity e_2 can be interpreted as the entity e_1 offering its characteristic features.

Conclusions

The presented similarity measure makes it possible to search for design solutions in the knowledge base to provide the basement for engineering solution synthesis tasks in the computer aided design systems. The obtained methodology ensures increase in the designer work efficiency at the cost of using both finished and similar to the current task solutions.

The further evolution of the developed methodology is an expansion of expressive means of the situations representation language for description of casual-effect relations between entities in a situation. In some cases it may be of interest to provide microsituations and entities degree expression importance.

References

1. *Леффингуэлл Д.*, Уидриг Д. Принципы работы с требованиями к программному обеспечению. Унифицированный подход. - М: Издательский дом "Вильямс", 2002. - 448 с.
2. *Петренко А.И.*, Семенков О.И. Основы построения систем автоматизированного проектирования. – Киев: Высшая школа. – 1984.
3. *Park JY*, Musen MA VM-in-Protege: A Study of Software Reuse // Proceedings of MEDINFO'98. - 1998. – pp. 112-115.
4. *Соммервил И.* Инженерия программного обеспечения. - М: Издательский дом "Вильямс", 2002. - 624 с. 5. *Гамма Э.*, Хелм Р., Джонсон Р., Влиссидес Дж. Приемы объектно-ориентированного проектирования. Паттерны проектирования. - СПб: Питер, 2001. - 368 с.
6. *Системный* анализ в экономике и организации производства. - Л: Политехника, 1991. - 398 с.
7. *Поспелов Д.А.* Ситуационное управление: теория и практика. - М: Наука, 1986. - 288 с.
8. *Гладун В.П.* Процессы формирования новых знаний. - София: Педагог 6, 1994. - 192 с.
9. *Beznosov Konstanin*. Architecture of Information Enterprises: Problems and Perspectives // Advanced Topics in Software Engineering - 1998. - p. 25
10. *Ерофеев А.А.*, Поляков А.О. Интеллектуальные системы управления. СПб: Издательство СПбГТУ, 1999. - 247с.
11. *Tversky A.*, Similarity, separability and the triangle inequality // Psychological Review. - 1982. - №89(4). - pp. 123-154.

Authors' Information

Alexander Ya. Kuzemin – Prof. of Information Department, Kharkov National University of Radio Electronics

Mikhail V. Sorochan – Post-graduate student of Kharkov National University of Radio Electronics, msorc@ua.fm

Igor V.Yanchevskiy – General Director of the Company “Computing Technique and Automation Means” (Russia)

Asanbek A.Torojev – General Director of the Joint Russian-Kirghiz Venture “Computing Technique and Automation Means”

MULTI-AGENT SYSTEMS IN THE HARVEST PROGNOSIS

Laura Ciocoiu, Cristian Paraschiv, and Dragoş Barbu

Abstract: The paper presents a case study of geo-monitoring a region consisting in the capturing and encoding of human expertise into a knowledge-based system. As soon as the maps have been processed, the data patterns are detected using knowledge-based agents for the harvest prognosis.

Keywords: data mining, topological maps, GIS, knowledge based agents, Model Based Reasoning

Introduction

The process of geo-monitoring a region needs to use knowledge-based systems as a resource for aiding the specialists and the people in achieving their objectives. The model design process represents, in fact, the transferring of human experience in monitoring into an *interactive model.* Knowledge about process of the geo-monitoring of a region (e.g. map interpretation, statistics methods, strategies, etc) is represented by models that refer to observable features and significance.

So, the model designer has an inference network for the reasoning process; following the initial design of a model, the evaluation techniques are used to assess it in order to extent it; so, there are identified specific parts of the model that should be modified and the priorities of these modifications should be recorded. Besides the way of the geo-monitoring a region for a number of subjects, this process is in an on-going change (interference of Internet knowledge, including the artificial intelligence elements).

Area Monitoring

The aim of this work is to present a case study for capturing and encoding human expertise in geo-monitoring a region into a knowledge-based system. The main goals of weather and geo-monitoring of a region involving the environmental preservation are: modelling of liquid flows of rivers, harvest prognosis, riverbed degradation (utilization of hydropower potential) such some snapshots of our site "Geo-monitoring of a region" **http://inteligent-agents.ici.ro/gis/**

After the pre-processing operation of the topological and lithographic maps, knowledge about the process of the geo-monitoring of a region (e.g. statistics methods, strategies, etc) is represented by models that refer to observable features and significance.

The prognosis of harvest is realized on the base of geographical data such as using knowledge based modelling (hierarchical tree of classes and method definition, FBS decomposition, agent cooperation)

Harvest Prediction

This tool uses the knowledge engineering concepts such as *function-behaviour relationships, intelligent agents definitions* for creating a complex database for geo-monitoring a region.

Knowledge Representation is based on Model Based Reasoning (MBR) for an intuitive functional description, which is concerned with the design knowledge and geo-monitoring actions. The Functional-Behaviour-Structure (FBS) modelling consists of recursively decomposition of the functions into sub functions using a catalogue to look up the *most appropriate functional element* that means a component or a set of components that perform a function. Functional reasoning adds functional concepts into functional hierarchy and uses *a Casual Process description* with Casual Functional representation. The tool uses shared knowledge bases where the knowledge is represented as classes, objects and rules.

