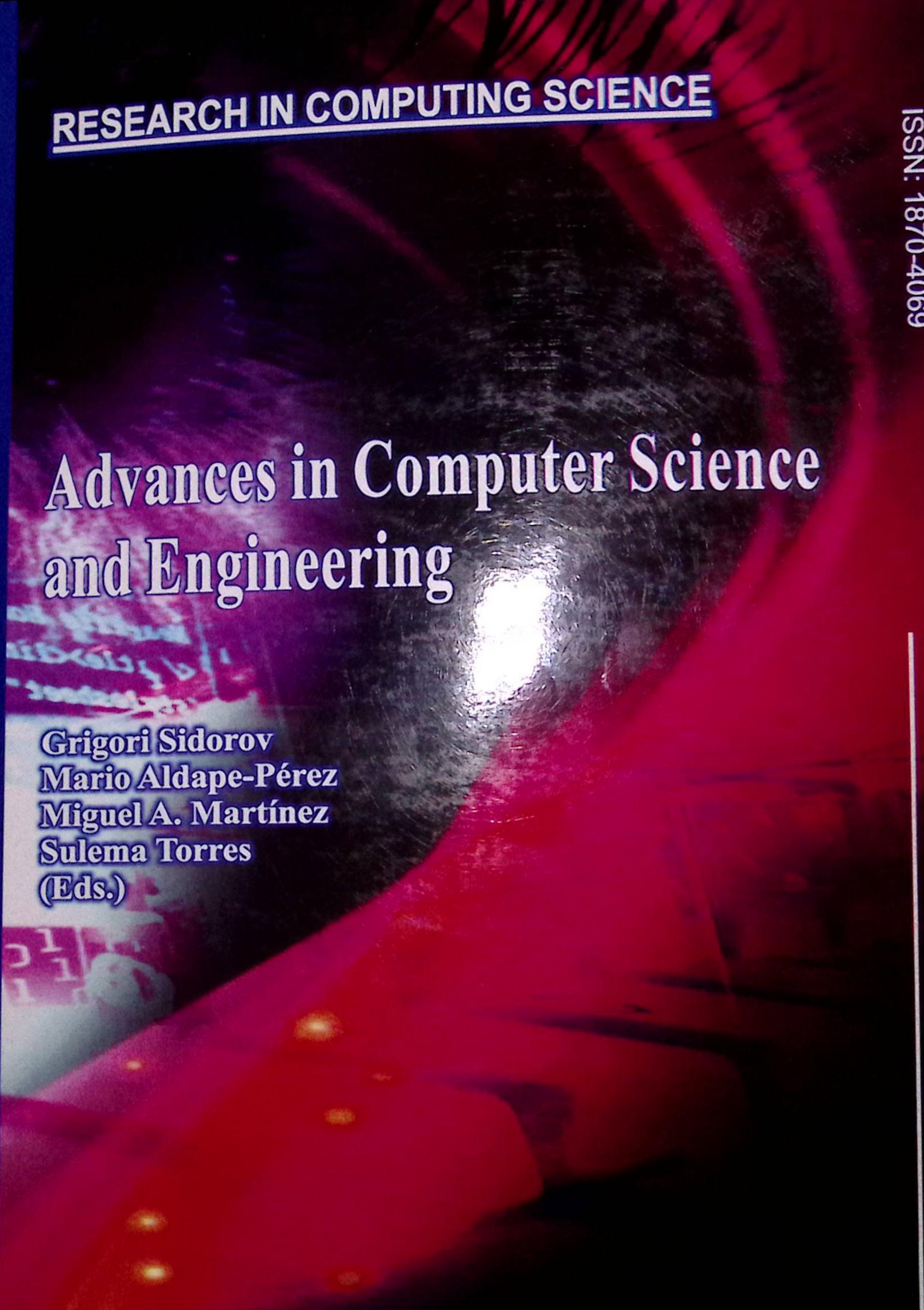
RESEARCH IN COMPUTING SCIENCE

ISSN: 1870-4069



Advances in Computer Science and Engineering

Grigori Sidorov
Mario Aldape-Pérez
Miguel A. Martínez
Sulema Torres
(Eds.)

Vol. 42

RCS
Research in Computing Science

Advances in Computer Science and Engineering

Research in Computing Science

Research in Computing Science es una publicación trimestral, de circulación internacional, editada por el Centro de Investigación en Computación del IPN, para dar a conocer los avances de investigación científica y desarrollo tecnológico de la comunidad científica internacional. **Volumen 42** Mayo, 2009. Tiraje: 500 ejemplares. *Certificado de Reserva de Derechos al Uso Exclusivo del Título* No. 04-2004-062613250000-102, expedido por el Instituto Nacional de Derecho de Autor. *Certificado de Licitud de Título* No. 12897, *Certificado de licitud de Contenido* No. 10470, expedidos por la Comisión Calificadora de Publicaciones y Revistas Ilustradas. El contenido de los artículos es responsabilidad exclusiva de sus respectivos autores. Queda prohibida la reproducción total o parcial, por cualquier medio, sin el permiso expreso del editor, excepto para uso personal o de estudio haciendo cita explícita en la primera página de cada documento. Impreso en la Ciudad de México, en los Talleres Gráficos del IPN – Dirección de Publicaciones, Tres Guerras 27, Centro Histórico, México, D.F. Distribuida por el Centro de Investigación en Computación, Av. Juan de Dios Bátiz S/N, Esq. Av. Miguel Othón de Mendizábal, Col. Nueva Industrial Vallejo, C.P. 07738, México, D.F. Tel. 57 29 60 00, ext. 56571.

Editor Responsable: *Juan Humberto Sossa Azuela, RFC SOAJ560723*

***Research in Computing Science** is published by the Center for Computing Research of IPN. **Volume 42**, May, 2009. Printing 500. The authors are responsible for the contents of their articles. All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without prior permission of Centre for Computing Research. Printed in Mexico City, May, 2009, in the IPN Graphic Workshop – Publication Office.*

Volume 42

Volumen 42

Advances in Computer Science and Engineering

Volume Editors:
Editores del Volumen

Grigori Sidorov
Mario Aldape-Perez
Miguel A. Martinez
Sulema Torres

Instituto Politécnico Nacional
Centro de Investigación en Computación
México 2009

ISSN: 1870-4069

Instituto Politécnico Nacional (IPN)
Centro de Investigación en Computación (CIC)
Av. Juan de Dios Bátiz s/n esq. M. Othón de Mendizábal
Unidad Profesional "Adolfo López Mateos", Zacatenco
07738, México D.F., México

http://www.ipn.mx
http://www.cic.ipn.mx

The editors and the Publisher of this journal have made their best effort in preparing this special issue, but make no warranty of any kind, expressed or implied, with regard to the information contained in this volume.

Indexed in LATINDEX and Periodica / Indexada en LATINDEX y Periodica

Printing: 500 / Tiraje: 500

Printed in Mexico / Impreso en México

Preface

Computer Science is a growing and very important area of modern science related to computer software and hardware design. Its development allows for computers being more and more fast, efficient and intelligent. This volume presents advances of investigation in several selected areas of Computer Science, mainly related to the field of the artificial intelligence.

This issue of the journal Research in Computing Science can be interesting for researchers in computer science, especially in areas related to artificial intelligence, and also for persons who are interested in the cutting edge themes of the computer science.

Papers for this volume were carefully selected by volume editors on the basis of the blind reviewing process performed by editorial board members and additional reviewers. The main criteria for paper selection were their originality and technical quality.

Totally, we received 75 papers by 217 authors from 9 countries that were submitted for evaluation; see Tables 1 and 2. Each submission was reviewed by three independent members of the editorial board of the volume or additional reviewers.

This volume contains revised versions of 31 accepted papers by 97 authors, selected for publication after thorough evaluation process. The acceptance rate is 41.3%. In Table 1, the number of papers per country is presented taking into account all authors of the paper. Table 2 presents the statistics of papers by topics according to the topics indicated by the authors. Note that one paper can correspond to more than one topic.

Table 1. Statistics of submissions and accepted papers by country / region

	Authors		Papers[1]	
Country/Region	Subm	Accp	Subm	Accp
Belgium	1	1	0.33	0.33
China	6	-	2	-
France	8	5	2.87	1.84
Korea, Republic of	2	2	1	1
Mexico	189	85	64.67	26
Spain	4	2	1.5	0.5
Tunisia	5	-	1.3	-
United Kingdom	1	1	0.33	0.33
Venezuela	1	1	1	1
Total:	*217*	*97*	*75*	*31*

[1] It is counted taking into account all authors. For example, for a paper by 3 authors: 2 from Mexico and 1 from UK, we add ⅔ to Mexico and ⅓ to UK.

Table 2 Statistics of submissions and accepted papers by topic

Topic	Submitted	Accepted
Algorithm Theory	4	1
Artificial Intelligence	20	8
Bioinformatics	6	2
Computer Architecture	1	1
Computer Vision	10	3
Control Systems	6	3
Data Mining	7	4
Database Systems	2	-
Digital Signal Processing	8	3
Digital Systems Design	2	2
Distributed Systems	5	2
Evolutionary Algorithms	8	4
Formal Languages	1	1
Fuzzy Logic	3	3
Geoprocessing	4	2
High-Performance Computing	2	-
Information Security	1	-
Knowledge Representation	8	3
Multi-Agent Systems	6	3
Natural Language Processing	8	2
Networks and connectivity	3	3
Neural Networks	4	2
Parallelism	3	-
Real Time systems	4	2
Robotics	7	3
Scientific Computing	6	1
Semantic Web	2	-
Software Engineering	7	1
Virtual Instrumentation	1	1
Other	19	7

The papers are structured into the following nine sections:

- Algorithm Theory,
- Natural Language Processing and Knowledge Representation,
- Pattern Recognition and Data Mining,
- Computer Vision,
- Multi-agent Systems and Simulation,
- Computer Networks,
- Digital Signal Processing,
- Computer Architecture and Digital Systems Design,
- Fuzzy Logic and Control.

This volume is a result of work and effort of many people. In the first place, we thank the authors of the papers included in this volume for the technical excellence of their papers that assures the high quality of this publication.

We also thank the members of the International Editorial Board of the volume and the additional reviewers for their hard work consisting in selection of the best papers out of many submissions that were received.

We would like to thank the personnel of the Center for Research in Computer Science of the National Polytechnic Institute: Virginia Contreras Hernández, Martín Moreno Ramírez, Oralia del Carmen Pérez Orozco and Ignacio García Araoz, for their indispensable help in the process of preparation of the volume.

The submission, reviewing, and selection process was conducted using the freely available system for journal and conference management *EasyChair*, www.EasyChair.org.

May, 2009

Grigori Sidorov
Mario Aldape-Pérez
Miguel A. Martínez
Sulema Torres

Table of Contents
Índice

Page/Pág.

Algorithm Theory

Natural Language Processing and Knowledge Representation

Pattern Recognition and Data Mining

Computer Vision

Multi-agent Systems and Simulation

Computer Networks

Digital Signal Processing

Computer Architecture and Digital Systems Design

Fuzzy Logic and Control

Algorithm Theory

Counting Edge Covers Sets for Estimating the Relevance of Communication Lines

Guillermo De Ita, Meliza Contreras and Pedro Bello

Faculty of Computer Science, Universidad Autónoma de Puebla
{deita,mcontreras,pbello}@cs.buap.mx

Abstract. Counting the number of edge covers on graphs is a #P-complete problem. We design efficient exact methods for computing the number of edge covers for acyclic graphs. Even more, we show that if a graph G does not contain intersectig cycles (any pair of cycles has not common edges) then its number of edge covers can be computed in linear time over the size of the graph G. This determines a border between efficient counting and exponential time procedures for counting the number of edge covers.
We also show how to apply the computing of the number of edge cover for estimating the relevance of the lines in a communication network, which is an important problem in the reliability analysis of a network.
Keywords: Counting Edge Covers, Combinatorial Algorithms, Reliability Analysis Network.

1 Introduction

Counting problems are not only mathematically interesting, but they arise in many applications. For example, if we want to know the probability that a propositional formula is true, or the probability that a graph remains connected given a probability of failure of an edge, we have to count to approximate such probabilities. Counting problems also arise naturally in Artificial Intelligence research. For example, some methods used in reasoning, such as computing 'degree of belief' and 'Bayesian belief networks' are computationally equivalent to counting the number of satisfying assignments to a propositional formula [8, 10]

Counting has become an important area in mathematics as well as in theoretical computer science, although it has received less attention than decision problems. Actually, there are few counting problems in graph theory that can be solved exactly in polynomial time, indeed an important line of research is to determine the class of graphs (or the class of restrictions) for which a counting problem could be solved in polynomial time.

An *edge cover* set of a graph G is a subset of edges covering all nodes of G. The problem of counting the number of edge cover sets of a graph, denoted as #Edge_Covers, is a #P-complete problem via the reduction from #Twice-SAT to #Edge_Covers [1].

Although the computation of #Edge_Covers is a hard problem, it is relevant to recognize the class of instances where this problem becomes an easy problem,

© G. Sidorov, M. Aldape, M. Martínez, S. Torres. (Eds.)
Advances in Computer Science and Engineering.
Research in Computing Science 42, 2009, pp. 3-14

Received 14/03/09
Accepted 27/04/09
Final version 05/05/09

that means, to identify the class of graphs where counting the number of its edge covers can be done in polynomial time. There is a scarce literature about the design of procedures for computing edge covers, and as far as we know, it is not known which is the largest polynomial class of graphs for the #Edge_Covers problem.

We address the computation of #Edge_Covers based on the topological structure of the graph. We focus on determining the topology of the graph G which allows to count the number of edge covers of G in an exactly and efficiently way. We show here that the #Edge_Covers problem can be computed in polynomial time for any acyclic graph.

We show the relevance to compute the number of edge cover for computing the relevance of the lines into a communication network. We show that the computation of the number of edge covers is an important value for estimating the 'strategic' value of each line of a network.

2 Preliminaries

A connected graph is a graph such that there exists a path between all pairs of vertices. If the graph is a directed graph, and there exists a path from each vertex to every other vertex, then it is a strongly connected graph.

A vertex cover of a graph $G = (V, E)$ is a subset $U \subseteq V$ that covers every edge of G; that is, every edge has at least one endpoint in U.

An edge cover, $\mathcal{E}$, for a connected graph $G = (V, E)$ is a subset of edges $\mathcal{E} \subseteq E$ which contains edges covering all vertex of G, that is, for each $u \in V$ there is a $v \in V$ such that $e = \{u, v\} \in \mathcal{E}$.

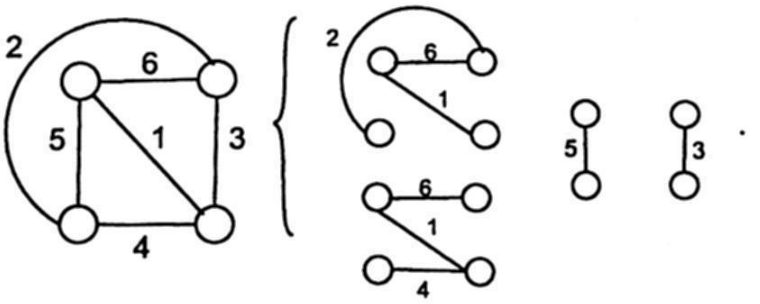


Fig. 1. Cases of edge covers

Given a connected graph $G = (V, E)$, let $C_\epsilon(G) = \{\epsilon \subseteq E : \epsilon$ is an edge cover of $G\}$ be the set of edge-covers sets that a graph G has. Let $NE(G) = |C\epsilon(G)|$ be the number of different edge-covers in a graph, and given any graph G, we denote the problem of computing the number $NE(G)$ as the #Edge_Covers problem.

The #Edge_Covers Counting Problem consists in given a connected graph, find the total number of edge covers of G. For example, in the figure 1 we show three different edge covers sets of a graph.

3 CEC: Counting Edge Covers

The value $NE(G)$ for any graph G, including the case when G is a disconnected graph, is obtained as: $NE(G) = \prod_{i=1}^{k} NE(G_i)$, where $G_i, i = 1, \ldots, k$ is the set of connected components of G. We should first determine the set of connected components of G, and this procedure can be done in linear time. Then, the time complexity of computing $NE(G)$ depends on the maximum time complexity of its connected components. Thus, we would consider just the different kinds of connected components in G, so from now on, when we mention a graph G we suppose that it consists of just one connected component.

We call *fixed edges* to the edges of a graph G that appear in all edge cover set of G. When an edge cover $\mathcal{E}$ of a graph is being built, we distinguish between two different states of a node; we say that a node u is *free* when it has not still been covered by any edge of $\mathcal{E}$, while if the node has already been covered we say that the node is *cover*.

In order to present the basic procedures for counting edge covers, we consider first the case when the graph is an acyclic graph, and we start analyzing the most simple topology of an acyclic graph. An initial procedure for counting edge covers is to apply a depth first search over the input graph. A basic a recursive procedure scheme for the depth-first search, is the following.

Algorithm 1 Procedure $dfs(G, v)$

```
Mark $v$ as discovered
for each vertex $w \in N(v)$ do
  if ($w$ is undiscovered) then
    $dfs(G, w)$
  end if
  Mark $v$ as finished
end for
```

Case A: Counting Edge Covers on Paths

Let $P_n = G = (V, E)$ be a path graph. We assume an order between vertices and edges in P_n, i.e. let $V = \{v_0, v_1, \ldots, v_n\}$ be the set of $n+1$ vertices and let $e_i = \{v_{i-1}, v_i\}$, $1 \leq i \leq n$ be the n edges of P_n.

Let $G_i = (V_i, E_i)$, $i = 0, \ldots, n$ be the subgraphs induced by the first i nodes of V, i.e. $G_0 = (\{v_0\}, \emptyset), G_1 = (\{v_0, v_1\}, \{e_1\}), G_2 = (\{v_0, v_1, v_2\}, \{e_1, e_2\}), \ldots, G_n = P_n = (V, E)$. $G_i, i = 0, \ldots, n$ is the family of induced subgraphs of G formed by the first i nodes of V. Let $\mathcal{CE}(G_i) = \{\mathcal{E} \subseteq E_i : \mathcal{E}$ is an edge cover of $G_i\}$ be the set of edge covers of each subgraph G_i, $i = 0, \ldots, n$.

We want to count the edge cover sets of P_n by considering the different ways of covering each node on the path. We travel by P_n in depth-first search and when a node is being visited, we consider the different ways to cover it. A path P_n has two special edges: e_1 and e_n which are fixed edges and for $i = 1, \ldots, n-1$, $\delta(v_i) = 2$, e_{i-1} and e_i are the two incident edges of v_i.

We associate with each edge $e_i, i = 1, \dots, n$ in the path, an ordered pair: (α_i, β_i) of integer numbers where α_i expresses the number of edge cover sets in $\mathcal{CE}(G_i)$ where the edge e_i appears in order to cover the node v_{i-1}, while β_i conveys the number of edge cover sets in $\mathcal{CE}(G_i)$ where the edge e_i does not appear, since v_{i-1} has been already covered (perhaps by the edge e_{i-1}).

Traversing by P_n in depth-first search, each pair (α_i, β_i) is computed in accordance with the type of edge $e_i, i = 1, \dots, n$ which is being visited. At the end of the search, the last pair (α_n, β_n) is computed, and such pair gives the value for $NE(P_n) = \alpha_n + \beta_n$.

The pair (1,0) is assigned to (α_1, β_1) since the edge e_1 is a fixed edge and e_1 has to appear in all edge cover of P_n. In general, any fixed edge e_p which starts a series (α_p, β_p) has $(1, 0)$ as initial pair.

If we know the pair $(\alpha_{i-1}, \beta_{i-1})$ for any $i < n$, then we know the number of times where the edge e_{i-1} appears or does not appear into the set of edge covers of G_{i-1}. When the edge e_i is being visited, the vertex v_{i-1} has to be covered considering for this its two incident edges: e_{i-1} and e_i. Any edge cover of $\mathcal{CE}(G_{i-1})$ containing the edge e_{i-1} (α_{i-1} cases) has already covered v_{i-1} and then the ocurrence of e_i is optional. But for the edge covers where e_{i-1} does not appear (β_{i-1} cases) the edge e_i must appear in order to cover to v_{i-1}. Then, the number of edge covers where e_i appears is $\alpha_{i-1} + \beta_{i-1}$ and just in α_{i-1} edge covers the edge e_i does not appear. Thus, we obtain the new pair (α_i, β_i) associated with the edge e_i, by applying the Fibonacci recurrence relation.

$$\alpha_i = \alpha_{i-1} + \beta_{i-1}; \quad \beta_i = \alpha_{i-1} \tag{1}$$

When the search arrives to the last edge e_n of the path, we have obtained the pair $(\alpha_{n-1}, \beta_{n-1})$ and since e_n is a fixed edge, then it has to appear in all edge covers of P_n, that means, $\alpha_n = \alpha_{n-1} + \beta_{n-1}$ and since e_n has not chance of not appearing in any edge cover of P_n then $\beta_n = 0$. We call *recurrence for processing fixed edges* to the recurrence:

$$\alpha_i = \alpha_{i-1} + \beta_{i-1}; \quad \beta_i = 0 \tag{2}$$

In this way, the pair associated with the last edge on a path is: $(\alpha_n, \beta_n) = (\alpha_{n-1} + \beta_{n-1}, 0)$. The series $(\alpha_i, \beta_i), i = 1, ..., n$ built based on the recurrences (1) and (2) allows to compute $NE(P_n)$ in linear time over the number of edges in P_n in accordance with the traversing of the path in depth-first search.

We call a *computing thread* or just a *thread* to a series of pairs $(\alpha_i, \beta_i), i = 1, ..., n$ used for computing the number of edge covers on a trajectory of n distinct and adjacent edges. The computation of each pair $(\alpha_i, \beta_i), i = 1, ..., n$ is done in accordance with the type of edge e_i on the trajectory. Notice that a trajectory could be a subgraph of a more complex graph.

In a path P_n, all nodes have degree 2 or 1. But, when the current edge e_i as well as the previous one e_{i-1} considered in a computing thread are incident to a node v_i which has been already covered or for which $\delta(v_i) > 2$, then the edge e_i can be taken into account or not since v_i has been covered previously, and then the pair (α_i, β_i) is computed from $(\alpha_{i-1}, \beta_{i-1})$ by the recurrence relation:

$$\alpha_i = \alpha_{i-1} + \beta_{i-1}; \quad \beta_i = \alpha_{i-1} + \beta_{i-1} \tag{3}$$

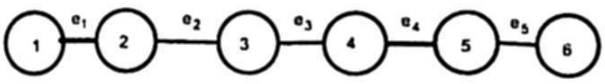
In the following examples, we denote with → the application of recurrence (1), with $\xrightarrow{o}$ the application of recurrence (3) and with ↦ the processing of fixed edges - recurrence (2).

The sequence: $0, 1, 1, 2, 3, 5, 8, 13, 21, 34, ...$, in which each number is the sum of the preceding two, is denoted as the Fibonacci series. The numbers in the sequence, known as the Fibonacci numbers, will be denoted by F_i and we formally define them as: $F_0 = 0$; $F_1 = 1$; $F_{i+2} = F_{i+1} + F_i, i \geq 0$. Each Fibonacci number can be bounded from above and from below by $\phi^{i-2} \geq F_i \geq \phi^{i-1}, i \geq 1$, where $\phi = \frac{1}{2} \cdot (1 + \sqrt{(5)})$ is known as the 'golden ratio'.

Example 1 *Let us consider the path P_5 (see figure 2). The computing thread for P_5 is: $(\alpha_i, \beta_i), i = 1, \dots, 5 : (1, 0) \rightarrow (1, 1) \rightarrow (2, 1) \rightarrow (3, 2) \mapsto (5, 0)$. Then $NE(P_5) = 5 + 0 = 5$.*

Theorem 1 *The number of edge cover sets of a path of n edges, is:*
$F_n = \texttt{ClosestInteger}\left[\frac{1}{\sqrt{5}}\left(\frac{1+\sqrt{5}}{2}\right)^n\right]$.

Proof. The thread $(\alpha_i, \beta_i), i = 1, \dots, n$ used for computing $NE(P_n)$, coincides with the Fibonacci numbers: $(F_1, F_0) \rightarrow (F_2, F_1) \rightarrow (F_3, F_2) \rightarrow (F_4, F_3) \dots \rightarrow (F_{n-1}, F_{n-2}) \mapsto (F_n, 0)$. Then, we infer that $(\alpha_i, \beta_i) = (F_i, F_{i-1})$ for $i = 1, \dots, n-1$ and $\alpha_n = F_n, \beta_n = 0$. And then $NE(P_n) = \alpha_n + \beta_n = F_n$.



$Edges: \quad e_1 \quad e_2 \quad e_3 \quad e_4 \quad e_5$
$(\alpha_i, \beta_i) : (1, 0) \rightarrow (1, 1) \rightarrow (2, 1) \rightarrow (3, 2) \mapsto (5, 0) = 5$

Fig. 2. Counting edge covers on a path

Case B: Counting Edge Covers on Trees

Let $T = (V, E)$ be a rooted tree. We distinguish each edge on the tree as follows: *root-edges*, which are the edges with one endpoint in the root node; *leaf-edges*, which are the edges with one endpoint in a leaf node of T. Given any intermediate node v of T, we call a *child-edge* of v to the edge connecting v with any of its children nodes, and the edge connecting v with its father node is called the *father-edge* of v. $NE(T)$ is computed traversing by T in post-oder and associating a pair (α_e, β_e) with each edge e of T, except for the leaf edges.

Algorithm #Edge_Covers_in_trees(T)
Input: A rooted tree T with root vertex v_r
Output: $NE(T)$
Procedure:

1. We reduce the input tree T to other tree T' by cutting all leaf nodes and leaf-edges from T, and by labeling as covered nodes all father nodes of the original leaf nodes of T.
2. Traversing by T' in post-order, a pair (α_e, β_e) is associated with each edge e in T'. Each pair is computed in the following way:
   (a) $(\alpha_e, \beta_e) = (1, 1)$ if e is a leaf-edge of T'.
   (b) when an internal node v is visited and it has a set of child-edges, e.g. $u_1, u_2, ..., u_k$ are the child-edges of v, as we have already visited all child-edge of v then each pair $(\alpha_{u_j}, \beta_{u_j})$, $j = 1, \ldots, k$ has been computed and associated to the child-edges. The set of child-edges of v is considered as just one child-edge e_u and its associated pair (α_u, β_u) is computed, as:

$$\alpha_u = \prod_{j=1}^{k}(\alpha_{u_j} + \beta_{u_j}) - \prod_{j=1}^{k}\beta_{u_j}\ ; \quad \beta_u = \prod_{j=1}^{k}\beta_{u_j} \tag{4}$$

   where α_u carries the number of different combinations of the child-edges of v for covering v, while β_u gives the number of combinations among the child-edges of v which do not cover to v. The case in which v has just one child-edge is consider in this case, with $\alpha_u = \alpha_{u_1}$ and $\beta_u = \beta_{u_1}$. The pair associated to the father-edge e_v of v is computed as follows:

$$(\alpha_v, \beta_v) = \begin{cases} (\alpha_u + \beta_u, \alpha_u) & \text{if } v \text{ is a free node or,} \\ (\alpha_u + \beta_u, \alpha_u + \beta_u) & \text{if } v \text{ is a cover node} \end{cases}$$

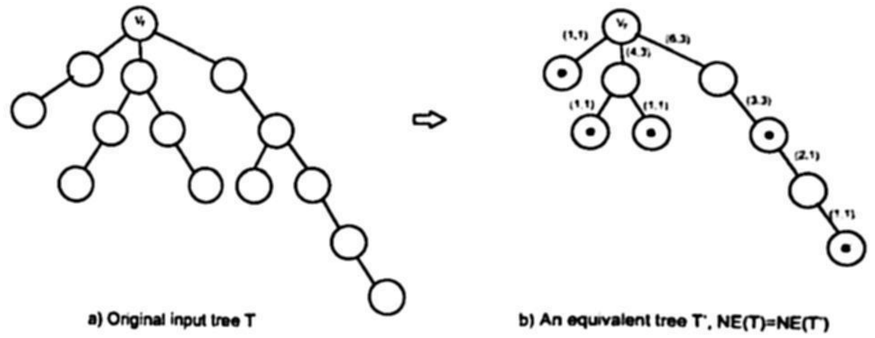
   This step (2) is iterated until it computes the pairs (α_e, β_e) for all edge e of T' and it stops when it arrives to the root node v_r.
3. Let $(\alpha_{u_r}, \beta_{u_r})$ be the pair associated with the root-edge of v_r. $NE(T)$ is computed according of the status of v_r; if v_r is a covered node then $NE(T) = \alpha_{u_r} + \beta_{u_r}$, otherwise $NE(T) = \alpha_{u_r}$.

This procedure returns $NE(T)$ in time $O(n+m)$ which is the necessary time for traversing T in post-order. Notice that the post-order search allows to give an evaluating order for each edge of T and at the same time, to compute the number of edge covers. We denote with $\succ$ the application of recurrence (4).

Example 2 *Let T be an input tree with root vertex v_r (see fig. 3a). T' is the reduced tree from T where its covered nodes are marked by a black point inside of the nodes (see fig. 3b). T' is traversing in post-order and each pair (α_e, β_e) is associated with each edge e of the tree. The pairs for the child-edges of v_r, are: (1,1). (4,3) and (6,3). Those three edges are combined in only one edge e_r by applying recurrence (4), the resulting pair (α_r, β_r) was computed as: $\alpha_r = (1+1)*(4+3)*(6+3) - 1*3*3 = 117$ and $\beta_r = 1*3*3 = 9$. Since v_r is the root node and it is free, then $NE(T) = \alpha_r = 117$.*

Case C: Counting Edge Covers on Simple Cycles

Let $C_n = (V, E)$ be a simple cycle with n edges. Let us order the nodes and edges of C_n, as: $V = \{v_1, \ldots, v_n\}$ and $E = \{e_1, \ldots, e_n\}, e_i = \{v_i, v_{i+1}\}, i = 1, \ldots, n-1$, $e_n = \{v_n, v_1\}$.



a) Original input tree T

b) An equivalent tree T', NE(T)=NE(T')

Fig. 3. Computing the number of edge covers for a tree

A computing thread L_p: $(\alpha_i, \beta_i), i = 1, \ldots, n$ is used for counting the edge covers of the path P_n contained in C_n. A depth-first search starts in the edge e_1 and the pair $(\alpha_1, \beta_1) = (1, 1)$ is associated with it, since e_1 is not a fixed edge.

Since all nodes in C_n have degree two, when a new edge is visited during the depth-first search the Fibonacci recurrence (1) is applied. Then, after n applications of recurrence (1), the pair $(\alpha_n, \beta_n) = (F_{n+1}, F_n)$, F_i being the i-th Fibonacci number, is obtained. Let $\mathcal{CE}(NC_n) = \{\mathcal{E} \subseteq E : \mathcal{E}$ is one of the edge sets counted through the thread $Lp\}$, then $NC_n = \mid \mathcal{CE}(NC_n) \mid = \alpha_n + \beta_n = F_{n+2}$.

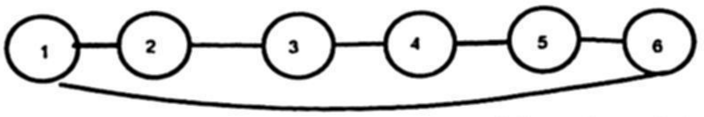
There are some edge sets $\mathcal{E} \in \mathcal{CE}(NC_n)$ where the edges e_1 and e_n do not appear, since the computation of the thread Lp allow these cases, although when the values $\beta_1 = 1$ and $\beta_n > 0$ represent not edge covers for C_n because they do not cover the node v_1. Then, in order to count only the edge covers of C_n, we have to substract from NC_n the edge sets that do not cover v_1.

Let $Y = \{\mathcal{E} \in \mathcal{CE}(NC_n) : e_1 \notin \mathcal{E} \wedge e_n \notin \mathcal{E}\}$ be the edge sets which cover all nodes of C_n except v_1, then $NE(C_n) = NC_n - |Y|$. In order to compute $|Y|$ a new computing thread $(\alpha'_i, \beta'_i), i = 1, ..., n$, denoted by C'_n, is built. C'_n starts with the pair $(\alpha'_1, \beta'_1) = (0, 1)$ since in this way we consider the edge sets where e_1 does not appear. After n applications of recurrence (1), we obtain as last pair of C'_n, $(\alpha'_n, \beta'_n) = (F_{n-1}, F_{n-2})$. For considering only the edge sets where neither e_1 nor e_n appear, (α'_n, β'_n) is taken as $(0, \beta'_n) = (0, F_{n-2})$, and $\mid Y \mid = F_{n-2}$. Then, $NE(C_n) = NC_n - |Y| = F_{n+2} - F_{n-2}$ and we deduce the following theorem.

Theorem 2 *The number of edge cover sets of a simple cycle C_n with n edges, expressed in terms of Fibonacci numbers, is: $NE(C_n) = F_{n+2} - F_{n-2}$.*

In the following examples, we denote with ' $\frown$ ' the binary operation between pairs: (α_n, β_n) and (α'_n, β'_n) - the final pairs of the two computing threads of a cycle - whose result is the pair: $(\alpha_n, \beta_n - \beta'_n)$ and which has to be assocciated with the last edge e_n of a cycle C_n. Notice that the computation of $NE(C_n)$ has a time complexity of $O(n)$ since we compute the two threads: Lp and C'_n in parallel while the depth-first search is applied.

Example 3 *Let C_6 be the simple cycle illustrated in figure 4. Applying theorem (2), we have that $NE(C_6) = F_{6+2} - F_{6-2} = F_8 - F_4 = 21 - 3 = 18$.*



$$\begin{array}{rcccccccccccc}
 & (\alpha_1, \beta_1) & \rightarrow & (\alpha_2, \beta_2) & \rightarrow & (\alpha_3, \beta_3) & \rightarrow & (\alpha_4, \beta_4) & \rightarrow & (\alpha_5, \beta_5) & \rightarrow & (\alpha_6, \beta_6) \\
Lp: & (1,1) & \rightarrow & (2,1) & \rightarrow & (3,2) & \rightarrow & (5,3) & \rightarrow & (8,5) & \rightarrow & (13,8) \\
C'_6: & (0,1) & \rightarrow & (1,0) & \rightarrow & (1,1) & \rightarrow & (2,1) & \rightarrow & (3,2) & \rightarrow & (5,3)
\end{array}$$

$$\Rightarrow (13,8) \curvearrowleft (5,3) = (13,5)$$

Fig. 4. Obtaining the number of edges covers of a cycle

4 Counting Edge Cover Sets for General Graphs

Let $G = (V, E)$ be a connected graph with $|V| = n$, $|E| = m$ and such that $\Delta(G) \geq 2$. Choose a node $v_r \in V$ (e.g. the node with minimum degree in V) for starting a depth-first search over G in order to build a spanning tree T_G where v_r is the root node.

The edges in T_G are called *tree edges*. While the edges in $(E - E(T_G))$ are called *back edges*. Given a back edge e, the union of the path in T_G between the endpoints of e and the same edge e forms a simple cycle, such cycle is called a *fundamental cycle* of G with respect to T_G. Then, each back edge embraces the maximum path contained in a fundamental cycle. Let $\mathcal{C} = \{C_1, \ldots, C_k\}$ be the set of fundamental cycles found during the depth first search of G. Given any pair of fundamental cycles C_i and C_j from $\mathcal{C}$, if C_i and C_j share edges, we call them *intersecting* cycles; otherwise, they are called *independent* cycles.

We call *critical point* of the graph G to each incident node v_p to a back edge. Given a critical point v_p of G the set of its incident edges $E_p \subset E$ is called a *critical edge set*. Before presenting our most general counting algorithm, let us show the method used for processing critical edge sets of a graph.

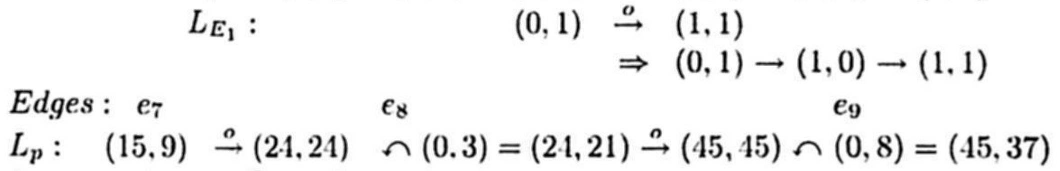
Case D: Processing Critical edge sets of a Graph
A main computing thread $Lp : (\alpha_i, \beta_i)$, $i = 1, \ldots, m$, is used for counting the edge cover sets of an input graph G. When an edge $e_i \in G$ is visited for the first time during the depth-first search (called the current edge), its associated pair (α_i, β_i) is computed according to the corresponding recurrence with the type of edge that e_i is in G. However, the series L_p counts all subsets of a critical edge set S of G including the case when all edges of S do not appear. Although for this latter case (for the empty subset of S) the critical point associated with S is not covered.

Then, when a critical point v_p is visited for fist time, we have to open a new computing thread in order to count the number of subsets considered by L_p and where all edges of E_p do not appear, and such number has to be subtracted from the current value of L_p. Notice that this case is a generalization of the processing of a back edge for simple cycles. The computing thread L_p is always active until the counting process finishes, while the subordinated threads are active until all edges of their corresponding critical edge set have been visited.

If a new critical point u_p is visited before visiting all edges of a previous critical edge set E_p, then some new auxiliary threads are created in order to count the number of subsets counted by the current computing threads and for which all edges of the new critical edge set E_u do not appear. In fact, for each active thread l_i (called a main thread of E_u) one auxiliary thread subordinated to l_i is created with initial pair $(0, \beta_u^{l_i})$ being $\beta_u^{l_i}$ equals to β_u in its respective main thread l_i.

When all edges of a critical edge set E_p had already visited, then its associated critical point v_p has been processed and the binary operator $\curvearrowleft$ is applied between the main and its corresponding subordinated thread. After applying $\curvearrowleft$, the subordinated threads created for proccesing E_p are closed and they stop to be active, keeping only the main threads of E_p.

We illustrate in figure 5, how to process critical edge sets. The initial graph G is formed by a cycle union a path on each endpoint of the back edge. The method used for processing the two critical edge sets of G is representative of the way of processing critical edge sets. For this graph, e_8 is the unique back edge. The two critical points in G are v_4 and v_8 which are the endpoints of the back edge. The two critical edge sets are: $E_1 = \{e_3, e_4, e_8\}$ and $E_2 = \{e_7, e_8, e_9\}$.



$$\begin{array}{llllllll}
Edges: & e_1 & e_2 & e_3 & & e_4 & e_5 & e_6 \\
L_p: & (1,0) \rightarrow & (1,1) \rightarrow & (2,1) & \xrightarrow{o} & (3,3) \rightarrow & (6,3) \rightarrow & (9,6) \\
L_{E_1}: & & & (0,1) & \xrightarrow{o} & (1,1) & & \\
 & & & & \Rightarrow & (0,1) \rightarrow & (1,0) \rightarrow & (1,1)
\end{array}$$

$$\begin{array}{lllllll}
Edges: & e_7 & & e_8 & e_9 & & e_{10} \\
L_p: & (15,9) & \xrightarrow{o} (24,24) & \curvearrowleft (0,3) = (24,21) & \xrightarrow{o} (45,45) & \curvearrowleft (0,8) = (45,37) & \longmapsto (82,0) \\
L_{E_1}: & (2,1) & \xrightarrow{o} (3,3) & \Rightarrow (0,3) & & & \\
e_7 \rightarrow L_p: & (0,9) & \xrightarrow{o} (9,9) & & & & \\
 & & \Rightarrow (0,9) & \curvearrowleft (0,1) = (0,8) & \xrightarrow{o} (8,8) \Rightarrow & (0,8) & \\
e_7 \rightarrow L_{E_1}: & (0,1) & \xrightarrow{o} (1,1) & \Rightarrow (0,1) & & &
\end{array}$$

Fig. 5. Computing NE(G) on a cycle combined with a path

The main thread L_p starts with $(\alpha_1, \beta_1) = (1, 0)$ since e_1 is a fixed edge. The recurrence (1) is applied for computing (α_2, β_2) and (α_3, β_3) since v_2 and v_3 have degree 2. Since e_3 is member of the critical edge set E_1, we have to count the edge sets not containing E_1 and which are being considered by L_p. Then, the auxiliary thread L_{E_1} is created and it is subordinated to L_p. The initial pair for L_{E_1} is $(\alpha'_3, \beta'_3) = (0, \beta_3) = (0, 1)$ since $\beta_3 = 1$ in L_p.

When e_4 is visited, recurrence (3) is applied over each active threads: L_p and L_{E_1}, since $\delta(v_4) > 2$. Although, as we want to count in L_{E_1} only the edge sets where e_3, e_4 (and in advance e_8) do not appear, then (α'_4, β'_4) is changed to $(0, \beta'_4)$, and in this case $(\alpha'_4, \beta'_4) = (0, 1)$.

Next, recurrence (1) is applied until arriving to edge e_7 since the nodes v_5, v_6 and v_7 have degree two. As e_7 is member of the critical edge set E_2, we must count the edge sets not containing E_2 and which are considered by the active threads. Then, two new auxiliary threads are created. The thread denoted by $e_7 \rightarrow L_p$ which is subordinated to L_p and $e_7 \rightarrow L_{E_1}$ which is subordinated to L_{E_1}. Their corresponding starting pairs are: $(0, 9)$ and (0,1) since 9 and 1 are the edge sets counted by L_p and L_{E_1} respectively, where e_7 does not appear.

When e_8 is visited, the recurrence (3) is applied over each active computing thread since $\delta(v_8) > 2$. Since all edges of E_1 have been visited, meaning that the first critical point of G has been processed, the operator $\curvearrowleft$ is applied between L_p and L_{E_1}, as well as between $e_7 \rightarrow L_p$ with $e_7 \rightarrow L_{E_1}$. After to apply the operator $\curvearrowleft$, the threads L_{E_1} and $e_7 \rightarrow L_{E_1}$ are closed and they stop to be active.

As the thread $e_7 \rightarrow L_p$ has been used for counting the sets not containing the edges of E_2, then its corresponding pair (α'_8, β'_8) is changed to $(0, \beta'_8)$. When e_9 is visited, the recurrence (3) as well as the operator $\curvearrowleft$ are applied between the two active threads since all edges of E_2 have already been visited, remaining only the main thread L_p. Finally, when e_{10} is visited, the recurrence (2) is applied and the last pair is $(\alpha_{10}, \beta_{10}) = (82, 0)$. And then, $NE(G) = \alpha_{10} + \beta_{10} = 82$.

One of the main techniques for counting objects on a graph has been the Markov chain Monte Carlo method [1, 2, 5, 3]. Although it is likely that this approximation technique provides efficient algorithms only for graphs with bounded maximum degree. For example, the Markov chain Monte Carlo procedures for counting the number of independent sets of a graph is likely to fail for graphs of maximum degree six or higher [5]. In our case, we are presenting deterministic and exact procedures for counting the number of edge covers of a graph based on the topological structure of the graph and without consideration of its maximum degree. Note that the basic procedures (A), (B) and (C) allow us to compute the number of edge cover sets for any graph with maximum degree 2.

5 Estimating the Relevance of Communication Lines

Complex networks, modeled as large graphs, received much attention during these last years. However, topological information on these networks is only available through intricate measurement procedures. Until recently, most studies assumed that these procedures eventually lead to samples large enough to be representative of the whole, at least concerning some key properties. An important application of counting edge covers is for estimating the degree of reliability in communication networks [10].

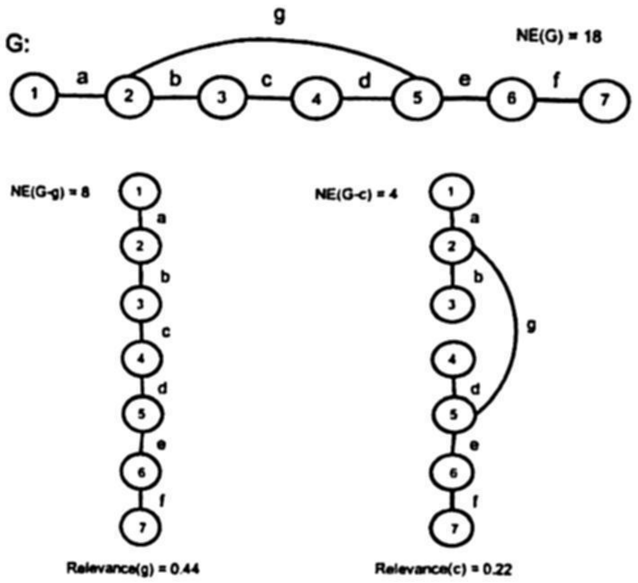


Fig. 6. Estimating the relevance of lines g and c

Table 1. Estimating the relevance of the lines of G

Without line	Edge covers	Relevance
a	15	0.17
b	7	0.61
c	4	0.77
d	6	0.66
e	8	0.55
f	10	0.44
g	8	0.55

For example, if we assume that the communication lines (edges) in a network G has the same 'failure probability' and those failures are independent of one another, we can measure different classes of reliability of the network, given that an edge $c \in G$ fails, according to what component of the network is considered. A way to estimate the 'relevance' of a line c in the network G is by applying the conditional probability $P_{c/G}$ which can be approximated by the fraction of the number of edge covers which are substracted when the edge c is removed (fails), that is, $P_{c/G} = 1 - \frac{NE(G-c)}{NE(G)}$. Thus, $P_{c/G}$ gives the strategic value of an edge c in a network G by estimating the relevance of the line. As c is any edge of G then $P_{c/G}$ could be used for estimating the relevance for any edge of G.

Then, the measure $P_{c/G}$ could be used for estimating the strategic value of any edge c of a graph G with respect to the other lines in G. In such a way that for greater values of $P_{c/G}$ means that the line c is most relevant for maintaining the connectivity of G, in case of failures, with respect to the other lines in G.

6 Conclusions

We determine different recurrence relations for counting, in incremental way, the number of edge cover sets of a graph according with each node and edge of the graph is being visited for the first time during a depth-first search. If the input graph has simple topologies, like: paths, trees, simple cycles or combination of the previous topologies, we can compute the number of edge covers in linear time in the size of the graph.

The class of graphs for which our novel efficient algorithms compute their number of edge cover sets determines a class of polynomial instances for counting the number of edge covers, and such class is a superclass of graphs of degree two without restrictions on the degree of the graphs, but rather, it depends on the topological structure of the graphs.

To know how to compute the number of edge covers is helpful for estimating the reliability of a communication network. For example, we have shown how to estimate the 'relevance' of any line c of the network based on the proportion of the number of edge covers where c does not appear with respect to the total number of edge covers in the network.

References

1. Bubley R., Dyer M., Graph Orientations with No Sink and an Approximation for a Hard Case of #SAT, *Proc. of the Eight Annual ACM-SIAM Symp. on Discrete Algorithms*, 1997, pp. 248-257.
2. Bubley R., Dyer M., Greenhill C., Jerrum M., On approximately counting colourings of small degree graphs, *SIAM Jour. on Computing*, 29, (1999), pp. 387-400.
3. Bubley R., Randomized Algorithms: Approximation, Generation, and Counting, Distinguished dissertations Springer, 2001.
4. Darwiche Adnan, On the Tractability of Counting Theory Models and its Application to Belief Revision and Truth Maintence, *Jour. of Applied Non-classical Logics*,11 (1-2).(2001), 11-34.
5. Dyer M., Greenhill C., Some #P-completeness Proofs for Colourings and Independent Sets, Research Report Series, University of Leeds, 1997.
6. Garey M., Johnson D., Computers and Intractability a Guide to the Theory of NP-Completeness, W.H. Freeman and Co., 1979.
7. Greenhill Catherine, The complexity of counting colourings and independent sets in sparse graphs and hypergraphs", *Computational Complexity*, 9(1): 52-72, 2000.
8. Roth D., On the hardness of approximate reasoning, *Artificial Intelligence 82*, (1996), pp. 273-302.
9. Tarjan R., Depth-First Search and Linear Graph Algorithms, *SIAM Journal on Computing*, Vol. 1, pp.146-160, 1972.
10. Vadhan Salil P., The Complexity of Counting in Sparse, Regular, and Planar Graphs, *SIAM Journal on Computing*, Vol. 31, No.2, pp. 398-427, 2001.

Natural Language Processing and Knowledge Representation

Controlling an SMS Transcription System using Heuristic and Empirical Criteria

Grégory Smits[1] and Christine Chardenon[2]

[1]GREYC-University of Caen, Caen France
greg.smits@gmail.com
[2]Orange Labs, Lannion France
christine.chardenon@orange-ftgroup.com

Abstract. Analysis modules which compose a linguistic process often have to cope with the problem of concurrent results generation. Control strategies aim at identifying the most relevant results among all generated ones. Using a generic control approach based on a multicriteria decision aid method, this article presents how empirical and heuristic criteria are combined to improve a SMS transcription system.

Keywords. Natural language processing, multicriteria decision aid, control strategy, sms transcription

1 Introduction

The use of communication devices like mobile phones or computers has lead to the emergence of new means of written spontaneous communication. Simple Message Service (SMS) or even forum on Internet have contributed to the development of a "SMS language". The transcription of text written in "SMS language" into a "standard" language like French is an important issue especially for application like text vocalisation or indexing.

TiLT [1] is a "generic" Natural Language Processing (NLP) toolbox that has been developed to answer different applicative needs, and which has already been applied to various tasks like query correction and indexation, coreference resolution, translation, abridging, ... This NLP system is based on the sequential application of analysis modules which are associated to linguistic resources.

Recently, this toolbox has been adapted to perform SMS transcription for French. This particular use of TiLT has brought concern on a recurrent problem that affects most of the NLP system: concurrent and erroneous results generation and propagation. Indeed, due to imprecisions in the linguistic resources, the inherent ambiguity of natural languages and the lack of complementarity of modular and sequential processes, indeterminations appear at different steps of the analysis process. These indeterminations are characterized by the generation of concurrent results. Some of these indeterminations are legitimate, when dealing with "natural" ambiguities or when decisive knowledge is not yet available

Received 08/03/09
Accepted 27/04/09
Final version 06/05/09

for the concerned stage of the analysis process, but most of them corresponds to incorrect interpretations.

To obtain valid final interpretations, it is necessary to control the respective relevance of the generated results using specific strategies. The goal of control strategies is to favour the most relevant results among all generated ones or symmetrically filter incorrect results.

Relying on theoretical works about knowledge base systems' control [2], [3, p. 26-43] has shown that the control of a NLP system can be considered as a decisional process where multiple heterogeneous comparison criteria have to be aggregated. This decisional formalisation of the control has conducted to an intersection between NLP and the MultiCriteria Decision Aid (MCDA) domain and more precisely outranking approaches, which propose an efficient and adapted methodology. Thus, a module dedicated to the results control based on an outranking approach has been developed and integrated as a central element of TiLT [3, p. 69-88].

This article is focused on the application of this outranking control strategy on the SMS transcription process.

Section 2 introduces the SMS transcription process and the problem of concurrent results generation. Section 3 presents the outranking control approach proposed by [3, p. 69-88] and the underlying module of control that has been integrated in TiLT's architecture. Section 4 describes the control strategy that has been defined for this particular case of SMS transcription and Section 5 gives an evaluation of this approach.

2 SMS Transcription

2.1 Related Work

SMS transcription or translation is still a recent problem and little work has been done on this topic. [4] uses a phrase-based statistical model to normalize SMS and then to translate English SMS in standard English. This task is sometimes compared to noisy text processing [5] but this approach does not take into account the particular aspects encountered in SMS. Commercial on-line software exists for French SMS translation (http://www.traducteur-sms.com or http://www.aidoforum.com/traducteur-sms.php), but it only proposes a rudimentary recognition of the most common SMS abbreviations without any linguistic processing.

[6] and [7] constitute the main references to French SMS transcription system. Both use a manually transcripted corpus of SMS proposed by the university of Louvain [8] to learn statistical language models.

Considering that such corpora were not available when we started working on SMS transcription, statistical methods were not conceivable and this is why we favoured a symbolic approach. Moreover, this applicative context constituted an interesting evaluation task for the TiLT toolbox.

2.2 TiLT for SMS Transcription

As shown in Fig. 1, the transcription process proposed by TiLT relies on the successive application of different analysis modules. First, the initial message is segmented using classical segmentation rules and specific ones like for smileys recognition. A French lexicon composed of 100 000 units enriched with 2 000 specific abbreviations (lol, msg, 2min, etc.) is used to lexically analyse each identified segment. Unknown forms are submitted to various correction and deduction strategies (typographic, morphologic, phonetic, etc.). The segments analysis generates a lattice of lexical units. This lattice is then submitted to the shallow parsing module, which regroups lexical units into chunks, and gives these chunks a syntactic label. This syntactic analysis makes use of grammatical rules, which specify constraints to be applied between chunks and internally between the lexical units which compose a chunk. Thus, the final transcription of the initial SMS corresponds to the succession of forms that has been syntactically validated.

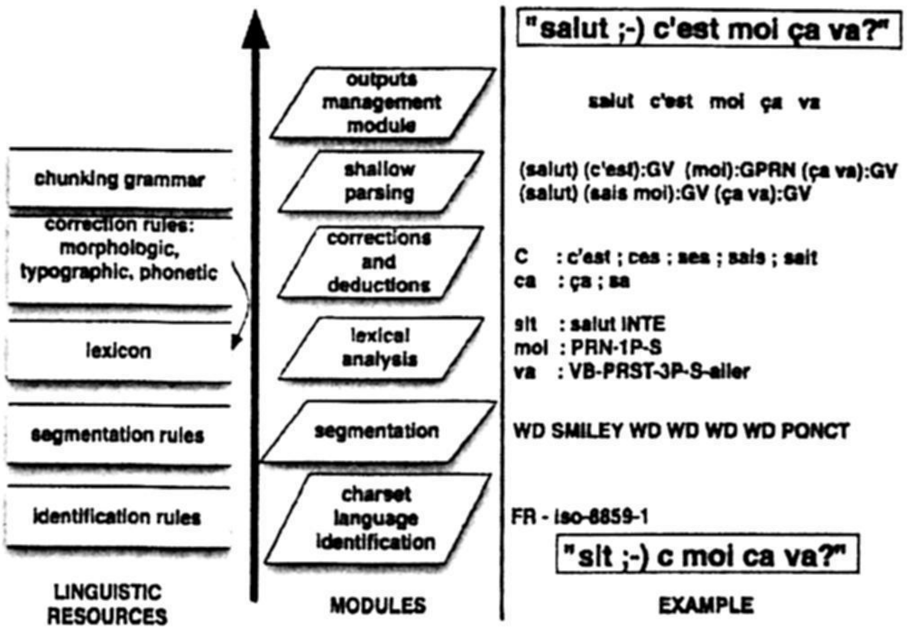

"salut ;-) c'est moi ça va?"
outputs management module
salut c'est moi ça va
chunking grammar
shallow parsing
(salut) (c'est):GV (moi):GPRN (ça va):GV
(salut) (sais moi):GV (ça va):GV
correction rules: morphologic, typographic, phonetic
corrections and deductions
C : c'est ; ces ; ses ; sais ; sait
ca : ça ; sa
lexicon
lexical analysis
slt : salut INTE
moi : PRN-1P-S
va : VB-PRST-3P-S-aller
segmentation rules
segmentation
WD SMILEY WD WD WD WD PONCT
identification rules
charset language identification
FR - iso-8859-1
"slt ;-) c moi ca va?"
LINGUISTIC RESOURCES
MODULES
EXAMPLE

Fig. 1. The initial symbolic transcription process

2.3 Indetermination

As most NLP systems, TiLT is faced with the problem of concurrent and erroneous results generation. This problematic phenomenon is even more present in such a context of spontaneous and atypical text processing.