Function-Behaviour-State/Structure Modeller supports functional design (functional knowledge decomposition and function modelling) that means the defining of *function-behaviour* and *behaviour-state (structure)* relationships, causal decomposition (hierarchies constructed from the *function viewpoint*), construction of behaviour network, behaviour simulation, evaluation.

Knowledge Representation consists of constructing Function Knowledge Base and Behavior Knowledge Base based on model-based reasoning permit to generate a lot of alternatives, some of them "the best" for this process.

The process of geo-monitoring a region is simulated by constructing the Intelligent Agents referring to collaboration/ communication facilities. So, this model supports multiagents that generate partial results, data dependencies and conflicts among them.

Due to the complexity of application, in the 'Harvest prediction' process, along with the GIS software, a Knowledge-based System is used. First of all, there are defined superclasses **geographical features, geological structure, soil type, seed type, improvements, rainfall** with their attributes and instances. Also, some special data structures and classes had to be defined such as **Harvest, Basic Tilling, Auxiliary tilling, Soil Maintenance, Crop, Plant, Plant Control, Plant Maintenance**. External parameters (geographical position, climate), soil and seed, with their attributes (type, treatment, history, period) and their methods (work, evolution) are considered. The goal is quality and quantity of the crop. Each class represents an intelligent agent; the agents communicate each other in order to solve the conflicts.

For example, for a plain region with the quantity of the rainfall satisfactory and treatment, the harvest is expected to increase with 20%, according to the rules of the model.

The interpretation of the output data. We make the classes *geographical features*, **rainfall** and **soil type**, which have the greatest influence on the harvest evolution. The most important class is *geographical features,* which determines the quantity of the rainfall and the type of the soil.

For the moment the model implementation is made for a few numbers of classes and rules. We intend to improve the algorithm to include additional classes and methods.

Conclusions and Future Trends

Collaborative efforts in artificial intelligence, statistics, and databases have been made in order to model the phenomena and to be a support of decision-making.

The digital geographic databases are simple representation of geographic knowledge at the level of basic geometric, topological and measurement constraints. Knowledge-based GIS attempts to build higher-level geographic knowledge into digital geographic databases for analysing complex phenomena. Geographic knowledge discovery is a source for knowledge-based GIS and intelligent analysis.

For the next work we intend to develop our application by map interpretation and information extraction, environmental features defining (e.g. inductive learning techniques and neural networks to classify and map soil types, mapping vegetation types), extracting patterns, creating the geographical database.

Acknowledgement

This Technical Report was made by a team from National Institute for Research and Development in Informatics, on September, 2003, under INTAS Project no. 397: "**Data Mining Technologies and Image Processing: Theory and Applications**" coordinated by Lappeenranta University of Technology, Finland.

We want to thank the team leader dr.mat. Angela Ionita, Deputy Director of RACAI, for her work and assistance during this period.

References

Bhandari, E. Colet, E., Parker, J., Pines, Z, Pratap, R., Pratap, R., Ramanujam, K., 1997, *Advanced scout: data mining and knowledge discovery in NBA data*. Data Mining and Knowledge Discovery, 1.

Brachman, R.J., Anand, T., 1996, *The Process of Knowledge Discovery in Databases*, in Advances in Knowledge Discovery and data Mining, U.M. Fayyad et al. (eds.) AAAI Press/ The MIT Press.

Buttenfield, B., Gahegan, M., Miller, H., Yuan, M., *Geospatial Data Mining and Knowledge Discovery.*

Fayyad, U. 1997. Editorial. *Data Mining and Knowledge Discovery*.

Fotheringham, A. S., Charlton, M. Brunsdon, C., 1997, *Two techniques for exploring non-stationarity in geographical data*. Geographical Systems.

Gahegan, M., 2000, *On the application of inductive machine learning tools to geographical analysis*, Geographical Analysis.

Gahegan, M. O'Brien, D. (1997). *A Strategy and Architecture for the Visualisation of Complex Geographical Datasets*. International Journal of Pattern Recognition and Artificial Intelligence, 11(2).

Gahegan, M., Takatsuka, M., Wheeler, M., Hardisty, F., 2000, *GeoVISTA Studio: A Geocomputational Workbench*. Proc. 4th Annual Conference on GeoComputation, UK, August 2000. URL: http://www.ashville.demon.co.uk/gc2000/.

Authors' information

Laura Ciocoiu – senior researcher. Main interests are in intelligent agents, multi-agent systems, multimedia systems and security tehnology

Dragoş Barbu – researcher. Main interests are in intelligent agents, multi-agent systems, neural networks

Cristian Paraschiv – programmer. Main interests are in security tehnology and dynamic Web portals.

National Institute for Research and Development in Informatics;
8-10 Averescu Avenue, 011455 Bucharest 1, ROMANIA
email: ciocoiu@ici.ro http://intelligent-agents.ici.ro

LOCAL GOALS DRIVEN HIERARCHICAL REINFORCEMENT LEARNING*

Arthur Pchelkin

Abstract: Efficient exploration is of fundamental importance for autonomous agents that learn to act. Previous approaches to exploration in reinforcement learning usually address exploration in the case when the environment is fully observable. In contrast, the current paper, like the previous paper [Pch2003], studies the case when the environment is only partially observable. One additional difficulty is considered – complex temporal dependencies. In order to overcome this additional difficulty a new hierarchical reinforcement learning algorithm is proposed. The learning algorithm exploits a very simple learning principle, similar to Q-learning, except the lookup table has one more variable – the currently selected goal. Additionally, the algorithm uses the idea of internal reward for achieving hard-to-reach states [Pch2003]. The proposed learning algorithm is experimentally investigated in partially observable maze problems where it shows a robust ability to learn a good policy.

Keywords: reinforcement learning, hierarchical behaviour, efficient exploration, POMDPs, non-Markov, local goals, internal reward, subgoal learning.

Introduction and Problem Statement

One of the directions in artificial intelligence is adaptive autonomous agents' research. This research direction started growing actively in 1985 [Maes95,Wil85], however, it was proposed to make researches in similar directions also before it [Bong75]. Reinforcement learning [Sut98,KLM96,Mit99] examines a question how an autonomous agent [Maes95] that senses and acts in its environment can learn to choose optimal actions to achieve its goals. The agent can perceive a set of distinct perceptions from its environment and has a set of actions that it can perform. At each discrete time step, the agent senses the current percept, and chooses an action to perform it. The environment responds by producing the succeeding state and the agent can perceive a new observation. If the agent achieves the goal state, the environment gives the agent a reward.

The set of actions allowed to the agent is fixed and defined before learning. The structure of the environment is unknown to the agent and is represented by a black box. This means that it has to obtain all knowledge helping to achieve the agent's goals only by itself, only by experimenting with the environment. The task of the agent is to perform sequences of actions, observe their consequences, and to learn a control policy.

Irreversible transactions: Efficient exploration plays a fundamental role [Thr92b] for autonomous agents that learn to act. In many reinforcement learning algorithms undirected exploration techniques are used. While undirected exploration techniques, *e.g.* random walk exploration, utilize no exploration-specific knowledge and ensure randomness into action selection, directed techniques rely on knowledge about the learning process itself, allowing for exploring in a more directed manner [Thr92a]. In many finite deterministic domains, any learning technique based on undirected exploration is inefficient in terms of learning time, *i.e.* learning time is expected to scale exponentially with the size of the state space [Whit91].

The reason for the difficult exploration by undirected techniques is the existence of irreversible transactions between states of the environment. Usually learning algorithms are being investigated in fully reversible domains [Sut98,KLM96,Mit99,McCallum95], *e.g.* the maze problem. In the maze problem each action has an opposite action, and the agent usually can easily undo its previous action by one step, *e.g.* the action "go left" can be undone by the action "go right". Because of this reason, each state of the environment becomes easily achievable. However, it could not be easily applied to the real case, *e.g.* if the goal of the agent is to build a house from blocks, one wrong random action may discharge all the previous work by destructing the built construct. That's why, the existence of irreversible transactions is not exclusion or the specifically invented difficulty but it is

* This research was partially supported by the Latvian Science Foundation under grant No.02-86d.

a real difficulty that was unfortunately ignored in many investigations making simulations in the artificially designed environments.

Usually in practice multiple situations are indistinguishable from immediate perceptual input. These multiple situations may require different responses from the agent. Usual reinforcement learning techniques, such as Q-learning [Wat89], can't be applied in partially observable domains. Due to this reason, McCallum has developed a learning algorithm "Utile Suffix Memory" (USM) [McCallum95] that is able to overcome the incomplete perception problem in order to learn a near-to-optimal policy in partially observable environments.

Efficient exploration in partially observable domains is a special difficulty [Chr92]. Previous approaches [Thr92a,Sut90] to exploration in reinforcement learning usually address exploration in the case when the environment is fully observable. In contrast, McCallum [McCallum97] considers efficient exploration applied to USM in the partially observable environment. In our study, we continued his research by applying to USM an exploration technique based on internal reward for hard-to-reach states [Pch2003], and our modification has outperformed the original algorithm in the case of difficult exploration.

Problem statement: Unfortunately, USM has one very important drawback – it is not able to make perceptual distinction by seeing too far future back in time. That's why, USM is not able to discover too complex temporal dependencies that include too many time steps. That was a motivation for the current paper to develop a reinforcement learning algorithm having an ability to overcome the combination of all the difficulties described above, *i.e.*:

- incomplete perception;
- irreversible transactions;
- complex temporal dependencies.

Automatic building of hierarchies: Complex temporal dependencies are usually solved by allowing the agent acting hierarchically [WS98,SS2000]. That's why, the present paper also considers hierarchical reinforcement learning as a key method for overcoming difficulties related with complex temporal dependencies.

There are many different models of hierarchical reinforcement learning [KLM96]. Considering different approaches of making hierarchical reinforcement learning, it is possible to distinguish two cases: (1) the use of structurally pre-determined domain-specific hierarchies and (2) automatic building of hierarchies. Most of hierarchical reinforcement learning algorithms are based on an assumption that fixed domain-specific knowledge about hierarchies is provided and can be exploited by the algorithm, *e.g.* [Die2000,PR97,Hum96] are only a small part of them. In contrast, in the current paper it is assumed that no prior domain-specific knowledge about subtask hierarchies is provided to the agent. Like [WS98,SS2000], the current problem statement is more difficult, but it is also more realistic.