In the example illustrated in Fig. 1, one can easily notice that most of the analysis modules that compose the transcription process are faced with concurrent and erroneous results. Indeed, the lexical analysis of ambiguous or ill-formed segments leads to concurrent lexical units. For example "C" can be phonetically corrected to "c'est", "ces", "ses", "sais", "sait", . . . , or even considered as the initial of a first name like "Céline", "Cécile", ... Considering these indeterminations, the lexical analysis module generates a lattice of lexical units. Despite the fact that the shallow parsing module aims at reducing the space of concurrent lexical units using syntactic constraints, indeterminations of syntactic groups and labelling remain. Thus, to generate a transcription for an initial SMS, a selection of the one best succession of lexical units has to be performed among the remaining concurrent lexical units.

Based on a first evaluation of this initial transcription process which concerns 9 000 messages coming from the corpus of Louvain [8], we have identified and quantified the different indetermination sources.

For one input segment, the lexical analysis module and its correction strategies generates an average of 15 concurrent lexical units. These lexical units can be factorized into 3 different morpho-syntactic categories. Considering one syntactic chunking, the shallow parser produces 2.5 concurrent syntactic labellings. Finally, the succesion of forms which composes the transcription result is chosen among 2.7 concurrent lexical units for each initial segment.

Faced with these indeterminations, strong heuristics have been integrated into the initial transcription system in order to select one best final transcription. The first one concerns the selection of one syntactic chunking. As recommanded by [9], the chunking having the largest chunks is preferred. The second one is materialized by a score which is associated to some lexical or morpho-syntactic features. Defined by experts, these scores are used to select the one best syntactic labelling, the one which regroups the lexical units having the highest scores, and is also used to perform a final selection of the syntactically validated lexical units which then compose the final transcription.

The first evaluation has also emphasized the fact that 25% of the SMS are not correctly transcribed although they are completely lexically covered.

Through an analysis of these errors [10], we have noticed that for 30% of these lexically covered SMS, the syntactic chunking is wrong. For about 35% of these SMS, the preferred syntactic labelling is not completely correct. Other errors are caused by inappropriate final selection of lexical units.

3 Controlling Analysis Processes

3.1 Related Works

Control strategies aim at identifying correct interpretations among all generated ones. Obviously, this objective can only be reached if distinctive information is available to evaluate the relative relevance of the concurrent results. Thus, a control strategy first relies on the integration or declaration of comparison criteria.

This problem has largely been addressed for the ranking (or reranking) of results generated by speech recognition systems [11]. Such strategies, mainly based on empirical knowledge, have also been applied to control the results of syntactic parser [12] [13], machine translation systems [14] or even natural language generation systems [15].

The use of additional and specific knowledge to evaluate the relative relevance of concurrent results has also been investigated for more specific NLP tasks like word sense disambiguation [16] [17], machine translation [14] or coreference resolution [18].

Nevertheless, it appears that control strategies have only concerned specific applicative contexts and there is no generic formalization or methodology for controlling a complete NLP system like TiLT.

3.2 A Decisional Approach of Control

[3] has proposed to consider this task as a decision process. Based on this formalization a generic control module has been implemented and in TiLT. Indeed, as decision problems [19], a control strategy relies on a first stage of concurrent results evaluation which is then use to identify the most preferred results.

Faced with the heterogeneity of the indetermination cases, it appears necessary to combine multiple criteria during the evaluation of the results. Moreover, contrary to most of the existing control strategies which rely on statistical methods, the approach proposed by [3] makes use of expert preferences in order to determine how the concurrent results have to be compared. This way, this approach can be applied when no representative corpus is available. This formalization has led to create an intersection between NLP and a domain specialized in the resolution of such problem: MultiCriteria Decision Aid (MCDA) and more precisely outranking approaches [20] which propose methods for aggregating incommensurable criteria.

3.3 A Generic Framework of Control based on Outranking

Let $R : \{r_1, r_2, ..., r_m\}$ be the set of concurrent results and $C : \{C_1, C_2, ..., C_n\}$ the considered comparison criteria. Each criteria constitutes an increasing function which is used to evaluate the results relevance. Thus, each result $r_i, i = 1..n$

is associated to a performances vector $\{g_1(r_i), g_2(r_i), ..., g_n(r_i)\}$ which represents its evaluation on each considered criterion. As it has been previously said, the main characteristic of outranking approaches is to rely on expert preferences which determine the way the results have to be compared according to their associated performances vectors. Such preferences express the importance and the uncertainty to grant to each criteria, and also modeled incompatibility situations when two results are compared. These preferences are materialised by:

$W : \{w_1, w_2, ..., w_n\}$ a weights vector,
$Q : \{q_1, q_2, ..., q_n\}$ indifference thresholds,
$P : \{p_1, p_2, ..., p_n\}$ preference thresholds,
$V : \{v_1, v_2, ..., v_n\}$ veto/incomparability thresholds,

Q and P express an imprecision margin when two performances are compared and V define incomparability limits.

The evaluation of the concurrent results relies on a pairwise comparison in order to establish outranking relations. A result r_1 outranks a result r_2, noted $r_1 \, S \, r_2$, if a sufficient majority of criteria validates the assertion of outranking (concordance measure $c(r_1, r_2)$) and if the minority that invalidates this assertion (discordance measure $d_k(r_1, r_2), k = 1..n$) is not too strong. The concordance measure $c(r_1, r_2)$ is based on partial concordance indices $c_k(r_1, r_2), k = 1..n$ computed for each criterion:

$$c_k(r_1, r_2) = \begin{cases} 0, & \text{if } g_k(r_2) - p_k \times g_k(r_1) \geq g_k(r_1) \\ ]0, 1[, & \text{if } g_k(r_1) * (1 + q_k) \leq g_k(r_2) - g_k(r_1) \leq g_k(r_1) * (1 + p_k) \\ 1, & \text{if } g_k(r_2) - q_k \times g_k(r_1) \leq g_k(r_1) \end{cases}$$

The concordance measure regroups partial concordance indices:

$$c(r_i, r_j) = \frac{1}{P} . \sum_{k=1}^{n} w_k . c_k(r_1, r_2)$$

where $P = \sum_{k=1}^{n} w_k$

The discordance is represented by partial discordance indices $d_k(r_1, r_2), k = 1..n$:

$$d_k(r_1, r_2) = \begin{cases} 1, & \text{if } g_k(r_2) - v_k \times g_k(r_1) \geq g_k(r_1) \\ ]0, 1[& \text{if } g_k(r_1) * (1 + p_k) < g_k(r_2) - g_k(r_1) < g_k(r_1) * (1 + v_k) \\ 0, & \text{if } g_k(r_2) - p_k \times g_k(r_1) \leq g_k(r_1) \end{cases}$$

A global credibility index $\sigma(r_1, r_2) \in [0, 1]$ is computed from $c(r_1, r_2)$ and $d_k(r_1, r_2), k = 1..n$ and repesents the credibility to grant to the outranking relation established between r_1 and r_2.

$$\sigma(r_1, r_2) = C(r_1, r_2) \prod_{k \in \overline{F}} \frac{1 - d_k(r_1, r_2)}{1 - C(r_1, r_2)}$$

The outranking relations established between pairs of concurrent results can then be interpreted in order to produce decision recommendation of three kinds:

ranking favoring results that outrank the largest number of other concurrent results with the highest credibility degrees, a partial pre-order can be computed to represent a ranking of the concurrent results,

selection results that outrank the largest number of other concurrent results with the highest credibility degrees without being outranked by other results constitute a set of favored results

classification compared with acceptability profiles which are associated to classes by experts, results can be affected to ordered classes of equivalence.

We suggest the interested reader to read [20] and [21] for more information about algorithms used to build these decision recommendations.

These recommendations are then interpreted by analysis modules in order to favor the most preferred results or to symmetrically filter the less relevant ones.

4 Controling the SMS Transcription Process

4.1 Criteria

Based on a manual analysis of the errors made by the initial transcription process, we have remarked that many recurrent and typical SMS patterns are not well transcripted, for example: "c bon" → "c'est bon", "a plus" → " plus", "comen sa va" → "comment ca va", ...

Through a manual analysis of the erroneous transcription generated by the initial process, we have noticed that frequent and simple words successions are not correctly transcripted.

So to improve this initial transcription process, especially for such recurrent lexical and syntactic patterns, we have integrated empirical criteria in order to favor the most frequent forms and successions of form. Thus, 20 000 SMS transcriptions of the Louvain corpus have been used to establish a frequency table of lemmatized and inflected forms.

This table is first used to associate to each candidate lexical unit its observed frequency and secondly, according to this frequencies table, the Viterbi algorithm [22] is applied on the lattice of concurrent lexical units to identify one best path of words bigrams.

Therefore, the initial heuristic criterion corresponding to an *a priori* definied quantitative preference for some morpho-syntactic categories is completed with the empirical criteria. Thus, each candidate lexical unit $r_i, i = 1..m$ is then evaluated on 4 criteria:

$g_1(r_i)$ preference score on the morpho-syntactic category of r_i

$g_2(r_i)$ the frequency of r_i's lemmatised form
$g_3(r_i)$ the frequency of r_i's inflected form
$g_4(r_i)$ a boolean criteria that is true if r_i belongs to the best word bigrams path

4.2 Control Strategy

Despite the fact that the previously enumerated criteria are associated to each generated lexical unit during the lexical analysis, it appears to be inefficient to set up a control stage directly at this stage of the transcription process. Indeed, the shallow parsing aims at reducing the size of the lexical lattice throught a validation of syntactic constraints. Thus, these criteria have been first used to select the one best syntactic labelling of syntactically validated lexical units, and then to select the one best final sequence of lexical units in order to establish the final transcription.

For these two control strategies, a preferences model (Sec. 3.3) favoring empirical criteria has been defined as illustrated by table 1:

Table 1. Preferences model

criteria	weight	ind. thresh.	pref. thresh.	veto thresh.
g_1	0.3	0.4	0.6	-
g_2	0.2	0.05	0.1	0.4
g_3	0.2	0.05	0.1	0.4
g_4	0.4	-	-	-

According to the performances vectors associated to the concurrent lexical items and to this preferences model, concurrent syntactic labelling have been ranked in order to identify the most preferred one. Considering this preferred syntactic labelling and the fact that lexical indeterminations can remain for each morpho-syntactic category, concurrent final lexical units are also ranked in order to determine for each category its most preferred lexical unit and so to establish the final transcription.

4.3 Example

Let us consider the message "si tu revil j anul tt" where the form "tt" can be corrected to {"tout", "tôt","toit","tête","tant",...}. Performances vectors associated to these concurrent lexical units are illustrated in table 2 and the preferences model of table 1 is used to compare these alternatives:

One can remark that the criteria concerning form frequencies and the belonging to the best words bigram path are concordant with the assertion "tout" S "tôt" and no criterion is discordant with it. So σ("tout", "tôt") $= 0.8$ and σ("tout", "tant") = 0.8 too. Moreover, as σ("tôt", "tant") $= 0.4$, these concurrent forms are ranked in the following descending preference order: "tout" $\succ$ "tôt" $\succ$ "tant".

Table 2. Performances vectors

form	$c_1(r_i)$	$c_2(r_i)$	$c_3(r_i)$	$c_4(r_i)$
tout	12	54	54	1
tt	15	23	23	0
tant	15	18	18	0
...	...	...	...	...

5 A First Attempt of Evaluation

During a first evaluation on 9 000 messages, the initial transcription process has obtained the results presented in Table 3 using the Jaccard and BLEU measures:

$$\text{Jaccard coeff.} = \frac{|R \bigcap S|}{|R \bigcup S|}$$

where R is the set of the forms proposed by the transcription process and S is the set of forms of the solution.

$$BLEU = BP.exp\frac{1}{N}\sum_{n=1}^{N} log(\frac{\text{nb. common n-grams}}{\text{nb. n-grams}})$$

where BP is a penalty which is imposed when forms are deleted from the initial message, $BP = min(1, exp(1 - \frac{nb.wordinsolution}{nb.wordsinhypothesis})$.
nb. n-grams = (message size$-N-1$) where N is the size of the largest considered n-gram.

Table 3. Evaluation of the initial transcription process

Jaccard	BLEU	nb. erroneous forms
0.745	0.712	31 248

Table 4 illustrates the results obtained with the controlled transcription process. The improvement that seems low at first glance has to be put into perspective with the identified progression margin. Indeed, a control strategy is efficient only if at least one of the concurrent results is correct. During the first evaluation, we have noticed that only 25% of the 9 000 messages were completly lexically covered and not well transcripted. So, considering this progression gap, the control strategy has lead to a diminution of 20% of the number of final erroneous forms.

Obviously, it would have been interesting to compare our symbolic transcription system with the statistical systems of [6] and [7]. Unfortunately, we do not know on which part of the corpus [6]'s system has been evaluated. Moreover,

Table 4. Evaluation of the controlled transcription process

Jaccard	BLEU	nb. erroneous forms
0.795	0.746	29 759

they use the word error rate as the evaluation measure which is a relevant measure for statistical language model but not for our symbolic system. Indeed, this measure takes into account the number of deleted and inserted words due to the (vocal or text) signal segmentation. As our system do not perform other segmentation than the one established by whitespace characters, no words are deleted nor inserted. [7] do not propose any quantitative evaluation.

6 Conclusion and Perspectives

To overcome the problem of concurrent and erroneous results generated by the different analysis modules which compose a symbolic SMS transcription process, we have proposed a control strategy which relies on an outranking approach. Such an outranking control method allows for the aggregation of heterogeneous (empirical and heuristic) criteria which are incommensurable. The evaluation of this control strategy on 9 000 messages has shown that 20% of the erroneous and lexically covered forms are well corrected.

For future work, we are experiencing the integration of a fifth criterion corresponding to the identification of one best trigrams path of morpho-syntactic categories in the lattice of lexical units. We think that this criterion will help for the identification of recurrent syntactic patterns.

Moreover, to improve the lexical coverage of our lexical analysis module, the correction and deduction strategies have to be reconsidered. Currently, only unknown forms, forms not present in our lexicon, are submitted to deduction and correction strategies. Obviously, a lot of forms in SMS messages are mispelled but correspond all the same to known forms. In the following example "il son fou", the form "son" should be written "sont", but as "son" correspond to a valid possessive pronoun, no correction and deduction strategies are applied and no correct lexical unit is present in the lattice of concurrent forms. Symmetrically, the improvement of the lexical coverage based on the systematic use of correction and deduction strategies will induce an important decrease of the system precision but will again justify the need for control strategies.

References

1. Guimier De Neef, E., Boualem, M., Chardenon, C., Filoche, P., Vinesse, J.: Natural language processing software tools and linguistic data developed by france télécom r&d. In: Indo European Conference on Multilingual Technologies (IECMT). (2002)

2. Bachimont, B.: Le contrôle dans les systèmes à base de connaissances. HERMES (1992)
3. Smits, G.: Une approche par surclassement pour le contrle d'un processus d'analyse linguistique. PhD thesis, Universit de Caen, Orange Labs (2008)
4. Aw, A., Zhang, M., Xiao, J., Su, J.: A phrase-based statistical model for sms text normalization. In: Proceedings of the COLING/ACL on Main conference poster sessions, Morristown, NJ, USA, Association for Computational Linguistics (2006) 33–40
5. Clark, A., Tim, I.: Pre-processing very noisy text. In: Proc. of Workshop on Shallow Processing of Large Corpora. (2003)
6. C.Kobus, Yvon, F., Damnati, G.: Transcrire les sms comme on reconnat la parole. In: Actes de la Confrence sur le Traitement Automatique des. (2008)
7. Beaufort, R., Roekhaut, S., Fairon, C.: Dfinition dun systme dalignement sms/franais standard laide dun filtre de composition. In: 9es Journes internationales dAnalyse statistique des Donnes Textuelles. (2008)
8. Fairon, C., Paumier, S.: A translated corpus of 30,000 French SMS. In: Proceedings of LREC2006. (2006)
9. Abney, S.: Parsing by chunks. Kluwer Academic (1991)
10. Guimier De Neef, E., A., Park, J.: TiLT correcteur de SMS : évaluation et bilan qualitatif. In: Actes de la conférence TALN. (2007)
11. Pusateri, E., Thong, J.V.: N-best List Generation using Word and Phoneme Recognition Fusion. In: Proceedings of the European Conference on Speech. (2001)
12. Collins, M., Koo, T.: Discriminative reranking for natural language parsing. Computational Linguistics (2005)
13. Charniak, E.: Coarse-to-fine n-best parsing and MaxEnt discriminative reranking. In: Proceedings of the 43rd Annual Meeting of the ACL. (2005)
14. Shen, L., Sarkar, A.. Och, F.: Discriminative reranking for machine translation. In: Proceedings of the Joint HLT and NAACL Conference. (2004)
15. Paiva, D., Evans, R.: Empirically-based Control of Natural Language Generation. In: Proceedings of the 43rd Annual Meeting of the ACL. (2005)
16. Rosso, P., Masulli, F., Buscaldi, D.: Word sense disambiguation combining conceptual distance, frequency and gloss. In: Proceedings of the International Conference on Natural Language Processing and Knowledge Engineering. (2003)
17. Dang, H., Palmer, M.: Combining Contextual Features for Word Sens Disambiguation. In: Proceedings of the SIGLEX/SENSEVAL Workshop on Word Sense Disambiguation. (2002)
18. Weissenbacher, D., Nazarenko, N.: A bayesian approach combining surface clues and linguistic knowledge : Application to the anaphora resolution problem. In: Proceedings of the Recent Advances in Natural Language Processing (RANLP'07). (2007)
19. Hansson, S.: Decision Theory: A Brief Introduction (1994)
20. Bouyssou, D.: Outranking approach. Encyclopedia of optimization (2001)
21. Mousseau, V., Slowinski, R., Zielniewitcz, P.: Electre Tri 2.0 Methodological guide and user's manual. Technical report, LAMSADE, Paris Dauphine (1999)
22. Forney, G.: The Viterbi algorithm. In: Proceedings of the IEEE. Volume 61. (1973)

A Concept-Suggestion Engine for Professional Multimedia Archives

Marco A. Palomino[1], Michael P. Oakes[1] and Tom Wuytack[2]

[1] University of Sunderland - Informatics Centre
St Peter's Way, Sunderland SR6 0DD, United Kingdom
{marco.palomino, michael.oakes}@sunderland.ac.uk

[2] Belga News Agency
Rue Frederic Pelletier 8b, 1030 Brussels, Belgium
wut@belga.be

Abstract. Choosing the optimal set of keywords to represent a search engine query is not a trivial task, and may involve an iterative process such as relevance feedback, repeated unaided attempts by the user or the automatic suggestion of additional terms, which the user may select or reject. This is particularly true of a multimedia search engine which searches on concepts as well as user-input terms, since the user is unlikely to be familiar with the full range of system-known concepts in advance. We propose two concept suggestion strategies: suggestion by semantic similarity, where the closest matching concept definitions in a glossary to the initial user input are found using the cosine similarity measure, and suggestion by normalised textual matching. where concepts are suggested if their headwords match the user input. Both methods were evaluated by comparing machine suggestions with concepts suggested by professional annotators, using the measures of micro- and macro- precision and recall. Although normalised text matching was the simpler technique, it performed much better on the recall-based measures, and only slightly less well on the precision-based measures.

Keywords: Information retrieval, multimedia, concepts, ground-truth, semantic similarity, cosine similarity.

1 Introduction

Popular multimedia search engines, such as *Google* [1], *YouTube* [2] and *Blinkx* [3], provide access to their repositories via text, as this is still the easiest way for their users to express their information needs. The indices of these search engines point to pictures, videos and some other resources on the Web based on their file names, surrounding text or transcripts. Regrettably, this results in disappointing performance when the visual content is not reflected in the associated text. Hence, a current trend in information retrieval consists of learning a lexicon of semantic concepts from multimedia examples and employing them as entry points when querying the Web [4].

*© G. Sidorov, M. Aldape, M. Martinez, S. Torres. (Eds.)
Advances in Computer Science and Engineering.
Research in Computing Science 42, 2009, pp. 29-40*

*Received 04/03/09
Accepted 27/04/09
Final version 04/05/09*

As part of the EU-funded *Video and Image Indexing and Retrieval in the Large-Scale (VITALAS)* project [5], which aims to produce a system dedicated to the intelligent access to professional multimedia archives, we have experimented with several techniques for the automatic suggestion of concepts derived from user-specified sample pictures and their textual captions. Currently, VITALAS employs a vocabulary of 525 concepts, whose entries vary from pure video format—like a detected *overlayed text*—settings and scenarios—like an *interview*—objects—like an *elephant*—or events—like a *celebration after scoring.*

The VITALAS concept vocabulary gives users semantic access to picture and video, allowing them to query on the presence or absence of content elements. However, selecting the right topic from a large vocabulary is a lengthy and resource-consuming task that should not be performed manually. Therefore, we have developed a *suggestion engine* that analyses textual captions contained in a multimedia archive and automatically derives the most relevant concepts for querying such an archive. The results yielded by our suggestion engine have been compared with the observations made by professional annotators, who reviewed the pictures and linked them to some of the concepts that best described them. Thus far, our evaluation indicates that simple textual matches between picture captions and the concept vocabulary delivers better suggestions than other more sophisticated approaches.

The remainder of this paper is structured as follows: Section 2 describes the multimedia collection that we employed in our study. Section 3 presents the concept vocabulary from which we derive our suggestions and elaborates on the acquisition of ground-truth annotation. Section 4 explains the different strategies that we propose to implement a suggestion engine. Section 5 reports on the evaluation of our results. Section 6 introduces the related work in this area of research, and Section 7 states our conclusions and plans for future work.

2 Belga's Multimedia Archive

In order to undertake our research, we made use of a large multimedia archive owned by the *Belga News Agency* [6]. Belga's archive covers Belgian and international news and current affairs—politics and economics, finance and social affairs, sports, culture and personalities. Although Belga's content is published in four different formats—text, pictures, audio and video—this paper concentrates on pictures and their associated textual captions, exclusively.

A *caption* is a free-text field whose content is created by photographers, and offers an explanation or designation accompanying a picture posted on Belga's website. Each group of photographers has its own conventions and styles to present information. As a consequence, certain captions include not only text relative to pictures, but also names of photographers, their initials and acronyms of press agencies, as well as the dates when the pictures were taken or published and some other ancillary information. Since none of these particulars are deleted before posting, we decided to keep them in our analysis too.

To ensure that we had enough material to carry out our work, Belga granted us access to a set of 1,727,159 pictures and captions that were published on its website between 22 June 2007 and 2 October 2007. Figure 1 displays an example of a typical picture and caption posted on Belga's website.

Fig. 1. Example of a Belga Picture and Caption: Italian Prime Minister, Romano Prodi (R), shakes hands with US President George W. Bush prior to their press conference at Chigi Palace in Rome, Italy, in June 2007. President Bush, who is in Rome as part of his trip through Europe, will also meet Silvio Berlusconi.

3 VITALAS Annotated Concept Vocabulary

Acquiring a suitable list of concepts for multimedia purposes is a major research challenge. Naphade *et al.* have built a multimedia ontology based on extensive analysis of video-archive query logs [7], and the *MediaMill Challenge* has employed this ontology as the main source for its 101 concept vocabulary [8].

As opposed to MediaMill, the VITALAS concept vocabulary is largely derived from the automatic extraction of keywords that characterise Belga's archive. Nevertheless, it has been refined manually over time. Originally, the vocabulary was derived from a comparison between Belga's captions and a model of general English language. The words that deviated from the model were very specific to the captions and thus made appropriate keywords to characterise the archive. Professional annotators evaluated the keywords and removed those that they considered unsuitable. The remaining keywords became the first entries in the VITALAS concept vocabulary. Later on, these entries were extended manually to guarantee that the vocabulary comprised as many categories as available in the news domain. Finally, we mined Belga's query logs programmatically to extract keywords that reflected the most important concepts from the users' perspective[1]. Some of these concepts were also added to the vocabulary.

[1] Belga gave us access to its query logs for exactly the same period when the collection of chosen captions was published on its website: 22 June 2007 – 2 October 2007.

Further details related to the VITALAS concept vocabulary have been published by Palomino *et al.* [9], and the entire concept vocabulary for the VITALAS project is available at the authors' website (`http://osiris.sunderland.ac.uk/~cs0mpl/VITALAS/`).

3.1 Ground-Truth Annotation

Recognising the importance of having a picture database containing ground-truth annotations parsed by humans, VITALAS selected 100,000 pictures published on Belga's website, and employed professional annotators to determine some of the concepts that best described them.

For annotation purposes, the presence of a concept was assumed to be binary: it was either visible in a picture or not—the location of the concept in the picture was not taken into account. A total of 1,000 pictures were annotated for each of the 525 concepts. However, the set of 1,000 pictures annotated for a particular concept were not necessarily the same as those annotated for any other concept.

Approximately 500 of the pictures annotated for a particular concept were chosen from the results of queries submitted by Belga users who had included that concept name. The rest of the pictures were chosen randomly. Hence, the first half of pictures is likely to contain positive samples, whereas the second one is likely to contain negative samples. Achieving this balance was vital for the evaluation of our experiments.

The ground-truth annotation also included the provision of a textual *definition*—or *description*—for each concept, together with relevant keywords and references to positive images. Table 1 shows an example of a VITALAS concept—*food*—accompanied by its description and reference to positive images. Table 1 also shows a picture that has been annotated positively as an image that does correspond with the concept *food.*

Table 1. Example of Disambiguation and Positively Annotated Picture

	Concept name: food **Concept description:** An image showing any substance reasonably expected to be ingested by a human or an animal for nutrition or pleasure. **Relevant keywords:** Cooking, meal. **Examples of positive images:** A picture of a table showing a served meal; a picture of dishes ready to be consumed; a picture of meat, fish, fruit or vegetables for sale in a market.

The VITALAS manual annotation process has yielded an incomplete, but reliable ground-truth for our concept vocabulary. Certainly, we would like to have all the pictures annotated for all of the concepts; yet, despite resource limitations, we have gathered a reasonably large subset of annotated pictures.

4 Concept Suggestion Strategies

In this paper, we evaluate two different approaches for suggesting to users the most relevant concepts related to a particular picture caption: suggestion based on the *semantic similarity* between the caption and the textual description of each concept; and suggestion based on *normalised-textual matching* between the caption and the concept vocabulary.

4.1 Suggestion by Semantic Similarity

As explained in subsection 3.1, each concept ω in the VITALAS concept vocabulary is associated with a textual description d_ω. Then, we can measure the semantic similarity between a caption and the textual description of each different concept. Both captions and descriptions are *normalised* before examining their semantic similarity: all text is converted to lower case, punctuation and numbers are removed, and extremely common and semantically non-selective words are deleted—the stop-word list that we are using was built by Salton and Buckley for the experimental *SMART* information retrieval system [10]. In addition, all text is *stemmed*, reducing inflectional and derivationally related forms of a word to a common base—the particular algorithm for stemming English words that we are using is *Porter's algorithm* [11].

We represent each concept description as a *vector*, whose entries correspond to unique normalised words. Since concept descriptions are written in natural language, word distribution corresponds, roughly, with *Zipf's law* [12]. Therefore, the vector space model proposed by Salton *et al.* [13] is appropriate for our semantic analysis. Specifically, with a collection of descriptions D, a concept description d_ω in D, and a caption q containing words t_i, we use the following implementation of the vector space model to compute the *cosine similarity* between caption q and concept description d_ω:

$$sim(q, d_\omega) = \frac{\sum\limits_{k \in (q \cap d_\omega)} tf_{kq} \cdot tf_{kd_\omega}}{\sqrt{\sum\limits_{k \in d_\omega} (tf_{kd_\omega})^2} \sqrt{\sum\limits_{k \in q} (tf_{kq})^2}},$$

where tf_{kq} is the frequency of the $k-th$ word contained in caption q, and tf_{kd_ω} is the frequency of the $k-th$ word contained in description d_ω.

It can be demonstrated that the resulting similarity between q and d_ω ranges from 0 meaning *no match*, to 1 meaning *complete match*, with in-between values indicating intermediate similarity [14]. Hence, we may chose a threshold and suggest concept ω as a possible *match* for q only if the similarity between q and d_ω is above the threshold.

4.2 Suggestion by Normalised Textual Matching

Considering the relatively small size of both captions and concept descriptions, it is computationally inexpensive to calculate their semantic similarity. Using an Intel© Xeon© CPU 5150 processor with 2GB of RAM, running under Microsoft Windows XP 2002 SP2, we can calculate the semantic similarity between a single caption and *all* of the 525 concept descriptions in less than a few hundred milliseconds. However, another strategy that we chose to assess due to its simplicity and extremely fast performance was the straight textual matching between the captions and the concept vocabulary.

As in the case of the semantic similarity, this second strategy begins by normalising the caption. Yet, in this case, we also normalised the concept vocabulary. As a second step, we look for exact matches between the words in the normalised caption and those available in the normalised vocabulary. For illustration purposes, Table 2 displays an example of a caption, its normalised version and the resulting matches with the concept vocabulary.

Table 2. Example of Normalised Textual Match

Original caption:	*Soccer Italy training—Italian forward Alessandro Del Piero of Juventus Turin practices his penalties during training at Wembley Stadium this afternoon, 11 February, before tomorrow's World Cup qualifying match against England.*
Normalised caption:	*soccer itali train italian forward alessandro piero juventu turin practic penalti train wemblei stadium afternoon tomorrow world cup qualifi match england*
Concept vocabulary:	abbey ... cup ... soccer ... stadium ...
Normalised vocabulary:	abbei ... cup ... soccer ... stadium ...
Matches:	cup, soccer, stadium

In the case of concept names made of more than one word—such as, davis_cup—different heuristics may be applied. We may look for precise matches of all the words contained in the concept name, which would limit the number of matches considerably, but would ensure that only captions referring explicitly to the concept name are matched.

A more relaxed approach would be to select one single word as the *headword*. For instance, we may say that davis is the headword for the concept davis_cup, and we will automatically associate all the matches of davis with this concept. This is the approach that we have pursued.

Due to space limitations, we cannot list the headwords for all of the concepts in the VITALAS vocabulary in these pages. Readers, however, are welcome to visit the authors' website for further details on this matter [15]. It should be observed that for certain concepts—such as `ac_milan_soccer`—two headwords have been selected—`milan` and `soccer`—though they do not necessarily have to appear together to provide a match—an appearance of either `milan` or `soccer` would provide a match for the concept `ac_milan_soccer`.

In the following section, we report on the use of different thresholds for this strategy. Of course, as we lower the threshold a larger number of false positives is suggested to the users. Nevertheless, *recall* also increases, which is more important than *precision* for concept suggestion, because users will benefit from being able to choose from a large variety of possible additional concepts and can easily reject unsuitable ones.

5 Evaluation

Precision and *recall* [16] are two standard measures used in information retrieval to evaluate performance. Precision and recall are defined in terms of a set of retrieved documents and a set of relevant documents. Given the particular characteristics of the multimedia archive that we have used, and the conditions of the ground-truth annotation that we have exploited, we have taken a modified version of the traditional definitions of precision and recall.

For the remainder of this paper, we refer to the *recall for caption q* ($\mathbb{R}_q$), and the *precision for caption q* ($\mathbb{P}_q$), as

$$\mathbb{R}_q \equiv \frac{\mid \Omega^q_A \cap \Omega^q_{MI} \mid}{\mid \Omega^q_A \mid},$$

$$\mathbb{P}_q \equiv \frac{\mid \Omega^q_A \cap \Omega^q_{MI} \mid}{\mid \Omega^q_{MI} \mid},$$

where Ω^q_A is the set of concepts that the annotators associated with the picture whose caption is q, and Ω^q_{MI} is the set of concepts that our automatic suggestion strategy proposed for the picture whose caption is q.

Averaging over the total number of pictures in the collection $\mathfrak{C}$, we made use of the following definitions for *micro-recall* ($\mu_{\mathbb{R}}$), *micro-precision* ($\mu_{\mathbb{P}}$), *macro-recall* ($\mathcal{M}_{\mathbb{R}}$) and *macro-precision* ($\mathcal{M}_{\mathbb{P}}$) [17],

$$\mu_{\mathbb{R}} \equiv \frac{\sum\limits_{q\in\mathfrak{C}} |\Omega^q_A \cap \Omega^q_{MI}|}{\sum\limits_{q\in\mathfrak{C}} |\Omega^q_A|} \quad , \qquad \mu_{\mathbb{P}} \equiv \frac{\sum\limits_{q\in\mathfrak{C}} |\Omega^q_A \cap \Omega^q_{MI}|}{\sum\limits_{q\in\mathfrak{C}} |\Omega^q_{MI}|}$$

$$\mathcal{M}_{\mathbb{R}} \equiv \frac{\sum\limits_{q\in\mathfrak{C}} \mathbb{R}_q}{|\mathfrak{C}|} \quad , \qquad \mathcal{M}_{\mathbb{P}} \equiv \frac{\sum\limits_{q\in\mathfrak{C}} \mathbb{P}_q}{|\mathfrak{C}|}.$$

Table 3 displays the values of each of the measures defined above for the two suggestion strategies described in Section 4—the highest values achieved for each measure are presented in bold font.

Table 3. Evaluation Results

Measure	Threshold	Semantic Similarity	Textual Match
μ_{R}			0.75
$\mu_{?}$			0.17
$\mathcal{M}_{\mathrm{R}}$			0.77
$\mathcal{M}_{?}$			0.21
μ_{R}	0.2	0.34	
$\mu_{?}$	0.2	**0.26**	
$\mathcal{M}_{\mathrm{R}}$	0.2	0.35	
$\mathcal{M}_{?}$	0.2	**0.35**	
μ_{R}	0.15	0.49	
$\mu_{?}$	0.15	0.17	
$\mathcal{M}_{\mathrm{R}}$	0.15	0.49	
$\mathcal{M}_{?}$	0.15	0.27	
μ_{R}	0.1	0.65	
$\mu_{?}$	0.1	0.08	
$\mathcal{M}_{\mathrm{R}}$	0.1	0.66	
$\mathcal{M}_{?}$	0.1	0.16	
μ_{R}	0.05	**0.83**	
$\mu_{?}$	0.05	0.02	
$\mathcal{M}_{\mathrm{R}}$	0.05	**0.84**	
$\mathcal{M}_{?}$	0.05	0.06	

Semantic similarity with a low threshold—0.05—performs better than normalised textual matching on the recall-based measures. High recall is more important than high precision for concept suggestion, since users will benefit from being able to choose from a range of possible additional concepts and can easily reject unsuitable ones.

Normalised textual matching performed better than semantic similarity when the threshold was greater than 0.05. Indeed, if we computed a *micro F-measure*, defined as $\mu_F \equiv \frac{2\mu_? \mu_{\mathrm{R}}}{\mu_? + \mu_{\mathrm{R}}}$, and a *macro F-measure*, defined as $\mathcal{M}_F \equiv \frac{2\mathcal{M}_? \mathcal{M}_{\mathrm{R}}}{\mathcal{M}_? + \mathcal{M}_{\mathrm{R}}}$, we would realise that the F-based measures for normalised textual matching—namely, $\mu_F = 0.28$ and $\mathcal{M}_F = 0.33$—are better than those of semantic similarity at any threshold.

Apart from its simplicity, one of the advantages of normalised textual matching is the fact that this technique is the fastest one to execute. Therefore, we envisage that an improved version of the normalised textual matching approach, where the annotators pick up the headwords, and extend them manually to better reflect concept relations, may yield very good results.

6 Related Work

The study published by Hoogs *et al.* [18] is among the first ones to add semantics to concept detection, by establishing links with a general-purpose ontology, which connected a limited set of visual attributes to *WordNet* [19]. However, combining low-level visual attributes with concepts in an ontology is a rather difficult task, due to the so-called *semantic gap* between them [20].

To cope with the demand for ground-truth, Lin *et al.* initiated a collaborative annotation effort for the *TRECVID 2003 benchmark* [21]. Using tools from Christel *et al.* [22] and Volkmer *et al.* [23], a common annotation effort was again made for the *TRECVID 2005 benchmark*, yielding a large set of annotated examples for 39 concepts taken from a predefined collection [7]. We have provided a larger compilation, increasing the concept vocabulary to 525 concepts, and getting 1,000 annotated pictures per concept.

The work reported by Snoek *et al.* [24] is closely related to ours. They did implement a concept suggestion strategy based on semantic similarity; yet, they made use of the *Lucene search engine* [25] as part of their implementation, and the goal of their study was different, as they attempted to obtain semantic descriptions and structure from WordNet. The results presented by Snoek *et al.* are not conclusive, but we may consider following their recommendations on *ontology querying* and *Resnik's measure of information content* [26] in future versions of our research.

7 Conclusions

We have described two methods of concept suggestion, with the aim of helping multimedia search engine users enhance their initial keyword queries with additional terms corresponding to system-known concepts, namely suggestion by semantic similarity and normalised textual matching. Although normalised textual matching was the simpler technique, it performed much better on the recall-based measures at most of the thresholds tried for the semantic similarity approach, and only slightly less well on the precision-based measures. As explained in Section 5, high recall is more important than high precision for query term suggestion, since the user will benefit from being able to choose from a range of possible additional concepts, and can easily reject unsuitable ones. However, in approaches such as the one proposed by Palomino *et al.* [27], where discovered additional concepts are automatically added to the query without prior user approval, precision is more important, since the inclusion of non-relevant concepts in the query can severely degrade performance.

Even though we have evaluated the quality of our concept selection using recall- and precision- based measures, we still need to measure the effect of our concept suggestion facilities on the overall search engine performance. Such evaluations are being carried out by our research partners at the *Institut National de l'Audiovisuel* [28], using recall, precision, and subjective measures of user satisfaction with the overall system.

7.1 Future Work: Similarity Matrix

In future versions of our suggestion engine, we are considering to determine the degree of association between every pair of concepts by means of a concept-to-concept *similarity matrix*. To produce each entry in this matrix, we plan to represent the relation between each concept and all of the captions in different vectors. The entries in these vectors would contain the frequency of appearance of the concept headwords in each caption of the collection. Then, the similarity of a pair of concepts would be given by the cosine similarity of their corresponding vectors. For instance, the most similar VITALAS concepts to `actress` would be `filming` (0.40), `film_festival_cannes` (0.39), `festival` (0.27), `academy_award` (0.15), `actor` (0.15) and `award` (0.15).

Once the similarity matrix is created, we can suggest concepts relevant to a particular picture by scanning its caption and searching for occurrences of the concept headwords on it. Afterwards, we derive from the similarity matrix the "similarity" between the appearing headwords and all the concepts in the vocabulary. As in the case of the semantic similarity approach, we define a threshold and suggest to the user only those concepts whose corresponding values are above the threshold.

By generating a similarity matrix automatically from Belga's specific documents, we expect to produce concept suggestions adapted to this particular domain much better than using the term relations available from a thesaurus or library classification.

Given that this additional suggestion strategy combines the features of the other two approaches proposed in this paper, we expect it to have a better performance. However, its implementation and evaluation are still under way.

Acknowledgments. This research was supported under the EU-funded VITALAS project—project number FP6-045389. The authors are very grateful to Belga News Agency for providing the data used to carry out their research, and to the following people for participating in the ground-truth annotation: Jonathan Dujardin, Jeroen Van Den Haute, Katrien Vanmechelen, Sini Cheang and Hecham Azarkan.

References

1. Google: *Web, Image and Video Search.* (2009) `http://www.google.com/`.
2. YouTube: *YouTube, LLC.* (2009) `http://www.youtube.com/`.

3. Blinkx: *Video Search Engine.* (2009) `http://www.blinkx.com/`.
4. Worring, M., Snoek, C.G.M., Huurnink, B., van Gemert, J.C., Koelma, D.C., de Rooij, O.: The MediaMill Large-Lexicon Concept Suggestion Engine. In *Proceedings of the 14th ACM International Conference on Multimedia*, Santa Barbara, CA, Association for Computing Machinery (October 2006) 785–786
5. VITALAS: *Video and Image Indexing and Retrieval in the Large Scale.* (2009) `http://www.vitalas.org/`.
6. Belga: *Belga News Agency.* (2009) `http://www.belga.be/`.
7. Naphade, M., Smith, J.R., Tesic, J., Chang, S.F., Hsu, W., Kennedy, L., Hauptmann, A., Curtis, J.: Large-Scale Concept Ontology for Multimedia. *IEEE MultiMedia* **13**(3) (2006) 86–91
8. Snoek, C.G.M., Worring, M., van Gemert, J.C., Geusebroek, J.M., Smeulders, A.W.M.: The Challenge Problem for Automated Detection of 101 Semantic Concepts in Multimedia. In *Proceedings of the 14th ACM International Conference on Multimedia*, Santa Barbara, CA, Association for Computing Machinery (October 2006) 421–430
9. Palomino, M.A., Oakes, M.P., Wuytack, T.: Automatic Extraction of Keywords for a Multimedia Search Engine Using the Chi-Square Test. In *Proceedings of the 9th Dutch-Belgian Information Retrieval Workshop (DIR 2009)*, Enschede, The Netherlands (February 2009) 3–10
10. Buckley, C.: Implementation of the SMART Information Retrieval System. Technical Report TR85-686, Computer Science Department, Cornell University, Ithaca, New York (May 1985)
11. Porter, M.: An Algorithm for Suffix Stripping. *Progam* **14**(3) (July 1980) 130–137
12. Gelbukh, A., Sidorov, G.: Zipf and Heaps Laws' Coefficients Depend on Language. In *Proceedings of the Conference on Intelligent Text Processing and Computational Linguistics*, Mexico City (February 2001) 332–335
13. Salton, G., Wong, A., Yang, C.: A Vector Space Model for Automatic Indexing. *Communications of the ACM* **18**(11) (November 1975) 613–620
14. Widdows, D.: Measuring Similarity and Distance. In *Geometry and Meaning*, CSLI Publications (November 2004)
15. Palomino, M.A.: *VITALAS Concept-Suggestion Engine.* (2009) `http://osiris.sunderland.ac.uk/~cs0mpl/VITALAS/`.
16. Belew, R.K.: *Finding Out About: A Cognitive Perspective on Search Engine Technology and the WWW.* Cambridge University Press, Cambridge, UK (February 2001)
17. Joachims, T.: *Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms.* Kluwer Academic Publishers (April 2002)
18. Hoogs, A., Rittscher, J., Stein, G., Schmiederer, J.: Video Content Annotation Using Visual Analysis and a Large Semantic Knowledgebase. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Madison, WI (June 2003) 327–334
19. Fellbaum, C.: *WordNet: An Electronic Lexical Database.* MIT Press, Cambridge, MA (May 1998)
20. Dorai, C.: Bridging the Semantic Gap in Content Management Systems: Computational Media Aesthetics. In *Proceedings of the International Conference on Computational Semiotics for Games and New Media*, Amsterdam, The Netherlands, Kluwer Academic Publishers (September 2001) 94–99
21. Lin, C.Y., Tseng, B.L., Smith, J.R.: Video Collaborative Annotation Forum: Establishing Ground-Truth Labels on Large Multimedia Datasets. In *Proceedings of the TRECVID 2003 Workshop*, Gaithersburg, MD (November 2003)

22. Christel, M., Kanade, T., Mauldin, M., Reddy, R., Sirbu, M., Stevens, S., Wactlar, H.: Informedia Digital Video Library. *Communications of the ACM* **38**(4) (April 1995) 57–58
23. Volkmer, T., Tahaghoghi, S., Thom, J.A.: Modelling Human Judgement of Digital Imagery for Multimedia Retrieval. *IEEE Transactions on Multimedia* **9**(5) (August 2007) 967–974
24. Snoek, C.G., Huurnink, B., Hollink, L., de Rijke, M., Schreiber, G., Worring, M.: Adding Semantics to Detectors for Video Retrieval. *IEEE Transactions on Multimedia* **9**(5) (August 2007) 975–986
25. Lucene: *The Lucene search engine.* (2009) `http://lucene.apache.org/`.
26. Resnik, P.: Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, Montreal, Canada (1995) 448–453
27. Palomino, M.A., Oakes, M.P., Xu, Y.: An Adaptive Method to Associate Pictures with Indexing Terms. In *Proceedings of the 2nd International Workshop on Adaptive Information Retrieval*, London, UK (October 2008) 38–43
28. INA: *Institut National de l'Audiovisuel.* (2009) `http://www.ina.fr/`.

Conceptual Model Based on *Change Relation* (*CR*)

Miguel Martinez, Serguei Levachkine, Marco Moreno,
Miguel Torres and Rolando Quintero

Intelligent Processing of Geographical Information Laboratory-CIC-National Polytechnic
Institute, Mexico City, Mexico
miguelrosales@sagitario.cic.ipn.mx
{sergei, marcomoreno, mtorres, quintero}@cic.ipn.mx

Abstract. This work is focused on modeling of geographic environment, particularly considering changes with respect to the existence of objects in time intervals. Modeling dynamic aspects provide additional semantic that is not considerate from approaches that only are focused in static objects. For example, in a traditional approach for modeling a rain, only are recorded results in each moment that are registered. While in own approach, floods in streets caused by rains and their effects are considered, providing additional semantic that can help to explain variations that exists in successive records. The aim is to develop a model based in "Changes" to describe the object's behavior in a dynamic geographic environment. We introduce *Change Relation* (*CR*) concept as a series of changes, which specify when a change occur by the existence of an event. In this paper, we explain necessary concepts to define the *Change Relations.*

1 Introduction

In the world there are not static objects, but all are dynamic in a certain degree. For decades, the scientific community on geographical information has a great interest in capturing these dynamic objects [14]. Recently, developed approaches have been based in events, where the events become the main component of modeling [22]. The events are abstractions of interconnected phenomena and activities in the real world. These abstractions, however not represent actions and behaviors isolated [2].

The happening of events depends on the fulfillment of certain conditions and involves some consequences in the modeling of world. This independence among events with the complex nature of geographical phenomena imposes additional requirements for the conceptualization of a temporary data model. This model requires more elaborate temporary structures that capture the complexity of geographic happenings as well as a formal representation of relations among such happenings. The inclusion of events in the data model provides a foundation to distinguishing particular semantic of movement based on patterns of events, providing the support to different types of events as well as providing a fundament to developing notification event systems.

© G. Sidorov, M. Aldape, M. Martínez, S. Torres. (Eds.)
Advances in Computer Science and Engineering.
Research in Computing Science 42, 2009, pp. 41-52

Received 23/03/09
Accepted 27/04/09
Final version 07/05/09

Traditionally, changes in geographic phenomena have been derived from a temporary reference frame. Temporal aspects of GIS have been researched from cartography perspective- [14][19], data models [10][17][21], and spatial databases [20][1]. Although to date, no model has been adopted to include time in a GIS.

A more specific approach about changes has considered the associated semantic with the change, as typically founded as part of many spatiotemporal processes including the appearance o disappearance of entities and production or transmission of entities [5][6].

Areas where a model that manage events can be useful in epidemiology, where provide response to changes in distribution of diseases and search for clues in pattern of disease occurrence who can aid in preventing spread of disease. This is an example that describes continuous changes.

The motivation of this work consists of recognizing the phenomena in geographic environment are dynamics, and require models that deal explicitly with the elements of these environments, which mainly are events. In this paper, we research if the dynamic component is performed trough their properties, relations, and involved events.

2 Backgrounds

A common method to capturing changes has been to rely with the sequence of snapshots or discrete samples in sequential moments in time [14][17]. The snapshots approach is a procedure used by cartographers based on animated sequences of maps [7]. Several techniques can be applied, such as playing with sequences of discrete samples at different speeds as pictures in a film, changing the length of a scene to affect the pace of an animation, or alter the order in the scenes presented. Researchers are interested in capturing the complexity of underlying processes, however, are often dissatisfied with the snapshots approach, because this approach ignore the events, each of which occurs separately [4]. Indeed, the changes that occur among snapshots are not explicitly stored; these must be determined by comparing spatial pattern of two successive statements. Another disadvantage is the storage of redundant information [14][18] that occur since the representation of localizations where not occur changes.

Recent works deal the evens as without duration, but not as continuous happenings, which seeks to deal with this problem.

This paper develops an approach based on "*Changes*" to modeling object movement in dynamic geographic domains, such as cars, aircraft, ships, etc. We consider, the "*Change*" as a set of related events, and an event as a unique happening.

3 Methodology

The main question that is done in GIS is Where is it?, then the main aspect is register the movement as a sequence of changes that capture the semantic of dynamism of such entities, such that a domain handler can infer the type of movements and movement patterns that occur. To provide support, such as timely intervention in dynamic domains to generate automatic system to notify warning events.

Dynamic aspects not only involve position change of objects, but also change in other properties. For this reason the *Change* concept is added to the conceptualization to represent if something happening in a time interval. This is a general concept, which is specified according to type of change occurred.

3.1 Spatial Relations in *Change* Concept

To define the changing spatial relations, firstly is needed to define which spatial relations are affected by some change. This variety of spatial relations can be grouped into three different categories:

- Topological relations, which are invariant under geometric transformations [8][9]; for instance, "the street *cross* the railroads" (see Fig. 1).

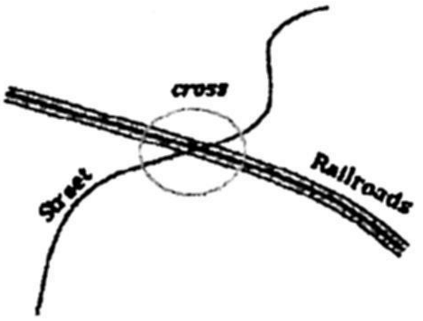


Fig. 1. The street *cross* the railroads.

- Metric relations in terms of distance and directions [16]; for instance, "the hill is just *north*" (see Fig. 2).

Fig. 2. The hill is just *north*.

- Relation concerning to partial or total order of spatial objects [13], as described by prepositions such as *in front of, behind of, over of, before of* [11][3][12]; for instance, "the car is *in front of* the house" (see Fig. 3).

Fig. 3. The car is *in front of* the house.

3.2 Change Relations (*CR*)

A Change Relation (*CR*) is the description of the change to the relations and properties of a geographic object in a state e. This *CR* represents the minimum part in the model. We can collect a set of *CR* to form a history of happenings. This is a chain of *CR*, where the first *CR* is the start point and the last *CR* is the end point in the history of a happening. In Fig. 4, we show the conceptualization of *Change* concept used to represent the *CR*.

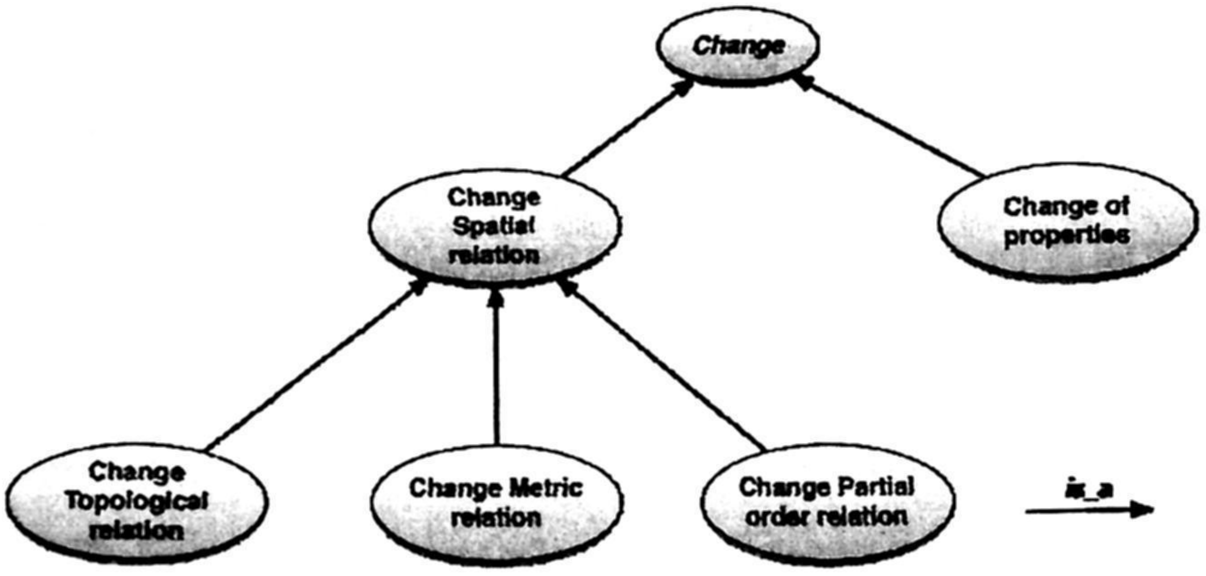


Fig. 4. Conceptualization of *Change* concept.

For instance, Fig. 5 represent the change from one state e_1 to a state e_2, where the object X is inside of object Y. Later in e_2, the object X crosses by the border of object Y. To represent this through chains of CR are as follows:

CR(Obj_X,Obj_Y, e_1)={*Change_CRelSpatial_CRelTopological_Inside => cross by*}

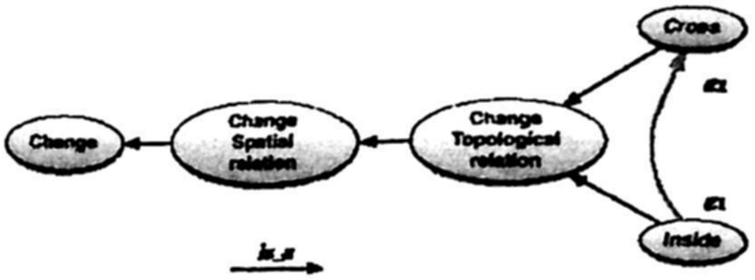


Fig. 5. Change of topological relation.

Where Obj_X is the object which is being compared with respect to Obj_Y. e_2 is the state that caused that the relation between the objects changing.

Changes can occur between a single object or between two or more objects. For the case where change an object property (see Fig. 6), we have:

$$CR\,(\text{Obj\_X,Val\_1}, e_1) = \{Change\_CProperty.Prop\text{-}a => Val\_2\}$$

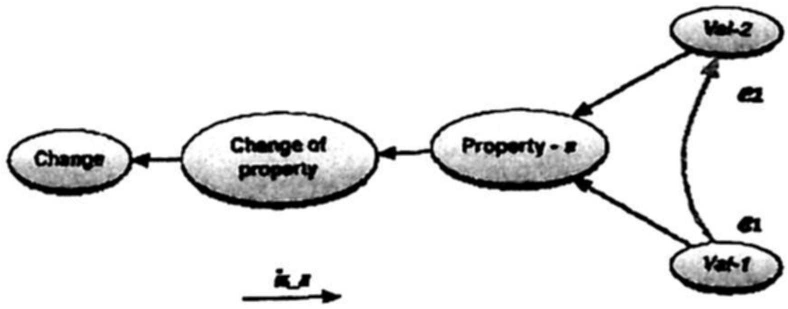


Fig. 6. Change of property value.

Where Obj_X is the object which one more of his properties change in value, and e_1 is the state in which the value changes.

Let's call as e_0 to the start state, for which every geographic object has a position, spatial relation with other geographic objects, and the value of their properties. Then, we can define this with chains of change.

$$CR\,(\text{Obj\_X, Val\_1}, e_0) = Prop\text{-}a => Val\_1÷$$

The CR to state e_0, is used to specify the properties values of the objects and the relation that exist among other geographical objects.

So, for instance, the change of value in a property for the object OBJ_X can be described with CR as:

$$\text{CR (Obj\_X, Val\_1}, e_0) = Prop\text{-}a => Val\_1;$$

$$\text{CR (Obj\_X, Val\_1}, e_1) = Change\_CProperty.Prop\text{-}a => Val\_2;$$

We can read these CR as: "the property a of the Obj_X *change* from Val_1 to Val_2".