Learning Algorithm

Training skills at local goals: In previous paper [Pch2003], it has been found that McCallum exploration techniques [McCallum95,McCallum97] may fail in the case when the environment isn't reversible, *e.g.* if there are one-direction ways. In the last case, it may be difficult to find a goal first time. USM uses Q-values to discover distinctions in the environment, but these Q-values are accessible only when the agent has reached the goal and has received the reward from the environment at least one time. Until this, the agent is unable to discover history distinctions and, thus, is unable to overcome incomplete perception problem. This problem has been solved [Pch2003] by giving the agent additional internal reward for state space exploration. Receiving additional internal reward for exploration USM was able to optimise its control policy not only for exploiting the environment but also for exploration in the same manner. It was relied on hypothesis that the perceptual distinctions discovered during exploration will help the agent to reach the goal state. In general, there is no principal difference between exploitation and exploration because in both cases the goal is to reach some special states of the world. In many cases distinctions needed for reaching the goal state are also needed for reaching some particular state. Simulation results in the maze domain had successfully confirmed the hypothesis.

The agent has the goal defined by its environment, let's call it the *global* goal (externally defined goal). However, any other state of the environment can also be considered as a goal, and it could be called a *local* goal (internally set goal). In the previous paper, while the agent was not able to reach the global goal, it was trying to reach some local goal to advance its skills of the environment control. Training skills at reaching local goals had helped the

agent to obtain skills sufficient for achieving the global goal. In the current paper, it is proposed to use the same idea of learning to reach local goals when the global goal is not achievable or it is hardly achievable.

Key ideas: The proposed learning algorithm is based on two ideas:

- hierarchical behaviour;
- training skills at local goals.

The agent needs hierarchical behaviour in order to overcome incomplete perception and complex temporal dependencies. Training skills at local goals are needed to make efficient exploration when the environment has irreversible transactions.

The learning algorithm has two parts. The first part selects the current main goal using three possible reasons: (1) a need to achieve the global goal, (2) a need to explore rarely observed perceptions and (3) a need to train skills at hard-to-reach goals. The task of the second part is choosing actions or selecting subgoals (like the calls of subroutines) in order to reach the current main goal. The selection of some subgoal can also be considered as abstract actions. In order to make the learning algorithm more simple, it is proposed to learn the primitive action selection policy and the abstract action selection policy in the same manner, applying the same principles.

Observations from the environment are used twice: (1) as the context information for action selecting and (2) as subgoals. In this sense, any main goal or any subgoal is a normal perception that has been observed but at the current moment the agent is trying to reach the state producing this observation. If the current observation is equal to the current goal, the goal is considered to be achieved. This means that the proposed learning algorithm is driven by local internal goals and their subgoals, and each goal or subgoal is a usual observation temporally considered in such role.

The agent also has the memory about successful cases and the agent adaptation rule could be described as follows: if in the context of observation p the selection of action a helps to achieve the current goal g (that can be a subgoal at the higher level), then next time the probability of the selection of the same action in the same context (observation p and the goal g) must be increased. This means that the learning algorithm doesn't exploit dynamic programming ideas about the estimation of distance to the goal, but it performs only pure reinforcement of successful actions.

To sum up, it should be noted that the agent does not only learn when to select what action, but also - when to select what subgoal. It means that the agent also must learn subgoal selecting policy.

Formal description of algorithm: The agent has the fixed set of actions A. The set of perceptions P is not directly given to the agent, instead, it is maintained all time and contains perceptions observed by the agent till the current time moment. Similarly, the counter $c(p)$ – the number of times the agent has observed percept p is maintained in order to provide directed exploration. Additionally, it is proposed to maintain the degree of difficulty $d[g]$ (initially equal to zero) for each local goal $g \in P$ that stores the total number of all failures minus the number of all successes at achieving goal g. Consequently, $D = \{ p \in P \mid d[p] > 0 \}$ can be considered as a set of difficult goals.

During learning process, the agent maintains its lookup table with real values $q[p,g,a]$, initially equal to 1, for each $p,g \in P$ and $a \in A \cup P$. This table doesn't store estimates of expected future discounted external rewards, *e.g.* as in Q-learning, instead, the value $q[p,g,a]$ stores the sum of all the internal rewards for performing action a or selecting a as a subgoal (if a is a perception) in the context of observation p and local goal g. The internal reward is not obtained from the environment, but the agent internally generates it for obtaining local goals or subgoals. To prevent the recursive calling of the same subgoal, it is proposed to define a set of the currently selected goals in the stack, noted by G.

It is also assumed that there is only one goal state in the environment, and the agent maintains a variable f storing initially *null*, or the observation of the goal state if the goal state has been achieved. The goal state can be recognized obtaining a positive reward from the environment.

The lookup table $q[p,g,a]$ is used for action selecting. However, this table is too big and it needs some kind of generalization on its values. For example, it may have an empty cell for some action a in the context of goal g and perception p, but at the same time it may have learned values for the same action in another context, and the last information also can be exploited in selecting of action a. For this purpose, we can define generalized value $Q(p,g,a)$ as follows:

$$Q(p,g,a) = \alpha_1 \cdot q[p,g,a] + \alpha_2 \cdot \frac{1}{|P|}\sum_{p \in P} q[p,g,a] + \alpha_3 \cdot \frac{1}{|P|^2}\sum_{p,g \in P} q[p,g,a]$$

Other notations: *t* – the current time moment, *random* – random value in interval [0;1).

The proposed hierarchical reinforcement learning algorithm can be described as follows:

```
Main:
  G = ∅
  repeat
    ExecuteAction(RandomAction); TryToReachGoal(GetMainGoal,λ)

GetMainGoal:
  if f ≠ null and (D = ∅ or random ≥ δ)  then result = f
  else if D ≠ ∅ and random < ½ then
         result = select g∈D with probability Pr(g) = 1
       else result = select g∈P with probability Pr(g) = 1/c(g)^η

TryToReachGoal(g,s):
  G = G∪{g} ; t0 = t ;  i = 0 ; E = ∅
  while CurrentObservation ≠ g and i < τ
    i = i+1
    p[i] = CurrentObservation
    V = { x∈A∪P | s ≥ 1/Q(p[i],g,x) & x∉G∪E∪{p[i]} }
    if V = ∅ then V = A
    a[i] = select x∈V with probability Pr(x) = Q(p[i],g,x)^β
    if a[i] ∈ A then ExecuteAction(a[i]) ; r[i] = SUCCESS
    else r[i] = TryToReachGoal(a[i], s - 1/Q(p[i],g,a[i]))
    if r[i] ≠ SUCCESS then E = E∪{a[i]}
  if CurrentObservation = g then
    while r[i] = SUCCES and i > 0 do
      q[p[i],g,a[i]] = q[p[i],g,a[i]] + 1/(t-t0)^γ
      i = i-1
  G = G -{g}
  if CurrentObservation = g then d[g] = d[g]-1 else d[g] = d[g]+1
  if CurrentObservation = g then result=SUCCESS else result=FAILURE
```

Notation *result* means the resulting value of a function, and notations t_0, *i*, *E*, *p*[*i*], *a*[*i*], *r*[*i*], *V* are local variables of function "TryToReachGoal". The algorithm has a series of parameters: α_1, α_2, α_3, β, η, δ, γ, λ, τ, and it is proposed to use the following settings: α_1=1, α_2=0.01, α_3=0.0001, β=5, η=5, δ=0.7, γ=0.1, λ=3, τ=4.

Simulation Results

The presented above learning algorithm has been tested using three different maze problems: maze1, maze2 and maze3 (see Fig.1). Each maze is a local perception grid world. The essence of this problem is searching for immovable goals in a maze. The agent's life consists of many trials: it is placed in a random empty cell, after which the agent has to find the goal (marked "G") searched with the least possible number of steps. Initially the agent has not any knowledge on the environment. Each trial can be considered as one problem solved by the agent. In the course of trials, the agent has to learn to quickly find this object.

The agent can move to nearest empty cells only (white or silver cells, but not black ones; some cells are specially highlighted with silver – it means that these cells have duplicated observations). Eight possible directions mean eight possible actions that the agent can execute. If the agent tries to move onto barrier, it stays at the same position. This creates many cycles in the environment, and makes the learning task more difficult. The agent can

perceive only the containment of nearest eight cells. So, there are different, but perceptually identical, world states.

Additionally, the cells can contain special symbols - arrows. These are normal empty cells, except the agent can move only in the direction defined by a corresponding arrow (in other case it stays in the same position). These arrows are needed to simulate discussed above irreversible transactions between the environment states.

To simulate complex temporal dependencies, there are presented two special cells: a door (marked "D") and a key (marked "K"). To be able to come into the cell with a door, the agent needs to visit the cell with a key before. After visiting the cell with a key, the agent is able to come into the cell with a door only once. If the agent is not able to come into the door, it stays at the same cell. The idea about the door and the key has been taken from the paper [WS98].