3.3 Intervals

To define an event is necessary to define the moment in this occur, we call this moment as ***interval***. An interval is defined as an element that which are composed by elements that are called ***instants***. An interval has two instants: start instant and final instant, and if the case a set of instants between the start and final instants. This set of instants form the interior of the interval.

An instant is the moment whose dimension is almost zero, therefore it is considered as instant, i.e., has no interior point. An instant is denoted by *i*. Basically, exists two types of intervals.

- Intervals with interior, i.e., there at least an instant between start and final instants. Which are called *continuous intervals* and is denoted as I_c.
- Intervals without interior, i.e., there is no instant between start and final instants. Which are called *instantaneous intervals* and is denoted as I_i.

A continuous interval is defined by instants the i_1 and i_2, where i_1 occur before that i_2, besides that there an instant such that is between both instants:

$$I_c = \{i_1, i_2 \exists i_x \mid i_1 < i_x < i_2\}$$

An instantaneous interval is defined by instants the i_1 and i_2, where both occur at same time, this is, there no other instant between i_1 and i_2.

$$I_i = \{i_1, i_2 \mid i_1 = i_2\}$$

Given these two types of intervals, it is possible to define relations between intervals. First, we define relations between continuous intervals, then, between instantaneous intervals and, finally, between two types of intervals.

3.4 Relations Between Continuous Intervals

The possible relations between continuous intervals are:

- *Equal*
- *EndEqual*
- *BegEqual*
- *Inside*
- *Cover*
- *Disjoint*
- *EndBegin*
- *Overlap*

All these relations are binaries, so it requires two intervals. The first relation is *Equal*. This relation defines the situation that given two intervals; both start and finish at same time. Given the intervals I_{c1} and I_{c2}, defined by i_1, i_2, i_3, i_4 intervals as follow:

$$I_{c1} = \{i_1, i_2 \,|\, i_1 < i_2\}$$
$$I_{c2} = \{i_3, i_4 \,|\, i_3 < i_4\}$$

then, we say that these intervals are equals if:

$$i_1 = i_3 \text{ and } i_2 = i_4$$

and we denote this relation as I_{c1} *Equal* I_{c2}, and show them in Fig. 7.

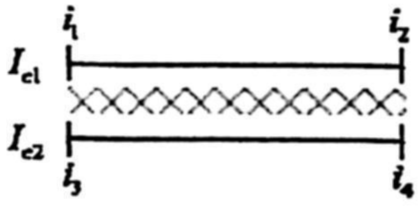


Fig. 7. I_{c1} *Equal* I_{c2}.

The relation *EndEqual* define the situation when two intervals finish at same time, but not at starting. Given the intervals I_{c1} and I_{c2}, we say that both finish at same time if:

$$i_1 < i_3 \text{ and } i_2 = i_4$$

and we denote this relation as I_{c1} *EndEqual* I_{c2}, and show them in Fig. 8.

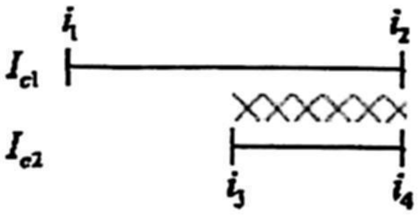


Fig. 8. I_{c1} *EndEqual* I_{c2}.

BegEqual is similar to the above, but in this case, both intervals start at same time and finish at different time. Given the intervals I_{c1} and I_{c2}, we say that both start at same time if:

$$i_1 = i_3 \text{ and } i_2 > i_4$$

and we denote this relation as I_{c1} *BegEqual* I_{c2}, and show them in Fig. 9.

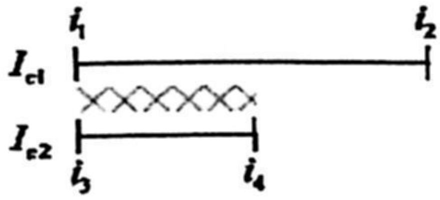


Fig. 9. I_{c1} *BegEqual* I_{c2}.

Next relation is named *Inside* and denote when an interval occur inside while another interval occur. Given the intervals I_{c1} and I_{c2}, we say that both start at same time if:

$$i_1 > i_3 \text{ and } i_2 > i_4$$

and we denote this relation as I_{c1} *Inside* I_{c2}, and show them in Fig. 10.

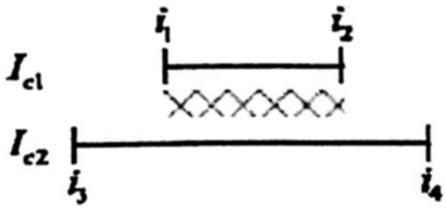


Fig. 10. I_{c1} *Inside* I_{c2}.

Cover relations is dual to the above, i.e. , this relation describe the inverse situation of *Inside*. Given the intervals I_{c1} and I_{c2}, we say that both start at same time if:

$$i_1 < i_3 \text{ and } i_2 > i_4$$

and we denote this relation as I_{c1} *Inside* I_{c2}, and show them in Fig. 10.

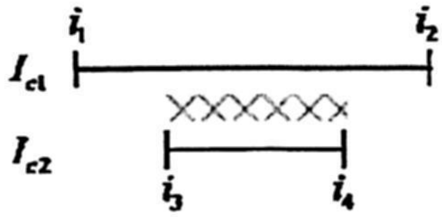


Fig. 11. I_{c1} *Inside* I_{c2}.

Next relation describes the situation where both intervals not have common parts. Given the intervals I_{c1} and I_{c2}, we say that both start at same time if:

$$i_1 < i_2,\ i_3 < i_4 \text{ and } i_2 < i_3$$

and we denote this relation as I_{c1} *Disjoint* I_{c2}, and show them in Fig. 12.

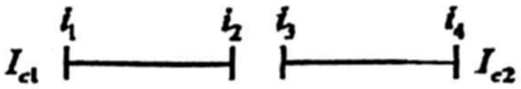


Fig. 12. I_{c1} *Disjoint* I_{c2}.

The situation where an interval start immediately when finishes other interval is names as *EndBeg*. Given the intervals I_{c1} and I_{c2}, we say that both start at same time if:

$$i_1 < i_2,\ i_3 < i_4 \text{ and } i_2 = i_3$$

Finally, last relations describe the situation where both intervals have common parts, but not are equals and no finish or start at same time. Given the intervals I_{c1} and I_{c2}, we say that both start at same time if:

$$i_1 < i_2,\ i_3 < i_4 \text{ and } i_2 > i_3$$

and we denote this relation as I_{c1} *EndBeg* I_{c2}, and show them in Fig. 13.

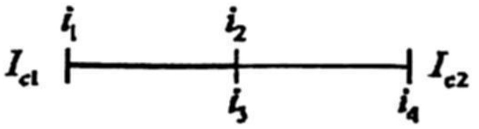


Fig. 13. I_{c1} *EndBeg* I_{c2}.

3.4 Relations Between Instantaneous Intervals

The possible relations between instantaneous intervals are just two. The first relations describe the situation where both intervals occur at same time. Then, Given the intervals I_{i1} and I_{i2}, we say that both start at same time if:

$$i_1 = i_2 = i_3 = i_4$$

and we denote this relation as I_{i1} *Equal* I_{i2}, and show them in Fig. 14.

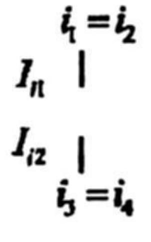


Fig. 14. I_{i1} *Equal* I_{i2}.

The second relation is the opposite of the above. This relation describes the situation when both intervals occur at different time. Given the intervals I_{i1} and I_{i2}, we say that both start at same time if:

$$i_1 = i_2 \text{ , } i_3 = i_4 \text{ and } i_2 < i_3$$

and we denote this relation as I_{i1} *Disjoint* I_{i2}, and show them in Fig. 15.

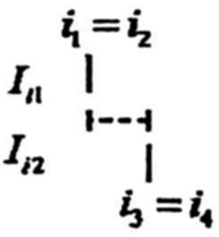


Fig. 15. I_{i1} *Disjoint* I_{i2}.

The next tables show the relations between different types of intervals. Table 1 shows relations between continuous intervals, Table 2 resume relations between instantaneous intervals and, finally, in Table 3 we resume relations that we believe could exists between continuous and instantaneous intervals.

Table 1. Relations between continuous intervals.

Relation	Meaning
Equal	Indicate given two intervals start and finish at same time.
EndEqual	Given two intervals, both finish at same time.
BegEqual	Given two intervals, both start at same time.
Inside/Cover	One interval is into other interval it is covered.
Disjoint	The intervals related have not common elements.
EndBegin	This relation indicates when the instant of an interval finishes, another starts immediately.
Overlap	Given two intervals, both share parts, but one finishes before other.

Table 2. Relations between instantaneous intervals.

Relation	Meaning
Equal	Indicate when both intervals occur at same time.
Disjoint	Indicate that both intervals occur at different time.

Table 3. Relations between continuous and instantaneous intervals.

Relation	Meaning
EndEqual	Indicate that an instantaneous interval occurs when finishes one continuous interval.
BegEqual	Indicate that an instantaneous interval occurs when starts one continuous interval.
Inside	One instantaneous interval occurs during the occurrence of one continuous interval.
Cover	One interval is into other interval it is covered.
Disjoint	The related intervals ~~related~~ ~~have not~~ does not have common elements.

4 Conclusions

It has worked to identify the elements that should be having the model to describe explicitly the changes that happening within a dynamic geographic environment. As part of these elements is to define the changes that occur on intervals, which can be of two sorts: continuous and instantaneous, resulting on relations between them. The next step is to define the relations that may exist between *Changes* and *CR*.

We believe that with these descriptions through of *CR*, we can make operations among *CR*s like union, intersection, and overlap. This can give us additional information about what happened in different descriptions at different times. As well as to make comparisons with patterns of set of *CR*'s and can detect differences among these to develop alert systems, for example.

As future work, is necessary to go up a conceptual level to define relations between events similarly to the relation between intervals. Also define granularities in temporal and conceptual terms.

Acknowledgements. The authors of this paper wish to thank the IPN, CIC, SIP and CONACYT for their support.

References

1. Al-Taha, K. and Barrera, R. Temporal data and GIS: And overview. In Proceedings of GIS/LIS '90, (Anaheim, CA: ASPRS/ACSM/AAG/URISA/AM/FM), pp. 244-254.
2. Campos, J. and Hornsby, K., Temporal constraints between cyclic geographic events, Proceedings of GeoInfo 2004, Campos do Jordao, Brazil, November 22-24, 2004.
3. Chang, S.K., Jungert, E. and Li, Y., "The Design of Pictorial Databases Based Upon the Theory of Symbolic Projections", Symposium on the Design and Implementation of Large Spatial Databases, Lecture Notes in Computer Science, Vol. 409, pages 303-323, Springer-Verlag, 1989.

4. Chrisman, N. Beyond the snapshot: changing the approach to change, error, and process. In Spatial and Temporal Reasoning in Geographic Information Systems, edited by M. Egenhofer and R. Golledge, Spatial Information Systems (New York, NY: Oxford University Press), pp. 85-93, 1998.
5. Claramunt, C. and Thériault, M. Managing time in GIS: an event-oriented approach. In Recent Advances in Temporal Databases, edited by J. Clifford and A. Tzunhilin, Berlin: Springer-Verlag, pp. 23-42, 1995.
6. Claramunt, C. and Thériault, M. Toward semantics for modeling spatio-temporal processes within GIS. In Proceedings of 7th International Symposium on Spatial Data Handling, edited by M. Kraak and M. Molenaar (Delft, NL, Taylor & Francis Ltd), pp. 47-63, 1996.
7. DiBiase, D., MacEachren, A., Krygier, J., and Reeves, C. Animation and he role of map design in scientific viualization Cartography and Geographic Information System, 19, 201-214 1992.
8. Egenhofer, M., "A Formal Definition of Binary Topological Relationships", Third International Conference on Foundations of Data Organization and Algorithms (FODO), Paris, France, Lecture Notes in Computer Science, Vol. 367, Springer-Verlag, pp. 457-472, June, 1989.
9. Egenhofer, M. and Herring, J., "A Mathematical Framework for the Definition of topological Relationships", Fourth International Symposium on Spatial Data Handling, pages 803-813, Zurich, Switzerland, 1990.
10. Frank, A., Qualitative temporal reasoning in GIS-ordered time scales. In Proceedings of Sixth International Symposium on Spatial Data Handling, edited by T. Waugh and R. Healey, Edinburgh, Scotland: pp. 410-431, 1994.
11. Freeman, J., "The modeling of Spatial Relations", Computer Graphics and Image Processing, 4:156-171, 1975.
12. Hernández, D., "Relative Representation of Spatial Knowledge: The 2-D Case", Cognitive and Linguistic Aspects of Geographic Space, Kluwer academic Publishers, Dordrecht (in press), 1991.
13. Kainz, S. F., "Logical consistency", Elements of Spatial Data Quality, pages 109-137, 1995.
14. Langran, G., Time in geographical information systems. London: Taylor and Francis, 1992.
15. Martínez, M.: Topological Descriptor to Topographic Maps, M. Sc. Thesis, Mexico, June 2006.
16. Peuquet, D. and Ci-Xiang, Z., "An algorithm to determine the directional relationship between arbitrarily-shaped polygons in the plane", Pattern Recognition 20(1): 65-74, 1987.
17. Peuquet, D. J., It's about time: A conceptual framework for the representation of temporal dynamics in geographic information systems. Annals of the Association of American Geographers, 84, 441-461, 1994.
18. Peuquet, D. and Wentz, E. An approach for time-based analysis of spatiotemporal data. In Proceedings of Sixth International Symposium of Spatial Data Handling, SDH '94, edited by T. C. Waugh and R. G. Healey (Edinburgh, Scotland: pp. 489-504, 1994.
19. Renolen, A., History graphs: Conceptual modeling of spatiotemporal data. In Proceedings of GIS Frontiers in Business and Science, (Brno, Czech Republic: International Cartographic Association), 1996.
20. Temporality in spatial databases. In Proceedings of GIS/LIS '88, (San Antonio, TX: ACSM/ASPRS/AAG/URISA), pp. 880-889.
21. Worboys, M. A unified model of spatial and temporal information. Computer Journal, 37, 26-34, 1994.
22. Worboys, M. F. and Hornsby, K., From objects to events: GEM, the geospatial event model, 2004.

A Tool for Designing Re-usable Process in Educational Software Development in Cooperative Environments

Garcia I. and Pacheco C.

Postgraduate Division, Technological University of the Mixtec Region
Huajuapan de León, Oaxaca (Mexico) *www.utm.mx*
ivan@mixteco.utm.mx, leninca@mixteco.utm.mx

Abstract. In the last few years, Educational Software has developed enormously, but a large part of this has been badly organized and poorly documented. Recent advances in the software technology can promote the cooperative learning that is a teaching strategy in which small teams, each composed by students of different levels of ability, use different learning activities to improve their understanding of a subject. How can we design Educational Software if we never learnt how to do it? This paper describes how the Technological University of the Mixtec Region is using a cooperative application to improve the quality of education offered to its students in the Educational Software design.

Keywords. Software reuse, educational software, cooperative learning, process reuse, process notation, process tailoring.

1 Introduction

As education and technology combine, the opportunities for teaching and learning are ever growing. However, the very rapid rate of change in the field of technology poses special problems for academic institutions, specifically for the engineering disciplines.

Nowadays, the software engineering and modern theories of learning converge in the construction of Educational Software (ES) to develop tools that define and implement educational objectives while preserving quality patterns. However, there are still gaps in our ability to assess whether a component or process meets the specified requirements or expectations and needs of students or groups of students. It is wrong to think that ES is less complex than commercial software, which has definitely received more attention from the field of software engineering.

ES covers a range of sub domains, types of systems, requirements and diverse idiosyncrasies which have been covered by the application of principles, specific methods and tools of software engineering in a specialized field. According to Van Schaik, *"clear and unambiguous emphasis on human learning and knowledge acquisition, differ from the educational software with other types of software"* [3]. Thus, ES is evidence of the technological impact on educational processes that has taken

© G. Sidorov, M. Aldape, M. Martínez, S. Torres. (Eds.)
Advances in Computer Science and Engineering.
Research in Computing Science 42, 2009, pp. 53-64

Received 06/03/09
Accepted 27/04/09
Final version 28/04/09

place in recent years, providing a valid alternative to students through an environment of generation (and regeneration) of knowledge.

From our perspective, ES development should incorporate a practical mechanism for designing and establishing effective activities for Instructional Design. The major objective is to facilitate and ensure the accomplishment of educational needs to a target audience, not forgetting to take into account its profile to edit the actual contents. It should also involve users identifying needs and/or specific problems and establishing mechanisms to provide adequate solid educational, communicative and computational principles [4]. However, a basic problem faced by the learning community is to determine how to develop and deliver quality content for learning experiences, while being able to compose, revise and update this content in an efficient way. This brings up the issue of reusability (content developed in one context being transferable to another context).

The profit of a high level of process reusability in software development is a sign of maturity in any discipline of engineering. Software engineering is no exception; in recent years reuse-based paradigms have contributed to the reduction of costs and schedules in the software industry. It is clear that in software engineering the compositional paradigm or component-based paradigm has dominated [20] [11] [17] [7] [10], but in the concrete case of ES, this situation is the opposite; and traditionally, the generative paradigm has been preserved [1] [13]. It is possible that reusability provides tangible benefits, but it is necessary to modify the actual process models in order to incorporate new activities, regarding to the compilation and coherent and centralized information maintenance of a specific domain.

Otherwise, we would stagnate in an ad hoc model where reusability would not generate more than meager benefits. One activity that has significant relevance in a model based on component reusability for software development is analysis and domain modeling. Domain analysis provides, among other products, a definition and a model itself that includes object identification, and common operations and relations among them. But, how to model this domain in an efficient way? The knowledge domain exists independently of the learner, and understanding is coming to know that which already exists. This knowledge can be learned, tested, and applied more or less independently of particular contexts. The use of software technology to support engaged learning goes hand in hand with the constructivism philosophy.

2 Cooperative Learning for Defining Reusable ES Processes

One method cited for accelerating process improvement in ES development is to replicate a standard, reusable process within other projects. However, the creation of a process that is usable on diverse projects is a difficult task. What is needed is an effective method to capture the common and variant elements of project-specific processes and create process definitions that can be applied in a variety of situations, i.e. that are reusable. But what is a reusable process? Feiler and Humphrey defined a process as *"a set of partially ordered steps intended to reach a goal"* [6].

A few years later, Hollenbach and Frakes defined process reuse as *"the usage of one process description in the creation of another process description. It is not multi-*

ple executions of the same process on a given project" [8]. These definitions could be applied to the educational systems context to create a type of graphical repository to produce and refine reusable processes and improve the quality of the final product at the same time.

A literature search shows that groundwork for SE process reuse exists. Research by Sheremetov et. al. [19] proposed the use of agents to expand the simple specification sequence of Instructional Management Systems integrated to SCORM v1.3. The research work of [21] conceptually describes modularity (granularity) of learning sequences, learning activities and actions, reusable learning objects and atoms, and reusable information atoms. Research by [15] relates a special type of labeled material called Intelligent Reusable Learning Components Object Oriented (IRLCOO), producing learning materials with interface and functionality standardized rich in multimedia, interactivity and feedback. In [2], Canales et. al. resumed the development of a Web-based Education System architecture that considers a diversity of requirements and provides the needed functionalities based on creating reusable components for content and evaluation tasks to reduce the complexity, change management, and reuse of learning.

We can observe that the creation of reusable processes is dependent on domain analysis. In spite of existing studies, not many domain analysis methods/tools for creating re-usable ES trough experiences of cooperative learning have been developed [14]. So, this paper presents a systematic and standard method for process reuse in educational systems projects using the "Learning by Doing" as strategy. The purpose of the process definition method is to create reusable ES processes within a repository which can tailor them to specific instructional and technical requirements in a cost effective manner. Reusable processes aid in transferring process knowledge between projects, reducing instructional costs, reducing effort, planning common educational projects and activities based on process data, and increasing quality in a continuously improving process oriented environment. The ES process life cycle contains the following steps:

- *Define reusable ES process from repository.* Even if a previous ES process does not exist, the student must start from here. The output is of one or more process descriptions, along with tailoring guidance and output products (see Process Manager in Section 3.1.2). The process descriptions are also integrated into the repository if they have been tested to ensure they are fit to (re)use.
- *Select process to adopt in new project.* Once the re-usable ES process is complete, the Process Manager is ready to deploy it in a real environment.
- *Tailor the ES reusable process on a project.* The re-usable ES process is tailored to meet the specific technical and instructional requirements and environment of an educational project.
- *Enact the process on the project.* The process is put into practice on the educational project. An assessment function evaluates the educational project to ensure that the new process is faithfully enacted.
- *Refine the process.* Based on the previous evaluation the process definition is then refined and inserted in the repository as a "good" process.

3 Building an Alternative Repository of ES Processes

A learning approach to ES development captures project-related information during the creation of individual ES artifacts which is then disseminated to subsequent educational projects to provide experience-based knowledge of development issues encountered at a repository of effective processes. These principles have been successfully applied in software engineering in [8] and [9], and we are trying to apply them in the ES context through an exploratory prototype, named ESPLib (Educational Software Process Library).

The initial ESPLib research focused on developing a tool to support the creation of domain repositories. In particular, we focused on defining precise processes to accomplish the instructional design. We assume the hypothesis that the most difficult aspect to students designing ES is the identification of the correct learning domain. They could be great programmers, capable and efficient, but they may not know how to represent in clear terms how to design the learning environment. ESPLib is composed of domains which are independent knowledge realms consisting of a set of activities and rules (reusable process) that define the context in which an ES process is applicable to a learning objective.

Our approach, illustrated in Figure 1, combines a rule-based system to match project characteristics to ESPLib processes and a deviation process to continuously update and improve new practices. At key points in a development procedure, students are taken through a set of questions designed to elicit educational project characteristics and match them to specific process elements in our schemas (or lifecycle models). This creates a customized process that goes through a review procedure validating the path taken through the rule-based system and assessing whether the processes assigned to the ES project are necessary and consistent, or are in need of further refinement or correction. The result is a modified process which should be followed.

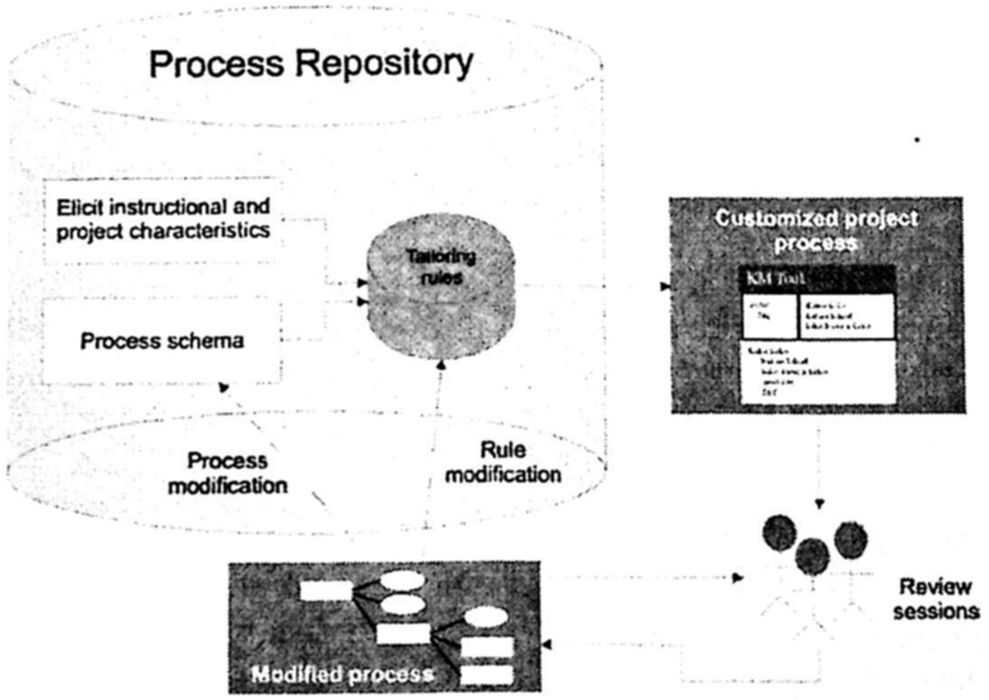


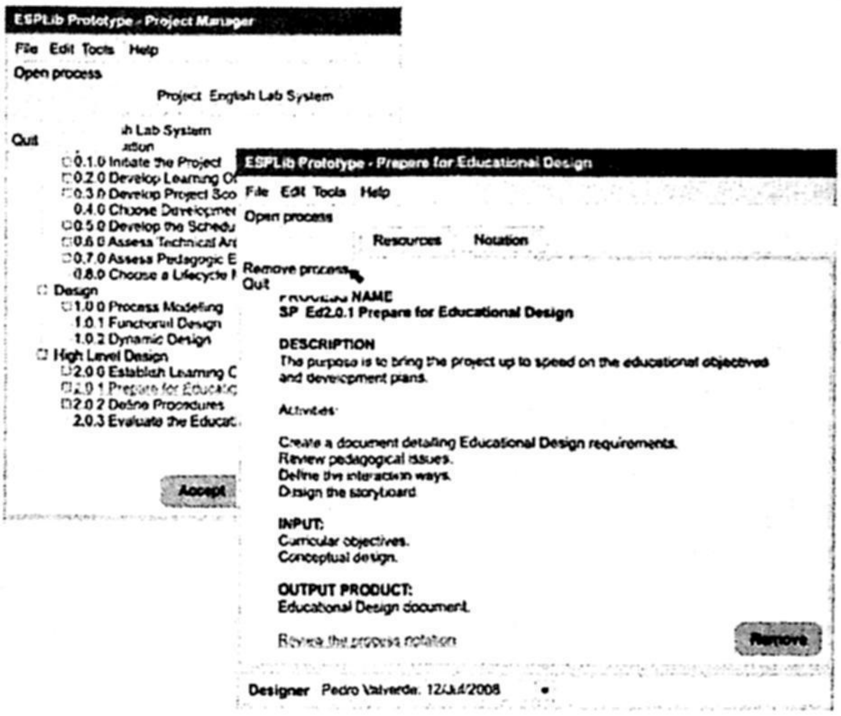
Figure 1. Tailoring and modifying ES process.

Cases are created to track conformance and document the development process for that ES project. In instances where deviations from the assigned process are requested, students must decide whether changes to rules and/or process elements are needed. In other words, a deviation from the current standard (defined by the rules and processes) sets a precedent that defines future actions under those circumstances. Thus it is clear that ESPLib implements the constructive theory of Jean Piaget: *you can learn if you can do it* [16].

The use of software technology can help the constructivist theory because it is possible to build a "Learning by Doing" environment that also combines the constructivist approach and the cooperative learning within a process and practice repository. The five elements in [5] define the cooperative learning: positive interdependence, individual accountability, face-to-face primitive interaction, appropriate use of collaborative skills, and group processing.

3.1 The ESPLib Tool

The main interfaces for ESPLib are shown in Figure 2. The Project Manager, shown to the left in Figure 2, displays a hierarchical arrangement of educational project activities.



PROCESS NAME
SP Ed2.0.1 Prepare for Educational Design

DESCRIPTION
The purpose is to bring the project up to speed on the educational objectives and development plans.

Activities:

Create a document detailing Educational Design requirements.
Review pedagogical issues.
Define the interaction ways.
Design the storyboard.

INPUT:
Curricular objectives.
Conceptual design.

OUTPUT PRODUCT:
Educational Design document.

Review the process notation

Designer Pedro Valverde. 12/Jul/2008

Figure 2. ESPLib Project Manager.

In the figure, a project named "English Lab System" has been chosen from the list of current projects. Each activity contains project-specific information, as shown in the window on the right of Figure 2, which was obtained by double-clicking on the activity named "Prepare for Educational Design" in the project hierarchy.

Processes are used in a rules-based decision support approach, and describe specific guides to establish development activities in ES projects. Guides consist of a description of a specific activity, the related activities, a description of input and output, and specific information about process notation. For example, in Figure 2, the process describes the specific activities of preparing educational design. This description was updated when the process was added to the repository of ESPLib

3.1.1 Creating Domains

As we said previously, ESPLib is composed of domains. Domains are independent knowledge realms that consist of a set of domain diagrams defining activities and domain conditions that define the context in which a process is applicable to a specific learning objective. Currently, educational projects belong to a single domain, which is chosen when a project is created in ESPLib. All subsequent project activities will use the activities and notations defined for that domain.

A domain defines the area of possible activities for educational projects within a given field (e.g. computer science, mathematics courses, electronic practices, learning a foreign language, etc.), structured in a work breakdown plan. They define standard activities that have proved useful for situations encountered by educational projects within that specific domain. The diagrams describe some of the problems or necessary steps that must be taken to ensure that the procedure is correctly followed, and can be updated by a project to reflect specific activities that this project follows.

The obtained descriptions are used to tailor the ES development process to the specific needs of different projects. Internally, ESPLib uses a simple forward chaining production mechanism implemented using an SQL database to represent alternative processes. The selection of a domain depends on a set of preconditions and events. Preconditions are represented by question/answer pairs. Questions are associated with the Questions tab of the Open or Add process Options. Questions are chosen by students to tailor the ESPLib's standard process to individual educational needs. These options address high-level project requirements which may influence which educational activities are chosen for the project.

When all preconditions of a domain evaluate to true, the Domain Manager "fires", causing a set of defined actions to be executed. The core actions in ESPLib can remove questions from the question stack, add a question to the question stack, or add a domain to a project. In addition, new diagram types can be created by using the Tool Option. Once the domain is identified, the student must describe and define the development process in formal terms.

With this multiple-choice questions ESPLib control the student's knowledge, on the processes repository, on the data database, and on the theory of software engineering.

3.1.2 Defining Notation

Process definitions are specified using the process definition component of the Process Manager and are based on a defined framework using notation.

Nowadays, in the context of ES, the research lines are focused on working with the specific components of systems, like content. There is an initiative that inculcates globalization of materials for its use in different learning programs and sessions through the use of "metadata". In this category, the proposal that most stands out is the Learning-Object Model (LOM) developed by IEEE-LTSC.

This model is composed of nine different records that identify general characteristics, lifecycles, metadata, technical and educational issues, relations with objects, and classification of resources [12]. ESPLib combines the LOM approach with the notation from Object Modeling Technique (OMT). OMT is one of the methodologies oriented to object analysis and design which is more mature and efficient than currently exists [18].

The great advantage of this methodology is its open nature (non-proprietary), which allows it to be in the public domain. This characteristic facilitates its evolution to match all current and future educational software needs. Process definitions are extended to include educational project information using the Process Manager component. Interactions between students and ESPLib occur through a graphical user interface (GUI). ESPLib's GUI consists of a main window containing a process display area, various pop-up dialogue windows which can be dismissed when no longer required, and a toolbox that can be displayed or hidden as required by selecting the Tool option. The GUI is illustrated in Figure 3 and is an example of an ESPLib's loaded process.

A process model must be defined using ESPLib before it can be saved. Process models are defined using the graphical notation and are constructed from a set of basic process elements which are displayed in the toolbox, shown to the right in Figure 3. ESPLib provides a notation that groups "process elements" in four types: phases, activities, transitions and documents.

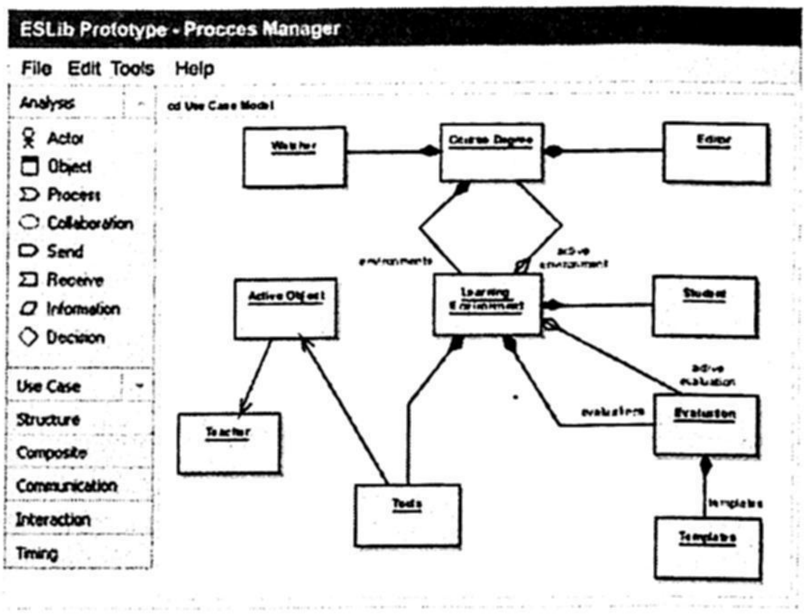


Figure 3. An Example of ES process in ESPLib.

A phase is a process element that is used to group other process elements, and can be used to build a hierarchical educational process definition by encapsulating other phases. It groups related activities together into a single unit and is often used as a synchronization mechanism. ESPLib offers two types of phases: a normal phase and a decision phase, called Phase and Decision respectively, in Figure 4.

A decision phase differs from a normal phase in that a Yes/No decision must be made as a result of undertaking this phase. A normal phase has no such requirement. An activity is a low level element which cannot be broken down further.

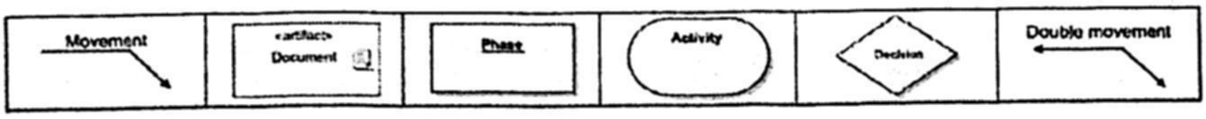


Figure 4. ESPLib notation.

A document process element is used to represent any artifact which is produced, including a source code and class documents, during the learning process. The final type of process element is a transition of which two types exist –a single directional arrow, called Movement, and a bi-directional arrow, called a Double Movement. Both types of transitions act as a connector between the elements of the process, with the transition from one process element to the next caused when the first is completed.

The single directional arrow is used to represent flow in one direction only towards the arrow head. The bidirectional arrow behaves the same as a single directional arrow in that information flows to the bold arrow head.

The difference between these two process elements is that a bidirectional arrow permits backtracking through a process if necessary. This could be triggered by omitting a learning objective or requirement in the previous analysis. A bidirectional arrow will allow an earlier phase to be re-executed to fix a problem. This arrow allows process definitions to be less cluttered than if explicit decision phases were required after every normal phase to determine if the previous phases were implemented correctly or if problems were identified. Each element has an expected duration and an indication of who is responsible for the execution of the process element. In addition to this, a document process element also contains the name of the document and where it is located in the learning process.

3.2 Creating and Modifying ESPLib's Processes

To define an ES process for use on a project, ESPLib allows an existing process model to be modified. Alternatively, the process can be defined starting with an empty process. Only one student of ESPLib is permitted to modify a software process at any one time, thus removing the possibility of multiple concurrent updates to the ES process definitions.

Process elements can be added to a process by selecting the required component in the toolbox and then drawing it into the process display using the mouse. Elements can be removed from a process by selecting the element and then choosing "cut" from the edit menu. When any modifications are made to a process the student is required to justify why the change was made. This is designed to provide some reasoning for the ES process that is to be followed for a particular project. The justification document for the process contains all the modifications, who made them, when they were made and the reason for the change. This file can be viewed by any student at any time.

Process models in ESPLib are defined hierarchically through the use of phases, reflecting the fact that a process consists of a series of levels. In ESPLib, it is possible to look inside a phase to see the elements that it contains, thus allowing activities in lower levels of the process hierarchy to be defined. Processes that do not conform to the defined process framework, which sets out rules which each process must abide by, are flagged as incomplete and may not be followed until they conform to the framework. The process framework is designed to assist in the description and modeling of educational software and is defined as part of ESPLib. This allows a process to be developed over a period of time. Storing process definitions allows previous models to be used as the basis for the construction of new process models.

ESPLib supports the dynamic modification of process definitions. This allows process instance to evolve throughout the lifetime of the educational project. This is especially important for projects of long duration. When a process definition is modified whilst it is being executed, the changes made to the process must still conform to the defined process framework. An example of a new process created in ESPLib is shown in Figure 5. Our students created a process to develop ES and help with foreign language classes in the English Lab. They modified an existing process (DIS12) and created an alternative process, the DIS13.

The Learning Environment phase was added because teachers considered that students might use computers, tapes, movies and books. The ESPLib's repository includes all the assets (templates and documents) collected to analyze a complete ES.

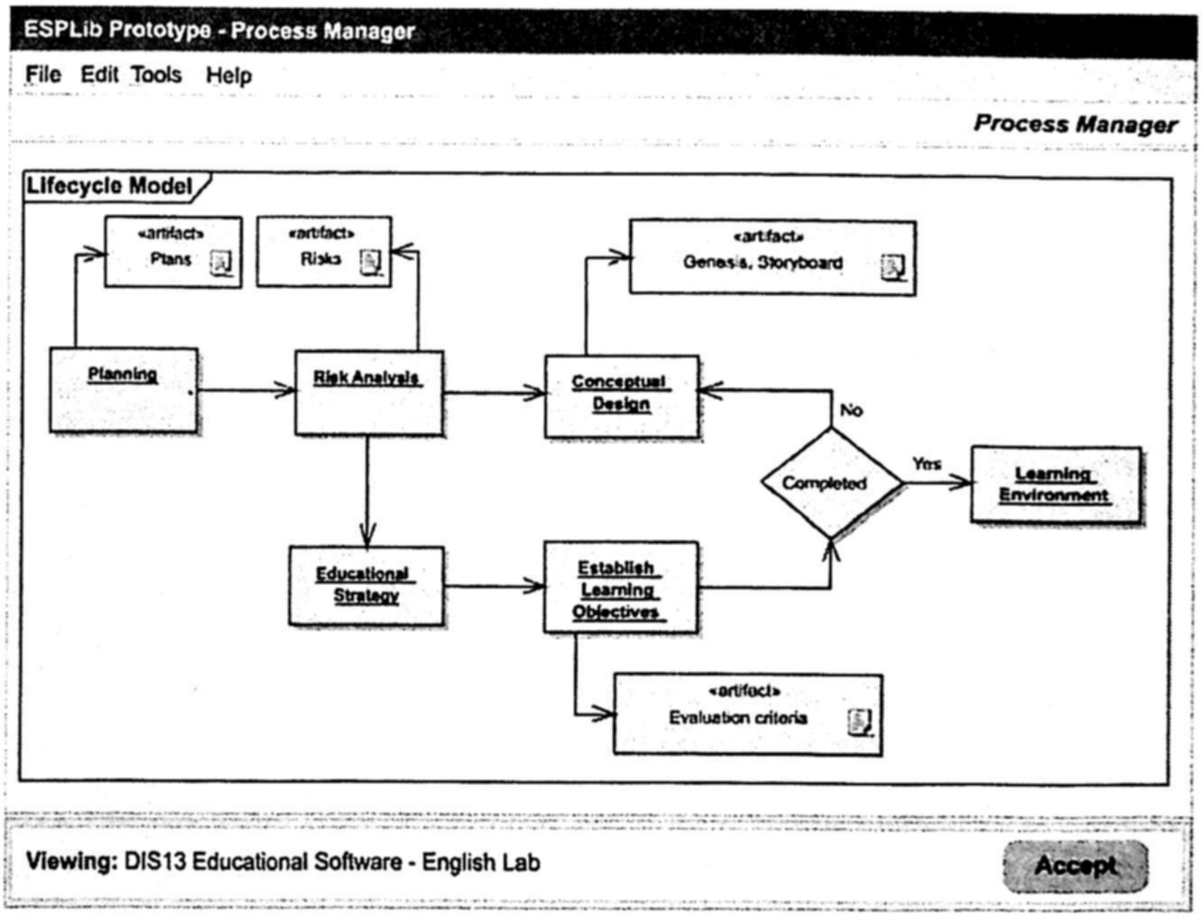


Figure 5. An ESPLib's process.

4 Evaluation and Future Work

As mentioned before, a tool for designing ES does not exist. There is a lot of information on how to improve the quality and performance of ES, but ESPLib attempts to "implement" that knowledge, in a real tool, to avoid the repetition of tasks already developed and typical errors. Traditional education is based upon a paradigm normally called the "knowledge reproduction model". This model is based on verbal lecture, drill and practice sessions, printed handouts, structured classroom activities, and office hours. In its pure form this model is grounded in the belief that knowledge is objective and the purpose of the teaching process is to transfer this static body of knowledge from its source to the student. The student, using this point of view, is seen as a passive learner "waiting to be filled" with knowledge, but the knowledge is not static. Learners have to actively interact with the learning environment and contents by browsing, searching, selecting, scanning and so on.

The general goal of ESPLib is to create educational knowledge management tools that are more proactive in delivering information to students than typical repositories. One way this can be accomplished is to create a tailorable process that provides domain-sensitive information to ES development efforts. The ESPLib tool and framework is flexible enough to bridge the gap between overly-restrictive ES development methodologies and ad-hoc practices to fit the needs of ES as they evolve.

The small number of process elements available for use in a process definition assists students understanding. Although only a small number of process elements exist, it is still possible to construct a process definition, such as the English Lab process, without the definition becoming overly cluttered.

It is possible in ESPLib to add additional attributes to a process element or to customize process elements. This is a valuable feature in a process modeling tool, allowing process definitions to be customized to the style used by any student. ESPLib allows dynamic modification of a process definition. This is a very valuable feature as process definitions are dynamic in nature and may change throughout the lifetime of a project. Another positive feature of ESPLib's process definitions is that they are easy to understand by all students due to the diagram notations. The process framework is not currently customizable making it impossible for any student to specialize the process rules to incorporate their specific policies. Storing the process definitions allows past process models to be examined and evaluated. This can be valuable when a process model is being chosen for use on an educational project because a successful process can be chosen.

It was clear from our brief study that the domain interface of ESPLib (Domain Manager) needs some improvements. In particular, there needs to be a method of finding answers to some behaviors (such as a case being added to a project) or other attributes.

Currently, students need to page through the questions one-by-one to find an answer with the desired precondition or action. Another improvement is related to the addition of a "content editor" module, the main idea being to relate each phase to a topic from the normal course and include (or modify) it according to the students needs.

5 Conclusions

Knowledge management for educational software development is more than just repositories and models. It also actively requires work between pedagogue and software specialists delivering information during the development process and ongoing process of capturing project experiences.

Using the knowledge management techniques can prevent the duplication of efforts, avoid repeating common mistakes, and help streamline the development process. The reviewed approaches have shown that the reuse process provides relevant useful and up-to-date information to improve the quality of ES. This mandates a strong tie between technologies and pedagogy in which using the technology must become part of routine work activities.

The contribution of this research pretends to define and implement a single tool to improve the analysis and design of ES using the cooperative learning approach, with the intention of successfully merging pedagogical and technical aspects equitably. As the repository of ESPLib grows through the principled evolution of the knowledge domain, it also becomes able to handle a wider range of domains, while evolving towards answers to problems that fit the student's technical and pedagogical context. The real question is not whether the repository of ESPLib is "correct" in some objective sense, but rather whether less mistakes are repeated and better ES solutions adopted when using the repository. Students who will work with the ESPLib' cooperative learning environment and who will have opportunities to work cooperatively with students who have different ability, ethnicity, gender, and so forth will be better able to build positively interdependent relationships than students who will have only an individualistic and a competitive learning.

References

1. Biggerstaff, T. "A perspective of Generative Reuse" Annals of Software Engineering. 5:169-226. 1998.
2. Canales, A., Peña, A., Peredo, R., Sosa, H., & Gutierrez, A. "Adaptive and intelligent web based education system: Towards an integral architecture and framework" Expert Systems with Applications 33: 1076–1089. 2007.
3. De Diana, I y Van Schaik, P. "Courseware Engineering Outlined: An overview of some research segues". ETTI 30(3): 191-211. 1993.
4. Díaz, M., Pérez, M., Grimmán, A. & Mendoza, Luis. "Proposal for Development of a Methodology for Educational Software under a systemic quality approach". Universidad Simón Bolivar. Venezuela. 2005. (in Spanish).
5. Felder, R. M. & Brent, R. "Cooperative Learning in Technical Courses: Procedures, Pitfalls, and Payoffs" ERIC Document Reproduction Service, ED 377038, 1994.
6. Feiler, P. & Humphrey, W. "Software process development and enactment: Concepts and Definitions" Software Engineering Institute CMU/SEI-92-TR-04. Pittsburgh, PA. September 1992.
7. Fiorini, S., Leite, J., & Lucena, C. "Process Reuse Architecture" Proceedings of the 13th International Conference CAISE 2001. Advanced Information Systems Engineering, Lecture Notes in Computer Science 2068:284-298. 2001.

8. Henninger, S., Lappala, K. & Raghavendran, A. "An Organizational Learning Approach to Domain Analysis", Proceeding of the 17th International Conference on Software Engineering, 95-104. 1995.
9. Henninger, S. "Case-Based Knowledge Management Tools for Software Development", Journal of Automated Software Engineering, 4: 319-340. 1997.
10. Henninger, S. "Tool Support for Experience-Based Methodologies" Proceedings of the 4th International Workshop on Learning Software Organizations, Lecture Notes in Computer Science 2640. 2003.
11. Hutchens, K., Oudshoorn, M. & Maciunas, K. "Web-Based Software Engineering Process Management" Proceedings of the Thirtieth Annual Hawaii International Conference on System Sciences (HICSS), 1: 676. 1997.
12. IEEE/LTSC. Institute of Electrical and Electronic Engineers, Inc./Learning Object Model. URL: http://www.ieee.org/lom
13. Okamoto, T. "The Model of Collaborative Learning and Technological Environment for Evoking Interactivity-Building of Knowledge". Proceedings of the IASTED International Conference, Computers and Advanced Technology in Education; Rhodes, Greece, 2003.
14. Peña, A., Sossa, J. "Web-based Education: A state of the art" UPIICSA and Computer Research Center IPN. 2005. (in Spanish).
15. Peredo, R., Ocaña, L. & Sheremetov, L. "Development of intelligent reusable learning objects for web-based education systems" Expert Systems with Applications. 28: 273-283. 2005.
16. Piaget, J. The Psychology of Intelligence. London Routledge. 2001.
17. Reis, R., Lima Reis, C. & Nunes, D. "APSEE-Reuse: A Case-Based Reasoning Model for Reuse and Classification of Software Process Assets" Proceedings of 7th International Workshop on Groupware (CRIWG'01). 2001.
18. Rumbaugh, J., Blaha, M., Premerlani, W. & Eddy, F. Modeling & Design Object Oriented. Prentice Hall. 1997
19. Sheremetov, L. & Peredo, R. "Development of Reusable Learning Materials for WBE using Intelligent Components & Agents", Technical Report, Instituto Mexicano del Petróleo and Laboratorio de Agentes del Centro de Investigación en Computación, México. 2002. (in Spanish).
20. Succi, G., Benedicenti, L., Predonzazi, P. & Vernazza, T. "Standardizing the Reuse of Software Processes" StandardView, 5(2): 74-83. June 1997.
21. Vladimir, U. & Maria, U. "Reusable learning objects approach to web-based education" International Journal of Computer and Applications, 25(3). 2003.

CLASS-W, a Grammar for Security System Development based on Environment Equipment Agents with Windows-Technology NT

Guadalupe Cota[1], Pedro Flores[1] and Joel Suárez[2]

[1] Departamento de Matemáticas, Universidad de Sonora, Hermosillo, Sonora, México, Código Postal 83000, lcota@hades.mat.uson.mx, pflores@hades.mat.uson.mx
[2] Centro de Investigación en Tecnologías de Información y Sistemas, Universidad Autónoma del Estado de Hidalgo, Pachuca de Soto, Hidalgo, México, CP 42000, jsuarez@reduaeh.uaeh.mx

Abstract. The issues regarding the development of security systems and "zero-day threats" are issues of concern to experts, individuals and public or private organizations, since new forms of intrusion used by malicious users or programs that are not recognized by protective tools constantly appear, and they are entered into the computer equipment to make them vulnerable. Taking this into account, and that no evidence was found on grammar for security systems development based on agents, this article describes an original proposal for the CLASS-W grammar, which provides mechanisms to coordinate joint work and feedback on operating system information about services, etc. For testing purposes, a system of agents has been developed to communicate and detect abnormal activity of ports and tasks that are loaded with the operating system.

Keywords: Grammar, Agent, Security, Windows.

1 Introduction

It is to be recognized that although there are software tools that attempt to resolve computer security problems, there is still the problem of the 'zero-day attack', the period which starts from the appearance of 'malware' and is not detected by the security tools database, which persists until the problem that is resolved. This vulnerability is exploited by viruses, worms, or users who seek to infiltrate into the computer systems to commit cyber crimes [1], [2] [3].

On the other hand, we must recognize that although various alternatives for solving this problem exists [4], [5], one of the most important related with it is Intrusion Detection Systems [6], that analyze network traffic and issue security alerts for detect possible "attacks", and nevertheless that these systems have been effective, they have serious drawbacks, both in its operation and in the complexity of their management, that, in addition to generating a large number of false alerts, and for analyzing the information traffic on the network, only based on records of the knowledge base they have, which is why they cannot avoid the damage that occur for unknown events, until the administrator detects them, finds a solution and updates the data, which can

Received 16/03/09
Accepted 27/04/09
Final version 06/05/09

be done in an indefinite period of time, which can range from one day to months or even years [7].

In this work it is suggested to use a different approach that consists of using CLASS_W (Content-Language Windows Security System) grammar, applying the agent of communication language (ACL) [8], defined as the standard by the Foundation for Intelligent Physical Agents (FIPA) [9], and implementing security levels using knowledge base based on the rules pre-defined by the computer network administrator.

For purposes of organization, this document is divided into six sections: the first corresponds to this introduction; the second, in view of the extensive of the grammar, we only mention the most relevant grammatical aspects of CLASS-W; the third describes the functional diagram of agents, which is represented by logic modules and profiles or roles specified; the fourth describes the form knowledge representation used by agents, and the fifth, is an example of the use of grammar using the agent platform JADE (Java Agent Development Framework) [10] and ACL.

2 Grammatical Aspects

CLASS-W has been designed based on the theory related to Context Free Grammar, which will be referenced hereafter as ***CFG*** [11], [12] and will be denoted by ***G*** as a quadruple $\boldsymbol{G} = (\boldsymbol{NT}, \boldsymbol{T}, \boldsymbol{P}, \boldsymbol{\sigma})$, where:

NT Not terminal symbols.

T Terminal symbols.

$\boldsymbol{\sigma}$ Initial symbol. ($\boldsymbol{\sigma} \in N$).

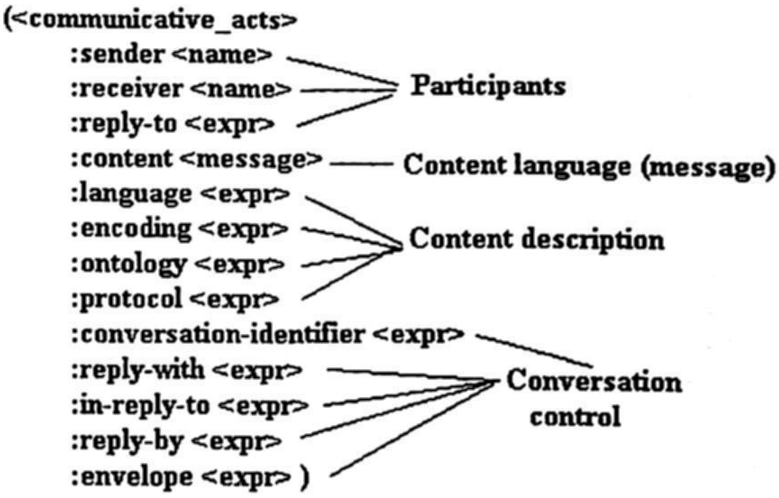
P Rules o grammaticals productions where:
$A \rightarrow a$ and $A \in NT$ and $a \in (NT \vee T)^{+}$.

To make the syntax description of ***CFG***, the Normal Form Extended Backus-Naur (***EBNF***) [13] will be used. The basic notation that is shown in table 1 is to specify the operations of a general nature such as the specification of the initial symbol, terminal and not terminal symbols, etc.

Table 1. Basic notation in *EBNF* for CLASS-W

Notation	Description
|	'Or'
{ t }	0 or more elements t
[t]	Optionality on t
,	'And'
< t >	No Terminal symbol t
::=	Involvement
....	Values to define
N	Maximum number of items

The language of content that will be generated by CLASS-W is implemented on the label ***'content'*** within the framework of the basic format ACL (see Figure 1), where communicative acts ***'sender'***, ***'receiver'*** and ***'content'*** are mandatory elements and the rest is optional.



```
(<communicative_acts>
     :sender <name>
     :receiver <name>
     :reply-to <expr>
     :content <message>
     :language <expr>
     :encoding <expr>
     :ontology <expr>
     :protocol <expr>
     :conversation-identifier <expr>
     :reply-with <expr>
     :in-reply-to <expr>
     :reply-by <expr>
     :envelope <expr> )
```

Figure 1. Basic format ACL.

The interpretations of the symbols of *CFG* that are extensive in CLASS-W are:

T Divided into a priori defined terminal symbols and terminal symbols which are referenced by the system security administrator to work on the security context of the operating system Windows-Technology NT, and the remaining will be stored in a database designed for this case (see Figure 3 and 5).

σ Refering to the scheme denoted by ACL ***'content'***, the initial symbol of CLASS-W is constituted by the non-terminal **<content>**

Below are specifications for the units of syntactic CLASS-W, which allow to write a program to control and define transactions sessions, instructions, alerts and safety record.

2.1 General Structure of a Program

```
<content>::= '(' < message > ')'
<message>::= '#'<key>'/' 'message''/'
          <transmitter_agent> '/' <type_agent> '/'
          <receiver_agent> '/' <type_conversacion> '/'
          <type_transaction> '/'
```

<type_transaction> is the non-terminal symbol, which is used to control the messages that are generated on the events.

2.2 Examples Transactions

Sessions:

```
<session>::='initial_s'<initial_session>
    |'close_s' <close_session> |'estab_s'
    <established_session> |'reject_s' <reject_session>
    |'error_s' <error_session>
<initial_session>::= '/' <agent>
<close_session>::= '/' <agent>
<reject_session>::=  '/' <agent>
<established_session>::= '/'<agent>'/' 'ref' '/' <agent>
<error_session>::=  <string>
<string>::= {<letter>}+
<letter>::= 'a'| ... |'z'|'A' | ... |'Z' |'0' |...|'9'
```

In the context of agents, this conversation can be represented according to the rules of protocol 'request protocol' FIPA [9] (see Figure 2) which is used when an agent asks another for an action and the target agent accepts, rejects or informs on the state of the communication. In cases 1, 4 and 5, when there are errors, rejections or not responses, the issuer must set a reasonable time to reschedule the request.

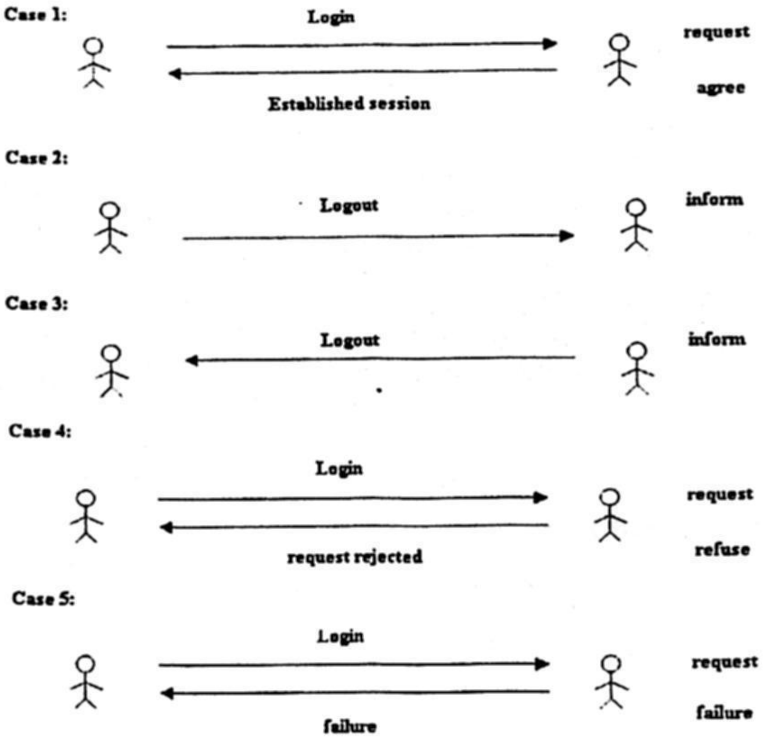


Figure 2. Session conversation represented according to the rules of request protocol FIPA.

Instructions:

```
<instruction> ::= 'stop'<stopr>|'active' <active>
    |'review' <review> |'enable' <enable>
```

```
|'disable'<disable> |'create' <create>
|'delete'<delete> |'modify'<modify>
|'update' <update> |'searchr' <search>
|'clasify' <clasify>
```

The instructions are related to actions that apply specifically on services, ports, applications, processes, policies, initial tasks, registry, network settings, etc, which are defined as ‘topics’ in the corresponding database.

Alerts:

```
<alert>::= 'alert''/'<id_alert>'/'<agent>
    '/'<level_security>'/'<ontology_topics>'/'
    <inf_ontology> '/'<action>
<inf_ontology>::= <list_processes> |<list_ports>
    |<list_services>|<list_task_initials>
    |<list_polices>|<list_changes_network>
<result_alert>::=<agent>'/'<id_alert> '/'
    <date>'/'<time>
```

The alerts are controlled by coordinator agents, who define the actions to be taken in cases that are considered harmful and could stand spread through computers on the network controlled by the agents system.

Definitions of Security:

```
<registry_security>::= <id_message>'/'<agent>
    '/'<extent>'/'<ontology_topics> '/'<date>
    '/'<time>'/'<level_security>'/'
    <data_part_extern > '/'{´,´<action>}+
<level_security>::= 0 |1|2|3|....|N
<action>::= In database.
```

where **<level_security>** can take the values of [1 ,...., N]: {low, medium, medium high, high, ..., N} (see Figure 3) .

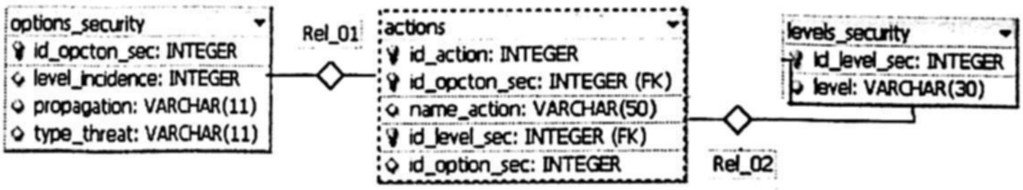


Figure 3. Relational schema for specifications of security options, levels and actions.

Interpreting the values as follows:

a. Type of communication [0-2]: {private, public, selective}
b. Incidence: {0-1, 2-10 and 10 onwards}
c. Propagation: {yes, no}
d. Type of threat: {internal, external}

3 Functional Diagram Agents

The overall operation of the system of agents is based on the detection of events that occur in their environment, collected through sensors, such as network cards and operating system resources [14] (see Figure 4), and is recorded on a blackboard [15] organized by topics, implemented in a database to feed the knowledge that is used in settings of security levels in the managed system.

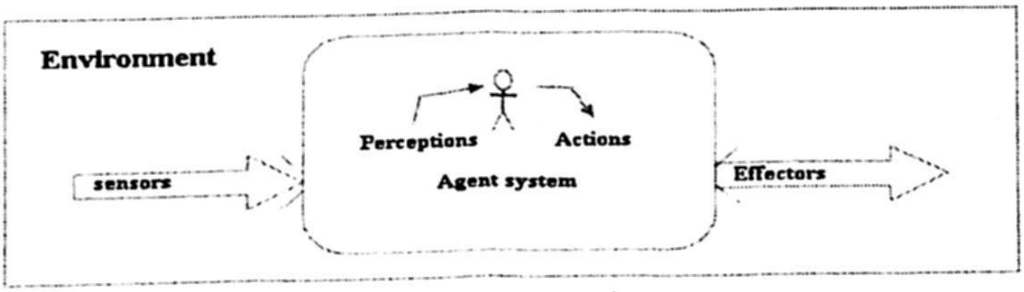


Figure 4. Agent system environment.

As one of the basic elements of CLASS-W is the blackboard, for which was created in database a group of tables (see Figure 5) and the grammar productions that are listed below:

```
<blackboard>::=<registry_blackboard>
    |<notify_blackboard> | <end_blackboard>
    |<wait_blackboard> | <operation_blackboard>
<registry_blackboard>::='registry_blackboard' '/'
    <agent> '/' <ontology_topics>
<notify_blackboard>::='notify_blackboard' '/'
    <ontology_topics>
<end_blackboard>::='end_blackboard' '/'<ontology_topics>
<operation_blackboard>::='id_blackboard' '/'<namber> '/'
    <agent> '/' <ontology_f>'/'<inf_aditional>'/'
    <type_aport> | 'id_blackboard' '/'
    <namber>'/'<agent> '/'<ontology_f>'/'<type_aport>
<type_aport>::='query' '/'<string>|'resp_yes' '/'<string>
    |'resp_no' '/'<string>|'resp_informative' '/'<string>
    |'resp_knowledge' '/'<string>|'proposal' '/'<string>
    |'intention' '/'<string>|'desire' '/'<string>
    |'data_aditional' '/'<string>
```

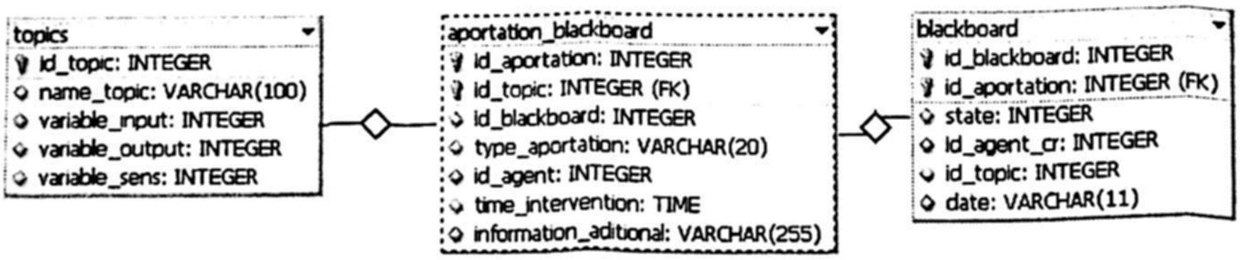


Figure 5. Relational schema for specifications to blackboard.

where the valued to be taken (attribute-table) can be:
(blackboard-state) [1-3]: {initiated, resolved, unfinished}
(blackboard-operation) [1-2]: {read, write}
(aportation_blackboard-type_aportation) [1-9]: {query, yes, no, response knowledge, information, proposal, intention, desire, additional data}
and
(topics-id_topic) [0 to N].
which will be provided by the administrator. The default values are: {processes, ports, services, startup tasks, policies, configuration, ..., N}.

3.1 Roles of Agents

In the hierarchy designed for the system of agents and implementation of CLASS-W, the entities of Administrator Multiagent System (AMS) and Directory Facilitator (DF) are mandatory elements in the scheme of FIPA [9] and JADE [10].

Below is a description of the type of agents that can be created under the specifications of CLASS-W, in addition to the already mentioned from the prior paragraph:

- 'Administrator'*.- Manages the overall operation of the agents system.
- 'Directory facilitator'*.- Manages records of agents and services.
- 'coordinator'.- Coordinates specialized work of multiagente system.
- 'network'.- Monitors performance of network.
- 'local'.- Responsible for monitoring internal function on equipment with different specification.

Note *: Mandatory elements in the scheme of FIPA agent.

4 Representation of Knowledge

As the logic is one of the main bases in the area of mathematics and computer science, and taking into account that the formalization of knowledge and the automation of the forms of reasoning are essential in many scientific development or technology areas, and has been very helpful, especially in the area of Artificial Intelligence [16]. For the implementation of W-CLASS, a section through the non-terminal <rule> is included, which allows representing knowledge using rules type Prolog in a functional scheme of system agents based on a cognitive profile that has the following characteristics [17]:

a) There is a mechanism to interact with each other to agents through a blackboard.
b) It implements a hierarchy of agents assigned to a role that defines their capabilities, limitations and assigned tasks to each of them.
c) It represents knowledge through the beliefs that an agent has over itself, relying on a knowledge base contained in a relational database.

d) It uses the knowledge recorded by the agents system to prevent or resolve problems that arise with the support in the analysis of a blackboard organized by topics and in the evaluation of rules contained in the logical modules, which has a relation to an instruction and an action or actions to be performed, when required.

The grammatical rules that correspond to this section are the following:

```
<programming_knowledge >::= ‘insert_rule’ ‘/’ <rule>
      |‘evaluate_rule’ ‘/’<rule>|‘search_rule’ ‘/’ <rule>
      |‘insert_condition ‘/’<condition>
      |‘evaluate_condition’ ‘/’<condition>
      |‘search_condition’ ‘/’<condition>
```

For the generation of rules for testing purposes, there is a module that has the following characteristics:

a. Translate a Prolog predicate to relational database including, in addition to information, the structure of predicates and relations.
b. Has been templates designed to generate Structured Query Language (SQL) [18], and interpret the validate of the information in the consultations undertaken.
c. Interact with the user through a visual interface that lets you enter and retrieve information from relational database designed for the module

5 Practical Examples Implemented on CLASS-W

A. The first example described in this section is built based on the behavior of port 135, which has been taken to this effect, because although it is used for services like Remote Procedure Call (RPC) and Message Queue Service (MQS), it has also been used by "worm" type programs to enter itself into systems with Windows environment [2], [3], [19]. In 2003, the worm W32/Lovsan.worm [20] and W32.Blaster.Worm [21, [22], have exploited the RPC vulnerability, which could allow remote code execution without authentication in affected system that received a specially crafted RPC request. In October 2008, the worm Conficker [23], is spreaded through networks at alarming rates via Microsoft Windows Server Service, exploiting RPC Handling Remote Code Execution Vulnerability and it was reported on March 2009 as Win32/Conficker.D.

The multiagent system that is created for the implementation of W-CLASS on JADE [10], in case analysis of blackboard on topic ‘ports’ is based in the following options:

- **‘Definition Rules’**.- Generation of script for introduced in database through a logical module the rule ***‘attackworm1’***, which is associated to port 135 and with a particular remote address network:
  Semantic predicates:
  `attackworm1(n_topic,n_port,n_connections,n_action).`

```
ontology_topic(n_topic).
port(n_port).
connections(n_connections).
action(naction).
```

Rules:

```
attackworm1(n_topic,n_port,n_connections,naction):-
ontology_topic(n_topic),port(n_port),
connections(n_connections), action(naction).
```

Facts:

```
ontology_topic("ports").port("135").
connections("1000"). action("stop").
action("review").
```

- **'Register of events'**.- The agents can register in blackboard events related with behavior of port 135 and address network associated.
- **'Evaluate rules'**.- The procedure for executing this consultation consists in:
  a. **'search rules'** previously defined through a logical module, which regulate use of the port that mentioned above
  b. **'Evaluate rules**' based on contributions that exist in the records of blackboard 'ports' with value 135:

  attackworm1("ports","135" , "1500", naction).

  If rule is true 'naction' returns values 'stop' and 'review', which refering to packets network of IP address registered in blackboard.
- **'Analyze blackboard'**.- Steps of the operation related to the topic 'ports':
  a. Search rules specified for the existing records on reported ports.
  b. Analyze state of 'alerts' on 'rules' specified by the administrator in the logic modules on the use of ports.
  c. View criteria to detect situations that may coincide with a pattern that is considered abnormal in relation to the use of port 135.
  d. Give instructions to the appropriate agents in order to execute certain actions previously defined and stipulated in the rules that have been set by the administrator in the logic modules, and these in turn recorded on the blackboard, their results are obtained from the same applications. An example of a message CLASS-W in order to send instructions after the evaluation rule ***'attackworm1'*** and review records of blackboard is:

Messages sent for coordinator agent ('CAgent01'):

```
"#0001/message/CAgent01/Coordinator/LAgent10/
private/instruction/stop/packets_network/135/"
"#0001/message/CAgent01/Coordinator/LAgent10/
private/instruction/review//135/"
```

Message sent for local agent ('LAgent10'):

```
"#0001/message/CAgent01/Coordinator/LAgent10/
private/response/agree/"
```

B. In order to present a second example based on references of virus and worms related with features that allow them to initiated automatically when the operating system is loaded, using the registry where are tasks of system startup.

The situation described in this section relates to the plan that can be implemented when a 'unknown' process has been registered in the task start and is considered to be a potential security problem. For purposes of illustration is used a current problem and some features of the Conficker Worm [23], which creates the files 'hloader_exe.exe' and 'hloader_dll.dll' in the system files folder ("c:\winnt", "c:\winnt\system32", "c:\windows\system"), inject this in the file 'explorer.exe' and recording with the following keys values:

```
HKEY_LOCAL_MACHINE/software/microsoft/windows/
currentversion/run/auto__hloader__key=c:\windows\system32
\hloader_exe.exe
HKEY_CURREN_USER/software/microsoft/windows/
currentversion/run/auto__hloader__key=c:\windows\system32
\hloader_exe.exe
```

Here is a scenario for the system of agents related to the topic 'startup_tasks', which can be implemented with CLASS-W:

a. Local agents have a registry of tasks authorized which will be loaded at system startup, and a list of processes that can run at any time.

b. Some local agents 'LAgent2', 'LAgent6' and 'LAgent10' have been found that the process 'hloader_exe.exe' is included as a startup task in the registry, and is not found on the list of authorized processes.

c. Local agents mentioned in the previous write on the blackboard, under the topic 'startup_tasks', its identification data and the process name 'hloader_exe.exe' and is classified as 'unkown task'

d. The coordinator agent 'CAgent01' revised the blackboard 'startup_tasks', and finds reported problem on 'hloader_exe.exe' and proceeds to do the following actions:

If there are rules associated with the process 'hloader_exe.exe' then the coordinator agent send instructions to the local agents which reported the problem. If not, the following instructions are sent to the local agents:

1. Check if the process 'hloader_exe.exe' is active, and if so, disable it until it receives a new instruction to confirms or revoke the action.
2. Remove strings located in the registry in startup tasks that are related to the file 'hloader_exe.exe'.
3. Check that the 'hloader_exe.exe' process is blocked and that it is not in execution.
4. Record a message in the administrator container with the following information: topic 'startup_tasks', process name 'hloader_exe.exe' which has been regarded as unknown, and 'preventive' instructions were sent to local agents who reported the problem.

When the administrator reviews the container, a decision should be made whether to create a rule with the actions that need to be implemented or deactivated the precautionary measures, and for the process named 'hloader_exe.exe' will be registered the respective indication on the blackboard 'instructions'.

Later, when the coordinator agent reviews the blackboard 'instructions', turns the respective actions.

A plan like the above can help to avoid the problem 'zero day', particularly relating to the unknown process that has been registered as startup tasks and can be a potential threat.

6 Conclusions

The security in computer networks is a topic that is emerging day by day due to the massive data management, though it has increased the productivity and wealth of all kinds of organisms or businesses, it has also been the means by which have led to significant financial losses, since as the countries make use of these type of services, the goal becomes more vulnerable to attacks based on the information. To counter and control this problem, it is necessary to adopt safety policies, based on computer technology, that provide the possibility of setting safety standards that allow to increase the level of confidence in the exchange of information through this medium.

Although the vulnerability of the systems cannot be eliminated completely, and therefore, the timely detection of security problems plays an important role, the "zero-day threats" are what most worry security experts.

Among the various alternatives that exist to control this problem, are those based on Intrusion Detection Systems that allow monitoring events that might be associated with malicious or harmful actions, but have the disadvantage of generating a large number of alerts false, and which do not prevent attacks that arise when these are unknown to the databases used to support the application.

With a different approach to the current, this presentation introduces the proposal to use a grammar such as CLASS-W as a tool to allow an expert, an administrator or developer of security systems, manage or develop system bases on agents that provide the possibility to harness and apply the experience taken in the area, in addition to using information obtained from the operating system based on Windows NT Technology, network resources, configuration, loaded applications start, the existence or absence of processes, use of ports, etc.

Finally, it should be added that although there are different types of grammars for management agents, we not found evidence of a grammar for development agent based systems in the area of security, and trust that the alternative is presented in this work, which may mark the way forward in developing such systems as the technology of agents has proven to be very efficient in other areas, and surpass other techniques or methodologies which are not achieved the same results.

References

1. McAfee Proven Security, Microsoft Word 0-Day Vulnerability III, http://vil.nai.com/vil/Content/v_vul27264.htm (2006).
2. Clark R.: How to cope with security radical changes in cyberspace, July 6, 2004, ID: 4198,
3. http://www.symantec.com/region/mx/enterprisesecurity/content/expert/LAM_4198.html. (in spanish).
4. Richard D. Pethia: Viruses and Worms: What Can We Do About Them?, Software Engineering Institute, Carnegie Mellon University, CERT® Coordination Center, Pittsburgh, Pennsylvania, USA, September 2003,
5. http://www.cert.org/congressional_testimony/Pethia-Testimony-9-10-2003/.
6. Bruce Schneier, SIMS: Solution, or Part of the Problem?, IEEE Security and Privacy, vol. 02, no. 5, pp. 88 (2004),
7. http://doi.ieeecomputersociety.org/10.1109/MSP.2004.83.
8. Cox, P. & Sheldon, T.: Windows 2000, Security Manual, pp. 18-24, 22-34, 44, 47, 54-81, 177, 410-424, 475, 675, 712-717. McGraw-Hill (2003). (in spanish).
9. The Universal.com.mx (Computer), is spreading the worm on YouTube, June 18, 2007, http://www.el-universal.com.mx/articulos/40642.html. (in spanish).
10. Snort (Intrusion Detection System), http://www.snort.org/about_snort/.
11. FIPA–Agent Communication Language Specifications,
12. http://www.fipa.org/repository/aclspecs.html.
13. FIPA – The Foundation for Intelligent Physicals Agents, http://www.fipa.org.
14. JADE – Java Agent Development Network, http://jade.tilab.com/.
15. Kelley Dean: Theory of Automata and Formal Languages, pp. 30, 105-170. Prentice Hall (1995). (in spanish).
16. Grune, Dick, Bal, Henri E., Jacobs, Ceriel J. H., and Langendoen, Koen G. Modern Compiler Design, pp. 548-596. McGraw-Hill/Interamericana de España, S.A.U, (2007). (in spanish).
17. V. Aho Alfred, Sethi Ravi & D. Ullman Jeffrey: Modern Compiler Design, pp. 25-82. Addison Wesley Longman, Pearson (1998). (in spanish).
18. Russell S. & Norving P.: Artificial Intelligence (A Modern Approach), pp. 36-45, 57, 110, 151, 157, 161, 163, 165, 166, 171, 196-201, 304, 419. Prentice-Hall (1995).
19. Carver N. and Lesser V.: The evolution of blackboard control architectures, pp. 19.23, Technical report, University of Massachusetts Amherst, October 1992. http://www.cs.siu.edu/~carver/ps-files/tr92-71.ps.gz.
20. Moore C. Robert. Logic and Representation. CSLI Lecture Notes nº 39. CSLI Publications, Stanford, California. 1995 (pp. 1-7, 11-14).
21. Pajares M. G. and Santos P. M.: Artificial Intelligence and Know Ingeniery, pp. 45-68, 95-106, 116-126. Alfaomega Grupo Editor, S.A. de C.V., México, D.F. (2006). (in spanish)
22. Groff James R, Weinbert Paul N.: SQL Reference Manual, pp. 95-276. McGrawHill (2002). (in spanish).
23. Cox, P. and Sheldon, T.: Windows 2000, Security Manual, pág. 18-24, 22-34, 44, 47, 475, 674, 712-717. McGraw-Hill (2003). (in spanish).
24. McAfee® Security AVERT (Anti-Virus Emergency Response Team),
25. http://www.nai.cl/es/security/virus/detail/v_100547.htm.
26. NOD32 Anti-Virus System, http://www.vsantivirus.com/gaobot-aa.htm.
27. Symantec Corporation,
28. http://www.symantec.com/region/mx/avcenter/data/la-w32.blaster.worm.html.
29. Conficker, http://conficker.com/.

Pattern Recognition and Data Mining

Delta Associative Memory: An Efficient Pattern Classifier

Mario Aldape-Pérez, Cornelio Yáñez-Márquez
and Amadeo José Argüelles-Cruz

Center for Computing Research, CIC
National Polytechnic Institute, IPN
Mexico City, Mexico
maldape@ieee.org; cyanez@cic.ipn.mx; jamadeo@cic.ipn.mx
http://www.aldape.org.mx

Abstract. The *Linear Associator* is one of the classical models of associative memories that can easily work as a binary pattern classifier if output patterns are appropriately chosen. However, this mathematical model undergoes fundamental patterns misclassification whenever fundamental input pattern cross-talk influence occurs. In this paper, a novel algorithm that overcomes *Linear Associator* weaknesses is proposed. The proposed algorithm computes a differential associative memory in order to obtain a dynamic threshold value for each unknown pattern to be classified. The efficiency and effectiveness of our approach is demonstrated through comparisons with other methods using real-world data.

Keywords: Associative Memories, Linear Associator, Machine Learning, Pattern Recognition.

1 Introduction

Pattern Recognition has existed for many years in a wide range of human activity, however, the general pattern recognition problem can be stated in the following form: Given a collection of objects belonging to a predefined set of classes and a set of measurements on these objects, identify the class of membership of each one of these objects by a suitable analysis of their measurements (features) [1]. At present, Pattern Recognition comprises a vast body of methods supporting the development of numerous applications in many different domains [2]. From an information theoretical perspective, an associative memory can be regarded as an extension of the neural computing approach for pattern recognition [3]. Furthermore, associative memories have a number of properties, including a rapid, compute efficient best-match and intrinsic noise tolerance that make them ideal for many applications. The majority of these applications are focused on problems related to pattern recognition tasks where the system takes in a pattern and emits another pattern [4]. As a consequence, associative memories have emerged as a practical technology with successful applications in countless pattern recognition tasks [5].

© G. Sidorov, M. Aldape, M. Martinez, S. Torres. (Eds.)
Advances in Computer Science and Engineering.
Research in Computing Science 42, 2009, pp. 79-90

Received 17/03/09
Accepted 27/04/09
Final version 07/05/09

The *Linear Associator*, which is one of the classical models of associative memories, is a heteroassociative memory that can easily work as a binary pattern classifier if output patterns are appropriately chosen. However, this mathematical model undergoes fundamental patterns misclassification whenever fundamental input pattern cross-talk influence occurs. In this paper, a novel algorithm that overcomes *Linear Associator* weaknesses is proposed. The experimental outcomes showed that fundamental input pattern cross-talk influence can be annulled by means of a dynamic threshold value which is computed for each unknown input pattern to be classified. As a consequence, classification rate is improved.

In the following section, a brief description of associative memories fundamentals is presented. In Section 3, *Linear Associator* technical details are presented. In Section 4, Delta Associative Memory mathematical foundations are presented. In Section 5, the main characteristics of the datasets that were used along the experimental phase are presented. A brief description of the algorithms used for comparison along the experimental phase is presented in Section 6 while in Section 7 some experimental results are shown using real-world data. Delta Associative Memory advantages, as well as a short conclusion will be discussed in Section 8.

2 Associative Memories

An associative memory $\mathbf{M}$ is a system that relates input patterns and output patterns as follows: $\mathbf{x} \rightarrow \boxed{\mathbf{M}} \leftarrow \mathbf{y}$ with $\mathbf{x}$ and $\mathbf{y}$ the input and output pattern vectors, respectively. Each input vector forms an association with its corresponding output vector. For each k integer and positive, the corresponding association will be denoted as: $(\mathbf{x}^k, \mathbf{y}^k)$. An associative memory $\mathbf{M}$ is represented by a matrix whose ij-th component is m_{ij}. Memory $\mathbf{M}$ is generated from an *a priori* finite set of known associations, called the fundamental set of associations. If μ is an index, the fundamental set is represented as: $\{(\mathbf{x}^\mu, \mathbf{y}^\mu) \mid \mu = 1, 2, ..., p\}$ with p as the cardinality of the set. The patterns that form the fundamental set are called fundamental patterns. If it holds that $\mathbf{x}^\mu = \mathbf{y}^\mu \; \forall \mu \in \{1, 2, ..., p\}$, $\mathbf{M}$ is autoassociative, otherwise it is heteroassociative; in this case, it is possible to establish that $\exists \mu \in \{1, 2, ..., p\}$ for which $\mathbf{x}^\mu \neq \mathbf{y}^\mu$. If we consider the fundamental set of patterns $\{(\mathbf{x}^\mu, \mathbf{y}^\mu) \mid \mu = 1, 2, ..., p\}$ where n and m are the dimensions of the input patterns and output patterns, respectively, it is said that $\mathbf{x}^\mu \in A^n$, $A = \{0, 1\}$ and $\mathbf{y}^\mu \in A^m$. Then the j-th component of an input pattern $\mathbf{x}^\mu$ is $x_j^\mu \in A$. Analogously, the i-th component of an output pattern $\mathbf{y}^\mu$ is represented as $y_i^\mu \in A$. Therefore the fundamental input and output patterns are represented as follows:

$$\mathbf{x}^\mu = \begin{pmatrix} x_1^\mu \\ x_2^\mu \\ \vdots \\ x_n^\mu \end{pmatrix} \in A^n \qquad \mathbf{y}^\mu = \begin{pmatrix} y_1^\mu \\ y_2^\mu \\ \vdots \\ y_m^\mu \end{pmatrix} \in A^m$$

A distorted version of a pattern $\mathbf{x}^k$ to be recalled will be denoted as $\widetilde{\mathbf{x}}^k$. An unknown input pattern to be recalled will be denoted as $\mathbf{x}^\omega$. If an unknown input pattern $\mathbf{x}^\omega$ with $\omega \in \{1, 2, ..., k, ..., p\}$ is fed to an associative memory $\mathbf{M}$, it happens that the output corresponds exactly to the associated pattern $\mathbf{y}^\omega$, it is said that recalling is correct [6].

3 Linear Associator

It is worth pointing out that James A. Anderson and Teuvo Kohonen obtained amazingly similar results known nowadays as *Linear Associator* [4,7].

3.1 Learning Phase

Let $\{(\mathbf{x}^\mu, \mathbf{y}^\mu) \mid \mu = 1, 2, ..., p\}$.be the fundamental set. In order to obtain an associative memory $\mathbf{M}$, the *Learning Phase* is done in two stages:

1. Consider each one of the p associations $(\mathbf{x}^\mu, \mathbf{y}^\mu)$, so an $m \times n$ matrix is obtained by $\mathbf{y}^\mu \cdot (\mathbf{x}^\mu)^t$

$$\mathbf{y}^\mu \cdot (\mathbf{x}^\mu)^t = \begin{pmatrix} y_1^\mu \\ y_2^\mu \\ \vdots \\ y_m^\mu \end{pmatrix} \cdot (x_1^\mu, x_2^\mu, ..., x_n^\mu) = \begin{pmatrix} y_1^\mu x_1^\mu & y_1^\mu x_2^\mu & \cdots & y_1^\mu x_j^\mu & \cdots & y_1^\mu x_n^\mu \\ \vdots & \vdots & & \vdots & & \vdots \\ y_i^\mu x_1^\mu & y_i^\mu x_2^\mu & \cdots & y_i^\mu x_j^\mu & \cdots & y_i^\mu x_n^\mu \\ \vdots & \vdots & & \vdots & & \vdots \\ y_m^\mu x_1^\mu & y_m^\mu x_2^\mu & \cdots & y_m^\mu x_j^\mu & \cdots & y_m^\mu x_n^\mu \end{pmatrix} \tag{1}$$

2. An associative memory $\mathbf{M}$ is obtained by adding all the p matrices

$$\mathbf{M} = \sum_{\mu=1}^{p} \mathbf{y}^\mu \cdot (\mathbf{x}^\mu)^t = [m_{ij}]_{m \times n} \tag{2}$$

in this way the ij-th component of an associative memory $\mathbf{M}$ is expressed as follows:

$$m_{ij} = \sum_{\mu=1}^{p} y_i^\mu x_j^\mu \tag{3}$$

3.2 Recalling Phase

The *Recalling Phase* for the *Linear Associator* is done by operating an associative memory $\mathbf{M}$ with an unknown input pattern $\mathbf{x}^\omega$, where $\omega \in \{1, 2, ..., k, ..., p\}$. Operate $\mathbf{M} \cdot \mathbf{x}^\omega$ as follows:

$$\mathbf{M} \cdot \mathbf{x}^\omega = \left[\sum_{\mu=1}^{p} \mathbf{y}^\mu \cdot (\mathbf{x}^\mu)^t \right] \cdot \mathbf{x}^\omega \tag{4}$$

Lets expand expression (4), which is:

$$\mathbf{M}\cdot\mathbf{x}^{\omega}=\left[\mathbf{y}^{1}\cdot\left(\mathbf{x}^{1}\right)^{t}+\mathbf{y}^{2}\cdot\left(\mathbf{x}^{2}\right)^{t}+\cdots+\mathbf{y}^{\omega}\cdot\left(\mathbf{x}^{\omega}\right)^{t}+\cdots+\mathbf{y}^{p}\cdot\left(\mathbf{x}^{p}\right)^{t}\right]\cdot\mathbf{x}^{\omega}$$
$$\mathbf{M}\cdot\mathbf{x}^{\omega}=\left[\mathbf{y}^{1}\cdot\left(\mathbf{x}^{1}\right)^{t}\right]\cdot\mathbf{x}^{\omega}+\cdots+\left[\mathbf{y}^{\omega}\cdot\left(\mathbf{x}^{\omega}\right)^{t}\right]\cdot\mathbf{x}^{\omega}+\cdots+\left[\mathbf{y}^{p}\cdot\left(\mathbf{x}^{p}\right)^{t}\right]\cdot\mathbf{x}^{\omega}$$
$$\mathbf{M}\cdot\mathbf{x}^{\omega}=\mathbf{y}^{1}\cdot\left[\left(\mathbf{x}^{1}\right)^{t}\cdot\mathbf{x}^{\omega}\right]+\cdots+\mathbf{y}^{\omega}\cdot\left[\left(\mathbf{x}^{\omega}\right)^{t}\cdot\mathbf{x}^{\omega}\right]+\cdots+\mathbf{y}^{p}\cdot\left[\left(\mathbf{x}^{p}\right)^{t}\cdot\mathbf{x}^{\omega}\right]$$

We obtain:

$$\mathbf{M}\cdot\mathbf{x}^{\omega}=\mathbf{y}^{\omega}\cdot\left[\left(\mathbf{x}^{\omega}\right)^{t}\cdot\mathbf{x}^{\omega}\right]+\sum_{\mu\neq\omega}\mathbf{y}^{\mu}\cdot\left[\left(\mathbf{x}^{\mu}\right)^{t}\cdot\mathbf{x}^{\omega}\right] \tag{5}$$

Expression (5) let us know about which restrictions have to be observed thus correct recalling is achieved. These restrictions are:

a) $\left[\left(\mathbf{x}^{\omega}\right)^{t}\cdot\mathbf{x}^{\omega}\right]=1$

b) $\left[\left(\mathbf{x}^{\mu}\right)^{t}\cdot\mathbf{x}^{\omega}\right]=0$ whenever $\mu\neq\omega$

Given an arbitrary chosen index ω , $\forall\omega\in\{1,2,...,k,...,p\}$, means that input pattern $\mathbf{x}^{\mu}$ should be orthonormal. This restriction is expressed as:

$$\left(\mathbf{x}^{\mu}\right)^{t}\cdot\mathbf{x}^{\omega}=\begin{cases}1 \text{ if } \mu=\omega\\ 0 \text{ if } \mu\neq\omega\end{cases} \tag{6}$$

If condition (6) is met, then a correct recalling is expected [6]. Therefore, expression (5) is expressed as:

$$\mathbf{M}\cdot\mathbf{x}^{\omega}=\mathbf{y}^{\omega}.$$

Nevertheless if orthonormality condition is not met, two situations appear:

· Factor $\left[\left(\mathbf{x}^{\omega}\right)^{t}\cdot\mathbf{x}^{\omega}\right]$ is not equal to 1

· Term $\sum_{\mu\neq\omega}\mathbf{y}^{\mu}\cdot\left[\left(\mathbf{x}^{\mu}\right)^{t}\cdot\mathbf{x}^{\omega}\right]$ is not equal to 0

This term is known as *cross-talk*, it represents some kind of noise that comes from input patterns interaction [4,7]. As a consequence correct recalling is not achieved, except in those cases where the number of stored patterns is rather small compared to n [5,6].

4 Delta Associative Memory

In this section, a novel algorithm that overcomes *Linear Associator* weaknesses is proposed. Due to the fact that an order relation between patterns implies an order relation between their characteristic set and vice versa [11], *cross-talk* influence can be annulled by means of a dynamic threshold value which is computed for each unknown input pattern to be classified. *Delta Associative Memory* algorithm applies the same learning phase as the *Linear Associator*, while a completely different recalling phase is proposed.

In what follows, let $\mathbf{M}$ be an associative memory whose ij-th component is denoted by m_{ij} and let ω be an index such that $\omega \in \{1, 2, ..., k, ..., p\}$. Let $\mathbf{x}^{\omega} \in \mathbb{R}^n$ be an unknown input pattern to be classified and let $m, n \in \mathbb{Z}^+$ be the dimension of the output patterns and input patterns, respectively.

Definition 1. ***Differential Associative Memory.*** *A Differential Associative Memory is denoted by Ψ^{ω}. Therefore, the ij-th component of Ψ^{ω}, denoted by ψ^{ω}_{ij}, is obtained according to the following rule:*

$$\psi^{\omega}_{ij} = |m_{ij} - x^{\omega}_j| \tag{7}$$

$\forall i \in \{1, 2, ..., m\}$, $\forall j \in \{1, 2, ..., n\}$.

Definition 2. ***Maximum threshold value.*** *The maximum threshold value, denoted by ζ^{ω}, is obtained according to the following rule:*

$$\zeta^{\omega} = \bigvee_{i=1}^{m} \left[\bigvee_{j=1}^{n} [\psi^{\omega}_{ij}] \right] \tag{8}$$

where $\bigvee$ is the maximum operator.

Definition 3. ***Minimum threshold value.*** *The minimum threshold value, denoted by α^{ω}, is obtained according to the following rule:*

$$\alpha^{\omega} = \bigwedge_{i=1}^{m} \left[\bigwedge_{j=1}^{n} [\psi^{\omega}_{ij}] \right] \tag{9}$$

where $\bigwedge$ is the minimum operator.

Definition 4. ***Delta Associative Memory (DAM).*** *Let θ^{ω} be the dynamic threshold value, such that $\alpha^{\omega} \leq \theta^{\omega} \leq \zeta^{\omega}$. A Delta Associative Memory is denoted by Δ^{ω}. Therefore, the ij-th component of Δ^{ω}, denoted by δ^{ω}_{ij}, is obtained according to the following rule:*

$$\delta^{\omega}_{ij} = \begin{cases} 1 & \textit{if} \quad |m_{ij} - x^{\omega}_j| \leq \theta^{\omega} \\ 0 & \textit{otherwise} \end{cases} \tag{10}$$

Definition 5. ***Positive contributions vector.*** *The Positive contributions vector is denoted by σ^{ω}. Therefore, the i-th component of σ^{ω}, denoted by σ^{ω}_i, is obtained according to the following rule:*

$$\sigma^{\omega}_i = \sum_{j=1}^{n} \delta^{\omega}_{ij} \tag{11}$$

Definition 6. ***Transition vector.*** *The Transition vector is denoted by τ^{ω}. Therefore, the i-th component of τ^{ω}, denoted by τ_i^{ω}, is obtained according to the following rule:*

$$\tau_i^{\omega} = \begin{cases} 1 & \textit{if} \quad \sigma_i^{\omega} = \bigvee_{h=1}^{m} \left[\sigma_h^{\omega}\right] \\ 0 & \textit{otherwise} \end{cases} \tag{12}$$

where $\bigvee$ is the maximum operator.

Definition 7. ***Unambiguously recalled class vector.*** *The Unambiguously recalled class vector, denoted by y^{ω}, is obtained according to the following rule:*

$$y^{\omega} = \begin{cases} \tau^{\omega} & \textit{if} \quad \sum_{i=1}^{m} \tau_i^{\omega} \leq 1 \\ 0 & \textit{otherwise} \end{cases} \tag{13}$$

4.1 Learning Phase

Generate a matrix $\mathbf{M}$ that will store the p associations of the fundamental set $\left\{\left(\mathbf{x}^1, \mathbf{y}^1\right), ..., \left(\mathbf{x}^p, \mathbf{y}^p\right)\right\}$, where $\mathbf{x}^{\mu} \in \mathbb{R}^n$ and $\mathbf{y}^{\mu} \in A^m \ \forall \mu \in \{1, 2, ..., p\}$. It is worth pointing out that there are m different classes. Therefore, each one of the input patterns belongs to a k class, $k \in \{1, 2, ..., m\}$, represented by a column vector $\mathbf{y}^{\mu}$, whose components will be assigned by $y_k^{\mu} = 1$, so $y_j^{\mu} = 0$ for $j = 1, 2..., k-1, k+1, ...m$; hence, the class statements are given in a 1-out-of-m-code [11], also known as *one-hot* codification [5,6].

```
Given:
    The fundamental set of associations $\{(\mathbf{x}^{\mu}, \mathbf{y}^{\mu}) \mid \mu = 1, 2, ..., p\}$ with
p as the cardinality of the set
Algorithm:
    Obtain $p$ matrices according to expression (1).
    for $\mu = 1$ to $p$ do
    {
        for $i = 1$ to $m$ do
        {
            for $j = 1$ to $n$ do
            {
            Compute $m_{ij}$ using expression (2).
            }
        }
    }
```

4.2 Classification Phase

Consists of finding the class which an unknown input pattern $\mathbf{x}^{\omega} \in \mathbb{R}^n$ belongs to. Finding the class means getting $\mathbf{y}^{\omega} \in A^m$ that corresponds to $\mathbf{x}^{\omega}$. If when an unknown input pattern $\mathbf{x}^{\omega}$ is fed to an associative memory $\mathbf{M}$, it happens that the output corresponds exactly to the associated pattern $\mathbf{y}^{\omega}$, it is said that classification is correct.

```
Given:
    An unknown input pattern $x^{\omega} \in \mathbb{R}^n$
Algorithm:
    Obtain a Differential Associative Memory $\Psi^{\omega}$, using (7)
    Compute the maximum threshold value $\zeta^{\omega}$, using (8)
    Compute the minimum threshold value $\alpha^{\omega}$, using (9)
    Initialize the dynamic threshold value $\theta^{\omega}$, that is, $\theta^{\omega} = \alpha^{\omega}$
    While $[(\sum_{i=1}^{m} \tau_i^{\omega} > 1)$ and $(\theta^{\omega} < \zeta^{\omega})]$ do
    {
        Obtain a Delta Associative Memory $\Delta^{\omega}$, using (10)
        Compute the positive contributions $\sigma^{\omega}$, using (11)
        Compute the transition vector $\tau^{\omega}$, using (12)
        Assign the recalled class vector $y^{\omega}$, using (13)
        Increment the dynamic threshold value $\theta^{\omega}$, that is, $\theta^{\omega} = \theta^{\omega} + 1$
    }
    Assign the unambiguously recalled class vector $y^{\omega}$, using (13)
```

5 Machine Learning Databases

Nowadays, the University of California Machine Learning Repository maintains 177 different datasets which can be used as test benches in order to obtain a fair performance comparison between different algorithms. In this section, a brief description of the datasets that were used along the experimental phase is presented.

5.1 Iris Data Set

This is perhaps the best known database to be found in the pattern recognition literature [8-10]. The data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant (Iris Setosa, Iris Versicolour, and Iris Virginica). One class is linearly separable from the other two; the latter are not linearly separable from each other. Each instance is conformed by 4 numerical features and a class tag.

5.2 Pima Indians Diabetes Database

This database was originally owned by the National Institute of Diabetes and Digestive and Kidney Diseases. This dataset was used to forecast the onset of diabetes mellitus. The diagnostic was investigated according to the World Health Organization criteria (i.e., if the 2 hour post-load plasma glucose was at least 200 mg/dl at any survey examination or if found during routine medical care). Each instance is conformed by 8 numerical features and a class tag.

5.3 Wisconsin Breast Cancer Database

This database was obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg. This dataset was assembled using clinical reports that arrive periodically. This database has been widely used for medical diagnosis applied to breast cytology. Each instance has one of 2 possible classes: benign or malignant. Each instance is conformed by 9 numerical features and a class tag.

5.4 Heart Disease Database

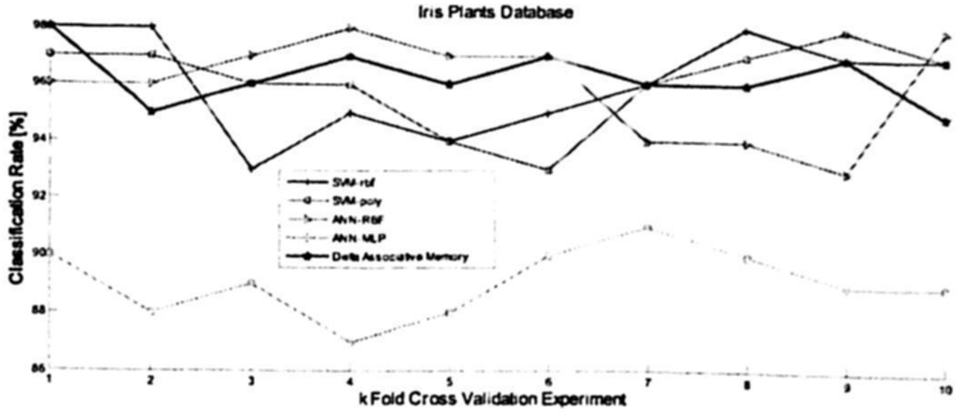
This database has been widely used in heart disease diagnosis. The data was collected by Robert Detrano, M.D. from V.A. Medical Center, Long Beach and Cleveland Clinic Foundation. The original dataset contains 76 attributes, but all published experiments refer to using a subset of 14 of them (13 numerical features and a class tag). The main purpose of this dataset is to forecast the presence of heart disease in the patient.

6 Algorithms Compared

Four of the most commonly used classification models were tested along the experimental phase. Support Vector Machines using polynomial kernel (SVMpoly) and radial basis kernel (SVMrbf) were tested using the OSU-SVM Matlab Toolbox Version 3.00 obtained from [15]. For SVMrbf, γ was tuned from the possible values $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10\}$ while the constrain penalty C was tuned from the values $\{10^{-1}, 1, 10, 10^{2}, 10^{3}\}$. The SVMpoly uses a degree from the range 2 to 6 and constrain penalty C from the set $\{0.05, 0.1, 0.5, 1, 5\}$. Two types of neural networks, the radial-basis network (ANNrbf) and the multi-layered perceptron (ANNmlp) were tested using the Matlab Toolbox implementation. The limit to the number of hidden neurons added by Matlab in training the ANNrbf was set to the lower of the number of instances and ten times the number of classes. The only parameter tuned was the spread of the transfer function, from 0.1 to 1.0 in increments of 0.1.

7 Experimental Results

Throughout the experimental phase, four databases taken from the UCI Machine Learning Repository were used (http://archive.ics.uci.edu/ml/). The main characteristics of these data sets have been expounded in section 5 and are summarized in Table 1. The experiments have been carried out as follows: Firstly, in order to obtain an associative memory **M**, the same number of input vectors for each class was randomly taken, which means that a balanced classifier is guaranteed [12]. Afterwards, in order to obtain reasonably unbiased performance estimates, the dataset was broken into k partitions (in our case $k = 10$). **Delta Associative Memory** behavior was evaluated using Stratified k Fold Cross Validation technique; as a result, the classification performance estimates are reasonably unbiased [12-14]. The classification performance for each one of the compared algorithms is shown using Cartesian coordinates in two dimensions (Figure 1 to Figure 4). The horizontal axis represents the k-th partition of the fundamental set that was evaluated; while the vertical axis represents the classification rate achieved for each one of the compared algorithms over the k-th partition of the fundamental set. Efficiency of the developed method was tested by performing the experiments 20 times, each time with different number of randomly taken input vectors. Performance of the classification phase is measured in terms of error rate [11], in this way, classifier accuracy is estimated through classification rate over unseen patterns [11-14].



Averaged classification results: Iris Plants Database

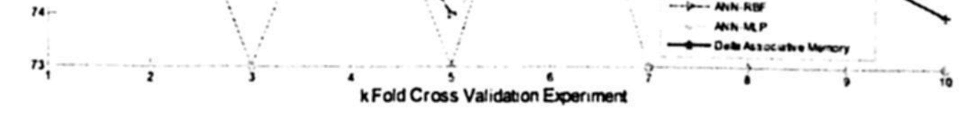
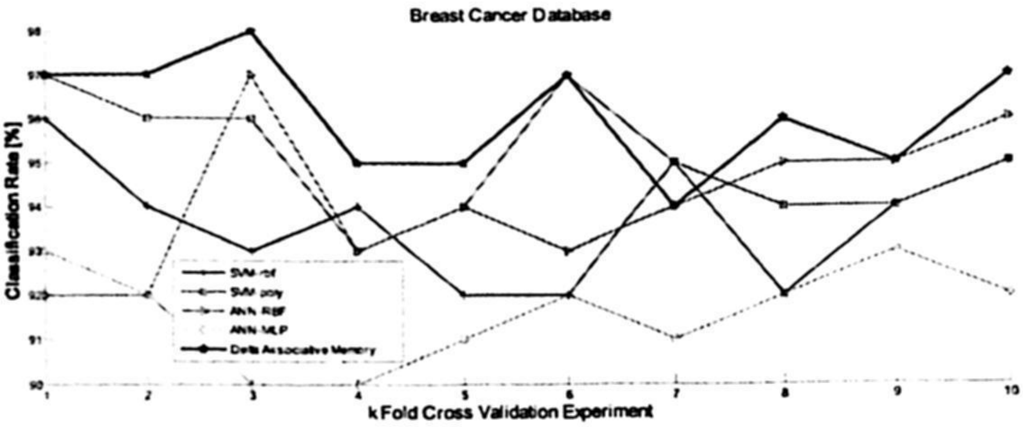


Table 1. Some characteristics of the data sets used in the experiments

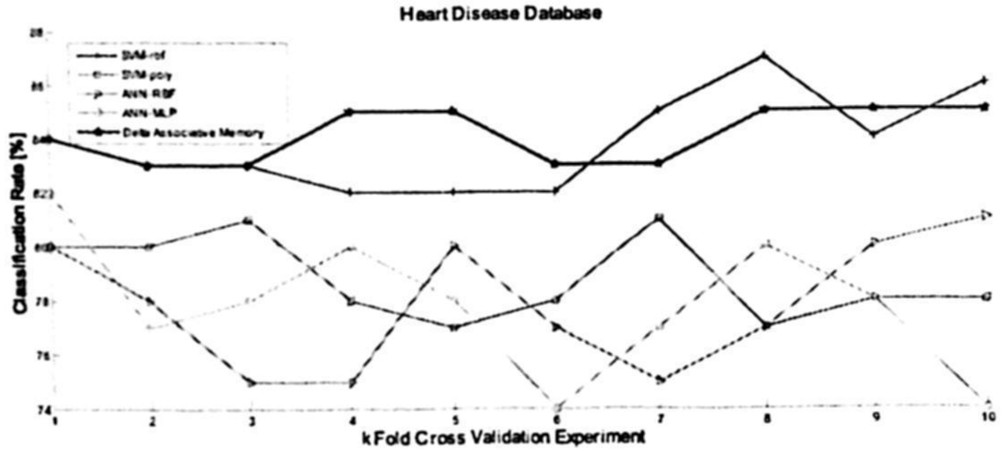
	Iris	Pima	Breast	Heart
Features	4	8	9	13
Classes	3	2	2	2
Patterns	150	768	699	270

Table 2. Mean Classification Rates

	SVM-rbf	SVM-poly	ANN-RBF	ANN-MLP	Delta AM
Iris	96.19%	96.19%	96.03%	89.18%	**96.27%**
Pima	76.97%	77.32%	76.91%	75.03%	**77.43%**
Breast	93.89%	95.16%	94.16%	91.64%	**96.08%**
Heart	83.85%	78.93%	77.78%	77.81%	**84.19%**



Averaged classification results: Breast Cancer Database



Averaged classification results: Heart Disease Database

8 Conclusions and Ongoing Research

In this paper a new algorithm which is based on one of the classical models of associative memories has been introduced. *Linear Associator* weaknesses are solved by means of a dynamic threshold value which is computed for each unknown input pattern to be classified. The algorithm computes a differential associative memory in order to obtain the maximum and minimum threshold values which represent the dynamic threshold upper and lower bounds, respectively. As a direct consequence, classification error rate is improved. Experimental results have shown that this algorithm is an efficient way to improve classifier accuracy.

Currently, we are exploring how to use the proposed approach for feature selection in machine learning and data mining. We are also working on a Xilinx Virtex II Pro FPGA implementation, based on recent mathematical results.

Acknowledgments The authors of the present paper would like to thank the following institutions for their economical support to develop this work: National Polytechnic Institute, Mexico (CIC, SIP, PIFI, COFAA), CONACyT and SNI.

References

1. Yáñez-Márquez, C., Felipe-Riverón, E. M., López-Yáñez, I., & Flores-Carapia, R. (2006). A Novel Approach to Automatic Color Matching. Lecture Notes in Computer Science (LNCS), 4225, 529-538.
2. Acevedo-Mosqueda, M. E., Yáñez-Márquez, C., & López-Yáñez, I. (2006). Alpha-Beta Bidirectional Associative Memories Based Translator. International Journal of Computer Science and Network Security (IJCSNS), 6, 190-194.
3. Yáñez-Márquez, C., Sánchez-Fernández, L. P., & López-Yáñez, I. (2006). Alpha-Beta Associative Memories for Gray Level Patterns. Lecture Notes in Computer Science (LNCS), 3971, 818-823.
4. Kohonen, T. (1972). Correlation Matrix Memories. IEEE Transactions on Computers, 21, 353-359.
5. Acevedo-Mosqueda, M. E., Yáñez-Márquez, C., & López-Yáñez, I. (2006). A New Model of BAM: Alpha-Beta Bidirectional Associative Memories. Lecture Notes in Computer Science (LNCS), 4263, 286-295.
6. Acevedo-Mosqueda, M. E., Yáñez-Márquez, C., & López-Yáñez, I. (2007). Alpha–Beta bidirectional associative memories: theory and applications. Neural Processing Letters, 26, 1-40.
7. Anderson, J.A., & Rosenfeld, E. (1990). Neurocomputing: Fundations of Research, Cambridge: MIT Press.
8. Márques de Sá, J.P. (2001). Pattern Recognition, Concepts, Methods and Application. Springer, Germany.
9. Webb, A. (1999). Statistical Pattern Recognition. Oxford University Press, USA.
10. Jain, A.K., Duin, R.P.W., & Jianchang Mao. (2000). Statistical pattern recognition: a review. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22, 1, 4-37.
11. Aldape-Pérez, M., Román-Godínez, I., & Camacho-Nieto, O. (2008). Thresholded Learning Matrix for Efficient Pattern Recalling. Lecture Notes In Computer Science (LNCS), 5197, 445-452.

12. Pal, S., & Mitra, S. (1999). Neuro-Fuzzy Pattern Recognition: Methods in Soft Computing . John Wiley & Sons, USA.

13. Hassoun, M.H. (1995). Fundamentals of Artificial Neural Networks. MIT Press, Cambridge.

14. Ritter, G. X., & Sussner, P. (1996). An Introduction to Morphological Neural Networks. in Proceedings of the 13th International Conference on Pattern Recognition, vol. IV, Track D 709-717.

15. OSU SVM Classifier Matlab Toolbox (ver 3.00) at http://www.ece.osu.edu/~maj/osu_svm/

Automatic Synthesis of Associative Memories by Genetic Programming, a First Approach

Juan Villegas-Cortez[1], J. Humberto Sossa A.[1], Carlos Avilés-Cruz[2] and Gustavo Olague Caballero[3]

[1] Centro de Investigación en Computación - IPN, Mexico D.F.
jvillegas@sagitario.cic.ipn.mx, hsossa@cic.ipn.mx

[2] Universidad Autonoma Metropolitana, Unidad Azcapotzalco, Mexico D.F., Mexico,
caviles@correo.azc.uam.mx

[3] Centro de Investigación Cientfica y de Educación Superior de Ensenada, BC., Mexico.
olague@cicese.mx

Abstract. Associative memories (AMs) are a particular case of artificial neural networks (ANNs), whose main purpose is a faster pattern association than traditional ANN models. This positive attribute has set a new area of research consisting in the application of AMs to specific pattern recognition problems, such as pattern recognition under mixed noise, or real valued patterns recall. Unfortunately every problem has its own difficulties and receive a different treatment, considering the pattern kind or its approach into the specific application scope considering the limited capacity of the AM classic model. Genetic Programming (GP) has proven great success for developping automatic programming, thus creating specialized solutions through the composition of terminals and functions using theory of evolution in order to generate a design based system for problem solving. This work uses GP for the automatic production of AMs in the solution of the classic AM operator for binary patterns problem. We present experimental results on two well-known problems that have been studied by the pattern recognition community (orthogonal and non orthogonal pattern sets association). Our methodology allows us to create novel designs that are different to traditional human-based designs. This work could be considered, as far as we know, as the first approach that applies GP to the synthesis of automatic AM.

1 Introduction

An associative memory (AM) is a special kind of ANN that allows to recall one output pattern from a given input one, such that it can work as a key that may be altered by some kind of noise (salt, pepper or mixed). Several models of AM have been developed during the last years; see: [11], [12], [17], [4], [19], [1], [18], most of them work out fine considering one kind of noise, and only one model works well for the case of mixed noise [23].

Thus, an association between input pattern, x, and output pattern, y, is denoted as (x^k, y^k) where k is the corresponding association. The associative

Received 18/03/09
Accepted 27/04/09
Final version 06/05/09

memory $\boldsymbol{M}$ is represented by a matrix whose components m_{ij} can be seen as the synapses of a neural network. The operator $\boldsymbol{M}$ is generated from an *a priori* set of finite known associations, named as the fundamental set of associations and is represented as $\{(x^k, y^k)|k = 1, ..., p\}$, where p is the number of associations. If $(x^k = y^k)\forall k = 1, \ldots, p$, then $\boldsymbol{M}$ is considered auto-associative, otherwise is hetero-associative. A distorted version of a pattern x to be restored will be denoted as $\widetilde{x}$. If $\boldsymbol{M}$ is fed with a distorted version of x^k and the output obtained is exactly y^k, the recalling feature is perfect. Regarding the components of m_{ij} belonging to $\boldsymbol{M}$, these are conformed by simple operations such as add, multiply, max, min and another defined operators. In order to preserve its simplicity the process of creating new structures for a AM turns out to be complex.