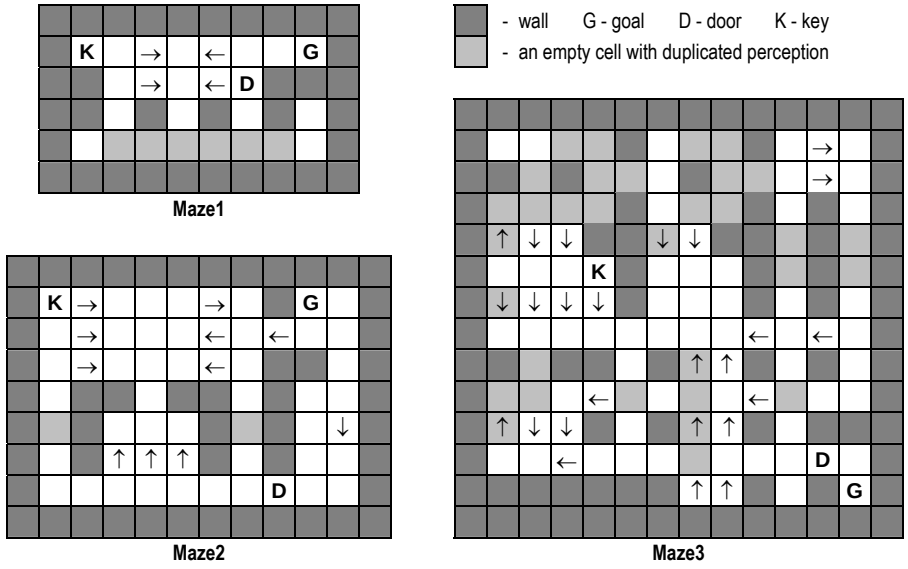


Figure 1. Three different maze problems: maze1, maze2 and maze3 for experiments

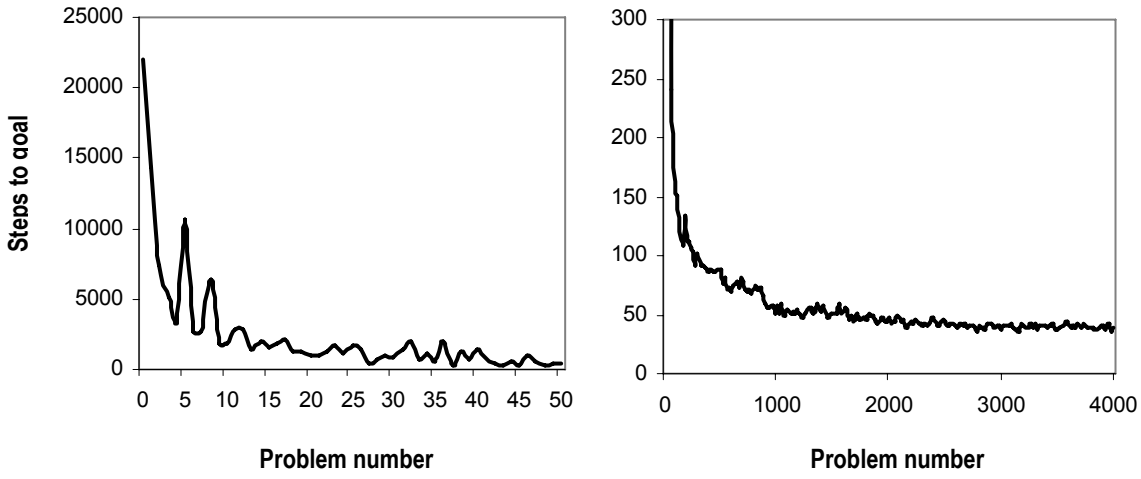


Figure 2. Experimental results in maze1

Results: Figures 2,3 and 4 show the dynamics of the number (see "steps to goals") of steps performed by the agent to reach the goal (a cell marked "G") in each trial during the learning process. Abscise "problem number" means the index of a goal searching trial. All experiments were repeated 10 times, and there are presented only averaged results.

Each simulation is represented by two graphs. The graph on the left shows the convergence of the number of steps to the goal in the beginning. Here, the learning algorithm shows its ability to roughly sketch out non-optimal but successful behavior. However, after this sketching the optimization of a policy continues very slowly. The second graph on the right shows this slow optimization. The same tendency could be noted on other two graphs.

Table 1 shows summary of all the experiments. It also contains additional information describing the selected maze problems. It should be noted that the number of random steps to the goal from a typical state is extremely large in maze2 and maze3. It is because of irreversible transactions between states in these environments.

The proposed learning algorithm shows very stable ability to form a good policy in each case. However, the experiments have also discovered the drawbacks of the algorithm: slow adaptation and non-optimality of resulting policy. The agent was able to form only a good policy, but not a theoretically optimal one.