Genetic Programming (GP) is a powerful bio-inspired technique, based on biological evolution theory, that pursues the automation of computer programming using Darwinian principles for the solution of real world and complex problems. It is considered as part of machine learning techniques, and it has also been used to optimize a population of programs using a fitness function that rates the goodness of every program within the population with the goal of solving a specific task. Its origins could be traced back to the application of evolutionary simulation by Barricelli in 1954 [2]. In the 1960's Rechenberg developed its applications for optimization methods, and then, in the 70's, together with his research team solved complex problems using evolutionary strategies [14]. Holland's research had great importance in the early 70's [3]. Smith (1980) and Cramer (1985) presented the first results using GP methodology. In 1981 Forsyth [7], reported little programs evolution in forensic science. Koza [6], is one of the most important exponents of GP and he has been a pioneer in GP applications developed for complex optimizations and has been continuously researching in solutions to several problems.

GP performs the evolution on computer programs (individuals) that are represented as tree structures. These trees are easily evaluated in a recursive way, every node has one function operator, and every terminal node has an operating element, as a result the use of GP in computer programming languages is favored. A set of generated individuals is considered as a generation, every individual represents a solution for a given problem that is meant to be solved. During evolution new individuals are achieved by means of both the selection of the best individuals (although the worst ones have also a little chance of being selected), and their mutual combination for creating new solutions using selection, crossover and mutation operators. After several generations or final convergence criterion based in fitness value, it is expected that the population might contain some good solutions for the addressed problem. Nowadays GP has a wide application covering several environments (see: [9], [8], [20], [21], [22]), its consequent great success comes from its capability of being accurately adapted to numerous problems. Although the main and more direct application is the generation of mathematical expressions, it has been also used in other fields such as rule generation, filter designing, 3D reconstruction models, photogrammetric network design, object recognition, interest point detectors, robotics etc., and finally in a more related field very close to our purposes: the automatic design of ANN [10].

In this paper we try to show the possibility of applying the GP methodology for an AM development and describe the problems inherent to the achievement of our final goal. It is organized as follows. In section 2 a brief description of the representative models of AM is presented. In section 3, we present our GP methodology for automatic synthesis of AM by means of GP. Section 4 shows our results. Finally, conclusions and suggestions for further research are presented in section 5.

2 Representative Models of AM

An AM can be classified into three basic models:

- The Linear Associator (based on the Kohonen correlation matrix [5]).
- The Hopfield Model [4] (based on recurrent type networks where the the units resulting as output are fed back as input units).
- The Bi-directional model (similar to the Linear Associator, but with bidirectional connections, i.e., $w_{ij} = w_{ji}$ for $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, n$).

Based on these primary proposals come other type of models that use mixed ideas in order to get better results for two problems: i) the salt or pepper noise case (the $\alpha\beta$ AM [24], and the Morphological Independent method [12]), and ii) the mixed noise case (The Median AM [1] and The Dynamic AM [23]). These last two proposals have covered gray levels and color images recall problems too. All the above models have one common feature, they try to reach the association with the use of simple operators, mixing the order of application of operators between patterns, or computing new characteristics describing the association. As a final result, the redesigning of new operators of AM is achieved through a process developed in two phases (association and recalling).

3 AM development with the use of GP

The development of AM has been a subject of investigation carried out by human experts, similar to the development of ANN. Further than that, our purpose is to face the problem of automatic synthesis of AM using GP. So far, there is only just one robust design of ANN using GP for a specific problem, Rivero [10], and a preliminary work about the automatic development of kernel pattern design for morphological AM [16], using Genetic Algorithms. Our proposal points towards to an automatic generation of AM, as opposed to the non automatic cases appearing in the reviewed works mentioned in section 2.

As a general rule, the task of AMs generation using evolutionary algorithms could follow the ANNs generation rule. First, the weight evolution starts from an ANN with an already determined topology [25], but in this case we know only one, simple association, using elementary arithmetic operators (having special care with the use of multiply operator with vectors); this is our starting point. Second, the evolution of architectures includes the generation of the topological structure, this means establishing the connectivity for every synapse, in our case, between the components of the input vector and the corresponding output vector. There is a one-to-one correspondence. The key is to keep in mind that the AM holds all the "aspects" of the association, having as input both pattern sets (source and target), during the association phase, and as output just only the target entity resulting from one source pattern, during the recalling phase.

3.1 Model

Our first proposal for a GP-development of AMs, takes into account the main feature of the Linear Associator model [5], that is, the *times*$\{*\}$ operator for performing local associations and the *sum*$\{+\}$ operator for global association; both operations conform the first stage for producing the association matrix M. The *times* operator between the M matrix and an input vector for the recalling process, comprises the second stage. We considered the possibility of taking new evolutionary operators in the local association in order to study all the possible aspects which had not been considered.

In this case, the nodes to be used are the following:

- Op_k. The evolutionary operator (which defines a M_i AM network behavior). It is generated as an individual in the form of a tree. This is the coded genotype.
- x_j and y_j. Nodes that belongs to the *Terminal Set T*. These input nodes are one entry of every pattern-vector of X and Y; so T is defined as: $T = \{x_j, y_j\}$, this is so in order to have a coded correlation input.
- The arithmetic operators, *Functions Set*, (F). They have been constructed considering the possible structures of individuals similar to reviewed models of AMs: $F = \{+, -, min, max, times\}$

The evolutionary process demands the assignation of a *fitness* value to every genotype. Such value is the result obtained after the evaluation of the network with the pattern set representing the problem to be carried out.

For the *Fitness Function*, first we considered the normalized correlation coefficient between the goal (g) and the source processed image (f) [13]. The objective fitness function f_A, known as *similarity* $(0 \leq f_A \leq 1)$, is defined as:

$$f_A = \frac{f \cdot g}{\sqrt{f \cdot f}\sqrt{g \cdot g}} \tag{1}$$

where f and g are two digital images of size $N \times N$, and $f \cdot g$ is given by the equation:

$$f \cdot g = \frac{1}{N} \cdot \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} f(i,j) \cdot g(i,j) \tag{2}$$

This function f_A intends to maximize the number of matching black pixels inside the images f and g, so this seems a reasonable choice for the fitness function. The optimum is found when $f_A = 1$, corresponding to the matching of all pixels. The worst case takes place for $f_A = 0$, when none of the pixels match.

Thus, we will use Eq. (1) as our fitness function, and it will be utilize for the evaluation of every generated individual m_i, applying the association between the source set $\bar{X}$, and the goal set $\bar{Y}$, both of them conform our fundamental set. The evaluation is carried out in three steps. First, we apply the *association* between $\bar{X}$ and $\bar{Y}$. Second, we perform the recalling taking the $\bar{X}$ set while keeping the recalled patterns as $\hat{\bar{Y}}$. Third and final, the computation of Eq. (1) between $\bar{Y}$ and $\hat{\bar{Y}}$, for all the local associations, is carried out in order to have a global fitness. This process is shown in Fig. 1.

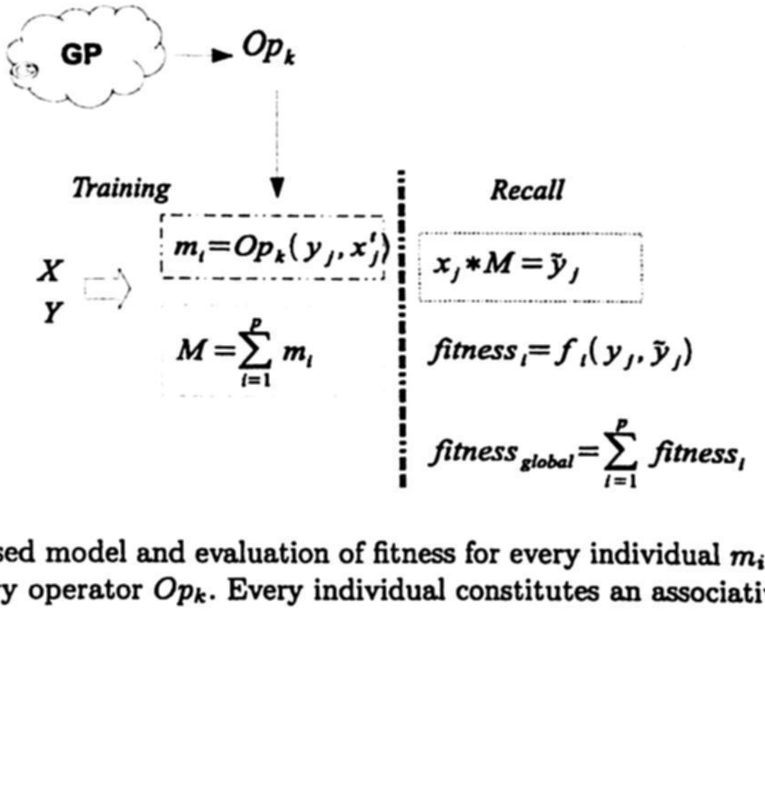


Fig. 1. Proposed model and evaluation of fitness for every individual m_i, created from its evolutionary operator Op_k. Every individual constitutes an associative rule.

3.2 GP Setup

All the experiments were implemented using Matlab with a GPLab toolbox ver. 3.0 [15]. According to the simplicity of the function set we performed several batch of runs consisting of 50 generations, each of them with 70 individuals. After every run we had one possible evolutionary solution for our problem.

The GP parameters used in our experiments are similar to those suggested by Koza [6], taking values of 0.7 for the crossover rate and 0.3 for the mutation rate, respectively. Mutation was based on the *ramped-half-and-half* initialization method, which was also used to initialize the population.

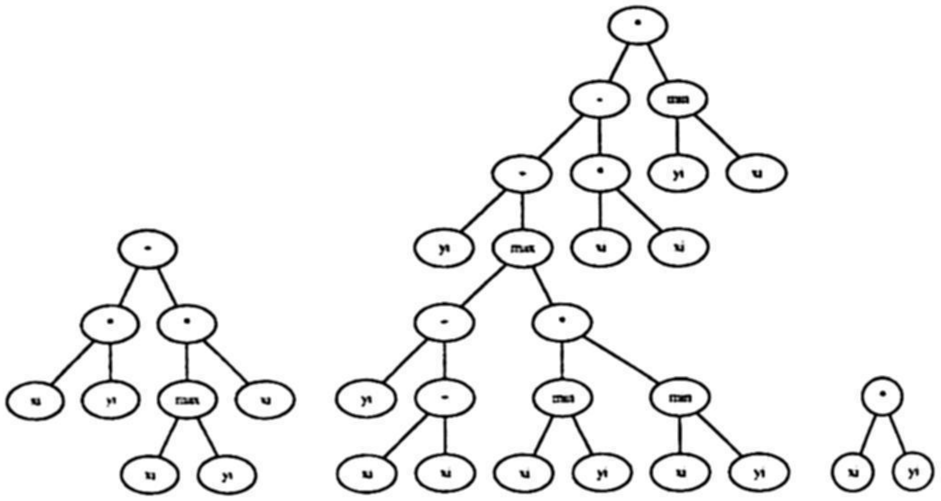
4 Results and Analysis

We applied the above methodology using simple vectors sets in binary sets, converting them to bi-polar notation (0 = −1, in order to reduce zeros for *times* operator) as:

$$X = \begin{bmatrix} -1 & -1 & -1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & -1 & -1 & -1 \end{bmatrix}, \quad Y = \begin{bmatrix} 1 & -1 & 1 & 1 \\ -1 & 1 & 1 & -1 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

where X and Y are the source and the goal vector sets, respectively. We tried to find an association between these sets in a hetero-associative memory rule. Every line per matrix is a vector such as $x_1 = [-1\ -1\ -1\ 1]$, it is associated with $y_1 = [1\ -1\ 1\ 1]$ and so on x_2 with y_2, etc.

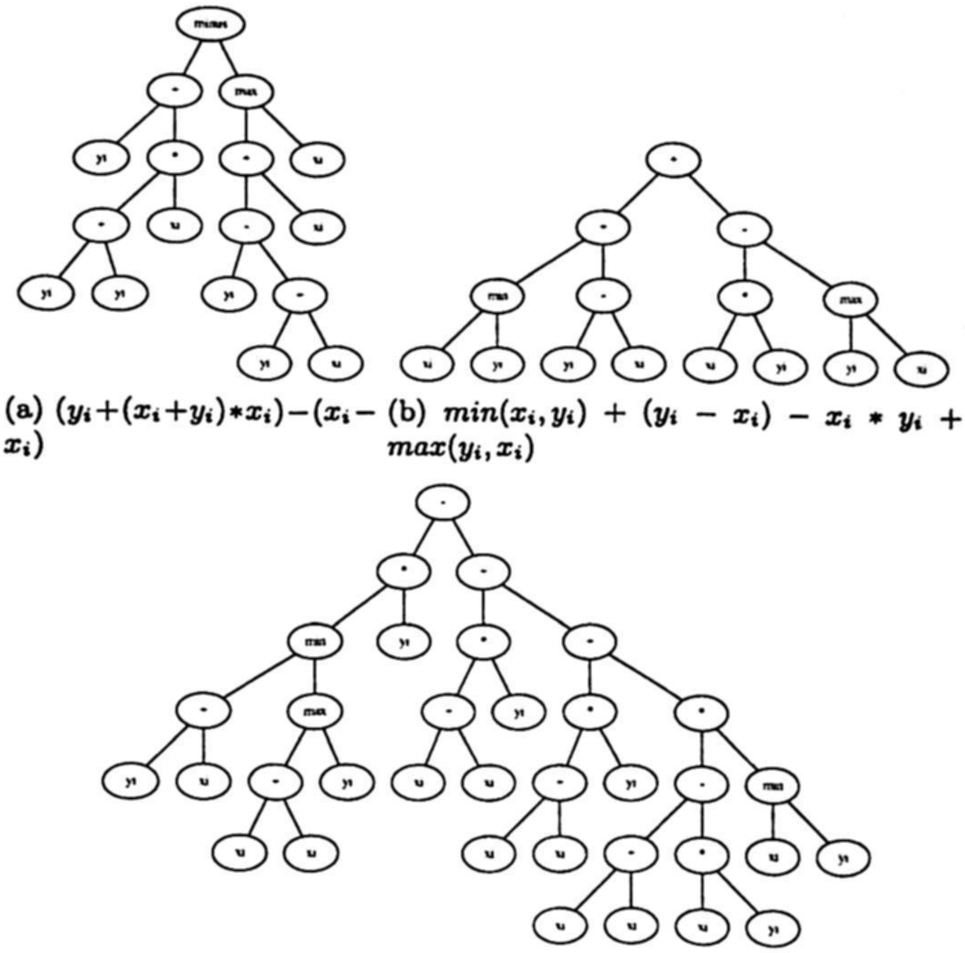
After the accomplishment of the batches described at the beginning of the previous section, under the described conditions, we got several individuals, as those shown in Fig. [2(a), 2(b)], with fitness values of 1. This last value is in concordance with the known result of the classic AM model depicted in Fig. 2(c), which is also one of the various evolutionary results of the same batch process. The selected individuals, shown in the Fig. 2, are just some of the total.



(a) $(x_i * y_i) + (max(x_i, y_i) * x_i)$ (b) $(y + max(y_i + 2x_i, min(x_i, y_i)^2) - {x_i}^2) * min(y_i, x)$ (c) $x_i * y_i$

Fig. 2. Synthesized individuals [(a), (b)] by GP model, and the classic operator (c).

These individuals present two different ways of association between two pattern sets and they make a perfect recall on the fundamental set. In order to test their robustness, we tested them over unknown pattern sets (random and orthogonal patterns generated for auto-associative and hetero-associative cases), taking the number of bits as the total number of vectors per test (ranging from 2 by 2 to 100 by 100, bits and vectors, respectively); all the former with the purpose



(a) $(y_i+(x_i+y_i)*x_i)-(x_i-x_i)$

(b) $min(x_i,y_i)+(y_i-x_i)-x_i*y_i+max(y_i,x_i)$

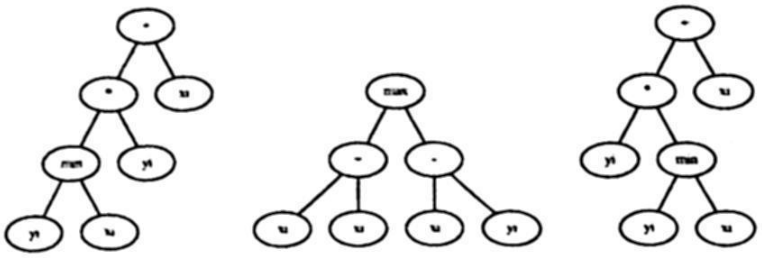
(c) $(min(y_i+x_i,max(x_i+x_i,y_i)))*y_i-((x_i+x_i)*y_i+(x_i+x_i)*y_i+(x_i+x_i-x_i*y_i)*min(x_i,y_i))$

Fig. 3. Synthesized individuals generated from orthogonal pattern sets under auto-associative relationship.

of reviewing the behavior on both cases, auto-associative and hetero-associative. The results are shown in Fig. 5.

Other two tests were performed with this methodology but this time applied for the synthesis of individuals of two more cases: i) the generation of individuals to perform a perfect recall over orthogonal pattern set in auto-associative mode (Fig. 3); and ii) another for random non orthogonal pattern set (Fig. 4). We tested these generated individuals with the same conditions and the same cases described in the previous paragraph. The results are shown in Fig. [6] and Fig. [7].

As we can see, these individuals fit perfectly for their corresponding training pattern sets, but when a new pattern set is associated with them, they do not work in a good manner. On the other hand, when the patterns are evolutionary orthogonal based on hetero-associative AM the individuals work perfectly (Fig. 7), as the classical model does.



(a) $(min(y_i, x_i) * y_i) + x_i$ (b) $max(x_i + x_i, x_i - y_i)$ (c) $(y_i * min(y_i, x_i)) + x_i$

Fig. 4. Synthesized individuals generated from orthogonal pattern sets under hetero-associative relationship.

5 Conclusions

The technique described in this report is suitable to developing simple AMs resulting from GP. These solutions are the outcome of an evolutionary process that explores all the possibilities of association using the described functions and terminals sets. All of this work was carried out by just one single evolutionary process for the local pattern pair association (x_i, y_i).

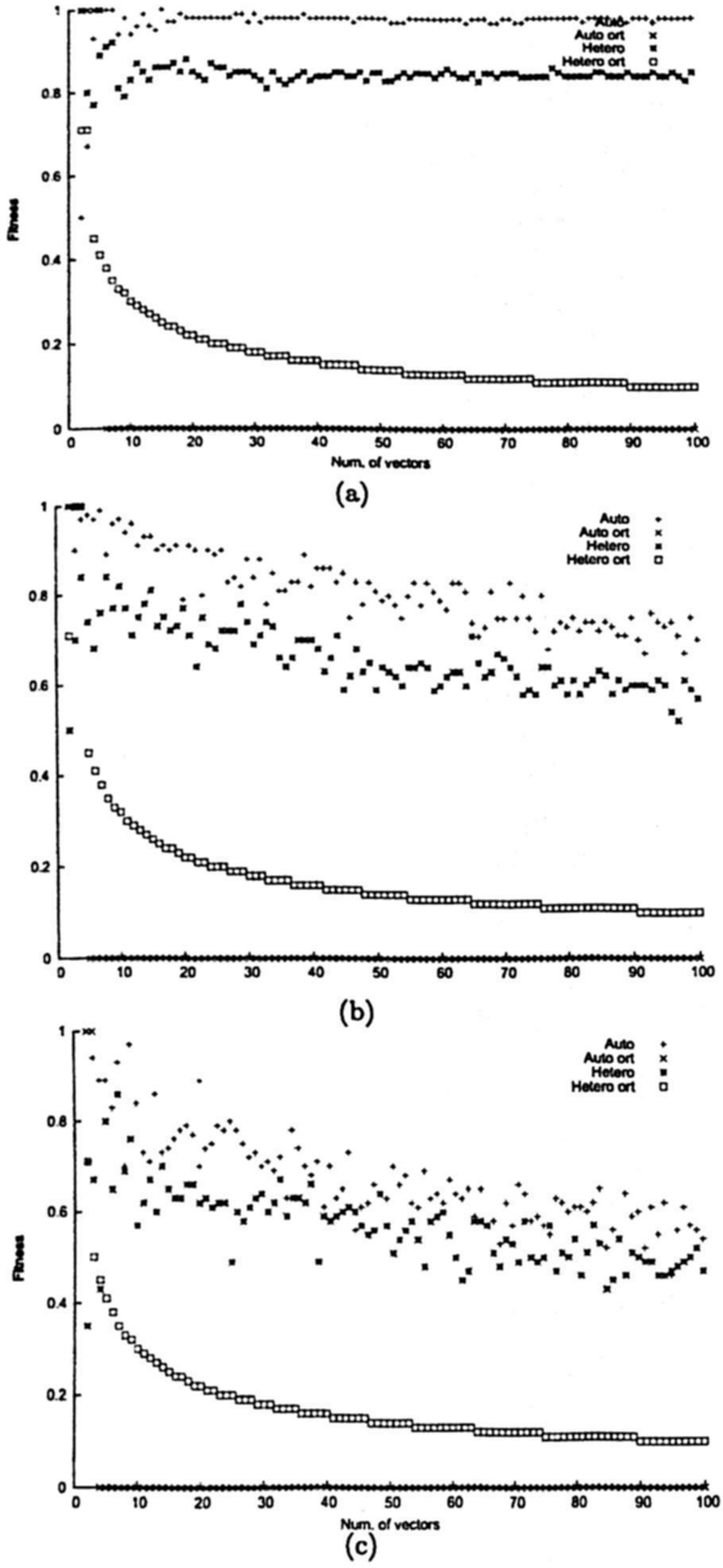
This system shows the following advantage: it is possible the development of AMs through the use of GP. With this model a great amount of possibilities could be explored given the same complexity to the proposed functions and terminal sets. Contrary to most AMs models working with two processes, the association and the recalling, this proposed model works only with the association stage, this could be compared to the Linear Associator AM branch.

Right now we are working in a new model of association taking into account the two processes of the training phase (local and global association), and the recalling process. Our future purpose is to introduce evolutionary operators for every substage (the local evolutionary association covered in this work, plus the global association, and the recalling), in order to have a common solution as the output of the cooperative co-evolution, considering three fundamental aspects of the co–evolutionary process: the individual evaluation, the population selection or variation, and the fitness kind per process. These aspects are shown in the dotted blocks of Fig. 1.

References

1. R. Barrón. Memorias asociativas y redes neuronales morfológicas para la recuperación de patrones. Tesis doctoral. CIC-IPN. 2006.
2. N. Barricelli, Esempi numerici di processi di evoluzione, Methodos, pp. 45-68. 1954
3. J. H. Holland, Adaptation in natural and artificial systems, University of Michigan Press, 1975.
4. J. J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. Proceedings of the National Academy of Sciences. 79: 2554-2558. 1982.

5. T. Kohonen. Correlation Matrix Memories. IEEE Transactions on Computers, Vol. C-21, pp. 353-359. 1972.
6. J. R. Koza. Genetic Programming: On the Programming of Computers by Means of Natural Selection. MIT Press. 1992.
7. R. Forsyth, "BEAGLE A Darwinian Approach to Pattern Recognition", Kybernetes, vol. 10, pp. 159–166, 1981.
8. B. Hernández, G. Olague, R. Hammoud, L. Trujillo and E. Romero. Visual learning of texture descriptors for facial expression recognition in thermal imagery. Comput. Vis. Image Underst., Vol. 106, pp. 258–269, 2007.
9. C. Perez B and G. Olague. Learning Invariant Region Descriptor Operators with Genetic Programming and the F-Measure. International Conference on Pattern Recognition. ICPR., 2008.
10. D. Rivero, J. Rabuñal and Alejandro Pazos. Automatic Desing of ANNs by Means of GP for Data Mining Tasks: Iris Flower Classification Problem. ICANNGA 07. Part I, Springer-Verlag LNCS 4431, pp. 276-285. 2007.
11. G. X. Ritter et al. Morphological associatives memories. IEEE Transactions on Neural Networks, vol. 9, no. 2, pp. 281-293, 1998.
12. G. X. Ritter, G. Urcid et al. Reconstruction of patterns from noisy inputs using morphological associative memories. International Journal of Mathematical Imaging and Vision. 19(2), pp. 95-111. 2003.
13. M. Quintana, R. Poli, E. Claridge. Morphological Algorithm Design for Binary Images Using Genetic Programming. Genetic Programming and Evolvable Machines. Vol 7, Issue 1. Pag: 81 - 102. 2006 .
14. I. Rechenberg. Evolutionsstrategie Optimierung technischer Systeme nach Prinzipien der biologischen Evolution (PhD thesis). Reprinted by Fromman-Holzboog. 1973.
15. S. Silva, J. Almeida. GPLAB - A Genetic Programming Toolbox for MATLAB. In Gregersen L (ed), Proceedings of the Nordic MATLAB Conference (NMC-2003), 273-278.
16. A. R. Silva Lavalle. Un Método de Algoritmos Genéticos para Optimización de Memorias Asociativas Morfológicas. Tésis. Univesidad de Puerto Rico. 2006.
17. H. Sossa, R. Barron, R. A. Vazquez. New associative memories for recall real-valued patterns. LNCS, 3287:195-202. 2004.
18. H. Sossa, R. Barrón. Extended $\alpha\beta$ associative memories. Revista Mexicana de Física. 53(1):10-20.
19. P. Sussner. Generalizing operations of binary auto-associative morphological memories using fuzzy set theory. Journal of mathematical imaging and vision, 19(2):81-93. 2003.
20. L. Trujillo, G. Olague. Automated Design of Image Operators that Detect Interest Points. Evolutionary Computation. MIT Press., vol. 16, no. 4, pp. 483-507., 2008.
21. L. Trujillo, G. Olague, F. Fernndez, E. Lutton. Behavior-based Speciation for Evolutionary Robotics. GECCO. Atlanta, GA, USA., 2008.
22. G. Olague, C. Puente. Honeybees as an Intelligent based Approach for 3D Reconstruction. ICPR. Hong Kong, China. 2006.
23. R. A. Vázquez and H. Sossa. A New Model of Associative Memories Network. ANNIP. Angers, France, May 9-12, 2007.
24. C. Yáñez M., J. L. Diaz de León S. Memorias asociativas basadas en relaciones de orden y operaciones binarias. Ph. D. Thesis abstract. Computación y Sistemas, Vol. 6, No. 4 pp. 300-311. 2003.
25. X. Yao. Evolving artificial neural networks. Proceedings of the IEEE, 87(9): 1423-1447, September 1999.

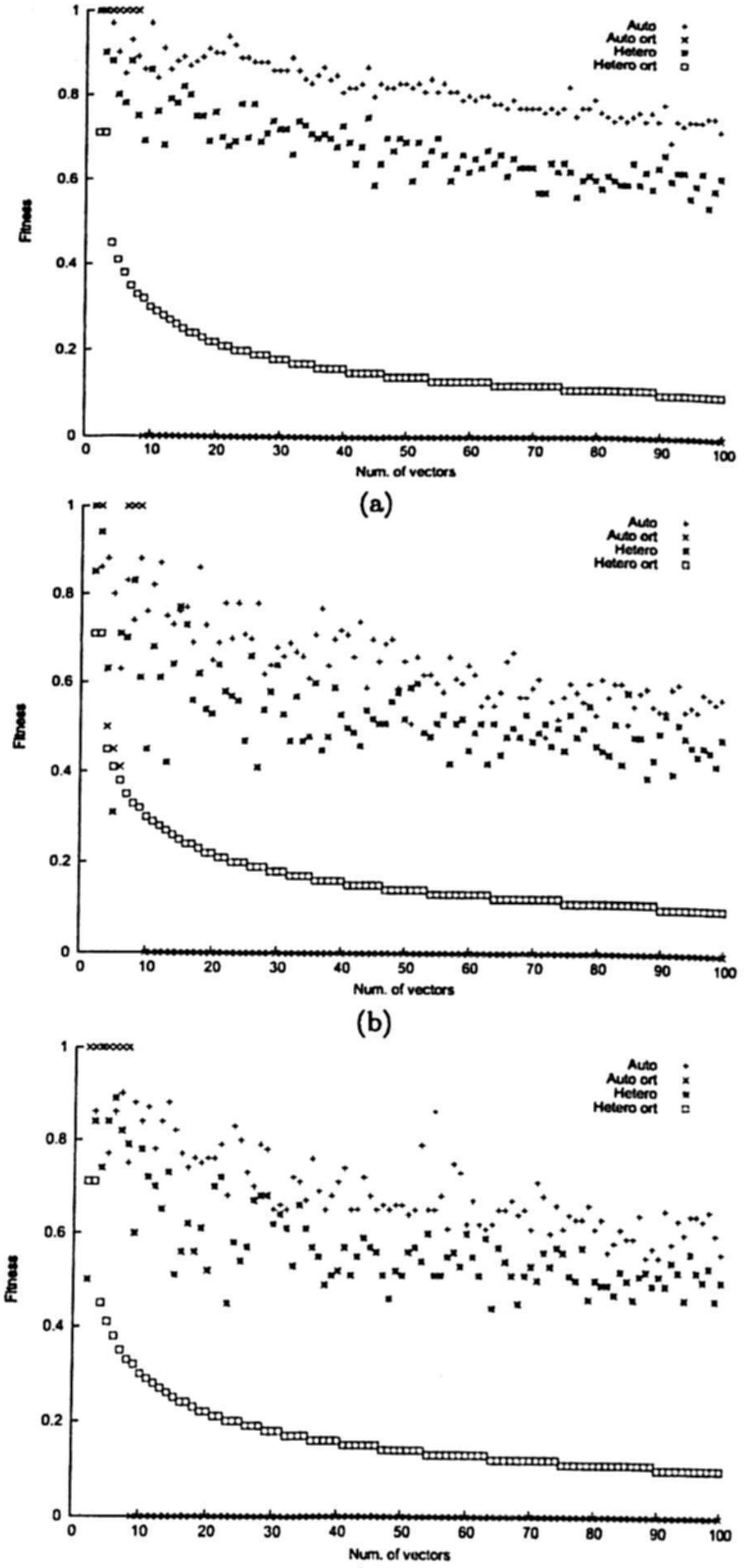


(a)

(b)

(c)

Fig. 5. Test of performance for evolutionary individuals generated from a random non orthogonal pattern sets under hetero-associative association.

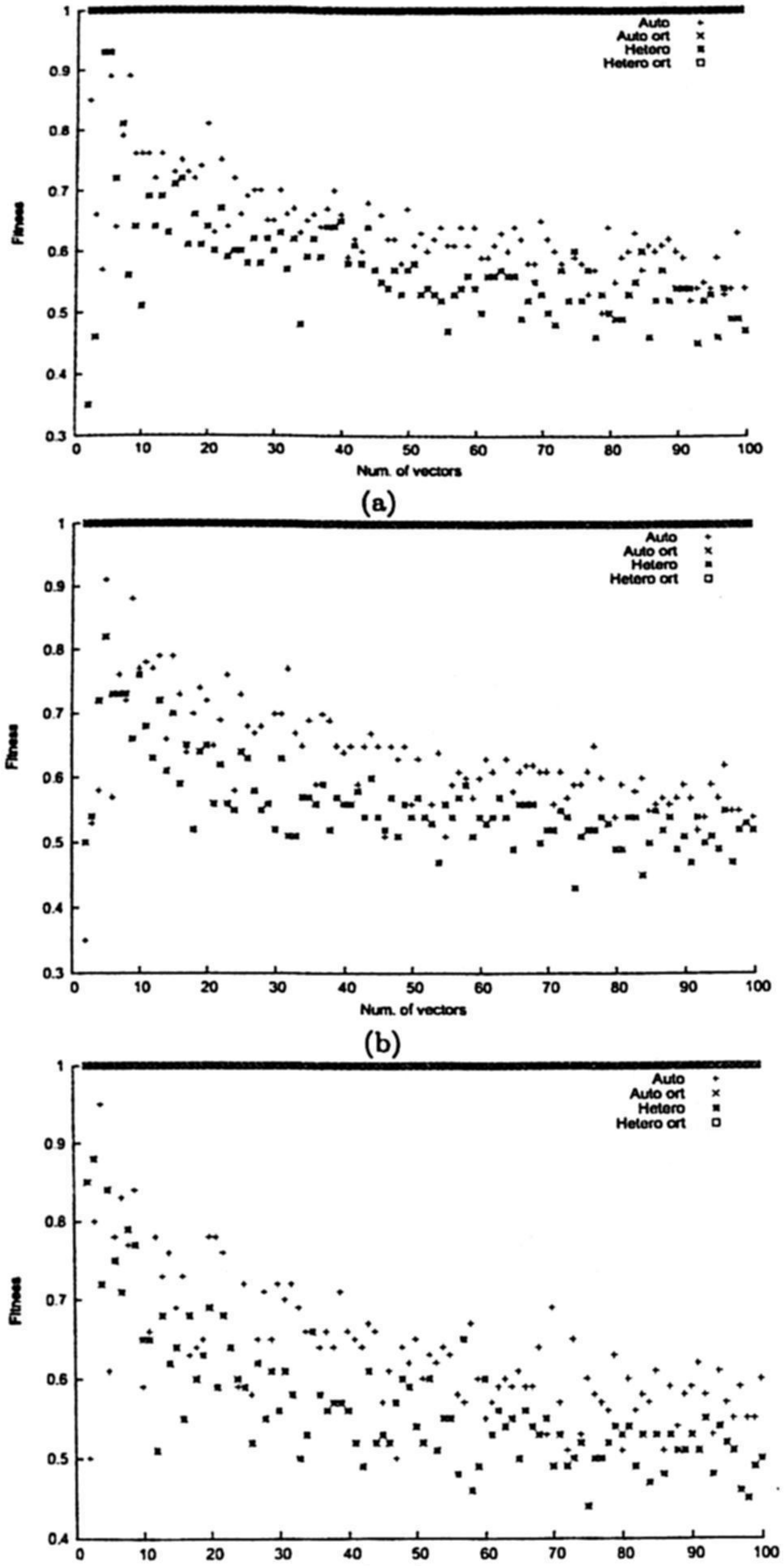


(a)

(b)

(c)

Fig. 6. Test of performance for evolutionary individuals generated from a random non orthogonal pattern sets under auto-associative association.



(a)

(b)

(c)

Fig. 7. Test of performance for evolutionary individuals generated from a random non orthogonal pattern sets under hetero-associative association.

Automatic Spatio-temporal Characterization of Criminal Activity. A New Algorithm and a Public Security Decision-support System

Víctor Manuel Martínez-Hernández[1], Salvador Godoy-Calderón[1],
Francisco Hiram Calvo Castro[1] and Marco A. Moreno-Armendáriz[2]

[1] Artificial Intelligence Laboratory
[2] Pattern Recognition Laboratory
Research Computer Center CIC-IPN
victorm@cic.ipn.mx, sgodoyc@cic.ipn.mx,
hcalvo@cic.ipn.mx, marco_moreno@cic.ipn.mx

Abstract. A new algorithm, using inductive learning techniques and capable of finding criminal activity trends is presented. This algorithm inductively constructs spatio-temporal characterizations of each type of criminal activity previously registered in a given study region. The program that executes the algorithm is also presented, along with some discussion about the advantages of its use. This kind of characterization provides valuable support for the public-security decision making process. It allows decision makers to better plan surveillance patrols and to concentrate efforts and/or resources on those areas (and times) of highest criminal incidence.

Keywords: crime analysis, inductive learning, crime patterns, crime prevention, public security, etc.

1 Introduction

Nowadays public security and crime fighting activities are some of the most important social concerns in every large city around the world [5]. Despite the enormous quantities of human and financial resources that Governments devote to these activities, the need for alternate mechanisms to increase the effectiveness and efficiency of the police forces daily work remains evident [7]. One of the main variables that severely limits both the effectiveness and efficiency of those activities is the reaction time to crime occurrences. Even the greatest efforts in creating and managing special "immediate reaction" forces, have shown that the marginal gain in this issue always turns out to be insufficient. And indeed it is insufficient to reduce the general crime incidence and to substantially modify the insecurity sensation among the citizens.

A better perspective of the situation can be achieved if the problem is ported to the scope of prevention instead of that of reaction. If public security authorities could determine, with an acceptable degree of precision, when and where criminal activities are characteristically taking place, a double benefit would be achieved. On one hand, it would be possible to concentrate all the logistic activities and resources necessary to prevent that specific kind of criminal activity in the geographic area and time frame

© G. Sidorov, M. Aldape, M. Martínez, S. Torres. (Eds.)
Advances in Computer Science and Engineering.
Research in Computing Science 42, 2009, pp. 103-111

Received 19/03/09
Accepted 27/04/09
Final version 07/05/09

forecasted. On the other hand, it would become possible to establish, in a dynamic but well supported manner, many of the parameters for public security daily activities, such as the design of specific patrolling itineraries, the allocation of resources and, of course the scheduling of security operations and of information and prevention campaigns by mass media.

As an additional asset for public security officials and politicians, the capability to learn the spatio-temporal trend in crime activities would provide an adequate foundation for evaluating their performance in decision making [7]. This foundation would be provided by the simple comparison between the foreseen trends and the countermeasures adopted and/or the outcome of the undertaken operations. All this would allow for a better determination of the annual budget devoted to public security, as well as its better allocation to the various entries related to this complex activity.

In this paper we present a new model and a strategy of criminal activity predicting by discovering its spatio-temporal trends within a specific time and zone under study [2]. The specific strategy proposed is based on an inductive learning process commonly used in pattern recognition for classification and clustering problems [1] [3]. The aforementioned model is one of the first steps in the process of building an intelligent automatic system to support decision-making processes in public security affairs.

2 Proposed Methodology

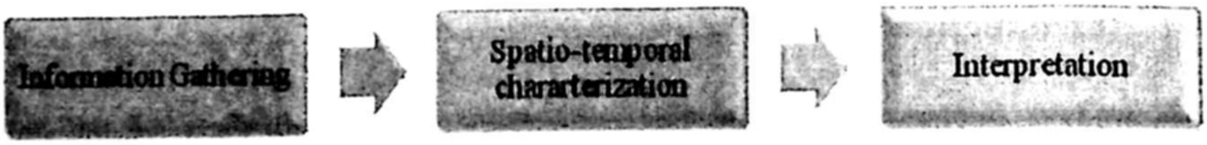


Fig. 1. Proposed Methodology

2.1 Information Gathering

It is necessary to gather information on crimes that occur in the region under study. The common sources of information are:

- Reports of Public Prosecutions.
- Citizen Alerts.

These reports are analyzed with the intention that each registered crime is transformed into a crime pattern, consisting of three components:

D= {Crime Act, Space, Time}

- Crime Act. Specifies the type of criminal activity that is registered (theft, rape, murder, etc.).

- Space. Specifies the place where the criminal act took place (surveillance sector, colony, street, etc.)
- Time. Indicates the time at which the crime occurred (year, month, day, hour, etc.).

2.2 Spatio-temporal Characterization

The learning process involves three specific phases: Aggregate crime patters into crime families, constructing two inductive definitions, one positive and one negative, for each crime family and, finally, constructing a neutral inductive definition for all types of criminal activity. Following, each of these stages is described:

2.2.1 Aggregate Crime Patterns into Crime Families

Once all registered crimes are converted into a crime patterns, the next step is to construct a classification of reference. To construct such sample, all patterns are clustered by the kind of preventive measures, material and logistic resources needed to prevent them. The choice of the number and nature of these families (classes) is entirely dependent on local policy and strategies. At the end of this stage, each class' patterns are ordered in time and space. Table 1 shows the resulting structure:

Table 1. classification of Reference

	$D_1 = (\{Space\}_1, \{Time\}_1)$	
.	.	
.	.	*Crime Family #1*
.	.	
	$D_p = (\{Space\}_p, \{Time\}_p)$	
	.	.
	.	.
	.	.
	$D_q = (\{Space\}_q, \{Time\}_q)$	
.	.	
.	.	*Crime Family #k*
.	.	
	$D_n = (\{Space\}_n, \{Time\}_n)$	

2.2.2 Constructing Inductive Definitions

At this stage it is necessary to determine the subsets of pattern features that will constitute the spatio-temporal properties to look for. A property is a subset of pattern features associated with specific values including both a spatial and temporal component:

$$P_i = \left\{ \begin{matrix} Property_{i_1} \\ < Value_{i_1} \end{matrix}, \ldots, \begin{matrix} Property_{i_n} \\ < Value_n \end{matrix} \right\}$$

Once these properties are defined, a query, looking for those properties, is run over the classification of reference. Those properties that appear more commonly in a class and less in others are considered to be characteristic of the class. The union of all the characteristic properties of a class (C_i) is called an Inductive Definition of that class, and has the following structure:

$$C_i = \{Property_{i_1}\} \cup \{Property_2\} \cup \ldots \cup \{Property_{i_m}\}$$

The algorithm includes the construction of two inductive definitions for each class and a neutral for all them. These definitions are constructed as follows:

a) **Positive Inductive Definition (C_i^{+}).**

1. Determine the characterization factor (F) for each spatio-temporal property (P) in each class (C_i):

$$F_{C_i}^{+}(P) = \frac{nP_{C_i}}{nP} \qquad\qquad 0 \leq F_{C_i}^{+}(P) \leq 1 \qquad\qquad (1)$$

Where nP_{C_i} is the number of occurrences of property P in class C_i, and nP is the total number of occurrences of property P in all classification of reference.

2. All spatio-temporal properties with characterization factor greater than a threshold $\beta_1$are considered to be included in the positive inductive definition for the class. The more close $F_{C_i}^{+}(P)$ is to 1, the more characteristic P is in class C_i.

b) **Negative Inductive Definition (C_i^{-}).**

All spatio-temporal properties with characterization factor less than a threshold β_2 and appearing in $\beta_2^{'}$ percent of the classes are included in the negative inductive definition for the class.

Threshold β_2 indicates the maximum characterization factor value a property must have to be considered as a negative characteristic. Threshold $\beta_2^{'}$ defines a minimum percentage of classes in which the property must appear to be considered as negative characteristic.

2.2.3 Building a Neutral Inductive definition.

a. Calculate the average characterization factor ($\hat{S}$) of each property (P) in all classes:

$$\hat{S}(P) = \frac{1}{k}\sum_{i=1}^{k} F_{C_i}^{+}(P) \qquad\qquad (2)$$

Where k is the number of classes.

b. For a property to be included in the neutral inductive definition it must satisfy the following condition:

P should have $|F^{+}_{C_i}(P) - \hat{F}(P)| \leq \beta_i$ in at least β'_i % of classes.

Once the spatio-temporal characterization is finished, positive, negative and neutral inductive definitions are integrated in the following manner:

Table 2. Inductive definitions

$C_i^{+} = \{Property_i\} \cup .. \cup \{Property_{i_{m_{i1}}}\}$ $C_i^{-} = \{Property_i\} \cup .. \cup \{Property_{i_{m_{ik}}}\}$	*Criminal Family #1*
$\vdots$	$\vdots$
$C_k^{+} = \{Property_k\} \cup .. \cup \{Property_{k_{m_{k1}}}\}$ $C_k^{-} = \{Property_k\} \cup .. \cup \{Property_{k_{m_{kn}}}\}$	*Criminal Family #k*
$L^{0} = \{Property_1\} \cup .. \cup \{Property_r\}$	

2.3 Interpretation

The set of inductive definitions have the following semantic:

a) The positive inductive definition of a crime family identifies places and moments where that specific criminal activity characteristically occurs (hotspot), therefore it suggests when and where it is necessary to increase surveillance and patrolling activities.

b) The negative inductive definition of a crime family identifies places and moments where that specific criminal activity characteristically does not occur, so, suggesting to re-assign resources from those places and moments to those hotspots identified in (a).

c) The neutral inductive definition has great importance in crime analysis; its properties identify those places and moments where the great majority of criminal activities take place although not in a characteristic manner. This definition suggests where and when it is necessary to implement more intensive surveillance and maybe to start preventive and informative mass media campaigns.

3 Experiments and Results

To test the proposed algorithm a set containing 16,995 registered crimes was used. This dataset was provided by the District of Cuautitlan Izcalli, Mexico and all crimes were committed in 195 neighborhoods in that district over a period of time from January 2004 to December 2008.

The registered crimes were converted into crime patterns and cluster into 17 crime families as shown in Table 3:

Table 4 shows the threshold used in the experimentation.

Table 3. classes and patrons used in the experimentation

Crime	Patterns
Housebreaking	393
Damage to private property	2635
Drug dealing	90
Illegal weapons	172
Homicide	435
Injuries	4890
Domestic violence	720
Burglary	477
Commerce and industry burglary	901
Robbery	2675
Assault	718
Public Transport mugging	84
Car parts theft	276
Smuggling	98
Car theft	2295
Kidnapping	89
Raping	47

Table 4. Threshold used in the experimentation.

Threshold	Value
β_1	0.65
β_2	0.2
β_2'	0.75
β_3	0.05
β_3'	0.30

Table 5 shows the number of positive and negative characteristic properties found for each crime family and the number of neighborhoods in which those criminal activities are characteristic.

Table 5. Positive and negative characteristic properties found for each crime family

Crime Family	Positive discriminating characteristic properties	Negative discriminating characteristic properties	Number of neighborhoods in which criminal activities are characteristic
Housebreaking	31	15	20/195
Properties Damage	263	5	73/195
Drugs dealing	18	25	4/195
Illegal Weapons	14	32	6/195
Homicide	49	8	28/195
Injuries	530	16	95/195
Domestic violence	56	12	38/195
Burglary	33	11	23/195
Industry burglary	56	10	36/195
Robbery	211	5	68/195
Assault	49	8	33/195
Transportation mugging	23	13	3/195
Car parts theft	46	11	20/195
Smuggling	24	2	7/195
Car theft	220	22	77/195
Kidnapping	8	2	6/195
Raping	4	4	4/195

The table 6 shows some spatio-temporal properties affected by Robbery.

Table 6. Characteristic properties affected by Robbery

Neighborhood	Month	Time
Bellavista	May	(15:01-18:00 hr)
	November	(15:01-18:00 hr)
Bosques de la Hacienda	June	(00:01-03:00 hr)
	September	(12:01-15:00 hr)
Bosques de Morelos	May	(09:01-12:00 hr)
	November	(09:01-12:00 hr)
Cofradia II	February	(21:01-00:00 hr)
Colonias del lago I	September	(00:01-03:00 hr)

Table 7 shows some spatio-temporal properties affected by some other crime families.

Table 7. Spatio-temporal properties affected by crime families

Neighborhood	Month	Time	Crime Families
Cumbria	January	(21:01-00:00 hr)	House Burglary Homicide Street Robbery Transportation mugging Car parts theft Car theft

4 Conclusions

The development of a new crime analysis model, based on inductive learning techniques is presented. The proposed model has some advantages over traditional criminal analysis: To mention some of them, algorithms of clustering are used in this kind of analysis in order to find those areas which are affected by a kind of criminal activity in an period of time (year, month, week, etc), so it is possible to use layers in geographical information systems (GIS) to represent the result of this analysis. However, criminal activities have to be analyzed by separated, so there is not a direct way to determine which criminal activity have more impact in areas with several criminal activities; the relation between criminal activities in an analyzed area is very important in order to implement preventive plans that reduce those criminal activities with more impact. The proposed model used characteristic spatio-temporal properties to decide which criminal activity have more impact in the analyzed area.

In the same context, traditional criminal analysis do not mention anything about how to use their result, when and where it is necessary to increase surveillance and patrolling activities, where and where it is possible to reduce surveillance and resource to attend those area which requires more resources, and finally which areas require special attention, because the great majority of criminal activities take place although not in a characteristic manner.

Other advantage of the proposed model is that the inductive definitions of each class, constructed by the spatio-temporal characterization procedure constitutes, by itself, valuable information for describing the criminal trends being studied; on the other hand, by adjusting the value of each threshold, it is possible to fine tune the level of precision we want to have in the inductive description of each class.

The type of inductive learning procedure presented in this paper is commonly used for supervised classification [1, 3] and conceptual clustering [5]. The proposed analysis method uses it in order to discover crime trends in large databases.

Finally, it is noteworthy that the proposed model opens new possibilities for the design and implementation of crime analysis systems where traditionally only statistical analysis had been used.

Reference

1. Baskakova L.V., Zhuravliov Y.I., "Modelo de algoritmos de Reconocimiento con Conjuntos de Representantes y sistemas de Conjuntos de Apoyo". Zh. Vivhislitielnoi Matematiki I Matematichskoi Fiziki. Tom 21, No.5, URSS, 1981.
2. Brachman R.J., Anand T., "The process of knowledge discovery in databases: A human centered approach". In U.M. Fayyad, G. Piatetsky-Shapiro, P.Smyth and R. Uthurusamy (editors). Advances in Knowledge Discovery and Data Mining, Chap.2. AAAI/MIT Press, 1996.
3. Diukova E.V., "About a parametric model of KORA-type Recognition algorithms". Soovshenia po prikladmoi matematiki. URSS, 1983.
4. Michalski R., Steep R.E. "A theory and Methodology of Inductive Learning". J. Carbonell, editor. Machine Learning: A Artificial Intelligence Approach. Chapter 11, Ed. Tioga, Palo Alto, California. 1984.
5. Oakley B. Robert, Goldberg M. Eliot, Dziedzic J. Michael, "Policing the New World Disorder: Peace Operations and Public Security". University Press of the Pacific. June 2002.
6. Quinlan J.R., "Discovering Rules by Induction from Large Collection of Examples". Michele, D. Editor, Expert System in the Microelectronics Age. Edinburgh University Press. 1979.
7. United Nations Publications. "Fighting crime in the twenty-first century: five regional meetings prepare for 1990 congress". 1990. Eighth UN Congress on the Prevention of Crime and the Treatment of Offenders. UN Chronicle. July 28, 2005.

An Adjusted Variable Neighborhood Search Algorithm Applied to the Geographical Clustering Problem

Beatriz Bernábe[1], Maria Osorio[2], Javier Ramírez[3],
José Espinosa[4] and Ricardo Aceves[5]

[1,2,4] BUAP, Benemérita Universidad Autónoma de Puebla,
Puebla, México
{Bernábe, Osorio, Espinosa}
Beatriz Bernabe, Facultad de Ciencias de la Computación
Maria Osorio, Facultad de Ingeniería Química
José Espinosa, Facultad de Ciencias Físico Matemáticas
[3] UAM, Universidad Autónoma Metropolitana. Departamento de Sistemas, México
[1,5] UNAM Universidad Nacional Autónoma de México, Departamento de Sistemas, México

Abstract. This paper describes the application of the Box-Behnken experimental design technique in the response surface methodology to find the best value for the parameters in the Variable Neighborhood Search algorithm for the geographical clustering problem. The solution of this problem demands a zone classification process where each zone is made of objects that best fulfill the objective, usually the minimum accumulated distance from the objects to the centroid in each zone; informally this process is named the geometric compactness. This well known application is an NP hard combinatorial problem. In this paper, we present the use of Variable Neighborhood Search VNS that has proven to be one of the best methods in the heuristic resolution of combinatory problems [9,10], but as a heuristic methodology, the conflict is centered in evaluating the quality of the solutions obtained and their corresponding parameters value [1].

Keywords: Geographical clustering Problem, Variable Neighborhood Search, Box-Behnken, Experimental Design.

1 Introduction

The geographic clustering problem (GCP) consists in the classification of objects in geographic units that fulfill a certain objective, mainly the geometric compactness [7,18,19]. The geographic units that have been considered correspond to AGEBs (Basic Geostatistic Areas) of the metropolitan zones in Toluca Valley MZTV [20].

The GCP problem belongs to the Territorial Design TD category and it is understood as the problem of grouping small geographic areas (basic areas) in greater geographical clusters called territories, in such a way that the acceptable grouping is the one that fulfills certain predetermined criteria [19]. The criteria or properties to fulfill in GCP problems depend on the space restrictions as continuity and geometric com-

© G. Sidorov, M. Aldape, M. Martínez, S. Torres. (Eds.)
Advances in Computer Science and Engineering.
Research in Computing Science 42, 2009, pp. 113-125

Received 15/03/09
Accepted 27/04/09
Final version 06/05/09

pactness [5,6,7,13,14,19]. The NP-hard condition of the GCP, implies solving a great number of geographic tasks that emphasizes the classification process directed toward the fulfillment of an objective.

Therefore, this problem is usually explained with a description oriented towards an optimization objective modeled mathematically as a cost function accompanied by the characteristics of the problem expressed as constraints. The NP nature of this classification problem, justify the utilization of a heuristic methodology to obtain a solution approximated to the optimal one [15] . The GCP is a special case of the classic clustering problem [7], but under the fulfillment of compactness, connectedness and/or homogeneity in some cases [19].

There are interesting works on classification under the criteria of minimal distances that have partially supported this paper [7,16] but they do not offer systematic methods to help to fit the parameters in a heuristic procedure according to the quality of the solutions offered [1,2,3]. In order to solve the problem, we selected PAM (Partitioning Around Mediods) [11] because is an exact and good partitioning algorithm that can be easily implemented and applied to the AGEBs, in order to obtain optimal solutions for small problems. We used the optimal solutions obtained and compared them with the solutions generated by Variable Neighborhood Search VNS for the GCP and used those results in the Box Bhenken experimental design developed for this paper.

To solve the GCP we developed our own partitioning algorithm that minimizes the distances between objects in order to obtain compactness between the AGEBs. However, the primary target of this research goes beyond revealing the solutions generated by the VNS, in this sense that we have applied a statistical methodology of Box Behnken experimental design to find the parameters that best directs the heuristic procedure to obtain high quality solutions because of its proximity to the optimal solution.

Since there are not clear methodologies to determine the right parameters for heuristic procedures as the VNS, our main contribution is centered exactly in this point, in the search and control of the statistical properties of the parameters in a VNS procedure, under a systematic process that allows us to observe the quality of the results for every combination in the design. In this context, this work presents a precise way to choose the correct parameters that lead to the generation of solutions of good quality.

This document is organized in 4 sections. The introduction is presented in section 1, in section 2 we present the Mathematical model for the geographical clustering problem and the solution and the Variable Neighborhood Search algorithm is presented in section 3. Section 4 describes the results obtained with the model and the validation of the parameters variation; and finally the conclusions are in section 5.

2 A Mathematical Model for the Geographic Clustering Problem

Many approaches have been used to solve the geographic clustering problem (CGP). The method utilized in this research to solve the AGEBs conglomerate design is similar to the method presented in [7], where the authors implemented a genetic algorithm for a similar zone design problem.

In the Geographic Clustering Problem solved here, the AGEBs are geographical units where each AGEB is separated by different distances of non uniform geometric structure, because the AGEBs are spatial data [14] and its geographical localization is given by latitude and longitude, that made easier the calculation of the distances between them. The AGEBs are clustered in a way that the AGEBs composing such groups are very close geographically, in order to minimize distances between them.

Basically, the strategy is to randomly choose AGEBs as centroids to identify the groups. Those AGEBs that are not centroids and have the shortest distance to a specific centroid-AGEB are members of a group or a cluster. This informal idea is the definition of geometrical compactness.

We did not define compactness in a formal way before, but the definition of compactness of geographic units is included in Definition 1:

Definition 1 **Compactness**

Let Z={1, 2, …, *n*} be the set of *n* objects to classify; the objective is to divide Z in *k* groups $G_1, G_2, \ldots, G_k$ with $k<n$, such that:

- $\bigcup_{i\ 1,k} G_i = Z$
- $G_i \cap G_j = \varnothing$ $\qquad i \neq j$
- $|G_i| \geq 1$ $\qquad i = 1,2, \ldots, k$

A group G_m with $|G_m| > 1$ is compact, if for every object $t \in G_m$ satisfies:

$$\underset{t \in Gm,\ i \neq t}{Min}\ d(t,\ i) < \underset{j \in Z\text{-}Gm}{Min}\ d(t, j)$$

A group G_m with $|G_m| = 1$ is compact only if its object t satisfies:

$$\underset{i \in Z - \{t\}}{Min}\ d(t,\ i) < \underset{j,l \in Gf,\ \forall f \neq m}{Min}\ d(j,\ l) \tag{1}$$

The neighborhood criterion between objects needed to achieve the compactness is given by the pairs of distances described in (1). Using this definition of compactness we will proceed to describe the model for the Geographic Clustering Problem (GCP).

2.1 Model for the Geographical Clustering Problem (GCP)

Data

UG= total of AGEBS

Let the initial set of n geographical units be

UG=$\{x_1, x_2, \ldots, x_n\}$, where

x_i is the i^{th} geographical unit (*i*=UG index)

k is the number of the zone (group)

The following variables are defined to refer to the different groups:
Z_i is the set of geographical units that belong to the i^{th} zone
n is the number of geographical units
C_t is the centroid
$d(i,j)$ is the euclidean distance from node i to node j (from one AGEB to another)

Constraints
$Z_i \neq \emptyset$ for $i = 1, \ldots, k$ (nonempty groups)
$Z_i \cap Z_j = \emptyset$ for $i \neq j$ (The same AGEBs cannot be in different groups)
$\bigcup_{t\ 1,k} G_t = U_g$ (The union of all the groups are all the AGEBSs

Objective Function Once the number of centroids (k) is decided (C_t with $t = 1, \ldots k$), the centroids will be randomly selected and the AGEBs will be assigned to the nearest centroids. Each AGEB i is assigned to the nearest centroid C_t. Then, for each AGEB i:

The objective function is the minimum of the sum of the distances between the centroids (for each k) and the AGEBs assigned to them. Each AGEB is assigned to the closest centroid (C_t).

$$\underset{t\ 1,\ldots,k}{Min}\ d(t, C_t)$$

For every k (where $k=1, \ldots, n$) the sum of the distances from every AGEBS assigned to each centroid is calculated and the minimum is selected. Therefore the objective function can be written as:

$$\underset{k=1,\ldots,nt}{Min}\left\{ Min\left\{ \sum_{t=1}^{k} \sum_{t \in c_t} d(i, c_t) \right\}\right\} \qquad (2)$$

3 The Variable Neighborhood Search (VNS)

The Variable Neighborhood Search (VNS) metaheuristic, proposed by Hansen and Mladenovic [9,10] is based in the observation that local minima tend to cluster in one or more areas of the searching space. Therefore when a local optimum is found, one can get advantage of its contained information. For example, the value of several variables may be equal or close to their values in the global optimum. Looking for better solutions, VNS starts exploring, first the nearby neighborhoods of its current solution, and gradually the more distant ones. There is a current solution *Sa* and a neighborhood of order *k* associated to each iteration of VNS. Two steps are executed in every iteration: first, the generation of a neighbor solution of *Sa*, named *Sp* ∈ *Nk* (*Sa*), and second, the application of a local search procedure on *Sp*, that leads to a new solution Sol. If Sol improves the current solution *Sa*, then the searching procedure will start now from G using $k = 1$. Otherwise, $k = k + 1$ and the procedure is repeated from Sa. The algorithm stops after a certain number of times that the complete exploration sequence $N_1;N_2; \ldots ;N_{kmax}$ is performed. The following algorithm shows how the solutions are obtained.

Two steps are executed in every iteration: first, the generation of a neighbor solution of Sa, named $Sp \in Nk\ (Sa)$,, and second, the application of a local search procedure on Sp, that leads to a new solution Sol. If Sol improves the current solution Sa, then the searching procedure will start now from G using $k = 1$. Otherwise, $k = k + 1$ and the procedure is repeated from Sa. The algorithm stops after a certain number of times that the complete exploration sequence $N_1;N_2; \ldots ;N_{kmax}$ is performed. The following algorithm shows how the solutions are obtained.

***Procedure** Variable Neighborhood Search (VNS)*

```
BEGIN
/*   Nk : k = 1, …, kmax, neighborhood structures */
/*   Sa   : current solution */
/*   Sp   : neighbor solution of  Sa */
/*   Sol: local optima solution */

REPEAT UNTIL (End) DO
     k ← 1
     REPEAT UNTIL (k ← kmax) DO
/*   Generate neighbor Sp of the kth neighborhood of
          Sa (Sp ∈ Nk (Sa))*/

          Sp ← GetNeighbor (Sa, Nk);
          Sol ← LocalSearch (Sp);
          IF (Sol is better than Sa) THEN
              Sa ← Sol;
          ELSE
              k ← k + 1
          ENDIf
     ENDDO
ENDDO
END
```

Partitioning algorithms match the objective function (2) in order to minimize the distance of the objects to their centroids. The geographic clustering algorithm is following the same objective with the use of VNS. The following pseudocode is commented in order to highlight the performance of both cycles and the search of the minimum distance in the objective function.

3.1 VNS Algorithm for the Geographical Clustering Problem (GCP)

Let n be the number of objects to classify
UG$_{ij}$ denotes that the object i is assigned to the centroid j for i=1,..,n; j=1,…,k
Let M={M$_1$,M$_2$,,M$_k$} be a solution of K centroids

MaxVNS /**maximum number of iterations to go over all the neighborhood search */*

MaxLS /**number of iterations of Local Search (LS) for each neighborhood */*

1. Initialization

```
/* Get an initial solution */
Generate initial random centroids $M = \{M_1, M_2, …, M_k\}$
/* Any AGEB can be a randomly obtained centroid */
BEGIN
       Current _cost ← Cost (M)
/* Another solution is generated and compared with the current so-
lution.  The  best solution is stored  */
cont←1
WHILE cont < MaxVNS DO
    BEGIN
       k-neighborhood ← 1
       /* Control variable */
       WHILE kneighborhood <> n DO
         BEGIN
              C ← Generates a random solution with a k-
                 neighborhood
              /* Gets a neighbor of k-neighborhood */
              Sol_neighborhood← LocalSearch (C);
              IF(Cost (Sol_neighborhood)<current_cost) THEN
                     M ← Sol_neighborhood;
              ELSE k-neighborhood ← k-neighborhood +1;
         ENDWHILE
/*Go to the next neighborhood only if the current
  solution(M) is not improved */
    ENDWHILE
       Cont ← cont+1
END
Return (M)  /* Solution with the minimum cost */
```

2. Cost Function (Sol)

```
/* Determine the quality of the solution  SOL, i.e. how much the
objective is minimized */

BEGIN
i ← 1
/* Initialize the first object */
 cost ← 0
WHILE (i ≤ n) DO
 BEGIN
/* For each object in Ug do */
       IF ($Ug_i$ is not a centroid) THEN
       BEGIN
           dmin ← dist($Sol_1$ , $Ug_i$ )
/* Represents the distance between the object and the $Sol_1$ (first
centroid where Sol represents the set of centroids). The distance
between each object and its nearest centroid is calculated */
           j ← 2
/* Go to the second centroid  */
              WHILE (j ≤ k) THEN
             BEGIN
                     IF (dist ($Sol_j$ , $Ug_i$) < dmin) THEN
```

```
/* Calculate the distance between the object i and the  $Sol_j$
   (another centroid) */
                              dmin ← dist ($Sol_j$ , $Ug_i$)
                         ENDIF
                         j ← j + 1
/* Go to the next centroid   */
                   ENDWHILE
                  cost ← cost + dmin
          ENDIF
        i ←i + 1
 ENDWHILE
Cost(Sol) ← cost
END
```

The Local Search (LS) algorithm improves the current solution searching in its neighborhood. It can finish finding a better solution or reaching the maximum number of iterations. The maximum number of iterations avoids cycling in the case that a better solution cannot be found.

4 Experimental Design

This section presents the necessary conditions to diminish the compactness between AGEBs, used as a function cost in the CGP solved with the VNS heuristic. The control variables used are the neighborhood structures (NS), the local search iterations (LS) and the number of groups (G) to consider. The quality of the results is systematically evaluated to identify the influence of the control parameters on the cost function and to model the dependency, exploring its influence in the obtaining of local optima solutions.

The experiments were performed in a computer with an Intel Centrino® processor, speed of 1.4 Ghz 768MB in RAM memory and 80 GB of Hard Disk memory. We tested 171 variables in 473 AGEBs. Each AGEB includes data of 55 blocks, in average.

An experimental design of answers' surface with a set of tests that deliberately change some variables with other remaining fixed in the system allowed us to observe the changes in the output variables and to explore the effects described in the previous paragraph.

4.1 Response Surfaces

The methodology of response surfaces is a combination of techniques of design and analysis of experiments that used in a sequential way, allow researchers, the determination of the operation conditions that produces solutions near to the optima [12].

A smooth complex function can come near locally (in "small" zones of the operation region) using polynomials of low order. If the zone where the local approach is made is "far" from the zone where the maximum is found, a polynomial of first order is a good approach. However, if the zone is "close" to the zone where the maximum is, it is necessary to use a polynomial of second order to describe the function.

A systematized analysis can be developed using a Box-Behnken design type. This type of design, due to its characteristics is easy to carry out for defined and adapted levels of the design parameters; besides, a rotary design with equal variance can be made for all points in the experiment that are equidistant to the center of the design region. On the other hand it is possible to make sequential experiments in zones that we pruned in order to study the individual effects of the control parameters and the combination of the synchronized effects [12].

Another advantage of this design is that the results can be modeled with a second order function and an analysis of the behavior of the cost function can be modeled using the methodology of the response surfaces.

A Box-Behnken design for five parameters of control is used in an experiment with 15 combinations and four central points. The results obtained with the heuristic method are used for choosing the levels of the parameters and the definition of the experimentation region. The parameter levels used in the experiment can be seen in Table 1. The nomenclature used in the tables 1 and 2, is: NS (neighborhood structures). LS (Local Search), G (Groups), FC (Cost Function).

The Box-Behnken's matrix of parameters design are: Factors = 3, Replicates = 1, Base runs = 15; Total runs = 15; Base blocks = 1; Total blocks = 1; Center points = 3. With these levels and parameters, 15 experimental combinations were tested.

In the test with Standard Order of 8 we observed that with 24 groups and parameters of LS = 530 and NS = 640 the cost function is 10.8371. This is the closest value to the optimal objective of 9.279 obtained with PAM. In contrast, PAM managed to get that solution in 27 hours and our VNS algorithm reduced considerably the computational time to 13 minutes, with 616529 iterations and 16 accepted solutions. The behavior of the objective value, represented with the cost function, according to the number of iterations for the VNS heuristic can be seen in Fig. 1.

Table 1. Levels used during the experimentation.

Parameter	High Level	Center Level	Low Level
LS	848	530	212
NS	640	400	160
G	24	18	12

Table 2. Experimental tests for BB.

Std Order	Groups	LS	NS	FC
15	18	530	400	12.6586
2	24	212	400	10.9011
4	24	848	400	10.8866
1	12	212	400	15.4535
5	12	530	160	15.3206
7	12	530	640	15.3221
14	18	530	400	12.5667
6	24	530	160	11.0177
13	18	530	400	12.5597
10	18	848	160	12.4411
11	18	212	640	12.6957
8	24	530	640	10.8371
3	12	848	400	15.1598
12	18	848	640	12.4726
9	18	212	160	12.8541

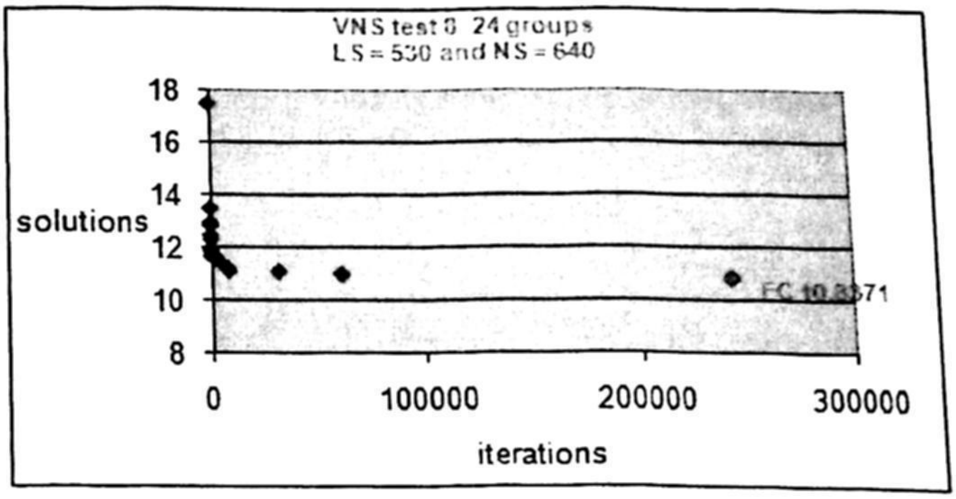


Fig. 1. Objective (Cost Function) vs. Number of Iterations for VNS Std. Order = 8, Groups =24, LS = 530, NS = 640, FC = 10.8371

This instance has been chosen as a representative example of the experiment designed, because it was found that 24 groups is an observed turning point in our multivariate statistical study.

4.2 Model for the Cost Function and Verification of the Experimental Model

Fig. 2 shows the residual plots for the experimental model. It can be concluded that the data behave normally and a second order model is well adapted.

The data obtained in the experiments were used in a second order regression model in order to get a prediction equation. The estimated regression coefficients are presented in Table 3.

Table 3 shows that the interaction effects between parameters are less important.

4.2 Predictions for the Cost Function

Figure 3 presents the response surface plots that show the effects of the variation of LS, NS and G on the cost function. It can be seen that the number of groups in a value of 24, generates a minimum in the function cost for high values in LS and NS.

An analysis of the surface plots allowed us to observe that the cost function tends to have reduced values for a greater number of groups, the NS value should be large and the LS value high. This analysis allowed us to limit the magnitude of the control parameters in the search of the minimum value in the cost function.

Contour plots were generated for regions in the surface plots with control parameters that generated cost function values near to the optimal.

It can be seen in Fig. 4 that the best cost function of 10.85, is reached for several values of BL and NS, fixing G in 24, and obtaining a contour of 10.8378, shown in the upper right of the graph. Otherwise for LS = 848, NS = 640 and G = 24 the objective function has a minimum value, as can be seen in Fig. 5. For 24 groups the optimal value obtained with PAM is 9.279 and the cost function obtained with the VNS is 10.8378 with 24 groups, LS = 848 and NS = 640 as an example. These parameters were used for the Regression with a Second Order model (See Fig. 5).

All solutions obtained by the VNS needed less than 13 minutes.

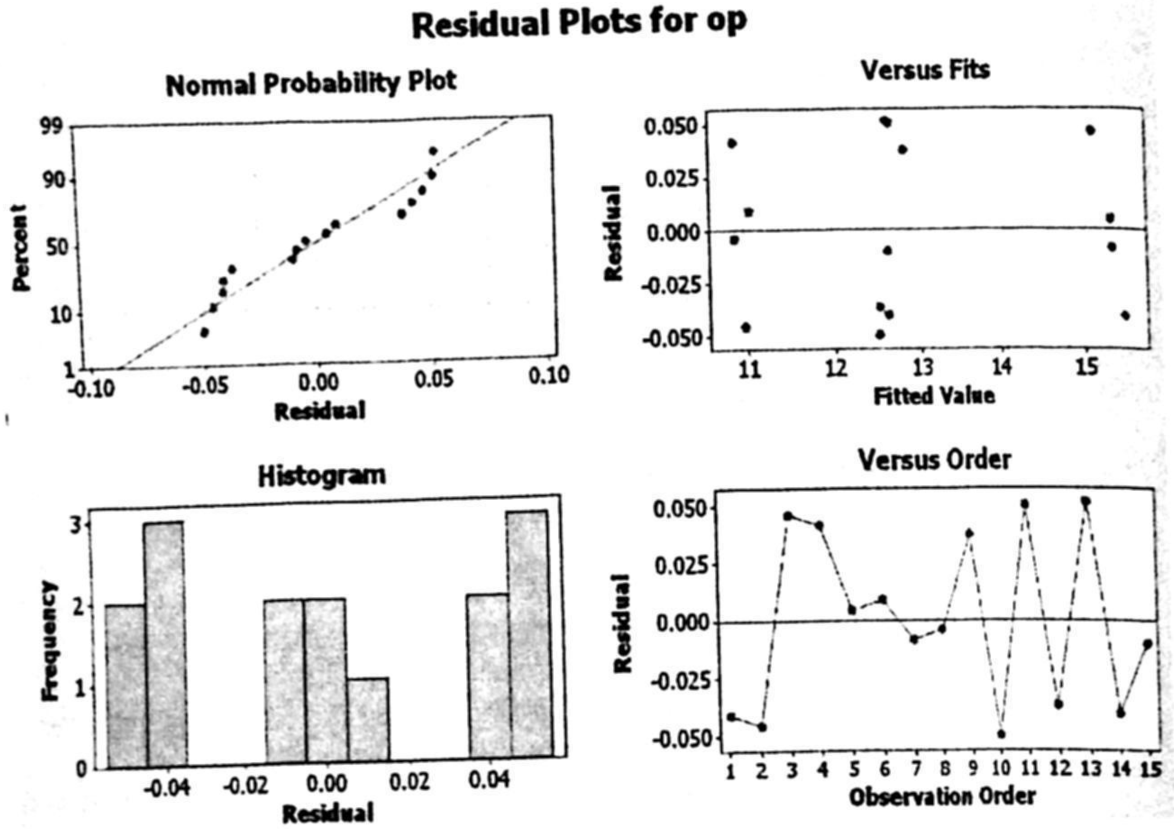


Fig. 2. Graphs of normal probability, residual and histogram of results

Table 3. Estimated Regression Coefficients

Term	Coef	SE Coef	T	P
Constant	24.1966	0.400207	60.460	0.000
G	-0.8701	0.035285	-24.658	0.000
LS	-0.0011	0.000489	-2.280	0.072
NS	-0.0002	0.000648	-0.280	0.791
G*G	0.0138	0.000912	15.175	0.000
LS*LS	-0.0000	0.000000	-0.315	0.766
NS*NS	0.0000	0.000001	0.573	0.592
G*LS	0.0000	0.000017	2.055	0.095
G*NS	-0.0000	0.000022	-1.443	0.209
LS*NS	0.0000	0.000000	1.505	0.193

S=0.06308, R-Sq=99.9%, R-Sq(adj)=99.9%

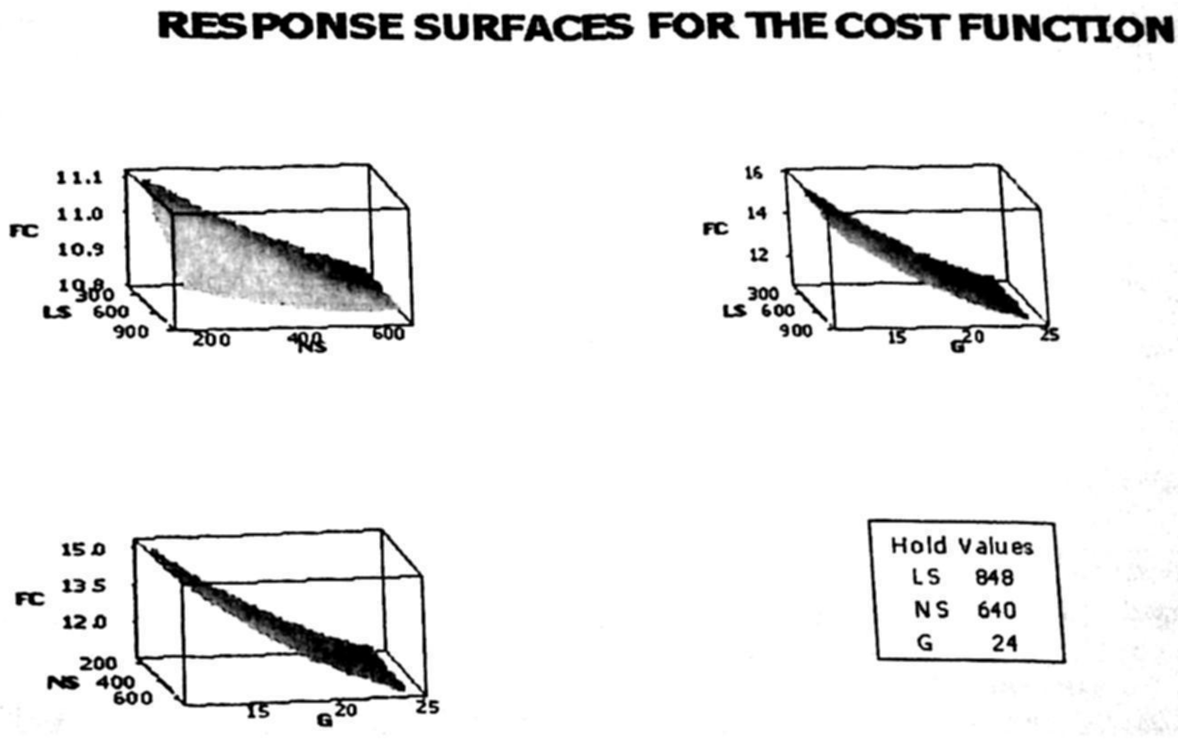


Fig. 3. Surface Plots for the Cost Function

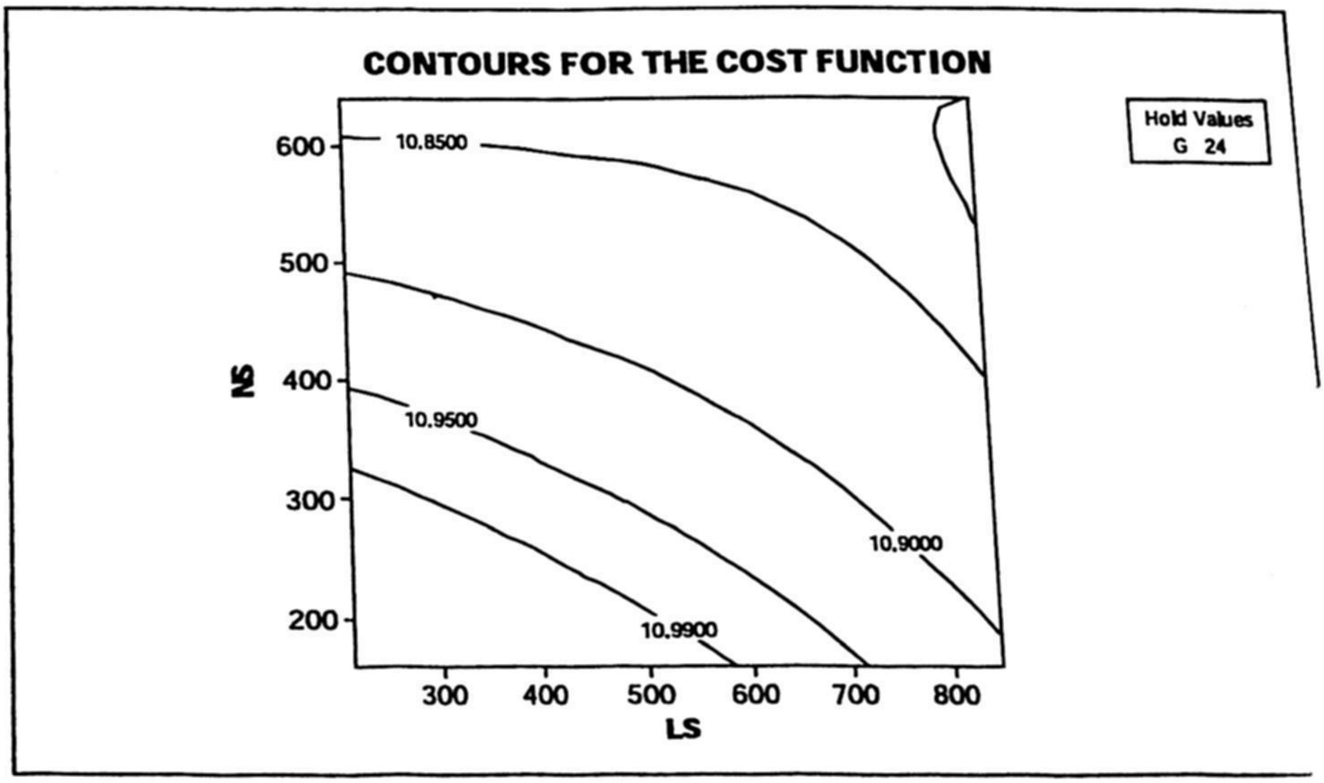


Fig. 4. Contour for 24 groups

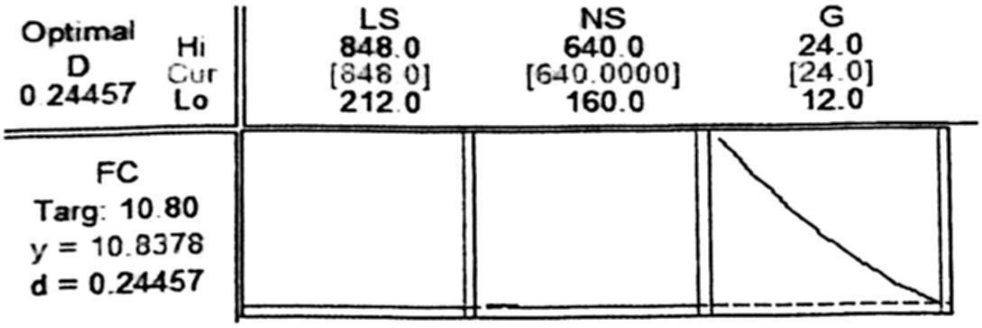


Fig. 5. Regression with the Second Order Model.

4 Conclusions

From the results obtained through all this work, we have found the VNS parameter values, used for solving the GCP that accompanied the best objective solutions.

a) In general, the number of groups is directly proportional to the quality of the solutions.

b) A value of NS close to 640 units, independently of the group size, will yield better values in the objective function.

c) The best objective values were found for LS values between 848 and 530.

The experiment used the results obtained with the empirical combinations, where 24 proved to be a good number of groups. For this reason, Tables 1 of section 4 used 24 groups. With these data development all the corresponding work tried to find a stationary point that could not be found. The future work pretends to extend the experiment, increasing the parameters value, leading to generate more instances that permit the experiment to be more extensive.

The experience obtained in the implementation of VNS for the GCP has been satisfactory in the sense that in few minutes the algorithm creates good quality solutions after 616529 iterations.

We have repeated the experiment using simulated annealing and got good solutions in a shorter time [4]. But the VNS solutions shown in this paper got better results. These results will allow us to perform a comparative work of both heuristics as a future research.

References

1. Barr, R. S., Golden, J.P., Resende, M. G., W.R. Stewart W.R.: Designing and Reporting on Computational Experiments with Heuristics Methods. Journal of Heuristics, Kluwer Academic Publisher (1995) 9–32.
2. Bernábe, L. B., López S.: Statistical Classificatory Analysis Applied to Population Zones. 8th. World Multiconference on Systemics, Cybernetics and Informatics (2004).

3. 3. Bernábe, L. B., A. Osorio, M. A., Duque J. C.: Clasificación Sobre Zonas Geográficas: Un Enfoque de Optimización Combinatoria para el Problema de Regionalización. XIII CLAIO Congreso Latino-Iberoamericano de Investigación Operativa (2006).
4. Bernábe, L. B., Ramírez R. J., Espinosa, R. J.: Evaluación de un algoritmo de recocido simulado con superficies de respuestas. Revista de Matemáticas Teoría y Aplicaciones, ISSN: 1409-2433 volume 16 number 1, (2009) 159-177
5. Duque, J. C.: Design of homogeneous territorial units. A Methodological Proposal and Applications. PhD dissertation, University of Barcelona (2004)
6. Duque, J. C., Ramos, R., Suriñach, J.: Supervised Regionalization Methods: A Survey. International Regional Science Review (2007) 195-220.
7. Fernando Bação, Victor Lobo, Marco Painho.: Applying genetic algorithms to zone design. Springer Verlag (2004)
8. Gordon, A. D.: A survey of constrained classification. Computational Statistics & Data Analysis (1996) 17–29.
9. Hansen P., Mladenovic, N.: Variable neighborhood search, Les Cahiers du GERAD (1996) 96-49.
10. Hansen P., Mladenovic, N.: Variable neighbourhood search. In Fred Glover and Gary A. Kochenberger, editors, Handbook of Metaheuristics, Kluwer (2003).
11. Kaufman, L., Rousseeuw, P.J.: 1987. Clustering by means of medoids. Statistical Data Analysis based on the L1 Norm, North-Holland, Amsterdam (1987) 405-416.
12. Montgomery D.: Design and Analisis of Experiments. Ed. Wiley 2ª Edition (1991)
13. Murtagh F.: A survey of algorithms for contiguity-constrained clustering and related problems. Computer Journal (1991) 82-88
14. Openshaw, S., Wymer C.: Classifying and regionalizing census data.In Census users handbook, ed. S. Openshaw, Cambridge, UK, GeoInformation International. (1995) 239-70
15. Pizza, E., Murilo, A., Trejos, J.: Nuevas técnicas de particionamiento en clasificación automática. Revista de Matemáticas Teoría y Aplicaciones, issn: 1409-2433 (1999) 51-66.
16. Romero D., Burguete J., Martinez E., Velasco J.: Parcelación del territorio nacional: Un enfoque de optimización combinatoria para la construcción de marcos de muestreo en hogares. INEGI (2004)
17. Rousseeuw, P.J., Hubert M., Struyf A.: Clustering in an object-oriented environment. Journal of Statistical Software (1997) 02-10
18. Zamora, A. E.: Implementación de un algoritmo compacto y homogéneo para la clasificación de zonas geográficas AGEBs bajo una interfaz gráfica. Tesis de Ingeniería en Ciencias de la Computación BUAP FCC, Asesor B. Bernábe (2006)
19. Zoltners, A., Sinha, P.: 1983, Towards a unified territory alignment: A review and model. Management Science, (1983) 1237-1256
20. http://www.inegi.gob.mx Instituto Nacional de Estadística, Geografía e Infomática (INEGI).

Induction of Decision Trees from Trained SVM Models using a TREPAN Based Approach

Douglas E. Torres D.

Decanato de Investigación y Postgrado DIP-UNEFA, Venezuela
Facultad de Ingeniería, Universidad Central de Venezuela
douglastd@cantv.net

Abstract. This paper describes the application of an Hybrid Intelligent System (HIS) to extract decision tress from a trained Support Vector Machine (SVM) model based on the TREPAN algorithm. TREPAN, a well-known technique developed originally to extract linguistic rules from a trained Artificial Neural Network, is modified to cope with SVM models. The proposed approach is tested on five data sets related to the medical domain, with excellent performance results.

1 Introduction

Support Vector Machines are efficient computing models that have shown excellent generalization performance in a variety of applications areas. However they have difficulty in explaining the results due to the lack of explanatory power and therefore are considered as "black-box " models. The same situation occurs with Neural Networks.

An important area of investigation that has been used in diverse application domains to develop interpretable expressions is based on the combination of different intelligent techniques such as neural networks, decision trees, systems based on fuzzy rules, reasoning based on cases, among others [1]. This operative synergy, called Hybrid Intelligent Systems (HIS) [2], seeks to improve the efficiency, reasoning power and comprehensibility of the integrand systems.

This paper presents, under the integrative perspective of HIS, an approach for the extraction of knowledge from SVM models. Several approaches have been proposed recently to obtain human interpretable expression, usually through rule extraction procedures from a trained SVM; see for example the works of Nuñez et al. [3], Fung et al. [4], Barakat and Diederich [5] and Zhanh et al [6]. In our study this knowledge is expressed through decision trees derived using a modification of TREPAN [7], an algorithm originally developed by Craven [8] to extract decision trees from a trained Artificial Neural Network (ANN).

The paper is organized as follows: Sec. 2 describes the SVM classifier. In Sec. 3 the TREPAN approach and its modifications are described, while Sec. 4 compares the results obtained by SVM and the modified TREPAN, on five publicly available data

Received 16/03/09
Accepted 27/04/09
Final version 08/05/09

sets and, finally Sec. 5 presents the conclusions.

2 Support Vector Machine

Support Vector Machines provide a novel approach to the two-category classification problem [9]. The methods have been successfully applied to a number of applications ranging from particle identification, face identification and text categorization to engine detection, bioinformatics and data base marketing. The approach is systematic and properly motivated by statistical learning theory [10].

Suppose we have N training data points $\{(\boldsymbol{X}_1, \boldsymbol{Y}_1), \ldots, (\boldsymbol{X}_N, \boldsymbol{Y}_N)\}$, where $\boldsymbol{X}_i$, i=1,.., N, is a vector of input variables and $\boldsymbol{Y}_i$ is the corresponding participation decision. Denote with S (resp. S) the convex hull of the points $\boldsymbol{X}_i$ with output +1 (resp. output –1). Thus, if S and S are linearly separable, we can think of constructing the optimal hyperplane $\boldsymbol{w} \cdot \boldsymbol{X}+b = 0$, which has maximum distance from these two convex hulls. The problem can be mathematically formulated as:

$$\begin{aligned} &\min_{\boldsymbol{w},b} \tfrac{1}{2}\, \boldsymbol{w}^{\mathrm{T}} \boldsymbol{w} \\ &\text{s.t.} \quad y_i(\boldsymbol{w} \cdot \boldsymbol{X} + b) \geq 1 \end{aligned} \tag{1}$$

where the quantities $\boldsymbol{w}$ and b are usually referred to as weight vector and bias [11]. This is a convex, quadratic programming problem in the unknowns ($\boldsymbol{w}$, b). It can be equivalently solved by searching for the values of the Lagrange multipliers α_i in the Wolfe dual problem. In this case we have $\boldsymbol{w} = \Sigma_i\, \alpha_i y_i\, \boldsymbol{X}_i$.

Only those points, which lie closest to the hyperplane, have $\alpha_i > 0$ and contribute to the above sum. These points are called support vectors and represent the essential information about the training set at hand.

Once we have found the optimal hyperplane, we simply determine on which side of the decision boundary a given test pattern $\boldsymbol{X}^*$ lies and assign the corresponding class label, using the function sgn ($\boldsymbol{w}{\cdot}\boldsymbol{X}^* + b$).

If the two convex hulls S and S are not linearly separable the optimal hyperplane can still be found by accepting a small number of misclassified points in the training set. A regularization factor C accounts for the trade off between training error and distance from S and S.

To adopt non-linear separating surfaces between the two classes, we can project the input vectors $\boldsymbol{X}_i$ into another high dimensional feature space through a proper mapping $\Phi(\cdot)$. If we employ the Wolfe dual problem to retrieve the optimal hyperplane in the projected space, it is not necessary to know the explicit form of the mapping Φ. We only need the inner product $K(\boldsymbol{X},\boldsymbol{X}') = \Phi(\boldsymbol{X}) \cdot \Phi(\boldsymbol{X}')$, which is usually called kernel function [12]. Different choices for the kernel function have been suggested; they must verify the Mercer's condition [13]; for example, the Gaussian radial basis function kernel: $K(\boldsymbol{X},\boldsymbol{X}') = \exp(-\gamma\|\boldsymbol{X}-\boldsymbol{X}'\|^2)$.

The need of properly choosing the kernel is a limitation of the support vector approach. In general, the SVM with lower complexity should be selected.

Several studies have recently reported on the application of Support Vector Machines (SVM) applied to medical databases [14,15,16].

3 Hybrid Intelligent Systems Models: The TREPAN algorithm

Hybrid Intelligent Systems are computational systems, which are based mainly on the integration of soft-computing techniques. This integration allows exploring their advantages in order to increase the overall system performance for a given task or to generate comprehensible representation of the knowledge [17]. With regard to the medical applications handled with hybrid intelligent systems, several studies reported in the literature concerning integration of soft-computing techniques as neural network and decision tress [18], evolutionary artificial neural networks [19], hierarchical soft computing [20].

In this paper the Extraction of Knowledge from an SVM model, which allows its validation and refinement, as well as the integration of connectionist and symbolic systems, is based on the TREPAN algorithm. TREPAN, originally developed to extract decision trees from a trained neural network, differs from other algorithms that extract information from neural networks in several ways [8]:

1. **The Oracle**. It is used to determine the class of each instance that is presented as a query. The Oracle is used for three different purposes: to determine the class labels for the network's training examples; to determine the class labels for the tree's leaves; and to select the splits that create each of the tree's internal nodes.
2. **Split types**. That is, the way the input space is partitioned. TREPAN forms trees that use M-of-N expressions for its splits, that is a Boolean expression specified by an integer threshold, m, and a set of n Boolean conditions. An M-of-N expression is satisfied when at least m of its n conditions are satisfied.
3. **Split Selection**. Split selection involves deciding how to partition the input space at a given internal node in the tree. TREPAN uses a special heuristic search process to build its splitting test.
4. **Tree expansion**. TREPAN grows trees using a best-first expansion that chooses the node where there is the greatest potential to increase the fidelity of the extracted tree to the network.
5. **Stopping Criteria**. Trepan uses local and global stopping criteria. A local criterion considers the state of only a single node to decide whether or not it should be made a leaf, and a global criterion considers the state of the entire tree to decide if the tree-growing process should stop.

As in any decision tree based approach, for example C4.5 [21], the amount of training data reaching each node decreases with the depth of the tree. TREPAN creates new training cases by sampling the distributions of the training data and uses the trained ANN as an oracle to answer queries during the learning process. TREPAN requires as input the weights and biases of the trained neural network and a training data set. As output it produces a decision tree that provides an approximation to the function represented by the ANN.

In this study, we used the TREPAN system developed by the Centre for Molecular Design at the University of Portsmouth [22] as part of the project: "Biological Data Mining: A comparison of neural networks and symbolic techniques". The program is a Matlab [23] implementation of the original TREPAN for classification problems and was successfully applied to neural networks in a variety of bioinformatics and chemoinformatics models. [24]. We replaced the original oracle based on an ANN trained model by an SVM model, trained with the Matlab Support Vector Machine Toolbox [25]. The modified TREPAN requires as input Lagrange multipliers, bias, a training data set, as well as kernel function and supports vector of the trained SVM. Figure 1 presents the data flow for the extraction of knowledge from an SVM trained model.

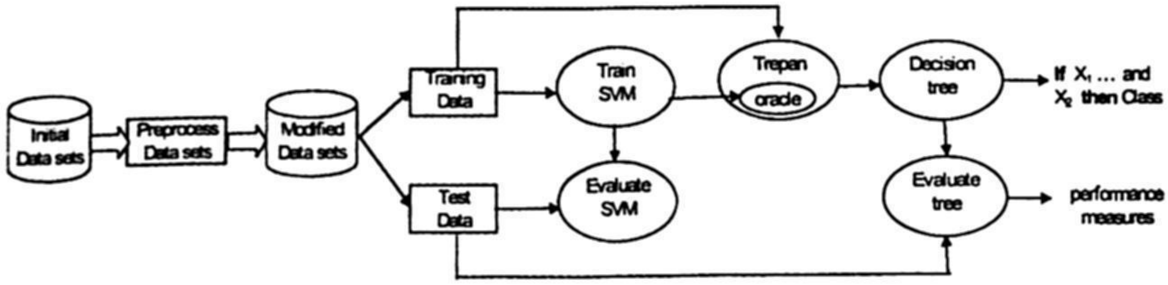


Fig. 1. Knowledge extraction data flow with TREPAN

As an example of the output generated by TREPAN, Figure 2 presents the decision tree extracted from the SVM model of the Haberman Survival data set (to be presented in the next section).

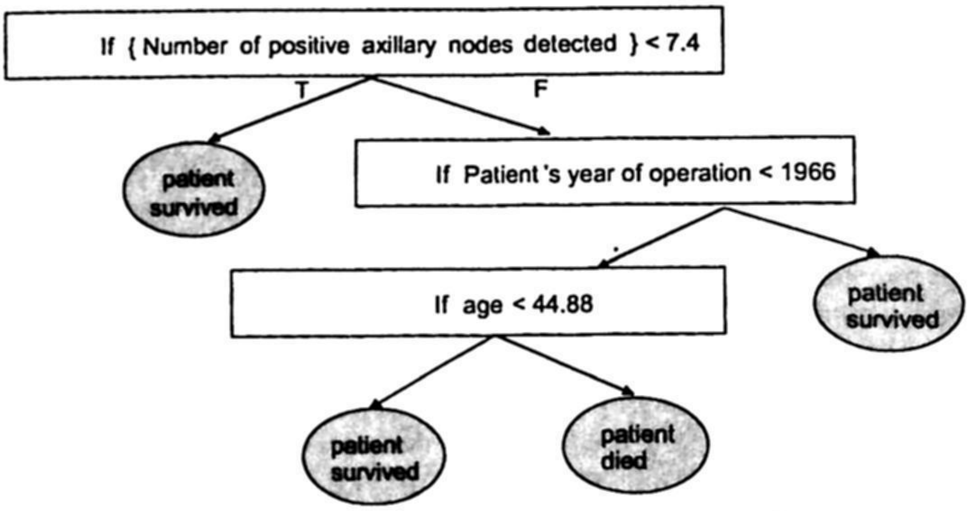

If { Number of positive axillary nodes detected } < 7.4
T
F
patient survived
If Patient 's year of operation < 1966
If age < 44.88
patient survived
patient survived
patient died

Fig. 2. Tree extracted by TREPAN for Haberman data set

Note that from the extracted tree, it is easy to obtain the following comprehensible rules:

- If Number of positive axillary nodes detected < 7.4 then class is patient survived 5 years or longer
- If Number of positive axillary nodes detected ≥ 7.4 and Patient's year of operation < 1966 and age < 44.88 then class is patient survived 5 years or longer

- If Number of positive axillary nodes detected ≥ 7.4 and Patient's year of operation < 1966 and age ≥ 44.88 then class is patient died within 5 year
- If Number of positive axillary nodes detected ≥ 7.4 and Patient's year of operation ≥ 1966 then class is patient survived 5 years or longer.

4 Experimental Results

4.1 Data Collection

We performed experiments on some commonly used data sets from the UCI repository [26] and Statlog project [27]. From the UCI Repository we selected the following data sets: Wisconsin Diagnostic Breast Cancer, Pima Indians Diabetes, Haberman Survival and Thyroid Gland. From the Statlog collection we selected the Heart disease. Table 1 shows the data sets characteristics: one of them correspond to multi-class data sets. A 10-fold cross-validation (CV) was performed. The data sets have been divided into 10 subsets of equal size. Each data sets was trained 10 times, each time leaving out one of the subsets from training and using only this omitted subset to evaluate the obtained model and average values are reported.

4.2 SVM Models

All data sets were trained using the Gaussian kernel and the Matlab Support Vector Machine Toolbox [25]. For all data sets we used the grid-search and cross-validation approach proposed by Hsu et al [28,29] using different kernel parameters in $\gamma = [2^4, 2^3, 2^2, ..., 2^{-10}]$ and $C = [2^{12}, 2^{11}, 2^{10}, .., 2^{-2}]$.

For multi-class data sets we used the one-against-all approach [30]. We built k SVM models, with k equal to the number of classes. The i-th model is trained with all of the examples in the i-th class with positive labels, and all other examples with negative labels. Average values of accuracy are reported.

Table 2 shows the performance of the trained SVM models during the testing phase along with the best parameters (γ,C) and the average number of support vectors derived. The performance is measured using sensitivity, specificity and accuracy indexes [31].

$$\text{sensitivit y} = \frac{\text{TP}}{\text{TP} + \text{FN}}; \quad \text{specificit y} = \frac{\text{TN}}{\text{TN} + \text{FP}} \qquad \textbf{(2)}$$
$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

where:

TP = Number of True Positive classified cases (the method correctly classifies)

TN = Number of True Negative classified cases (the method correctly classifies)
FP = Number of False Positive classified cases (the method labels a case as positive while it is a negative)
FN = Number of False Negative classified cases (the method labels a case as negative while it is a positive).

4.3 TREPAN Results

As previously described the trained SVM models were evaluated by TREPAN. Table 3 presents the average performance results for the testing phase for the data sets considered along with the average number of rules for the induced trees and the relative variation with respect to SVM accuracy. Table 3 shows that the Fidelity index (that is the percentage of predictions made by the tree extracted by TREPAN that agree with the predictions made by SVM), is over 90 % for all data sets, except Heart disease data set. The accuracy of the induced trees is very similar to the one obtained by SVM, with an average error of only 3.59 %. In others study the extracted rule from set has better generalization performance than the trained model, has been also reported in [24,33] for other data sets.

Table 1. Data sets and their characteristics

Data set	Number instances	Number Attributes	Attributes Type	Classes
Heart disease	270	13	Real, Nominal	2
Wisconsin Diagnostic Breast Cancer	569	32	Real	2
Pima Indians Diabetes	768	8	Real	2
Haberman Survival	306	4	Real	2
Thyroid Gland	215	5	Real	3

Table 2. Average performance results for SVM (Testing phase)

Data set	(γ, C)	Average Support Vectors	Sensitivity %	Specificity %	Accuracy %
Heart disease	$(2^{-6}, 2^{6})$	93.6	88.31	87.93	88.15
Wisconsin Diagnostic Breast Cancer	$(2^{-2}, 2^{3})$	69.1	98.09	98.06	98.07
Pima Indians Diabetes	$(2^{-2}, 2^{1})$	373.2	74.15	79.4	77.99
Haberman Survival	$(2^{-3}, 2^{3})$	147.3	60.61	77.66	75.82
Thyroid Gland	$(2^{-0}, 2^{4})$	19.3	98.36	95.41	97.36

Table 3. Average performance results for TREPAN (Testing phase)

Dataset	Fidelity %	Average Number of rules	Sensitivity %	Specificity %	Accuracy %	Relative Variation %
Heart disease	86.67	7.0	84.21	81.36	82.96	-5.89
Wisconsin Diagnostic Breast Cancer	94.55	4.5	92.31	94.46	93.67	-4.49
Pima Indians Diabetes	91.66	12.2	70.14	78.46	76.17	-2.33
Haberman Survival	94.6	3.8	62.07	77.26	75.82	-0.16
Thyroid Gland	93.87	3.9	94.19	88.37	92.25	-5.25

Figures 3-6 show the trees extracted by TREPAN for the datasets: Heart Disease, Wisconsin Diagnostic Breast Cancer, Pima Indians Diabetes and Thyroid Gland, respectively. The trees extracted correspond to the fold with the highest value of accuracy. These examples are related to a single class. For example, Fig. 3 shows the tree extracted by TREPAN for data set Heart disease. From here, a set of rules can be generated to classify absence or presence of heart disease.

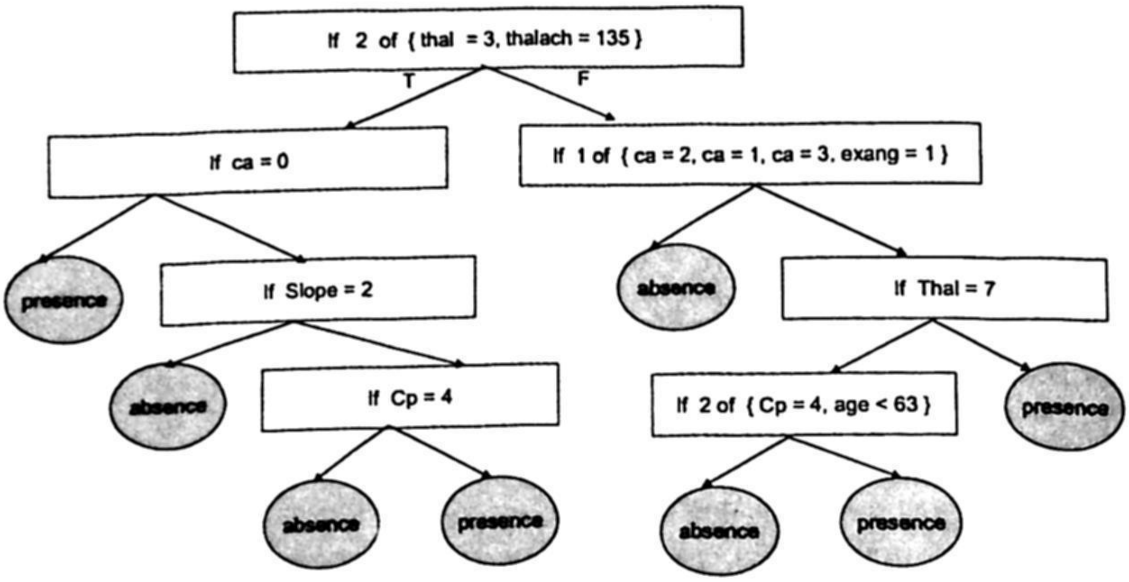

If 2 of { thal = 3, thalach = 135 }
T
F
If ca = 0
If 1 of { ca = 2, ca = 1, ca = 3, exang = 1 }
presence
If Slope = 2
absence
If Thal = 7
absence
If Cp = 4
If 2 of { Cp = 4, age < 63 }
presence
absence
presence
absence
presence

Fig. 3 Tree extracted by TREPAN for data set Heart disease

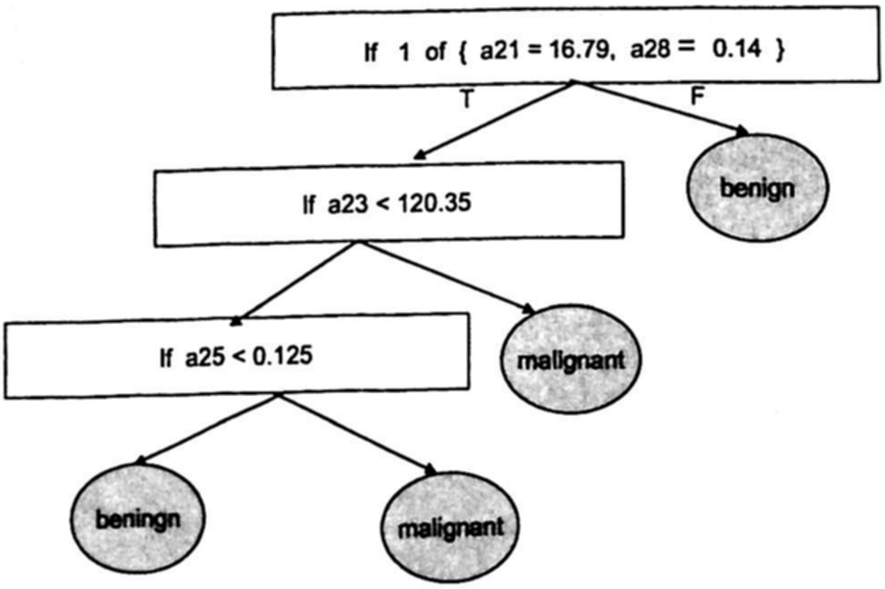


Fig. 4 Tree extracted by TREPAN for data set Wisconsin Diagnostic Breast Cancer

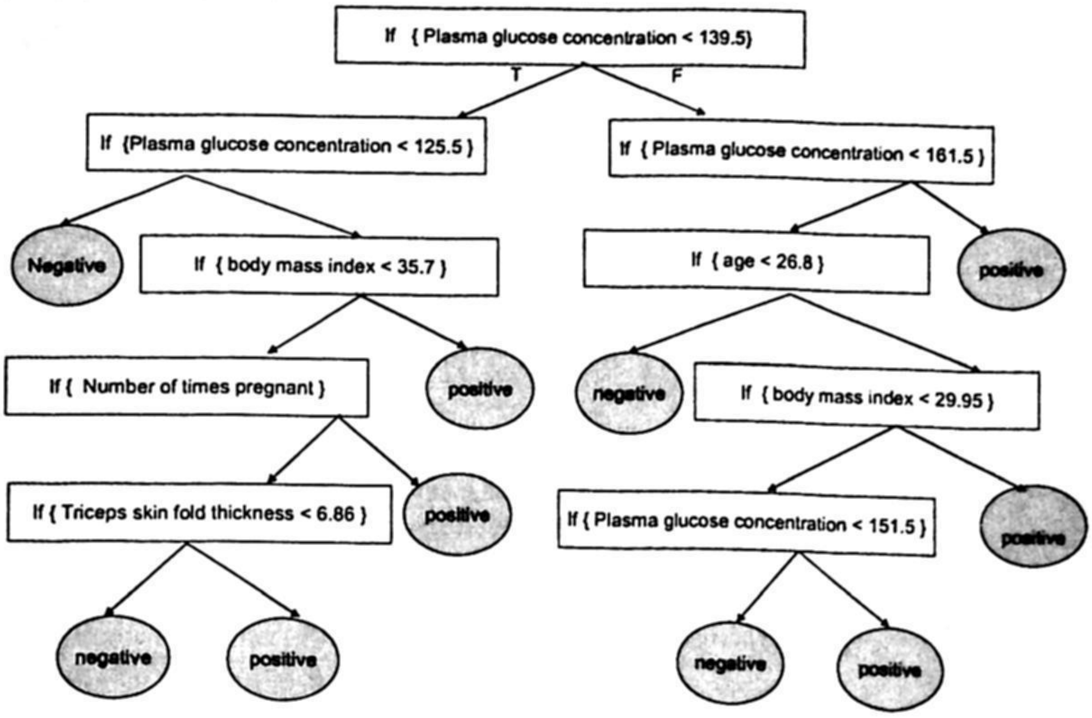

If { Plasma glucose concentration < 139.5}
T
F
If {Plasma glucose concentration < 125.5 }
If { Plasma glucose concentration < 161.5 }
Negative
If { body mass index < 35.7 }
If { age < 26.8 }
positive
If { Number of times pregnant }
positive
negative
If { body mass index < 29.95 }
If { Triceps skin fold thickness < 6.86 }
positive
If { Plasma glucose concentration < 151.5 }
positive
negative
positive
negative
positive

Fig. 5 Tree extracted by TREPAN for data set Pima Indians Diabetes

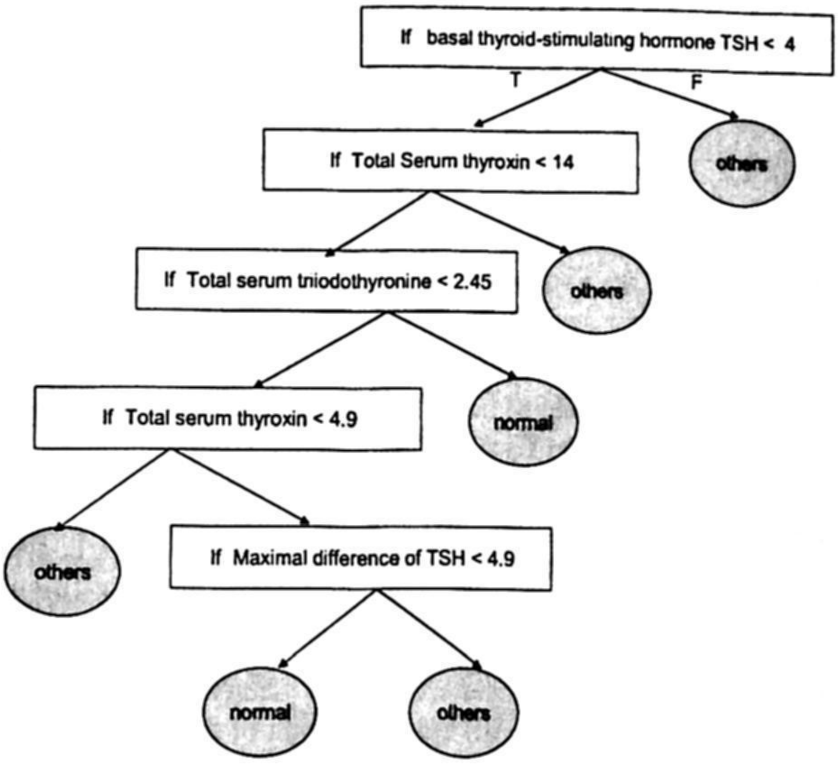

If basal thyroid-stimulating hormone TSH < 4
T
F
If Total Serum thyroxin < 14
others
If Total serum tniodothyronine < 2.45
others
If Total serum thyroxin < 4.9
normal
others
If Maximal difference of TSH < 4.9
normal
others

Fig. 6 Tree extracted by TREPAN for data set New Thyroid Gland

Table 4 shows the average accuracy rates for the data sets considered (for the testing phase), of SVM, TREPAN and C4.5, a well-known decision tree algorithm [21].