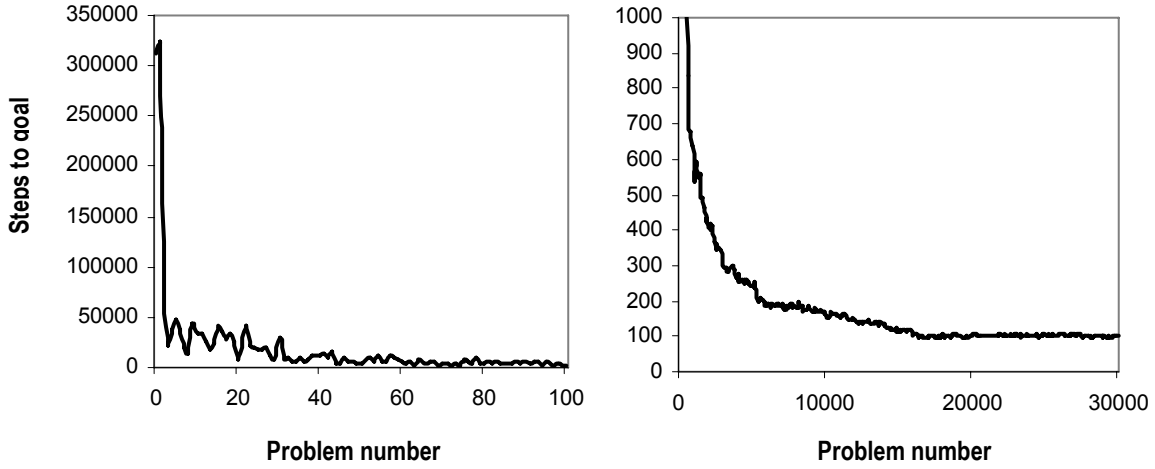


Figure 3. Experimental results in maze2

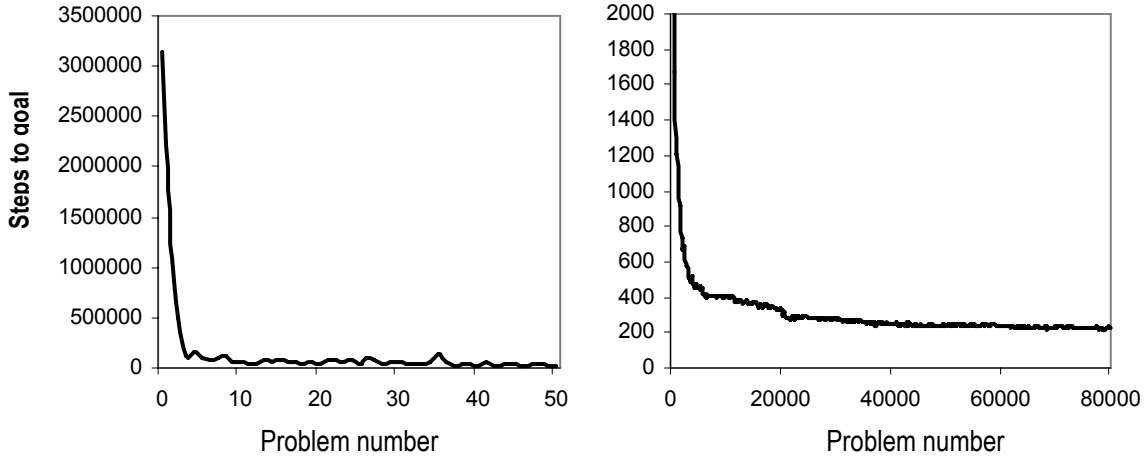


Figure 4. Experimental results in maze3

Table 1. Results summary

		Maze1	Maze2	Maze3
Environment description	The number of perceptions	21	55	89
	The number of positions in the maze	25	56	
	The number of positions in the maze with not unique observations	6	2	32
	The number of states	45	97	208
	The number of belief states	51	98	225
	The number of random steps to the goal from a typical state	$4.1 \cdot 10^4$	$2.3 \cdot 10^9$	$4.3 \cdot 10^{10}$
	The theoretically optimal number of steps to the goal	12	25	29
Experimental results (on average)	The resulting average steps to the goal	39	97	234
	The time of the first goal achievement	$2.2 \cdot 10^4$	$3 \cdot 10^5$	$3.1 \cdot 10^6$
	The total number of solved problems	$2*10^3$	$1.7 \cdot 10^4$	$2 \cdot 10^4$
	The total learning time needed to converge to a good policy	$2.6 \cdot 10^5$	$6.4 \cdot 10^6$	$2 \cdot 10^7$
	During learning the number of steps to the goal was reduced X times	$5.6 \cdot 10^2$	$3 \cdot 10^3$	$1.3 \cdot 10^4$
	The resulting number of steps to the goal is X times bigger than the optimal one	3.25	3.88	8.07

Comparison to Other Algorithms

Developing a new algorithm, it is important to analyze its place in the context of other algorithms serving for similar purposes. For this aim, it was decided to compare the proposed learning algorithms with other reinforcement learning algorithms. In order to make such comparison, several criteria were proposed in Table 1. These criteria were used for the comparison presented in Table 2.

It should be noted that the presented evaluation of algorithms is very rough and could not be considered as a fully proved comparison of different algorithms. However, it can help to describe features of the developed algorithm. The advantages of the proposed algorithm are the simple implementation and its stable ability to form a good policy in the extremely complex case when the environment has irreversible transactions and complex temporal dependencies. However, it has also drawbacks: non-optimality and slow adaptation.

Table 2. The description of criteria

Criterion	**Description**
SI	The implementation is simple.
CT	The computational time taken at each time step is small.
LT	The ability to learn the policy fast.
OPT	The resulting policy is very close to the optimal one.
PO	The ability to overcome the incomplete perception in partly observable environments.
EF	The ability to perform efficient exploration when the environment has many irreversible transactions.
CD	The ability to learn complex temporal dependencies.
PD	The ability to discover perceptual distinctions.
SA	The architecture is not comprehensive or composite.
UL	The number of hierarchy levels is not limited.

Table 3. Comparison to other algorithms

Learning algorithm	**SI**	**CT**	**LT**	**OPT**	**PO**	**EF**	**CD**	**PD**	**SA**	**UL**
Q-learning [Wat89]	+	+	+/-	+	-	-	-	-	+	-
SSS algorithm [SS2000]	+/-	+	-	+/-	+	-	+/-	-	-	+
HQ-learning [WS98]	+/-	+	-	+/-	+	-	+	-	-	-
USM algorithm [McCallum95]	-	-	+	+	+	-	-	+	+	-
USM + "internal reward for hard-to-reach states" [Pch2003]	-	-	+	+/-	+	+	-	+	+	-
The proposed algorithm	+	+	-	-	+	+	+	-	+	+

Conclusion

In this paper a new hierarchical reinforcement learning algorithm was presented that doesn't exploit any domain-specific knowledge about subtask hierarchy, but automatically builds useful hierarchies. The algorithm was developed with purpose to overcome the combination of three, previously known in reinforcement learning, difficulties: (1) incomplete perception, (2) irreversible transactions and (3) complex temporal dependencies. The key idea of the algorithm is to exploit the observation from the environment not only as context information for action selecting, but also as local, internally selected, goals and subgoals. This makes the agent to be hierarchical reinforcement learner, driven by local goals, that has a native ability of efficient exploration.

The proposed learning algorithm was experimentally investigated in different and very complex maze problems, showing very stable ability to form a good policy in each case. However, the experiments have also discovered the drawbacks of the algorithm: slow adaptation and non-optimality of resulting policy (the agent was able to form only a good policy, but not a theoretically optimal one).