Table 4. Average Accuracy Rate (%) for SVM, TREPAN and C4.5

Data set	SVM	TREPAN	C4.5
Heart disease	88.15	82.96	78.52
Wisconsin Diagnostic Breast Cancer	98.07	93.67	94.73
Pima Indians Diabetes	77.99	76.17	75.39
Haberman Survival	75.82	75.82	71.9
Thyroid Gland	97.36	92.25	92.09

In order to empirically evaluate the accuracy of the models, a statistical method, as suggested by Mitchell [32] was used. Using a 95 % confidence level, the statistical test shows that average accuracy for SVM and TREPAN on Pima Indians Diabetes and Haberman Survival data sets are equal, while SVM outperforms TREPAN on the other data sets. However, TREPAN has a better explanation capability.

5 Conclusions

In this paper we have presented an approach for extracting rules from SVM trained models. The proposed approach, based on TREPAN, a modification of the well know TREPAN algorithm, was tested on five commonly data sets publicly available. The results showed that the proposed approach produces a classification system with performance indexes as accurate as the trained SVM models but providing a set of trees, which allow a better understanding of the data set under analysis.

Acknowledgements. The author are grateful to the anonymous referees for their comments and valuable suggestions on the earlier version of the paper.

References

1. Jain L.C, Martin N. M. (1998): Fusion of Neural Networks, Fuzzy Sets, and Genetic Algorithms, Industrial Applications. Ed. CRC Press.
2. Jacobsen H.A. (1998): A Generic Architecture for Hybrid Intelligent Systems. IEEE Fuzzy Systems IEEE Fuzzy Systems. Anchorage, Alaska.
3. Núñez H., Angulo C., Català A. (2002): Rule-extraction from Support Vector Machines, ESANN'2002 Proceedings – The European Symposium on Artificial Neural Networks, Bruges (Belgium), ISBN 2-930307-02-1, pp. 107-112.
4. Fung G., Sandilya S., Rao B.(2004): Rule extraction from Linear Support Vector Machines, Computer Aided Diagnosis & Therapy Solutions, Siemens Medical Solutions, Submitted.
5. Barakat N., Diederich J.(2004): Learning-based Rule-Extraction from support Vector Machines Performance On Benchmark Data Sets. In Kasabov, N. and chan, Z. s. H. Eds. Conference on Neuro-computing and Evolving Intelligence (KEDRI 2004), Auckland, New Zealand.
6. Zhang Y., Su Y., Jia T. and Chu J. (2005): Rule Extraction from Trained Support Vector Machines. PAKDD, LNAI 3518, Springer-Verlarg, Berlin
7. http://www.biostat.wisc.edu/~craven/
8. Craven M.W. (1996): Extracting Comprehensible Models from Trained Neural Networks. Ph.D. Thesis. University of Wisconsin-Madison.
9. Cristianini N., Shawe-Taylor J. (2000): An Introduction to Support Vector Machines. Cambridge University Press, Cambridge.
10. Vapnik V. (1998): Statistical Learning Theory, John Wiley & Sons.
11. Cortes C., Vapnik V. (1995): support-vector network. Machine Learning, 20:273-297
12. Guyon I, Stork G (2000) Linear Discriminant and Support Vector Classifiers In: Advances in Large Margin Classifiers ed. By Smola A.J. et al. The MIT Press Cambridge, MA.
13. Campbell C. (2000): An Introduction to Kernel Methods, In R.J. Howlett and L.C. Jain, editors, Radial Basic Function Networks: Design and Applications, Springer Verlag, Berlin, 2000, 31.
14. Valentini G., Muselli M., Ruffino F. (2004): Cancer recognition with bagged ensembles of support vector machines. Neurocomputing, 56: 461-466.
15. Bertoni A., Folgieri R., Valentini G. (2005): Bio-molecular cancer prediction with random subspace ensembles of support vector machines. Neurocomputing, 63: 535-539.
16. Cardoso J., Pinto J., Cardoso M. (2005): Modelling ordinal relations with SVMs: An application to objective aesthetic evaluation of breast cancer conservative treatment. Neural Networks, 18: 808-817.
17. http://www.comp.nus.edu.sg/~pris/HybridSystems/DescriptionDetailed1.html
18. Jerez J., Gómez J., Ramos G., Muñoz J., Alba E. (2003): A combined neural network and decision trees model for prognosis of breast cancer relapse. Artificial Intelligence in Medicine, 27: 45-63.
19. Kim K-J, Cho S-B. (2004): Prediction of colon cancer using an evolutionary neural network, 61: 361-379.

20. Yeh J-S., Cheng C-H. (2005): Using hierarchical soft method to discriminate microcyte anemia. Expert Systems with Applications, 29: 515-524.
21. Quinlan J.R. (1993): Programs for Machine Learning, Morgan Kaufmann Publishers.
22. Trepan – Matlab. Available at http:///www.cmdmport.ac.uk/biomine
23. MathWorks (2002) *Matlab 6.5* R13.
24. Hudson B., Whitley D., Ford M., Browne A. (2003): Biological Data Mining: A comparison of Neural Network and Symbolic Techniques. Technical Report. Centre for Molecular Design, University of Portsmouth. Available at: http://www.cmd.port.ac.uk/biomine
25. Cawley G. C. MATLAB Support Vector Machine Toolbox v. 0.54. University of East Anglia School of Information Systems, Norwich, U.K. Available at: http://theoval.sys.uea.ac.uk/~gcc/svm/toolbox
26. C.L.Blake and C.J.Merz: UCI Repository of Machine Learning Databases. University of California, Department of Information and computers Science, Irvine, CA. available at: http://www.ics.edu/~mlearn/MLRepository.html
27. Statlog. http://www.liacc.up.pt/ML/statlog/databases.html
28. Hsu C.-W, Lin C.J. (2002): A Comparison of Methods for Multi-class Support Vector Machines. IEEE Transaction on Neural Networks, 13: 415-425.
29. Hsu C.-W., Chang C.-C., Lin C.-J. (2004): A Practical Guide to Support Vector Classification. Department of Computer Science and Information Engineering, National Taiwan University. Available at http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf.
30. Anguita D., Ridella S., Sterpi D. (2004): A new Method for Multiclass Support Vector Machines Proc. of the IEEE Int. Joint Conf. on Neural Networks, IJCNN 2004, Budapest, Hungary, July 2004
31. Veropoulos K., Campbell C., Cristianini N. (199): Controlling the Sensitivity of Support Vector Machines, Proceedings of the IJCAI99.
32. Mitchell T. M. (1997): Machine Learning , McGraw Hill, New York.
33. Torres D., Rocco C. (2005): Extracting Trees from Trained SVM Models using TREPAN Based Approach. Hybrid Intelligent System, IEEE Computer Society, RJ, Brasil.

Computer Vision

A Methodology to Determine the Resolution of the Photograph Printed in the Mexican ID-card

Samuel Sanchez[1], Edgardo Riveron[2] and Salvador Godoy-Calderón[1]

[1,2] Patter Recognition Laboratory
[1]ssislas@yahoo.com.mx, [2]edgardo@cic.ipn.mx
[3] Artificial Intelligence Laboratory
[3]sgodoyc@gmail.com
Research Computer Center, National Polytechnic Institute,
Juan de Dios Bátiz s/n, esq. Miguel Othón de Mendizábal, Nueva Industrial Vallejo,
C. P. 07738, GAM

Abstract. The following paper shows a methodology to determine if the picture appearing on the ID-card is a digital 800 dots per inch (dpi) image. This "non-official" document is used as the main identification card in Mexico. For this methodology the basic concepts of mathematical morphology, such as dilation and erosion and the top-hat transformation are used to extract characteristics by using a previously defined structure element on the digital photograph. Subsequently the threshold is applied and the image is labeled to count the number of black dots contained in the resulting image. This procedure tells us if the picture printed on the ID-card is authentic. The analysis is carried out on the photograph scanned at 1200 dpi.

1 Introduction

The ID-card with photograph is a document issued by the Instituto Federal Electoral (IFE) (Federal Electoral Institute), which is autonomous. The Id-card is used to vote in local and federal elections in Mexico. Despite the fact that this Id-card is NOT an official document, it has become the main identification for Mexicans. It is the only Id-card issued to 18 year old and older citizens and it is also used for bank and cash transactions, and even to allow entry to night clubs or to sell cigarettes or alcoholic beverages to bearers in Mexico.

The ID-card has many shields and safety measures against forgery [1] [2]. One of the security measures mentioned in this paper relates to the photograph printed on the ID-card, which has some characteristics to discourage forgery.

Two kinds of photos are printed on the ID-card: the analogical photo, obtained with a conventional Polaroid [3] camera and a digital photo in the back, with an 800 dots per inch (dpi) resolution. This difference is due to the fact that before 2001 the IFE did not have the technology to print photos on plastic cards, so they used snapshots from a commercial Polaroid (analogical image) camera (Fig. 1a). Subsequently, the IFE decided to print the photo digitally, with 800 dpi resolution (Fig. 1b) [4].

Received 20/03/09
Accepted 27/04/09
Final version 07/05/09

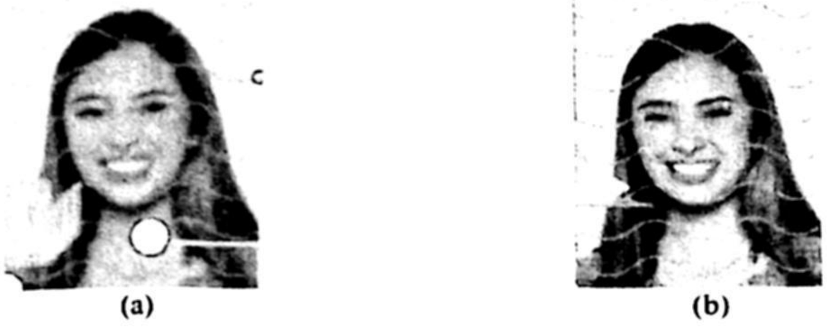


Fig. 1. Id-card Photos: a) Analogical, b) Digital with 800 dpi resolution.

2 Main Objective

The ID-card contains several security measures that make the document secure and impossible to duplicate. Nowadays there are no systems in Mexico capable of identifying a forged ID-card. Therefore, we are looking for a methodology that will allow us to authenticate ID-card photos which should be printed with an 800 dpi resolution as their special characteristic. With this information and other features, we will be able to decide on the authenticity of the Id-card.

3 Spatial Resolution

Intuitively, spatial resolution is a measure of the smallest discernible detail in an image. Quantitatively, *spatial resolution* can be stated in a number of ways, with *line pairs per unit* distance, and *dots (pixels) per unit distance* being among the most common measures. Suppose that we draw a chart with alternating black and white lines, each of width W units (*W can be less than 1*). The width of a *line pair* is thus $2W$, and there are $1/2W$ line pairs per unit distance. For example, if the width of a line is 0.1mm, there are 5 line pairs per unit distance (mm). A widely used definition of image resolution is the largest number of *discernible* line pairs per unit distance (e. g., 100 line pairs per mm). Dots per unit distance are a measure of image resolution used commonly in the printing and publishing industries. In the U. S., this measure is usually expressed as *dots per inch* (dpi). As an idea of the difference in quality, newspapers are printed with a 75dpi resolution, magazines with 133dpi, glossy brochures with 175dpi, and some books with 2400dpi [7].

4 Morphological Image Processing

Mathematic morphology is based on shape and geometry. Morphological operations simplify the image and preserve the main forms of objects. Mathematical morphology uses point set theory as well as the results of integral geometry and topology [9].

4.1 Top-Hat Transformation

One of the main applications of these transformations is removing objects from an image by using a structuring element in the opening or closing operations over objects that do not fit in it. The top-hat transform is used for clear objects located on a dark background; for this reason, the name *top-hat* is used frequently when we refer to this transformation. A common use of top-hat transformations is correcting the effects of non-homogeneous illumination [7].

The *top-hat transformation* of a gray-scale image *f* is defined as the original image,*f, minus its opening* (Eq. 1)*:*

$$Thatf=f-(f \circ b) \tag{1}$$

Where: $(f \circ b)$ is the morphological opening.

The *opening* of a set *A* by a structuring element *B, denoted A or B*, is defined as (Eq. 2):

$$f \circ b=(f\,\theta b)\oplus b \tag{2}$$

Thus, the opening *A* by *B* is the erosion of *A* by *B*, followed by the dilation of the result by *the* same structuring element *B*. The *Opening* generally smoothes the contour of an object, breaks narrow isthmuses and eliminates thin protrusions [12].
The *Erosion* is defined as (Eq. 3):

$$f\,\theta b(x,y)= mins,t\in b\{f(x+s,y+t)\} \tag{3}$$

The *Dilation* is defined as (Eq. 4):

$$f\oplus b(x,y)= maxs,t\in b\{f(x-s,y-t)\} \tag{4}$$

5 Methodology

Fig. 2 shows the flowchart of the methodology used to verify the spatial resolution of the photos printed on ID-cards. First we digitalize the ID-card with a resolution over the printed ID-card photo (in this case 1200 dpi) obtaining the area of interest, which is the photo (a). When we have the photo we must determine if it was analogically or digitally printed, by using the methodology employed in [13] (b). Later, if the photo happens to be digital, we cut a small portion of the image; the size of the area is determined by equation 5.

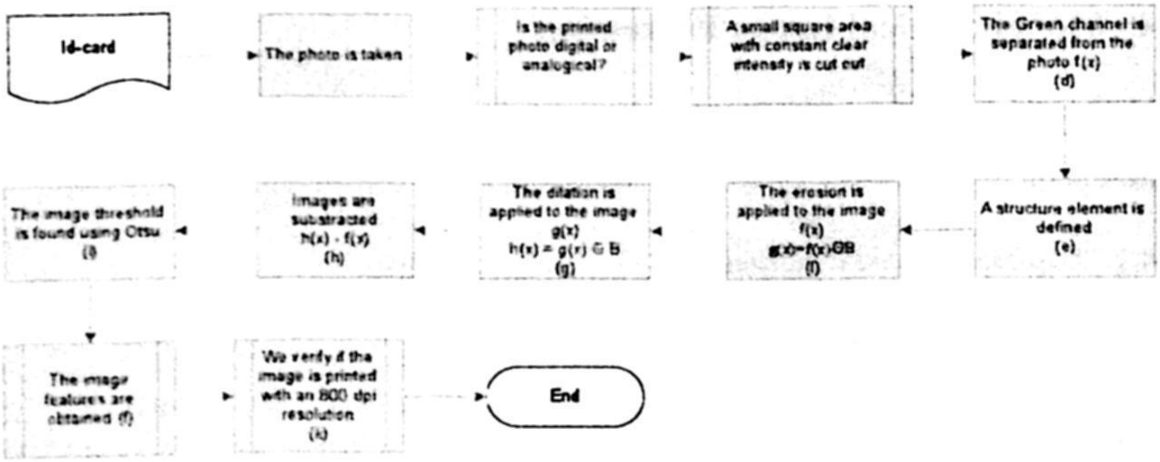

Id-card
The photo is taken
Is the printed photo digital or analogical?
A small square area with constant clear intensity is cut out
The Green channel is separated from the photo f(x) (d)
A structure element is defined (e)
The erosion is applied to the image f(x) g(x)=f(x)⊖B (f)
The dilation is applied to the image g(x) h(x) = g(x) ⊕ B (g)
Images are substracted h(x) - f(x) (h)
The image threshold is found using Otsu (i)
The image features are obtained (j)
We verify if the image is printed with an 800 dpi resolution (k)
End

Fig. 2. Block diagram of the methodology used.

$$\textit{Window size= 116Scanned Resolution} \qquad (5)$$

Where the *Window size*, is the size in pixels of the area to be analyzed in the photo. ***116*** corresponds to a sixteenth part of one inch, that is the area that must be verified; the Scanned *Resolution* is the value in pixels of the spatial resolution with which the image was digitalized.

In this case images where scanned in a 1200 dots per inch spatial resolution, so:

$$\textit{Window Size= 116(1200)}$$

Therefore:

$$\textit{Window Size=75 Pixels}$$

Subsequently to the cut (c) the Green channel of the image is selected (d). Then a square *structural* element must be defined (e) which size depends on the size of the printed dot. The greater the spatial resolution of the printed photo, the smaller the area of each one of the printed dots in the image, as we can see in Fig. 3.

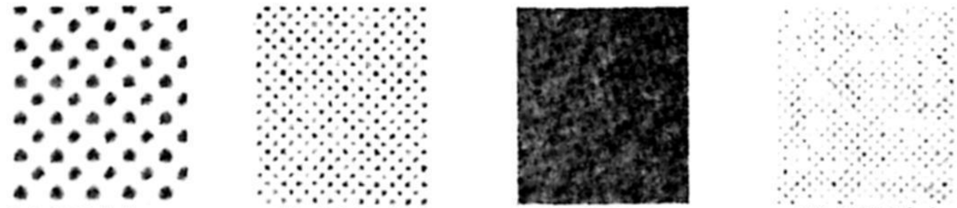

(a) 300dpi (b) 600dpi (c) 800dpi (d) 1200dpi

Fig. 3. Square image with dots in a different spatial resolution.

Then, the erosion is applied to the image, followed by the dilation with the same structure element (opening). Then the original image is subtracted with the result image (top-hat), so we can see the points where the ink has been printed in more quantity. Next, we threshold the image using Otsu [6] and we label the objects to finally get the features of the image; in this case, we label the points using different colors [8], then we count the total colors that appeared in the labeling image using equation 6, obtained experimentally; this equation allows us not to repeat the same color twice.

$$Color=64R+25+16G+25+2(B+25) \tag{6}$$

Where *R* is the color value in the Red channel, *G* is the color value in the Green channel and *B* is the color value in the Blue channel.

Finally, equation 7, which was determined empirically, is used to find out if the resolution corresponds to the 800 dots per inch in the area of the image and then the photo may be considered to be genuine.

$$Resolution= 2Points*161*2*Scanned\ ResolutionPrinted\ Resolution \tag{7}$$

Where:

Points is the number of the points that appeared inside the area which are approximately the same area; number *16* corresponds to the sixteenth part of an inch, that is the area that we analyzed; number 2 corresponds to the two points, one black and one white, that implicitly integrate spatial resolution; *Scanned Resolution* refers to the resolution of the digitized image; *Printed Resolution* means the assumed resolution of the original digital image.

In this case, the image was digitalized with a 1200 dpi spatial resolution and the printed image must be in the 800 dots per inch range according to the secure measure with which the photo was printed. Therefore, the formula is:

$$Resolution= 2Points*161*2*32 \tag{8}$$

So, the formula is reduced to:

$$Resolution= 2Points*48 \tag{9}$$

When the number of dots printed in the image through the Otsu algorithm is known, the resolution is obtained with formula 9 and we verify that it corresponds to the 800 dots per inch of the printed digital image.

6 Experimentation and Results

To corroborate the methodology, we made some experiments using genuine ID-cards digitalized with 1200 dots per inch (dpi) using an HP scanner, 367Q HP Scanjet along with some other printed images with different spatial resolutions: 300, 600, 800 and 1200 dpi.

Fig. 4. Digital photo from an ID-card.

From each image selected for the experiment we applied the previously described methodology (obtained from a clear area of the photo) and we applied the morphological transformations to get the printed dots and determine the spatial resolution of the print.

Fig. 5. Image of the square area to be analized.

In the methodology we used a 7 x 7 pixels square structure element in a disc form with reference point in [3,3] as we can see in Fig. 6.

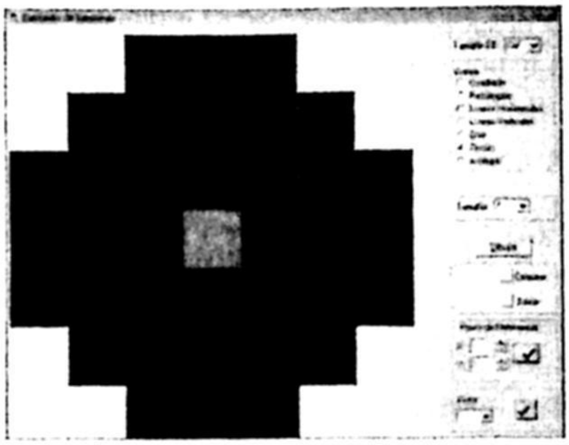


Fig. 6 Disc-shaped structuring element (SE).

We separated the image in the Green channel, then we eroded the image with the established SE, and we observe the resulting black points (Fig. 7), where the area in which more ink was printed, appeared smaller due to erosion:

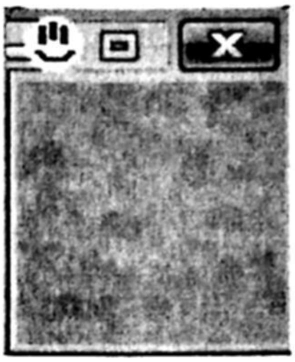


Fig. 7. Green channel eroded

Then, we dilated the last image (Fig. 7) with the same structure element. In the resulting image it can be observed that the black points dilated and manifested themselves in a more clear manner (fig 8). This image represents the opening of the image in Green channel.

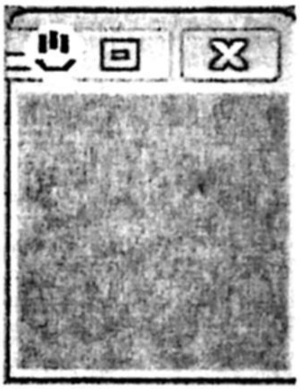


Fig. 8. Dilated image.

Subsequently, we applied the top-hat transformation using the subtraction of the original image and its opening, getting as result Fig. 9, where the points which were printed in the digital photo can be observed:

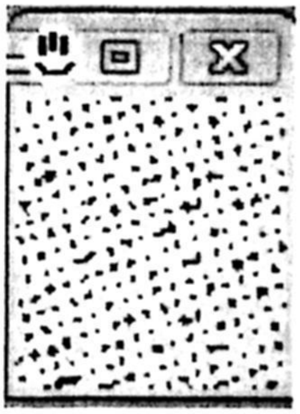


Fig. 9. Subtraction of the original image and his opening.

We threshold the image using Otsu's algorithm:

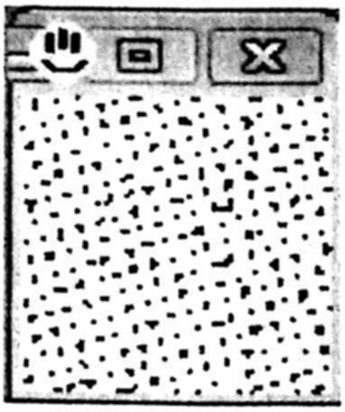


Fig. 10. Threshold image.

We labeled the points of the image and counted each color appearing in the image:

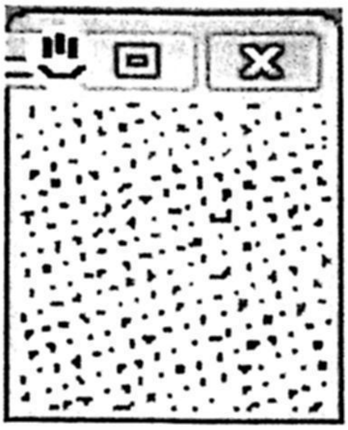
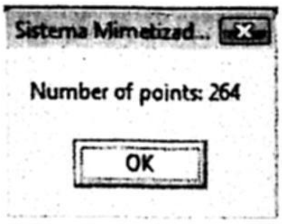


Fig. 11. Labeling image.

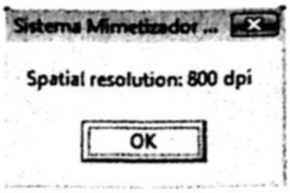
For this case, the result of the counter was:



Number of points: 264

Fig. 12. Counter.

Then, we applied equation (9) and we got the following result: **779.91** dots per inch. The difference with 800 dots per inch is because when we used the morphological operation on the image, the image loses some details. Some points that are in the original image, because of their size, tend to disappear during the initial opening.



Spatial resolution: 800 dpi

Fig. 13. Final counter.

This same procedure was applied to 16 different id-cards, and we got the following results (table 1):

Table 1. Experimentation results

Sample	Points	Resolution (Formula) (dpi)	Resolution (Round) (dpi)
1	211	697.24	700
2	246	752.85	800
3	246	752.85	800
4	247	754.38	800
5	256	768.00	800
6	256	768.00	800
7	259	772.49	800
8	259	772.49	800
9	260	773.98	800
10	264	779.91	800
11	267	784.33	800
12	269	787.26	800
13	269	787.26	800
14	271	790.18	800
15	272	791.64	800
16	278	800.32	800

In sample #1, the result is 697.24 dpi due to the dark intensity tones in the sampled image, as we can see in Fig. 14. It must be remembered that the previously described methodology does not work for dark intensities for they don't permit correctly getting the number of printed points (recall the top-hat definition). It is necessary to use clear zones that allow the correct transformation.

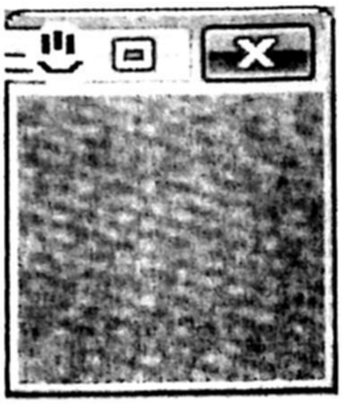


Fig. 14. Image from an area with dark tones.

In the rest of samples we did not get exactly 800 dots per inch (only in the sample 16) because generally the image loses some points when the morphological operations are applied. Afterwards, we made some experiments on analogical images, with the aim to check that the proposed methodology works correctly when the images are not digital images.

The original image of the green channel with 75 x 75 pixels (Fig. 15a) we apply the erosion (Fig. 15b) with the same disc shaped structuring element of 7 pixels diameter and then we apply the dilation (Fig. 15c); after that we subtract the resulting image (opening) from the original image (Fig. 15d) and we threshold it by using Otsu (Fig. 15e); then we label the image (Fig. 15f) and count the number of points with different colors in the image.

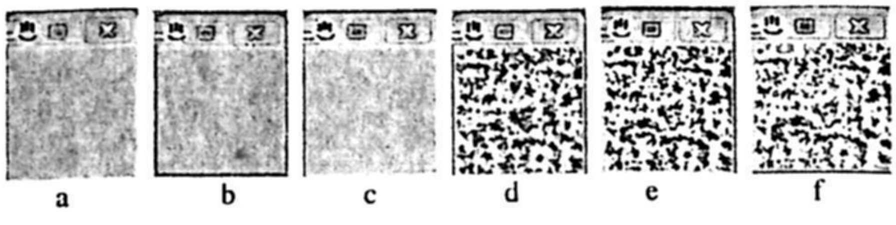


Fig. 15. Analysis of an analogical image.

The result was: *Number of points*: 182; *Resolution*: Analogical Image.

The same procedure was applied in a similar way to the following printed samples with a spatial resolution of 300, 600 and 1200 dpi. The results were:

a) For a digital image printed at 300dpi (Fig. 16):

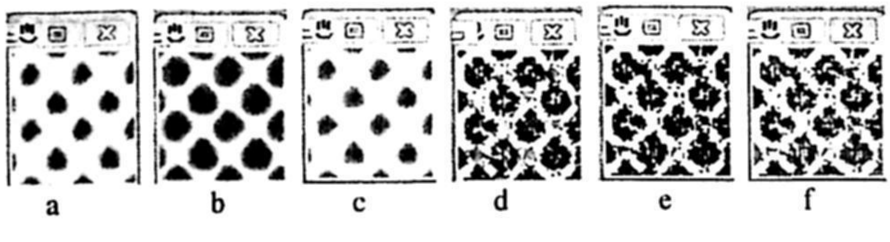


Fig. 16. Analysis of the digital image printed at 300 dpi.

Number of points: 49; *Resolution*: 336dpi; *Rounded* 300dpi.

b) For a digital image printed at 600dpi (Fig. 17):

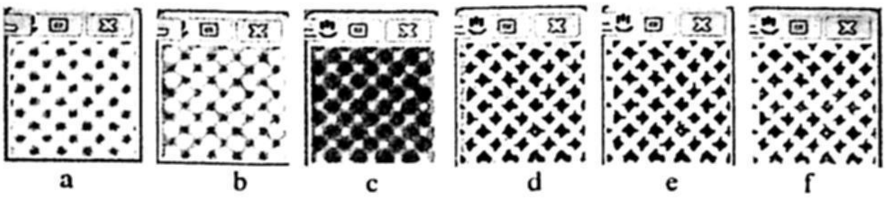


Fig. 17. Analysis of the digital image printed at 600 dpi.

Number of points: 65; *Resolution*: 386.99dpi; *Rounded*: 400dpi.

c) For a printed digital image in 1200dpi (Fig. 18):

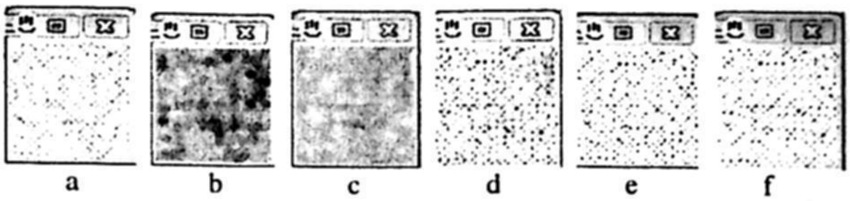


Fig. 18. Analysis of the digital image printed at 1200 dpi.

Number of points: 396; *Resolution*: 955.16dpi; *Rounded*: 1000dpi

As we can note, the results do not approach to the expected spatial resolution because images were printed at a different resolution that we wanted to determinate. Nevertheless, the results obtained in these cases allow us to asses that the images were not printed at 800dpi.

7 Conclusions

We can asses that the methodology proposed in this paper is a reliable way to determine the spatial resolution of the photo printed at 800dpi in the ID-card. This methodology is based on basic morphological operations such as dilation, erosion and the *top-hat* transformation. The *top-hat* transformation must be applied on clear areas of the image, in order to avoid errors on the resolution calculation. Subsequently, we threshold the image using the Otsu method and we label the resulting image in order to count the number of points with different colors in the image. The area size of the analysis is defined by the scanned resolution multiplied by the sixteenth part. The size and shape of the structuring element plays an important role in the methodology; the greater the spatial resolution, smaller must be the size of the structuring element; in the other hand, the smaller is the spatial resolution the greater must be the size of the structuring element. Once we have counted the points in the thresholded image. The methodology does not allow getting exactly the final resolution because when the morphological operation is applied on the image, some details is lost, that is some points present in the original image, because of their size, tend to disappear after the initial opening. When the final result is closest to 800dpi, we can asses that the photo was printed at that resolution.

Acknowledgments. The authors of this paper wish to thank to the Center for Computing Research (CIC), Mexico; Postgraduate and Research Secretary (SIP), Mexico, and National Polytechnic Institute (IPN) Mexico, for their support.

References

1. Hasta en doce ocasiones han repuesto la credencial del IFE, en http://noticias.vanguardia.com.mx/d_i_308095_t_Hasta-en-12-ocasiones-han-repuesto-la-credencial-del-IFE.htm
2. Elementos y características de la credencial para votar con fotografía, en http://www.ife.org.mx/documentos/DERFE/RFE2/cred/CredencialVotar_anverso.swf
3. Diario Oficial de la Federación, México, 30 de septiembre de 1992.
4. Diario Oficial de la Federación, México, 31 de enero de 2001.
5. Código Penal Federal, México, 2000.
6. N. Otsu, "A Threshold Selection Method from Gray-Level Histograms", IEEE Transactions on Systems, Man and Cybernetics, 9(1):62-66. 1979.
7. R. C. Gonzalez y R. E. Woods, Digital Image Processing, 3rd Edition, Prentice Hall 2008.
8. Sossa Azuela Juan Humberto, Rasgos Descriptores para el Reconocimiento de Objetos, Instituto Politécnico Nacional, 2006.
9. Pajares Gonzalo, De la Cruz Jesús, Visión por computador, Imágenes digitales y aplicaciones, Alfaomega Grupo Editor, 2004.
10. Adobe Photoshop CS2, versión 9.0, Adobe Systems Inc., 2005.
11. Umbaugh, S. E. Computer Vision and Image Processing: A Practical approach using CVIP tools. Prentice-Hall, EngleWood Cliffs: NJ, 1998.
12. Soille P. Morphological Image Analysis, Principles and Applications. Springer, Berlin, 1998.
13. Sanchez Samuel, Felipe Edgardo, Godoy Salvador. A morphological-based method to determine the kind of media (analogical or digital) used to print id-photographs, 17th International Conference on Computing, CIC-IPN, Mexico 2008.

Segmenting Blood Vessels in Retinal Images using a Entropic Thresholding Scheme

Fabiola M. Villalobos-Castaldi, Edgardo M. Felipe-Riverón and Cecilia Albortante-Morato

Center for Computing Research, National Polytechnic Institute,
Av. Juan de Dios Batiz and Miguel Othon de Mendizabal, Mexico, D.F.,
P. O. 07738. México, Phone: 5729 6000/56515.
fvillalobosb07@sagitario.cic.ipn.mx, edgardo@cic.ipn.mx, amoratob07@sagitario.cic.ipn.mx

Abstract. In this paper it is presented a fast and automatic method to segment and extract blood vessels from retinal images. The proposed method is based on the second local entropy and in the grey level co-occurrence matrix (GLCM). It is very helpful for the computation of a threshold the information that the GLCM has about the contours and the spatial distribution of the grey levels. Therefore, the algorithm is designed to have flexibility in the definition of the blood vessel contours, since it is able to fit the contours. Using the information of the GLCM a statistic feature is calculated which acts as a threshold value. The average time required for the propose method is 6 seconds. To asses the ability and speed of the proposed method, the experimental results are compared with the state-of-the-art results obtained by another methods.

1 Introduction

The retina is the only place in humans where blood vessels can be directly visualized in a non-invasively in vivo way. Ophthalmic photography is a highly specialized form of medical imaging dedicated to the study and treatment of disorders of the eye. There are two common procedures to perform such photography: (a) angiography (Figure 1.a)) and (b) fundus photography (Figure 1.b)). Angiography is the imaging of vessels, and the resulting pictures are angiograms. Angiography of the retina of the eye requires the injection of a small amount of dye into a vein in the patient's arm. The dye travels through the blood stream and is photographed using special cameras and colored light in the meantime it travels through the vessels of retina. Fluorescein Angiography and Indocyanine Green (ICG) are the two main types of such procedure. Figure 1a.shows examples of ophthalmic photography.

Alternatively, when performing for ophthalmic fundus photography for diagnostic purposes, the pupil is dilated with eye drops and a special camera called a *fundus camera* is used to focus on the fundus. The resulting images are detailed and revealing, showing the optic nerve through which visual 'signals' are transmitted to the brain and the retinal vessels which supply nutrition and oxygen to the tissue.

Figure 1.1b shows an example of a digital fundus photograph, where the contrast between the blood vessels and retinal background is not as good as it is in fluorescein

Received 23/03/09
Accepted 27/04/09
Final version 08/05/09

angiogram. Fundus photographs are usually taken using a green filter ('red-free') to acquire images of retinal blood vessels, since green light is absorbed by blood and appear in the fundus photograph darker than the background and the retinal nerve fiber layer.

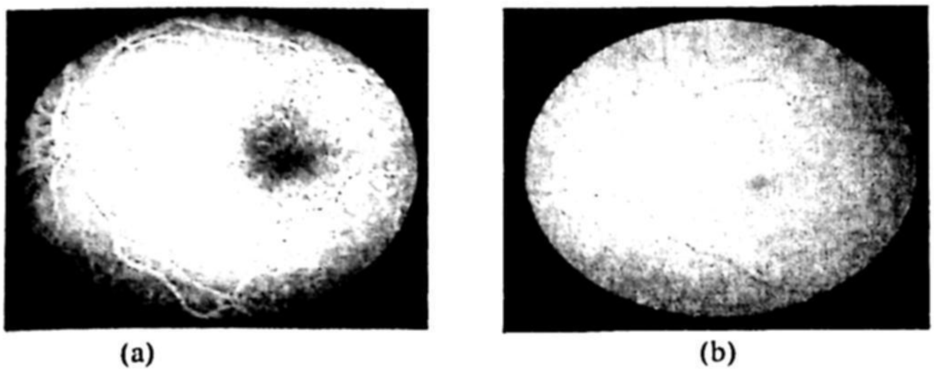


Fig.1. Examples of ophthalmic photography (a) Fluorescein angiogram (b) Digital fundus photograph.

Fluorescein angiography permits to detect and quantify changes in the blood vessels geometry more accurately than in fundus photography due to the high contrast between the blood vessel and the background retinal layer; it is bloody and sometimes unsuitable for certain people because of allergic reactions; thus fundus photography is more widely used in clinics. Despite the high resolution of photographs in fundus photography, the contrast between the blood vessels and retinal background tends to be poorer than that in angiograms.

Retinal or fundus images provide information about the blood supply system to the retina. Ocular fundus image assessment has been widely used by the medical community for diagnosing vascular and non vascular pathologies. Inspection of the retinal vasculature may reveal hypertension, diabetes, arteriosclerosis, cardiovascular disease and stroke [13]. For example, central retinal artery occlusion usually causes generalized constriction of retinal arteries. While central retinal vein occlusion typically produces dilated tortuous veins, arteriosclerosis can cause arteries to acquire a copper or silver color. Hypertension may result also in focal constriction of retinal arteries, and diabetes can generate new blood vessels (neovascularization). Among the features in ocular fundus image, the structure of retinal blood vessels plays an important role in revealing the state of retinal diseases. In addition, blood vessels can also serve as landmarks for image-guided laser treatment of choroidal neovascularization.

1.1 Fundus Image Analysis

Although the underlying mechanisms for some eye diseases are not fully understood, its progress can be prevented by early diagnosis and treatment. A number of steps is necessary to carry out such analysis. In general, it is necessary first to segment the blood vessels from the fundus image. Accurate blood vessel segmentation is fundamental in the analysis of fundus images since further analysis usually depends on the accuracy of this segmentation. It allows a quantitative measurement of the geometrical changes of arteries, tortuosity or lengths, and provides the localization of landmark points such as bifurcations needed for image registration [8]. It is therefore desirable to provide ways of automating the process of the analysis of fundus images, using computerized image analysis, so as to provide at least preliminary screening information and also as an aid to diagnosis to assist the clinician in the analysis of difficult cases. Visually, vessels in the fundus image appear as dark lines on a relatively uniform bright background. Various methods are known for segmenting the fundus image. The aim of vasculature segmentation is to take a fundus image as input and create a new binary image of the complete vasculature as output. This output image contains the mask of the vascular structure. Only when the vascular mask has been created and the vessels have been detected, the analysis process can be done. Therefore, accurate vasculature segmentation is fundamentally important in the analysis of fundus images, as further analysis usually depends on the accuracy of this segmentation.

Segmentation methods vary depending on the imaging modality, application domain, method being automatic or semi-automatic, and other specific factors. While some methods employ purely intensity-based pattern recognition techniques such as thresholding followed by connected component analysis, other methods apply explicit vessel models to extract the vessel contours. Depending on the image quality and general image artifacts such as noise, segmentation methods may require image preprocessing before applying the segmentation algorithm. Other methods apply post-processing to overcome the problems of so called over-segmentation.

1.2 Characteristics of the Vasculature

The vasculature has a number of characteristics that the image processor can exploit in developing a segmentation technique [6], [14], [3].

- The vessel cross-sectional gray level profile approximates a Gaussian shape.
- The vasculatures is piecewise linear, it can be represented by many connected line segments.
- The direction and gray level of a vessel do not change abruptly, they are continuous.
- The vasculature is tree-like; all vessels are connected to all other vessels and they all originate within a single area, the optic disc.
- Arterial vessels and venous vessels do not cross themselves independently.
- Every vessel crosses are between the arterial and venous branches.

Some factors that hinder vascular segmentation are:

- Vessels are obviously not all the same size, shape or color.
- The contrast can sometimes be low; the vessel color can be close to that of the background.
- Some background features have similar attributes as vessels.
- Vessels crossings and bifurcations can confuse some techniques.
- The optic disk can be wrongly segmented as a vessel.

1.3 Vasculature Segmentation Methods

Advances in vascular imaging technology have provided radiologist non-invasive imaging modalities that can give accurate vascular information, which helps the physician to define the character and extent of a vascular disease, aiding diagnosis and prognosis (Kirbas C., 2004). As stated previously, accurate vascular extraction is the primary task in automated ophthalmic image analysis. The existing vessel extraction techniques and algorithms can be classified into five main categories as follows:

1. Pattern recognition techniques,
2. Tracking-based approaches,
3. Artificial intelligence-based approaches,
4. Model-based approaches and,
5. Others

2 The Proposed Method

The proposed method in this paper uses one of the basic approaches to edge detection: the enhancement/thresholding method to achieve a fast algorithm for automated detection of blood vessels in retinal images.

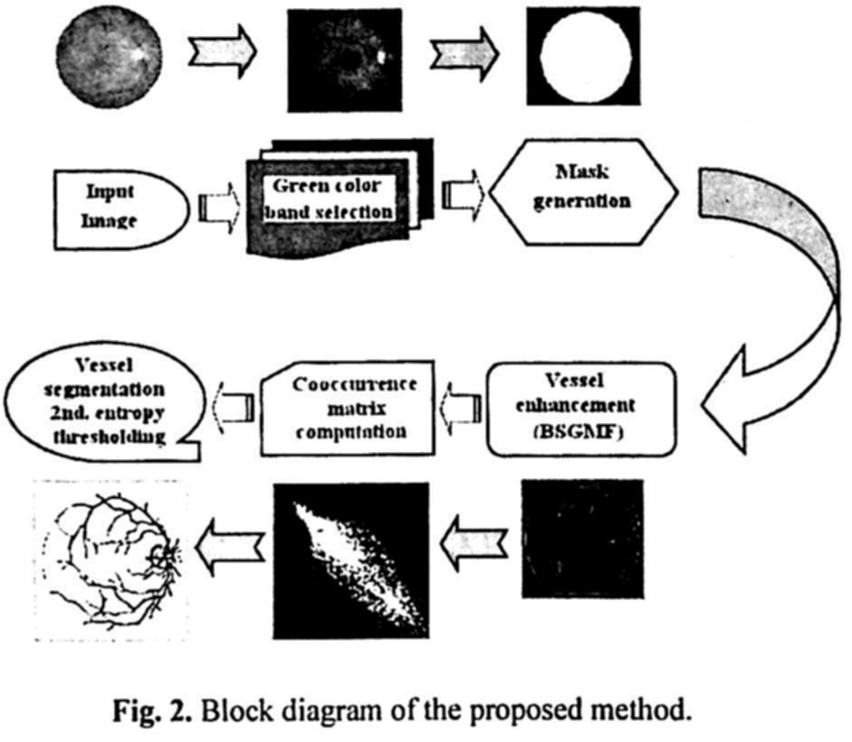

Input Image
Green color band selection
Mask generation
Vessel segmentation 2nd. entropy thresholding
Coocurrence matrix computation
Vessel enhancement (BSGMF)

Fig. 2. Block diagram of the proposed method.

As depicted in Figure 2, our method follows four main steps: 1) automatic mask generation to avoid processing of the black border and corners present in images; 2) application of a matched filter to enhance the vessels edges; 3) computation of the co-occurrence matrix; and 4) automatic vessel network segmentation using the second order entropy.

2.1 Mask Generation

Mask generation aims at labeling pixels belonging to the fundus Region of Interest (ROI) in the entire image. Pixels outside that ROI are those belonging to the dark surrounding region in the image. Those pixels are not strictly dark (zero intensity value) and the need to discard them for subsequent processing stages are necessary. The mask generation uses a thresholding with a free parameter empirically chosen such that pixels with intensity value above that threshold are considered to belong to the ROI. The threshold is applied in the green color band of the image (Figure 3).

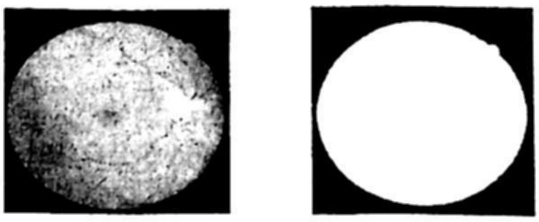


Fig. 3. Example of mask generation

The algorithm is robust enough to allow automatic mask generation in images of low visual quality, which is the principal issue in this step. All images of our dataset were correctly masked with this process.

2.2 Enhancement of Blood Vessels by Belt-Shape Gaussian Matched Filters

It can be noted that the retina vessels can be represented by piecewise linear segments with Gaussian-shaped cross sections. A matched filter is constructed for the detection of the vessel edge segments searching in all possible directions. Gray-level values of a blood vessel cross section can be approximated by a Gaussian curve (Eq. 1):

$$f(x,y) = A\left[1 \pm K\exp\left(-\frac{d^2}{2\sigma^2}\right)\right] \qquad \textbf{(1)}$$

Where d is the perpendicular distance between the pixel at (x,y) and the centerline of the blood vessel, σ defines the spread of the intensity profile, A is the gray-level intensity of the local background, and K is a constant used to account for the reflectance of the blood vessel relative to its neighborhood.

Based on Eq. (1), Chaudhuri [2] derived a 2-D Gaussian θ angle matched filter kernel given by (Eq. 2):

$$K_\theta(x,y) = \pm\exp\left(-\frac{u^2}{2\sigma^2}\right) \qquad \textbf{(2)}$$

Where $(u,v)^T$ is a new coordinate of $(x,y)^T$ after that $(x,y)^T$ is rotated by a θ angle. In other words, if we let $R(\theta)$ be the rotation matrix specified by θ, then:

$$R(\theta)=\begin{bmatrix}\cos\theta & -\sin\theta\\ \sin\theta & \cos\theta\end{bmatrix} \tag{3}$$

And $(u,v)^T=(x,y)^T R(\theta)$. Chaudhuri et al., 1989, used Eq. (2) to design 12 2-D Gaussian θ angle-matched filter kernels $\{K_\theta(x,y)\}_{i=0}^{11}$ to cover 12 orientations of a blood vessel in angles $\theta_0=0°$, $\theta_1=15°$, ... , $\theta_{11}=165°$, where the difference between two consecutive angles is 15 deg., i.e., $\theta_{i+1}-\theta_i=15°$.

One drawback of the approach of Chaudhuri et al., 1989, is the high computational complexity. It requires separate implementations of 12 kernels $\{K_\theta(x,y)\}_{i=0}^{11}$. To mitigate these problems, a bell-shaped Gaussian matched filter (BSGMF) was developed to cover all 12 orientations where designed kernel was given by [9] (Eq. 4).

$$K(x,y)=\pm\exp\left(-\frac{x^2+y^2}{2\sigma^2}\right) \tag{4}$$

With the tail truncated at $x^2+y^2=3\sigma^2$. Equation (4) is not a low-pass filter, but rather a bell-shaped Gaussian filter. As noted in [2], it was an effective filter to detect bell-shaped objects. Since the new kernel has a Gaussian shape along every direction that makes it looks like a bell, it is named the BSGMF.

2.3 Computation of the Co-occurrence Matrix

Given a digital image of size MxN with *L* gray levels denoted by G = {0,1,.....,*L*-1}, let *f*(*x,y*) be the gray level of the pixel at the spatial location (*x,y*). Then the image can be represented by a matrix $F=[f(x,y)]_{MxN}$. A co-occurrence matrix of an image is an LxL square matrix, denoted by $W=[t_{ij}]_{LxL}$, where t_{ij} is the number of transitions from gray level value *i* to gray level value *j* defined as follows [9], [10]:

$$t_{ij}=\sum_{m=1}^{M}\sum_{n=1}^{N}\delta_{mn} \tag{5}$$

With

$$\delta_{mn}=1$$

If

$$\begin{cases} f(m,n)=i, and, f(m,n+1)=j\\ and\ /\ or\\ f(m,n)=i, and, f(m+1,n)=j \end{cases} \tag{6}$$

$$=0 \text{ Otherwise}$$

Each entry in the matrix t_{ij} (Eq. 5) gives the number of times the pixel gray level *j* follows the gray level *i* in some pattern. Depending upon different patterns, then different definitions of co-occurrence matrix are possible. It has been reported in [4] and [10] that consideration of both horizontal and vertical transitions allows all the edges to participate in the threshold selection. Note that the co-occurrence matrix just defined in (6) considers only the pixel transition to the right as well as to below since it has been found that including the pixels on the left and above the transition does not provide significant information and improvement.

Normalizing the total number of transitions in the co-occurrence matrix, a desired transition probability from gray level *i* to *j* is obtained by

$$p_{ij} = \frac{t_{ij}}{\sum_{k=0}^{L-1}\sum_{l=0}^{L-1} t_{kl}} \tag{7}$$

If t, $0 \leq t \leq L$-1, is a threshold, then "t" partitions the co-occurrence matrix into four quadrants, namely, A, B, C and D as seen in Figure 4.

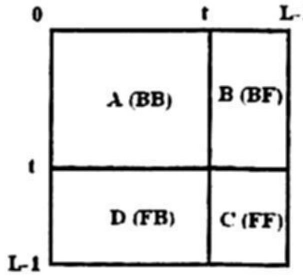


Fig. 4. Quadrants of the co-occurrence matrix.

These four quadrants can be further grouped into two classes. If we assume that pixels with gray levels above the threshold are assigned to the foreground (vessel), and those with gray levels equal to or below the threshold are assigned to the background, quadrants *A* and *C* correspond to local transitions within background (B), denoted by BB and foreground (F), denoted by FF, respectively, as shown in Figure 4.

Similarly, quadrants *B* and *D* represent transitions across boundaries between background and foreground, thus, they can be denoted by BF and FB, respectively. Thus, if we let $G_0 = \{0,\ldots, t\}$ and $G_1 = \{t+1, \ldots., L\text{-}1\}$, the four quadrants *A*(BB), *B*(BF), *D*(FB), and *C*(FF) are determined by the gray-level ranges G_0xG_0, G_0xG_1. G_1xG_0. G_1xG_1, respectively. Then the probabilities associated with each quadrant can be obtained by:

$$P_A^t = \sum_{i=0}^{t}\sum_{j=0}^{t} p_{ij} \qquad P_B^t = \sum_{i=0}^{t}\sum_{j=t+1}^{L-1} p_{ij} \tag{8}$$

Normalizing the probabilities within each individual quadrant, such that the sum of the probabilities of each quadrant equals one, we get the cell probabilities for different quadrants. In the case of the first quadrant (Eq. 9 and Eq.10):

$$P_{ij}^{A}=\frac{p_{ij}}{P_{A}}=\frac{t_{ij}\Big/\left(\sum_{i=0}^{L-1}t_{ij}\right)}{\sum_{i=0}^{t}\sum_{j=0}^{t}t_{ij}\Big/\sum_{i=0}^{L-1}\sum_{j=0}^{L-1}t_{ij}} \tag{9}$$

$$=\frac{t_{ij}}{\sum_{i=0}^{t}\sum_{j=0}^{t}t_{ij}} \tag{10}$$

$$\text{For } 0\le i\le t, 0\le j\le t$$

Similarly,

$$P_{ij}^{B}=\frac{p_{ij}}{P_{B}}=\frac{t_{ij}}{\sum_{i=0}^{t}\sum_{j=t+1}^{L-1}t_{ij}} \tag{11}$$

$$0\le i\le t, t+1\le j\le L-1$$

And

$$P_{ij}^{C}=\frac{p_{ij}}{P_{C}}=\frac{t_{ij}}{\sum_{i=t+1}^{L-1}\sum_{j=t+1}^{L-1}t_{ij}} \tag{12}$$

$$t+1\le i\le L-1, t+1\le j\le L-1$$

$$P_{ij}^{D}=\frac{p_{ij}}{P_{D}}=\frac{t_{ij}}{\sum_{i=t+1}^{L-1}\sum_{j=0}^{t}t_{ij}} \tag{13}$$

$$t+1\le i\le L-1, 0\le j\le t$$

2.4 Segmentation

Segmentation is the technique and procedure used to divide an image into non-overlapping different regions according to their particular characteristics. The pixel values in the same segmented region have similar attributes, but the pixel values between different regions have dissimilar attributes. Thresholding is a method commonly used in segmentation and is the foundation of other segmentation methods [10]. In recent years, in order to reduce the information loss during the segmentation process, an information theory method has been introduced into experimental segmentation.

For example, based on 1D entropy method, an image can be divided into two parts according to the threshold value obtained from the 1D entropy. Once the entropy of each one of the two parts is equal or approximately equal, the value of 1D entropy will reach the maximum and the automatic segmentation will be realized. Compared to 1D entropy, 2D entropy can reflect not only the information of pixel gray value, but also the statistical rule of two pair of pixels between fixed-position in the image. Thus, the 2D entropy has been widely utilized. The first step of 2D segmentation is to build the 2D histogram, which can be commonly established by using the gray-level gradient co-occurrence matrix.

2.4.1 1D Entropy

Let X be a discrete random variable and its probability distribution $p_i = \{X = x_i\}$, $i = 1, 2, \ldots., n$, then the 1D entropy can be defined as follows (Eq. 14):

$$H(X) = -\sum_{i=1}^{n} p_i \log p_i \tag{14}$$

Where $\{x_1, x_2, ..., x_n\}$ is a sample of random variable X, and the base of logarithm could be 2, 3 or 10.

2.4.2 2D Entropy

Now, the second order local entropy of the background can be defined as:

$$H_A^{(2)}(t) = -\frac{1}{2}\sum_{i=0}^{t}\sum_{j=0}^{t} P_{ij}^{A} \log_2 P_{ij}^{A} \tag{15}$$

Similarly the second order entropy of the object can be written as:

$$H_C^{(2)}(t) = -\frac{1}{2}\sum_{i=t+1}^{L-1}\sum_{j=t+1}^{L-1} P_{ij}^{C} \log_2 P_{ij}^{C} \tag{16}$$

Hence the total second order local entropy of the object and the background can be written as:

$$H_T^{(2)}(t) = H_A^{(2)}(t) + H_C^{(2)}(t) \tag{17}$$

The gray level corresponding to the maximum of $H_T^{(2)}(t)$ gives the threshold for object-background segmentation.

$$(t) = Arg\left\{\max_{0 \le t \le L-1} H_T^{(2)}(t)\right\} \tag{18}$$

3 Experimental Results

In this paper we used the images included in the well-known DRIVE database to asses the performance of the proposed method [11].

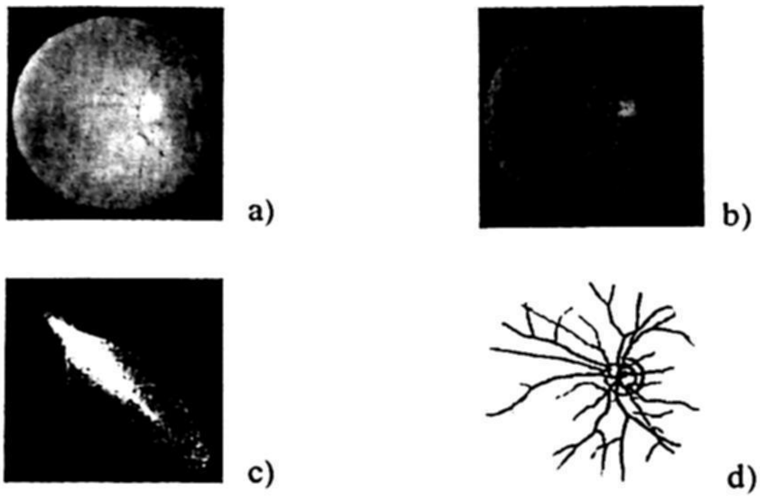


Fig. 5. Steps of the segmentation process for a typical image. a) Original image. b) Green color band used for the enhancement process. c) Co-occurrence matrix obtained and the corresponding threshold (Red cross). d) Segmented vessel and the elapsed time to get it.