Bibliography

[Bong75] Бонгард, М.М., Лосев, И.С., Смирнов М.С. Проект модели организации поведения – Животное // Моделирование обучения и поведения. – М.: Наука, 1975.

[Chr92] Chrisman, L. Reinforcement Learning with Perceptual Aliasing: The Perceptual Distinctions Approach, In Tenth National Conference on Artificial Intelligence, 1992, pp.183-188.

[Die2000] Dietterich, T.G. Hierarchical Reinforcement Learning with the MAXQ Value Function Decomposition, Journal of Artificial Intelligence Research, 13, pp.227-303, 2000.

[KLM96] Kaelbling, L.P., Littman, L.M., Moore, A.W. Reinforcement learning: a survey, Journal of Artificial Intelligence Research, Vol. 4, 1996, pp.237-285.

[Hum96] Humphrys, M. W-learning: a simple RL-based society of mind. Technical report 362, University of Cambridge, Computer Laboratory, 1996.

[Maes95] Maes P., Modeling Adaptive Autonomous Agents, Artificial Life: An overview, MIT Press, 1995.

[McCallum95] McCallum, R.A., Reinforcement learning with selective perception and hidden state (Ph.D. dissertation). Department of Computer Science, University of Rochester, Rochester, NY, 1995.

[McCallum97] McCallum, R.A., "Efficient Exploration in Reinforcement Learning with Hidden State", AAAI Fall Symposium on "Model-directed Autonomous Systems", 1997.

[Mit99] Mitchel, T.H., Machine learning, The McGraw-Hill Companies. Inc., 1999.

[Pch2003] Pchelkin A. Efficient exploration in reinforcement learning based on Utile Suffix Memory, journal "Informatica", Lithuanian Academy of Sciences, Vol.14., 2003.

[PR97] Parr, R., Russul, S. Reinforcement learning with hierarchies of machines. Advances in Neural Information Processing Systems 9. MIT Press, Cambridge, MA.

[Sut98] Sutton, R.S., Barto, A.G. Reinforcement learning: An introduction. Cambridge, MA: MIT Press, 1998.

[SS2000] Sun, R., Sessions, C. Self-segmentation of sequences: automatic formation of hierarchies of sequential behaviors. IEEE Transactions on Systems, Man, and Cybernetics: Part B Cybernetics, 30(3), 2000.

[Thr92a] Thrun, S.B., Efficient exploration in reinforcement learning, Technical Report CMU-CS-92-102, Carnegie Mellon University, Pittsburgh, PA 15213, 1992.

[Thr92b] Thrun, S.B., The role of exploration in learning control. In DA White & DA Sofge, editors, Handbook of Intelligent Control: Neural, Fuzzy and Adaptive Approaches. New York, NY: Van Nostrand Reinhold, 1992.

[Wat89] Watkins, C. Learning with Delayed Rewards. Ph.D Thesis, Cambrige University, Cambrige, UK.

[Whit91] Whitehead, S.D. Complexity and cooperation in Q-learning, In Proceedings of the Eighth International Workshop on Machine Learning, 1991, pp.363-367.

[Wil85] Wilson S.W. Knowledge growth in an artificial animal, Proceeding of the First International Conference on Genetic Algoritms and Their Applications. Hillsdale, New Jersey: Lawrence Erlbaum Associates, 1985, pp.16-23.

[WS98] Wiering, M., Schmidhuber, J. HQ-learning, Adaptive Behavior, vol. 6.2, pp. 219-246, 1998.

Author Information

Arthur Pchelkin – PhD. Student**;** Institute of Information Technology, Technical University of Riga; 1, Kalku St., Riga LV-1658, Latvia; e-mail: arturp@inbox.lv

In memoriam

Bulgarian and world computer science lost a prominent colleague:

Dimitar Petrov Shishkov

22th January 1939, Varna – 8th March 2004, Sofia

D. Shishkov graduated mathematics at Sofia University in 1962. In the last year of his studies he started a specialization as a programmer at the Joint Institute of Nuclear Research – Dubna which lasted three years. Then he worked at the Institute of Mathematics for two years. In 1966 D. Shishkov together with a group of experts transferred to the newly created Central Laboratory for Information Technologies. In 1976 he defended his PhD dissertation. He has been an associate professor in computer science at Sofia University since 1985 and a professor in computer science since 2000.

His scientific interests and results were in the fields of computer architectures, computational linguistics, artificial intelligence, numerical methods, data structures, etc.

He was remarkable with his teaching activities.

D. Shishkov was the creator of high-quality software for the first Bulgarian electronic calculator “ELKA” – one of the first calculators in the world as well as for the series of next calculators and for specialized minicomputers.

He was the initiator of the international project “Computerization of the natural languages”.

He was a member of a range of international scientific organizations. Among his numerous activities was the organization of the I-st Programming competition in 1979.

D. Shishkov was the initiator of sport dancing in Bulgaria (1967) and founder of the first sport-dancing high school education in the world.

D. Shishkov was a highly accomplished person with a diversity of interests, with a developed social responsibility and accuracy in his work.

In 1996 D. Shishkov was awarded with the International Prize ITHEA for outstanding achievements in the field of Information Theories and Applications.

We are grateful to D. Shishkov for the chance to work together with him for establishment and development of IJ ITA.

TABLE OF CONTENTS