The DRIVE database contains 40 color images, which were captured in digital form from a Canon CR5 non-mydriatic 3CCD camera at 45° field of view (FOV). The images are of size 565 × 584 pixels, 8 bit per color channel. The images have been divided into 2 sets. A training set and a test set. Each one contains 20 color retina images. Each set also contains the corresponding segmented images, which were graded by two experts (called first observer and second observer). Images in the training set were graded only once, and were named from 21_manual1 to 40_manual1. The images in the test set were graded by the two specialists, and the segmented images were named as 1_manual1 to 20_manual1, and 1_manual2 to 20_manual2. The total amount of segmented images in the database is 60, 20 in the training set and 40 in the test set.

3.1 Execution Time

Finally, we focus the analyses on the time that the proposed algorithm takes out to accomplish the whole process for blood vessel segmentation. On a Pentium 4, CPU 1.73GHz, with a MATLAB 7.4.0 (R2007a) implementation, it takes as average 6 seconds to obtain the image with the segmented vessels. Table 1 shows the elapsed time comparison amongst our method with some state-of-the-art results obtained from [2], [12], [1] and [5]. The execution time required for each step is illustrated in Table 1.

Table 1. Comparison of the execution times.

Method	Execution Time
Manual	2 hr
Yiming Wang [12]	7.0 min
Soares JVB [5]	3.0 min
T. Chanwimaluang [1]	2.5 min
S. Chaudhuri [2]	1.0 min
Villalobos and Felipe	**6.2 s**

Table 2. Execution time of each step.

No. of the step	*Step*	*Execution time (seconds)*
1	Green color band separation	**0.1453**
2	Enhancement	**5.5400**
3	Mask generation	**0.2000**
4	Vessel segmentation	**0.3200**
	Total	**6.2053**

As it can be seen, the most time-consuming process is the enhancement step where the BSGMF is applied, and the segmentation step based on the second local entropy thresholding. Figure 6 illustrates the results of some of the segmented vessel images and the elapsed time required to get each of them.

4 Conclusions

In this paper we presented a pixel processing approach for the fast automatic detection and extraction of retinal vessels from retinal fundus images using a thresholding method based on the second local entropy and in the gray level co-occurrence matrix (GLCM). Using one of the approaches of the edge detection, the enhancement/thresholding, the proposed method reduced the time required for extracting and segmenting the retinal vessel of fundus images without reducing the accuracy. In most of the methods, the time is not a normal relevant aspect to evaluate, in spite its importance when we deal with a huge image data base. That is why we focus also in this aspect. The average time required to segment the complete vessels is 6 seconds. Based on these results we consider that our method offers a good alternative for those applications where the time of analysis plays an important role.

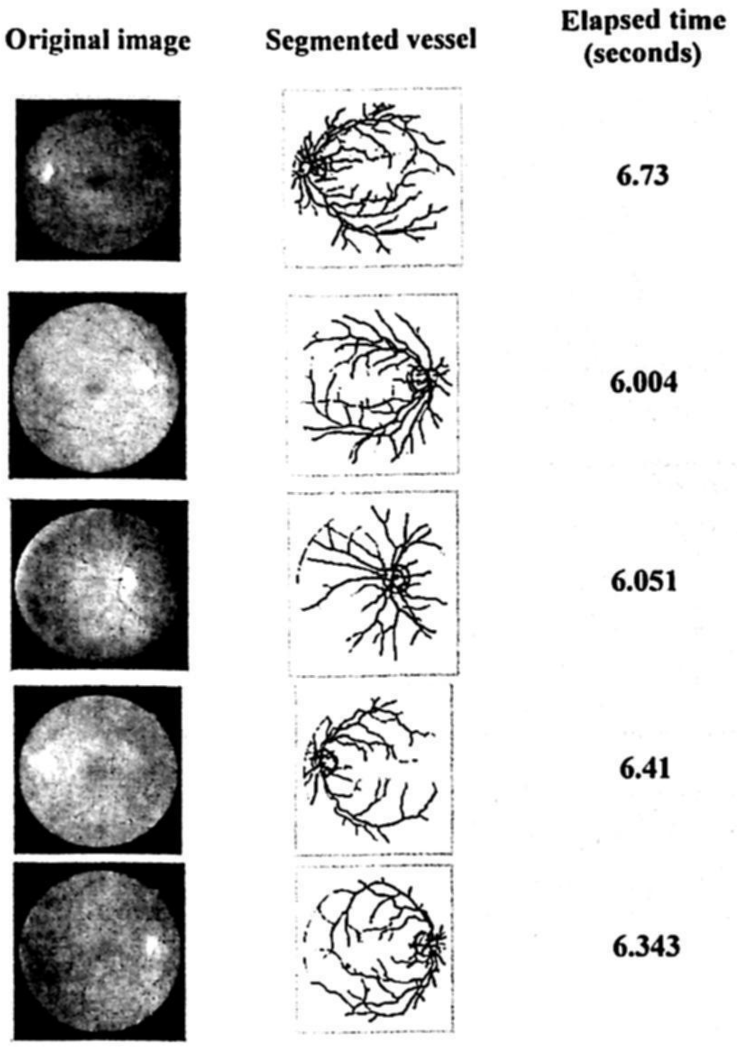


Fig. 6. Original image, the segmented vessel image and the elapsed time for some images.

Acknowledgements. The authors of this paper wish to thank the Computing Research Center (CIC), Mexico; Postgraduate and Research Secretary (SIP), Mexico, under research grant 20082213, and National Polytechnic Institute (IPN), Mexico, for their support.

References

1. Chanwimaluang T. and Fan G., (2003), An efficient blood vessel detection algorithm for retinal images using local entropy thresholding. In: *Proc. of the IEEE Intl. Symp. on Circuits and Systems.*
2. Chaudhuri S., Chatterjee S., Katz N., Nelson N., and Goldbaum M., (1989), Detection of Blood Vessels in Retinal Images Using Two-Dimensional Matched Filters, IEEE Transactions on Medical Imaging, 8(3):263–269.
3. Matsopoulos G. K., Mouravliansky N.A., Delibasis K.K. and Nikita K.S., (1999), Automatic retinal image registration Scheme using global optimization techniques, IEEE Trans. Information technology in biomedicine, vol. 3.

4. Pal N.R., Pal S.K., (1989), Entropic thresholding, Signal Process 16, 97-10.
5. Soares JVB, Leandro JJG, Cesar RM, Jelinek HF, and Cree MJ., (2006), Retinal vessel segmentation using the 2-D Gabor wavelet and supervised classification, IEEE Transactions on Medical Imaging, 25:1214–1222.
6. Zana F. and Klein J. C., (1997), Robust Segmentation of Vessels from Retinal Angiography. In International Conference on Digital Signal Processing, pages 1087–1091, Santorini, Greece.
7. Kirbas C. and Quek (2004), A review of vessel extraction techniques and algorithms, ACM Comput. Surv., 36(2):81–121.
8. Badrinath Roysam, Kenneth H. Fritzsche, Charles V. Stewart, (2002), Doctoral Dissertation, Computer Vision Algorithms for Retinal Vessel Width Change Detection and Quantification, Department of Electrical, Computer, and Systems Engineering Rensselaer Polytechnic Institute.
9. Ching-Wen Yang, Dye-Jyun Ma, Shuenn-Ching Chao, Chuin-Mu Wang, Chia-Hsin Wen, Chien-Shun Lo, Pau-Choo Chung, Chein-I Chang, (2000), Computer-aided diagnostic detection system of venous beading in retinal images, Society of Photo-Optical Instrumentation Engineers. 39(5) 1293–1303.
10. Zhang Y F, and Zhang Y, (2006), Another Method of Building 2D Entropy to Realize Automatic Segmentation, International Symposium on Instrumentation Science and Technology; Journal of Physics: Conference Series 48, 303–307.
11. Wang L. and Bhalerao A., (2003), Model Based Segmentation for Retinal Fundus Images. In Proc. of Scandinavian Conference on Image Analysis (SCIA).
12. Niemeijer M., Staal M., J. van Ginneken J., van, Loog M B. B. and. Abraamoff M. D., (2004), Comparative study of retinal vessel segmentation methods on a new publicly available database, Sciences Institute at http://www.isi.uu.nl/Research/Databases.

Robust Edge Detection Algorithm using Ant Colony System

Inkyeom Kim and Minyoung Yun

Department of Information and Communications Engineering,
Sungkyul University, Anyang, Korea
{kik, alabama }@sungkyul.edu

Abstract. The Ant Colony System (ACS) is easily applicable to the traveling salesman problem (TSP) and it has demonstrated good performance on TSP. Recently, ACS has been emerged as the useful tool for the pattern recognition, feature extraction, and edge detection. The edge detection is widely utilized in the area of document analysis, character recognition, and face recognition. However, the conventional operator-based edge detection approaches require additional post processing steps for the application. In the present study, in order to overcome this shortcoming, we have proposed the new ACS-based edge detection algorithm which has the capabilities to detect finer edges as well as to extract connected edges. The experimental results indicate that this proposed algorithm has the excellent performance in terms of robustness and flexibility.

1 Introduction

The Ant Colony System (ACS) is a meta-heuristic algorithm based on the foraging behavior of ant colonies. Real ants are capable of finding the shortest path from a food source to their nest by exploiting pheromone information. Since its development by Dorigo et al.[1], ACS has been applied to complex combinatorial optimization problems such as the traveling salesman problem (TSP)[2,3], the quadratic assignment problem (QAP), and many discrete optimization problems such as vehicle routing [4], sequential ordering, graph coloring, and routing in communication networks. Recent applications of ACS include pattern recognition, image extraction, and edge detection [5,6,7,11,12].

A significant amount of research is being currently conducted to develop an edge detection algorithm which can be applied to detect and localize the boundaries of objects in an image. Since the detected edges are widely applicable to the areas of document analysis, character recognition, and face recognition, it is crucial for the edge detection algorithm to efficiently and clearly detect the edges. Well-known conventional edge detection operators include the Sobel operator using the gradient-based method, the Laplacian operator based on the second derivation method, and the Canny operator which is the most widely used

© G. Sidorov, M. Aldape, M. Martínez, S. Torres. (Eds.)
Advances in Computer Science and Engineering.
Research in Computing Science 42, 2009, pp. 167-176

Received 06/03/09
Accepted 27/04/09
Final version 08/05/09

edge detection operator. Even though operator based edge detection methods can detect edges clearly, they require an additional step for next-stage image processing.

In the present study, we propose a new ACS-based edge detection algorithm which has the ability to detect finer edges as well as to extract connected edges. Since application of the edge detection algorithm is highly sensitive to the thickness and range of the edges, the edge detection algorithm for actual applications must be flexible in order to accurately represent the thickness and range of the edges. In this regard, the proposed edge detection algorithm is flexible and robust for the application of ACS.

2 Edge Detection Using ACS

Nezamabadi-pour, Saryazdi, and Rashedi[7] applied ACS for detecting edges in digital images. In their approach, the image is considered as a two-dimensional graph where each pixel is denoted as a vertex. Ants move from pixel to pixel and mark the visited pixel with pheromone. In the initial stage, m ants are placed randomly on each pixel. The intensity of the all pixels is set to 0.0001. Ants statistically choose one of their eight-neighboring pixels with the probability described in equation (1). The probability of moving the kth ant from vertex (r, s) to vertex (i, j) is expressed as:

$$p^k_{(r,s),(i,j)} = \begin{cases} \frac{(\tau_{(i,j)})^\alpha(\eta_{(i,j)})^\beta}{\Sigma_u\Sigma_v(\tau_{(u,v)})^\alpha(\eta_{(u,v)})^\beta}, & \text{if } (i,j) \text{ and } (u,v) \in \text{admissible nodes} \\ & r-1 \leq i, u \leq r+1, s-1 \leq j, v \leq s+1 \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

The heuristic information $\eta_{i,j}$ of pixel (i, j) is defined by the following formulation.

$$\eta_{i,j} = \frac{1}{I_{Max}} \times Max \begin{bmatrix} |I(i-1,j-1) - I(i+1,j+1)|, \\ |I(i-1,j+1) - I(i+1,j-1)|, \\ |I(i,j-1) - I(i,j+1)|, \\ |I(i-1,j) - I(i+1,j)| \end{bmatrix} \tag{2}$$

After each step, the pheromone is updated by the following equation:

$$\tau_{(i,j)}(new) = (1-\rho)\tau_{(i,j)}(old) + \Delta\tau_{(i,j)} \tag{3}$$

where

$\Delta\tau_{(i,j)} = \Sigma^m_{k=1}\Delta\tau^k_{(i,j)}$, and,

$$\Delta\tau^k_{(i,j)} = \begin{cases} \eta_{(i,j)}, & \text{if } \eta_{(i,j)} \geq b \text{ and } k\text{th ant displaces} \\ 0, & \text{otherwise} \end{cases}$$

Here b is a threshold value. If a pixel is not chosen by the ants, its intensity of pheromone decreases exponentially. To avoid stagnation in the searching process, the minimum of pheromone intensity is limited by τ_{min}. Since $\tau_{min} \geq 0$, the probability of choosing a specific pixel can not be zero.

3 Proposed Edge Detection Algorithm

In the proposed edge detection algorithm, edges in the digital image are searched by using the ACS which was developed by Dorigo et. al [2]. The present algorithm detects edges by utilizing the accumulated pheromone on the edge area. The pheromone, τ_{ij}, accumulated at the pixel (i, j) represents the measure of the probability for moving from the current pixel (i, j) to other pixels. Heuristic information of the digital image is obtained from the relationship, $\eta_{ij} = d_{ij}$ and d_{ij} which denotes the intensity difference between the current pixel (i, j) and the neighboring pixels. τ_{ij} and η_{ij} are stored at the matrices of pheromone and heuristic information, respectively. Each ant initially begins the search process at a randomly chosen pixel. Among the unvisited pixels, a next searching pixel is selected according to the values of τ_{ij} and η_{ij} . Each ant updates the pheromone for the visited pixel at every step and all ants update the pheromone one more time for the all the visited pixels if the searching step reaches the predefined number of steps.

If the digital image has a resolution of $M \times N$ pixels, m ants are randomly placed at m pixels in the initial stage. Figure 1 illustrates the edge detection algorithm proposed in this study.

```
algorithm: Proposed ACS for Edge Detection {

  Initialize Data;
  while (not terminate) {
      place $m$ ants at $M \times N$ pixels;
      repeat (for each ant)
          apply search construction rule to find edges;
          apply local pheromone updating rule;
      until (construct a solution)
      apply global pheromone updating rule;
      apply evaporation rule;
  }
}
```

Fig. 1. Procedure of the proposed edge detection algorithm

3.1 Search Rule

In this proposed algorithm, when ant k is moved from pixel (i, j), a next pixel (l, h) is chosen by the pseudo-random proportional rule described in the following equation.

$$(l,h) = \begin{cases} \arg\max J, & \text{if } q \leq q_0; \\ \arg\max J, & \text{if } q \geq q_0 \& J \leq q_1; \\ \arg\min J, & \text{if } q \geq q_0 \& J > q_1; \end{cases} \tag{4}$$

where q is a random variable uniformly distributed in [0.1], q_0 $(0 \leq q_0 \leq 1)$ is a parameter for the random search, and J is a random variable selected by the probability distribution. When the probability for selecting the next pixel exceeds the specified value, a variable, q_1 is applied for searching the homogeneous region. The probability for moving ant k from pixel (i, j) to pixel (l, h) is expressed below.

$$p^k_{(ij),(lh)} = \frac{[\tau_{ij}]^\alpha [\eta_{ij}]^\beta}{\sum_{l \in N^k_i} \sum_{h \in N^k_j} [\tau_{(i+l)(j+h)}]^\alpha [\eta_{(i+l)(j+h)}]^\beta} \quad \text{if } \; l \in N^k_i \& h \in N^k_j \tag{5}$$

Here α and β are parameters for determining the importance of the pheromone τ_{ij} and the relative importance of heuristic information η_{ij}, respectively. N^k_i and N^k_j denote the set of unvisited pixels.

According to equation (5), the selection probability of a pixel (l, h) from a pixel (i, j) is determined by pheromone τ_{ij} and heuristic information η_{ij}. η_{ij} is computed by equation (2) adopted from the previous study [7]. Each ant selects a next visiting pixel (l, h) which has a larger amount of pheromone and the greatest gray-level difference among all pixels in the search. If $\beta = 0$, probability for selecting a next visiting pixel depends only on the pheromone level, τ_{ij}. If $\alpha = 0$, ants are dependent only on the heuristic information, η_{ij}. To avoid such conditions, α and β must generally satisfy the condition where $\alpha \geq 1$ and $\beta \geq 1$.

3.2 Local Pheromone Update

Whenever an ant visits a pixel (i, j), the pheromone level for the pixel is updated by applying the following equation.

$$\tau_{ij} = (1 - \xi)\tau_{ij} + \xi\eta_{ij} \tag{6}$$

Here η_{ij} denotes the difference between the current pixel (i, j) and the neighboring pixels in an 8-neighbor-based search, and ξ is a variable with the range, $0 < \xi < 1$. According to a Dorigo et al. [4], the best performance is achieved at $\xi = 0.1$. By applying equation (6), the pheromone level τ_{ij} at any pixel (i, j)

previously visited by is progressively reduced for each successive visit. Consequently, any previously visited pixel has a much lower probability to be selected by the following ants. Since this procedure increases the probability that ants will select unvisited pixels, it prevents stagnation which is a result of repeated visits to pixels not along the best path.

3.3 Global Pheromone Update

After each ant completes the specified steps, the pheromone levels for a set of visited pixels H^{sp} are updated by the following equation.

$$\tau_{ij} = (1-\rho)\tau_{ij} + \rho\Delta\tau_{ij}^{sp}, \qquad \forall(i,j) \in H^{sp} \tag{7}$$

where $\Delta\tau_{ij}^{sp}$ is the amount of pheromone to be added to pixel (i,j) and it is defined below as:

$$\Delta\tau_{ij}^{sp} = \sum_{k=1}^{m} \Delta\tau_{ij}^{k}, \tag{8}$$

$$\Delta\tau_{ij}^{k} = \begin{cases} \eta_{ij}, & \text{if } (i,j) \text{ visited by ant } k \\ 0, & \text{otherwise} \end{cases}$$

Here $\Delta\tau_{ij}^{k}$ represents the heuristic information for a pixel (i,j) visited by an ant k and the parameter ρ denotes the pheromone evaporation rate. In general, the best performance is obtained when $\rho = 0.1$.

3.4 Pheromone Evaporation

Immediately after having visited a pixel (i,j) during the edge detection process, the pheromone is reduced by the local pheromone update rule. On the other hand, the pheromone levels for all pixels visited by an ant are increased according to the global pheromone update rule. Even though the pheromone is deposited and reduced repeatedly due to global and local updates during the search process, it is still possible that pheromone at the certain pixels could be gradually accumulated over time due to the increased searching activity of ants. This accumulation of pheromone could cause stagnation. This situation is resolved by applying pheromone evaporation over the entire search range. The pheromone evaporation effect is imposed by the following equation.

$$\tau_{ij} = (1-\rho)\tau_{ij}, \qquad \forall(i,j) \in I \tag{9}$$

Here I is a total image and evaporation rate ρ has a value 0.1.

4 Experimental Results and Discussion

The experiments on the proposed algorithm are performed on an Enterprise RedHat 2.1. For each test, we chose the parameters which yielded the optimal solution in the previous experiments [2,3,8]. The optimal values of the special parameters for processing the digital images are determined through extensive experiments. In the present study, the problem parameters are:

$$\xi = 0.1, \rho = 0.1, \alpha = 2, \beta = 3, q_0 = 0.9, q_1 = 0.25, m = 1000, SR = 64.$$

The initial value of pheromone is set to $\tau_0 = 1/(SR \times I_{max})$ and SR denotes one cycle with 64 steps. The parameter t represents the working time which is required to complete a cycle. After completing a search process of one cycle (64 steps), t is increased by increments of 1 and this search process is continuously performed until t reaches 100.

Since the pheromone levels for all pixels are identical at the initial state, the edges are not clearly detected at the early stage of the searching process. It is assumed that the detected edges will become progressively clear over time. In the moving criteria based on ACS, ants move to the next pixel by using the heuristic information. However, if the moving criterion is determined only by the heuristic information, well detected edges could not be obtained for the general situations. In order to circumvent this drawback, the present approach is set to the search point by utilizing q_0 and q_1 appeared in equation (4). If only q_1 is used for the search criterion, it is quite possible for the ants to search only for background area. However, if q_1 is simultaneously used with q_0 supporting the randomness, the quality of detected edges can be substantially improved.

Figure 2 shows the resulting edge detection images from a search by 1000 ants from 1 to 16 cycles. It can be clearly seen that the fundamental edges of the image are realistically detected. Since the background area is searched only once, the background image is removed due to the pheromone evaporation process. On the other hand, a number of ants visit the edge region where the pheromone is gradually accumulated at the proximity of the edge area. Even if the certain edges are visible in the early stage of search, these edges could diminish due to the pheromone evaporation during the search process. As shown in Figure 2(e) and (f), the experimentally captured images at $t = 8$ and $t = 16$ are nearly identical except for the ridge of nose. This and other differences occur due to pheromone deposition and evaporation during the search. Moreover, the noticeable differences are not observed for the detected edges at $t \geq 10$ and $t = 8$. If the edge detection process relies heavily on the heuristic information, it is quite possible to lose the information for the much finer edges. In the generalized ACS algorithm, even if the parameter β influences the heuristic information, the overall results are not affected by β. However, in the digital image, if β becomes large, stagnation could occur because ants cannot search for the proper edges.

Figure 3 shows the enlarged results obtained for three different values of α and β. Figure 3(a) displays the result when $\alpha = 1, \beta = 3$. At this value, the

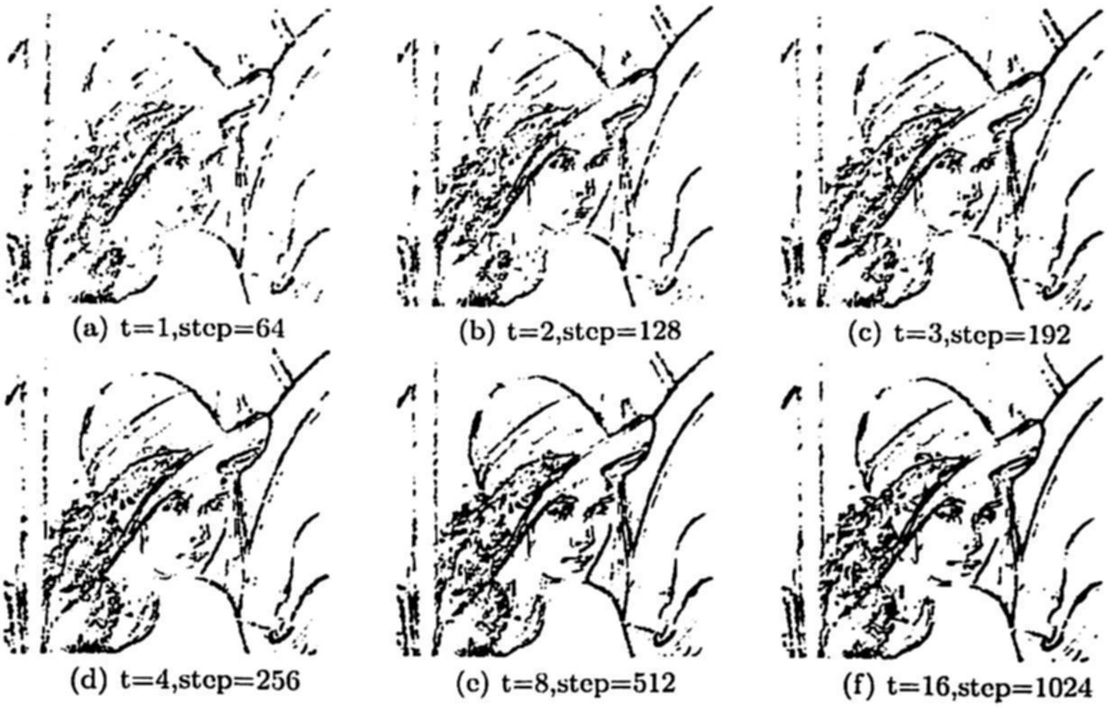
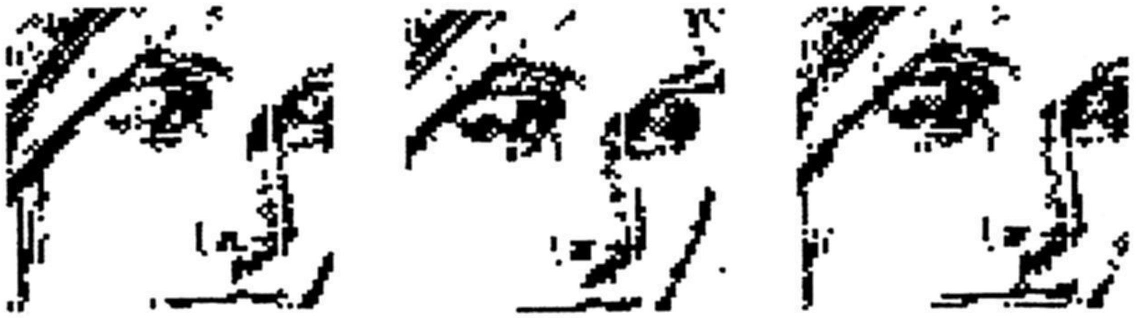

(a) t=1,step=64
(b) t=2,step=128
(c) t=3,step=192
(d) t=4,step=256
(e) t=8,step=512
(f) t=16,step=1024

Fig. 2. Results of edge detection by cycle *t*

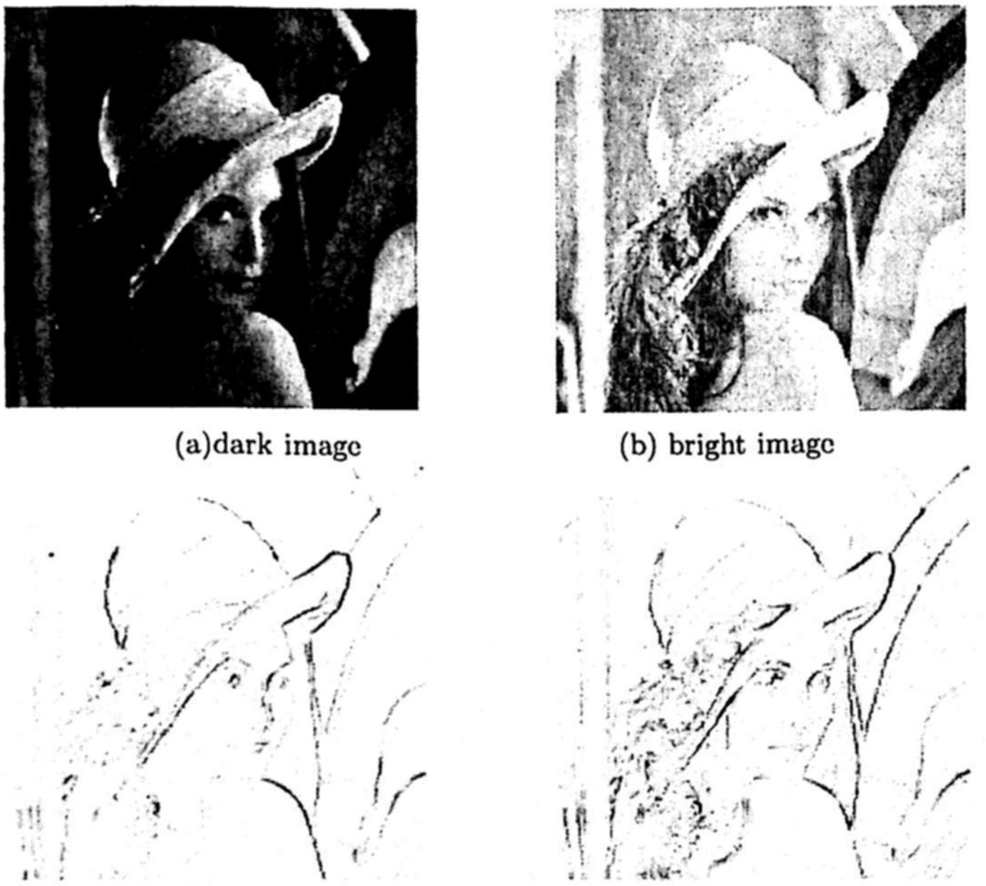
edge detection process is highly dependent on the heuristic information and the corresponding search results include many edges. However, many edges involved in the search results are not always meaningful. Figures 3(b) and 3(c) presents the results obtained for $\alpha = 3, \beta = 2$ and $\alpha = 2, \beta = 3$. As indicated in Figures 3(b) and 3(c), even if the heuristic information is very influential at the large values of β, the edges could be adequately detected by by using an optimal value for α. In terms of edge detection, noticeable differences exist around at the eyes and the ridge of nose. When α is larger than β, the edges are weakly detected on the ridge of the nose and the detected edges close to the nose are widely spread apart, as presented in Figure 3(b). The wide spacing of edges is mainly due to the fact that pheromone levels have a greater influence than the heuristic information on search behavior when α is larger than β. In this study, the optimum values for parameters α and β are obtained by comparing the experimental result having the best edge detection with the edge information of the original image. The experimental results indicate that $\alpha = 2$ and $\beta = 3$ produce the best performance for the edge detection as shown in the Figure 3.

Figure 4 shows the robustness of the proposed method for edge detection. It is shown the results of the detected edges in darker and brighter image than original Lena. In dark image, the overall results is good except the around of the shawl which has so many black region. In case of brighter image, it is shown that more specific edges were detected. Because of the high gray level, it is detected finer edges than in dark image in the shawl around. However, the detected edges compared the original Lena image with the dark and bright images show satisfied results.



(a) alpha=1 beta=3 (b) alpha=3 beta=2 (c) alpha=2 beta=3

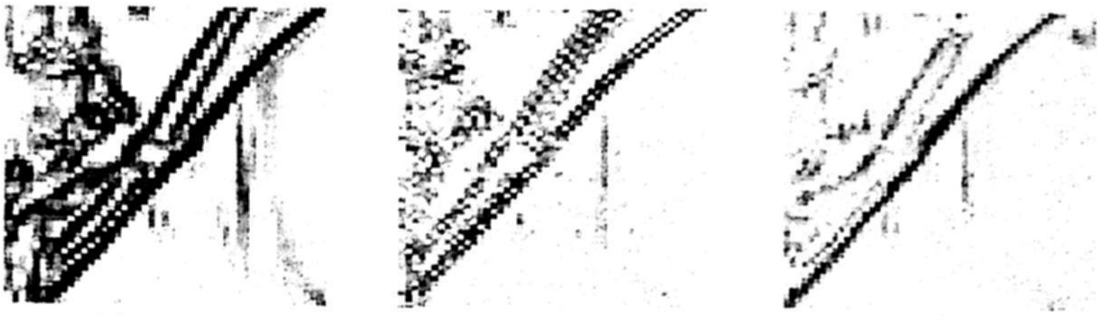
Fig. 3. Edge detection results for various values of α and β



(a)dark image (b) bright image

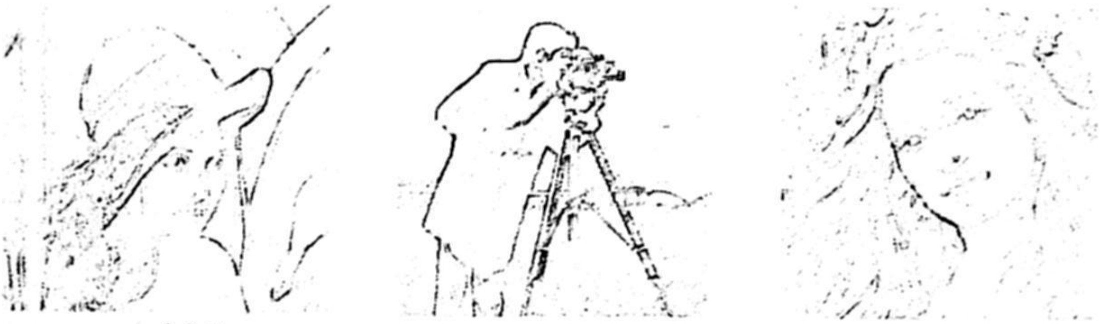
(c) Edge Result of dark image (d) Edge Result of bright image

Fig. 4. Results of edge detection in dark and bright Image



(a) Sobel (b) Laplacian (c) Proposed

Fig. 5. Enlarged results for comparing edge detection



(a) Lena (b) cameraman (c) face

Fig. 6. Results of edge detection using the proposed algorithm

It is compared proposed method with famous edge detection methods of Sobel and Laplacian. It is used 3x3 operators for Sobel and Laplacian. Figure 5 show the results of the enlarged edges of the Lena by Sobel, Laplacian and proposed edge detection method. The result of Sobel show too thick edges. Laplacian presents finer thin edges but edges is spread and not connected. However, the results of the proposed methods are shown finer thin and connected edges. The proposed method presents a robust and flexible results for the edge detection.

Finally, the proposed algorithm is applied to capture the images of Lena, a Cameramen, and a Face. As displayed in Figure 6, these experimental results clearly indicate that the present ACS-based algorithm is quite capable of detecting the edges of various complex images when the proper values for α and β are utilized. Moreover, in the case of Lena, as using the combination of q_0 and q_1, the proposed algorithm successfully detects the finer edges around her eyes and the shawl that were not captured in a previous study [7].

5 Conclusion

In this study, we have proposed an ACS-based algorithm which can accurately detect the edges of the digital images. In the conventional method for the image processing, edge detection is performed for fixed values for the image pixels or by using the known operators. Thus, the conventional operator-based edge detection approaches require additional post processing steps for image processing applications. In order to overcome this shortcoming, we have proposed a new ACS-based edge detection algorithm which has the capabilities to detect finer edges as well as to extract connected edges. The proposed algorithm is able to detect the edges with the combination of pheromone and heuristic information which are produced by ants in random fashion. In the present edge detection procedure, the gray levels around edges in the image are gradually changed, because gray levels are not obtained from any criterion based on the fixed values. The experimental results clearly indicate that the present ACS-based algorithm utilizing the proper values of parameters, q_0 and q_1, is quite capable of detecting edges of complex images. Moreover, in the case of Lena, the proposed algorithm successfully detects the finer edges on eyes and the shawl which were not captured in the previous study. The developed ACS-based algorithm can also be used for other image processes including image segmentation and image recognition.

References

1. M. Dorigo, V. Maniezzo and A. Colorni. The Ant System: Optimization by a colony of cooperating agents. *IEEE Transactions on Systems, Man, and Cybernetics-Part B*, 26(1) : 1-13, 1996.
2. M. Dorigo and L. M. Gambardella. Ant Colonies for the Travelling Salesman Problem. *BioSystems*, 43:73-81, 1997.
3. M. Yun and I. Kim. Improved Ant Colony System Using Subpath Information for the Traveling Salesman Problem. *Research on Computer Science*, Vol. 16, pp. 185-194, 2005.
4. M. Dorigo, G. D. Caro and L. M. Gambardella. Ant Algorithms for Discrete Optimization. *Artificial Life*, 5(3):137-172, 1999.
5. V. Ramos and F. Almeida. Artificial ant colonies in digital image habitats: a mass behaviour effect study on pattern recognition. *Proceedings of ANTS'2000 - International workshop on ant algorithms*, 113-116, 2000.
6. V. Ramos, F. Muge and P. Pina. Self-organized data and image retrieval as a consequence of inter-dynamic synergistic relationships in artificial ant colonies. *Hybrid Intelligence Systems*, 87, 2002.
7. H. Nezamadadi-pour, S. Saryazdi and E. Rashedi. Edge detection using ant algorithm. *Soft Computing*, August, 2005.
8. R. Gonzales and R. E. Woods. Digital Image Processing. Second Edition, Prentice Hall, 2002.
9. M. Dorigo and T. Stutzle. Ant Colony Optimization. MIT Press, 2003.
10. E. Bonabeau, M. Dorigo and G. Theraulaz. Swarm Intelligence: From Natural to Artificial Systems. Oxford University Press, 1999.
11. A. Rezaee. Extracting Edge of Images with Ant Colony. *Journal of Electrical Engineering*, 59(1):57-59, 2008.
12. Y. Yu and L. Guo. Ant Colony search for Edge Extraction in Noise Image. *Journal of Electronics and Information Technology*, 30 (6): 1271-1275, 2008.

A Parametric Method Applied to Phase Recovery from a Fringe Pattern based on a Particle Swarm Optimization

J.F. Jimenez, F.J. Cuevas, J.H. Sossa, N. Cruz and L.E. Gomez

Centro de Investigación en Computación-IPN,
Unidad Profesional Adolfo-López Mateos, Av. Juan de Dios Bátiz s/n
and M. Othón de Mendizábal, Zacatenco, México, DF. 07738, Mexico
Centro de Investigaciones en Óptica A.C.
Loma del Bosque #115, Col. Lomas del Campestre
C.P. 37150, León Gto. México
email: jfvielma@cio.mx, fjcuevas@cio.mx, hsossa@cic.ipn.mx,
nareli@cic.ipn.mx, lenis54@yahoo.com.

Abstract. A parametric method to carry out fringe pattern demodulation by means of a particle swarm optimization is presented. The phase is approximated by the parametric estimation of an nth-grade polynomial so that no further unwrapping is required. On the other hand, a diferent parametric function can be chosen according to the prior knowledge of the phase behavior. A particles swarm is codified with the parameters of the function that estimates the phase. A fitness function is established to evaluate the particles, which considers: (a) the closeness between the observed fringes and the recovered fringes, (b) the phase smoothness, (c) the prior knowledge of the object as its shape and size. The swarm of particles evolves until a fitness average threshold is obtained. The method can demodulate noisy fringe patterns and even a one-image closed-fringe pattern successfully.

Keywords: Phase retrieval; Fringe analysis; Optical metrology; Particle Swarm Optimization.

1 Introduction

In optical metrology, a fringe pattern (interferogram) can be represented using the following mathematical expression:

$$I(x,y) = a(x,y) + b(x,y) \times \cos\left(\omega_x x + \omega_y y + \phi(x,y) + n(x,y)\right) \qquad (1)$$

where *x, y* are integer values representing indexes of the pixel location in the fringe image, *a(x,y)* is the background illumination, *b(x,y)* is the amplitude modulation and is $\phi(x,y)$ the phase term related to the physical quantity being measured. ω_x and ω_y are the angular carrier frequency in directions *x* and *y*. The term $n(x,y)$ is an additive phase noise. The purpose of any interferometric technique is to determine the phase term, which is related to the physical quantity, being

Received 08/03/09
Accepted 27/04/09
Final version 09/05/09

measured. One way to calculate the phase term $\phi(x, y)$ is by using the phase-shifting technique (PST) [1–5], which needs at least three phase-shifted interferograms. The phase shift among interferograms must be known and experimentally controlled. This technique can be used when mechanical conditions are met throughout the interferometric experiment.

On the other hand, when the stability conditions mentioned are not covered, there are many techniques to estimate the phase term from a single fringe pattern, such as: the Fourier method [6,7], the Synchronous method [8] and the phase locked loop method (PLL) [9], among others. However, these techniques work well only if the analyzed interferogram has a carrier frequency, a narrow bandwidth and the signal has low noise. Moreover, these methods fail for phase calculation of a closed-fringe pattern. Additionally, the Fourier and Synchronous methods estimate the phase wrapped because of the arctangent function used in the phase calculation, so an additional unwrapping process is required. The unwrapping process is difficult when the fringe pattern includes high amplitude noise, which causes differences greater than 2π radians between adjacent pixels [10–12]. In the PLL technique, the phase is estimated by following the phase changes of the input signal by varying the phase of a computer-simulated oscillator (VCO) such that the phase error between the fringe pattern and VCO's signal vanishes.

Recently, regularization [13–15] and neural networks techniques [16,17] have been used to work with fringe patterns, which contain a narrow bandwidth and noise. The regularization technique establishes a cost function that constrains the estimated phase using two considerations: (a) fidelity between the estimated function and the observed fringe pattern and (b) smoothness of the modulated phase field. For example, in [13,18], the cost function considers a neighborhood of the pixel under analysis to fit a plane. The plane is determined by using gradient descent technique, which takes the partial derivatives of the cost function with reference local phase and carrier frequency into consideration. Its main drawback is that it can easily fall on a local minimum due to use of local gradients in the case of interferograms generated by phase fields that contain many minimal and/or maximal. In a recent work, Servin et al. [18] proposed the fringe-follower regularization phase tracker technique (FFRPT), where a scanning strategy is used to avoid the last drawback. A disadvantage of this is that a low-pass filtering and a binary threshold operation are required. These operations depend on the form of the particular fringe pattern image. Additionally, the FFRPT could be affected by noise presence due to the local consideration (taking a small neighborhood) to fit a plane to each central pixel in the image.

In the neural network technique, a multi-layer neural network (MLNN) is trained by using a set of fringe patterns and a set of phase gradients provided from calibrated objects. After the MLNN has been trained, the phase gradient is estimated in the MLNN output when the fringe patterns (interferograms) are presented in the MLNN input. The drawback of this technique is the requirement of a set of training fringe images and their related phase measurements.

In this work, we propose a technique to determine the phase $\phi(x, y)$, from a fringe pattern with a narrow bandwidth and/or noise, by parametric estimation of a global non-linear function instead of local planes in each site *(x,y)* as it was proposed in [13,18]. A particle swarm optimization (PSO) [19–21] is used to fit the best non-linear function to the phase from the full image not a small neighbourhood as in regularization techniques. The PSO technique was selected to optimize the cost function instead of gradient descent technique since non-convergence problems are presented when it is used in a non-linear function fitting. PSO strategy reduces the possibility of falling in a local optimum. When a noisy closed fringe pattern is demodulated, neither a low-pass filter nor a binarizing operator is required. On the other hand, regularization techniques need both of them.

2 PSO Applied to Phase Recovery

As described by Eberhart and Kennedy, the PSO algorithm is an adaptive algorithm based on a social-psychological metaphor; a population of individuals (referred to as particles) adapts by returning stochastically toward previously successful regions.

The fringe demodulation problem is a difficult problem to solve when the noise in the fringe pattern is high, since many solutions are possible even for a single noiseless fringe pattern. Besides, the complexity of the problem is increased when a carrier frequency does not exist (closed fringes are presented).

Given that for a closed fringe interferogram there are multiple phase functions for the same pattern, the problem is stated as an ill-posed problem in the Hadamard sense, since a unique solution cannot be obtained [22]. It is clear that image of a fringe pattern $I(x,y)$ will not change if $\phi(x,y)$ in Eq. (1) is replaced with another phase function $\tilde{\phi}(x,y)$ given by

$$\tilde{\phi}(x,y)=\begin{cases}-\phi(x,y)+2\pi & (x,y)\in R,\\ \phi(x,y) & (x,y)\notin R\end{cases} \tag{2}$$

where R is an arbitrary region and k is an integer. In this work, a PSO is presented to carry out the optimization process, where a parametric estimation of a non-linear function is proposed to fit the phase of a fringe pattern. Then, PSO technique fit a global non-linear function instead of a local plane to each pixel just like it is made in regularization techniques [13,18]. The fitting function is chosen depending on the prior knowledge of the demodulation problem as object shape, carrier frequency, pupil size, etc. When no prior information about the shape of $\phi(x,y)$ is known, a polynomial fitting is recommended. In this paper, authors have used a polynomial fitting to show how the method works.

The purpose in any application of PSO is to evolve a particle swarm of size P (which codifies P possible solutions to the problem) using update velocity and position of each particle, with the goal of optimizing a fitness function adequate to the problem to solve.

In phase demodulation from fringe patterns, the phase data can be approximated by the selection from one of several fitting functions, such as:

$$f_1(a,x,y)=a_0+a_1x+a_2y+a_3x^2+a_4xy+a_5y^2+\ldots+a_{\frac{(n+1)(n+2)}{2}}y^n, \tag{3}$$

$$f_2(a,x,y)=\sum_{i=1}^{N}a_i \exp\left[-\frac{(x-x_i)^2+(y-y_i)^2}{\sigma}\right] \tag{4}$$

or

$$f_3(a,x,y)=\sum_{j=0}^{n}\sum_{k=-j}^{j}a_{jk}R_j^{|k|}(p)e^{ik\theta} \tag{5}$$

among others. Parameter a is the coefficient function vector or matrix depending on the selected function, which will be estimated by the PSO (whose solution is codified inside of the particle). The function is selected on the basis of background experiment information. For example, if the phase to be calculated is smooth a loworder two-dimensional polynomial may be fitted. In this work, the fitness function U, which is used to evaluate the pth particle a^p in the swarm, is given by

$$\begin{aligned} U(a^p)=\alpha-\sum_{y=1}^{R-1}\sum_{x=1}^{C-1}\{&(I_N(x,y)\\ &-\cos(\omega_x x+\omega_y y+f(a^p,x,y)))^2\\ &+\lambda[(f(a^p,x,y)-f(a^p,x-1,y))^2\\ &+(f(a^p,x,y)-f(a^p,x,y-1))^2]\}m(x,y), \end{aligned} \tag{6}$$

where x, y are integer values representing indexes of the pixel location in the fringe image. Superindex p is an integer index value between 1 and P, which indicates the number of chromosome in the population. $I_N(x,y)$ is the normalized version of the detected irradiance at point (x,y). The data were normalized in the range $[-1,1]$. ω_x and ω_y are the angular carrier frequencies in directions x and y. The function $f(\cdot)$ is the selected fitting function to carry out the phase approximation. $R\times C$ is the image resolution where fringe intensity values are known and λ is a smoothness weight factor (it should be clear for the reader that a higher value of parameter λ implies a smoother function to be fitted). The binary mask $m(x,y)$ is a field which defines the valid area in the fringe pattern. The parameter a can be set to the maximum value of the second term (in negative sum term) at Eq. (6) in the first chromosome population, which is given by

$$
\begin{aligned}
\alpha = \max_{p}\Bigg\{\sum_{y=1}^{R-1}\sum_{x=1}^{C-1} \{(I_N(x,y) \\
-\cos\left(\omega_x x+\omega_y y+f\left(a^p,x,y\right)\right)\right)^2 \\
+\lambda\Big[\left(f\left(a^p,x,y\right)-f\left(a^p,x-1,y\right)\right)^2 \\
+\left(f\left(a^p,x,y\right)-f\left(a^p,x,y-1\right)\right)^2\Big]\ \}m(x,y),
\end{aligned}
\tag{7}
$$

then parameter α is used to convert the proposal from minimal to maximal optimization since a fitness function in a PSO is considered to be a nonnegative figure of merit and profit [19].

The first term (in negative sum term) at Eq. (6) attempts to keep the local fringe model close to the observed irradiances in least-squares sense. The second term (in negative sum term) at Eq. (6) is a local discrete difference, which enforces the assumption of smoothness and continuity of the detected phase.

2.1 Particles

At the beginning of a PSO, a set of random solutions are codified in a particle swarm of size P. Each particle a is formed by the parameter function vector (possible solution) and chained string such as

$$
a=\left[a_0 \vdots a_1 \vdots a_2 \vdots \ldots \vdots a_n\right] \tag{8}
$$

Each dimension a_i is a random real number in a defined search range $(\min(a_i), \max(a_i))$ (the userdefined maximum and minimum of a_i). These values can be initialized using prior knowledge (e.g. in the polynomial case, components x and y are related to the interferogram tilt so if a closed fringe is presented, then these values are near 0). Every dimension is generated as

$$
a_i = random\left(\min\left(a_i\right), \max\left(a_i\right)\right) \tag{9}
$$

The next iterations of particles, positions and velocities are adjusted, and the function is evaluated with the new coordinates at each time-step.

2.2 Particle Velocity and Position Update

During each generation each particle is accelerated toward the particle's previous best position and the global best position. At each iteration a new velocity value for each particle is calculated based on its current velocity, the distance from its previous best position, and the distance from the global best position. The new velocity value is then used to calculate the next position of the particle in the search space. This process is then iterated a set number of times, or until a minimum error is achieved.

In the inertia version of the algorithm an inertia weight, reduced linearly each generation, is multiplied by the current velocity and the other two components are weighted randomly to produce a new velocity value for this particle, this in turn affects the next position of the particle during the next generation. Thus, the governing equations are:

$$v_{id}(t+1) = \omega \cdot v_{id} + c_1 \cdot \varphi_1 \cdot (P_{lid} - a_{id}(t)) + c_2 \cdot \varphi_2 \cdot (P_{gd} - a_{id}(t)) \tag{10}$$

$$a_{id}(t+1) = a_{id}(t) + v_{id}(t+1) \tag{11}$$

where a_i is particle i's position vector, v_i is particle i's velocity vector, c_1 and c_2 are positive constants, are called acceleration coefficients, φ_1 and φ_2 are random positive numbers between 0 and 1. Some researchers have found out that setting c_1 and c_2 equal to 2 gets the best overall performance, where as ω is called inertia weight. P_l is the local best solution found so far by the i-th particle, while P_g represents the positional coordinates of the fittest particle found so far in the entire community. Once the iterations are terminated, most of the particles are expected to converge to a small radius surrounding the global optima of the search space.

2.3 PSO Convergence

The PSO convergence mainly depends on the population size. It should be clear that if we increase the population size, more chromosomes will search the global optimum and a best solution will be found in a minor number of iterations, although the processing time can be increased [19,20]. A good rule of thumb for swarm size is to choose as large a population size as computer system limitations and time constraints allow.

To stop the PSO process, different convergence measures can be employed. In this paper, we have used a relative comparison between the fitness function value of the *gbest* particle in the swarm and value a, which is the maximum possible value to get in Eq. (6). Then, we can establish a relative evaluation of uncertainty to stop the PSO as

$$\left| \frac{\alpha - U(a^{\bullet})}{\alpha} \right| \leq \varepsilon, \tag{12}$$

where $U(a^{\bullet})$is the fitness function value of the *gbest* particle a in the swarm in the current iteration, and ε is the relative error tolerance. Additionally, we can stop the process in a specified number of iterations, if Eq. (12) is not satisfied.

3 Experiments

The parametric method using a PSO was applied to calculate phase from three different kinds of fringe patterns: shadow moiré closed fringe pattern. We use a particles swarm size equal to 100, inertia a number in the range $[0.1,0.9]$, velocity a number in the range $[0.0001,0.0009]$. In each particle, the coded coefficients of a fourthgrade polynomial were included. The following polynomial was coded in each particle:

$$\begin{aligned} p_4(x,y) = {} & a_0 + a_1x + a_2y + a_3x^2 + a_4xy \\ & + a_5y^2 + a_6x^3 + a_7x^2y + a_8xy^2 \\ & + a_9y^3 + a_{10}x^4 + a_{11}x^3y + a_{12}x^2y^2 \\ & + a_{13}xy^3 + a_{14}y^4 \end{aligned} \tag{13}$$

so that 15 coefficients were configured in each particle inside swarm to be evolved.

3.1 Close Fringe Pattern

A low contrasted noisy closed fringe pattern was generated in the computer using the following expression:

$$I(x,y) = 127 + 63\cos(P_4(x,y) + \eta(x,y)). \tag{14}$$

where

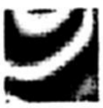
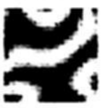
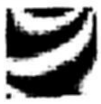
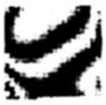
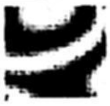
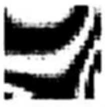
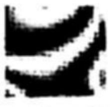
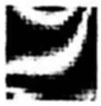
$$\begin{aligned} p_4(x,y) = {} & -0.7316x - 0.2801y + 0.0065x^2 \\ & + 0.00036xy - 0.0372y^2 \\ & + 0.00212x^3 + 0.000272x^2y \\ & + 0.001xy^2 - 0.002y^3 \\ & + 0.000012x^4 + 0.00015x^3y \\ & + 0.00023x^2y^2 + 0.00011xy^3 \\ & + 0.000086y^4 \end{aligned} \tag{15}$$

and $\eta(x,y)$ is the uniform additive noise in the range $[-2radians, 2radians]$. Additionally, the fringe pattern was generated with a low resolution of 60×60. In this case, we use a parameter search range of $[-1,1]$. The swarm of particles was evolved until the number of iterations and relative error tolerance ε was 0.05 in Eq. (12). This condition was achieved in a time of 60s on a AMD Phemon X4-2.5 GHz computer. The fringe pattern and the contour phase field of the computer generated interferogram are shown in Fig. 1. The PSO technique was used to recover the phase from the fringe pattern. The fringe pattern and the phase estimated by PSO is shown in Fig. 1. The normalized RMS error was 0.12 radians and the peak-to-valley error was 0.94 radians. The test as shown in Table 1, and shows best particle for the testers shows in Table 2.

Table 1. Table of parameters of inertial and velocity

	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.0001	2.870	3.432	3.612	3.505	3.839	3.277	2.916	2.777	2.395
0.0002	3.007	3.044	3.210	3.083	2.725	2.680	1.688	1.801	2.366
0.0003	1.665	1.875	2.565	2.559	1.576	1.708	**1.151**	1.945	2.469
0.0004	2.17	**1.738**	2.777	1.912	**1.290**	2.171	1.806	**0.567**	1.946
0.0005	1.883	1.860	2.838	1.686	1.701	2.063	1.969	0.791	1.792
0.0006	2.106	2.134	2.900	**1.086**	2.318	1.705	1.645	1.399	2.343
0.0007	1.928	1.993	**0.853**	1.168	2.019	2.270	1.772	1.428	1.828
0.0008	**0.893**	1.938	1.350	1.531	2.019	2.632	1.373	1.373	2.260
0.0009	1.536	1.911	1.436	1.773	2.407	**0.313**	1.902	0.779	**1.523**

Table 2. Shows of the best particles

Inertia	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Velocity	0.0008	0.0004	0.0007	0.0006	0.0004	0.0009	0.0003	0.0004	0.0009
									

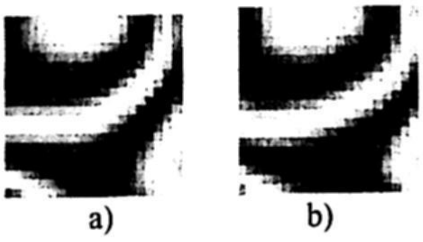


Fig. 1. (a) Fringe pattern and (b) phase field obtained by using PSO technique

4 Conclusions

A PSO was applied to recover the modulating phase from closed and noisy fringe patterns. A fitness function, which considers the prior knowledge of the object being tested, is established to approximate the phase data. In this work a fourthgrade polynomial was chosen to fit the phase.

A swarm of particles was generated to carry out the optimization process. Each particle was formed by a codified string of polynomial coefficients. Then, the swarm of particles was evolved using velocity, position and inertial

The PSO technique works successfully where other techniques fail (Synchronous and Fourier methods). This is the case when a noisy, wide bandwidth and/or closed fringe pattern is demodulated. Regularization techniques can be used in these cases but PSO technique has the advantage that the cost function does not depend upon the existence of derivatives and restrictive requirements of continuity (gradient descent methods). Since the PSO works with a swarm of possible solutions instead of a single solution, it avoids falling in a local optimum. Additionally, no filters and no binarizing operators are required, in contrast with the fringe-follower regularized phase tracker technique.

The PSO has the advantage that if the user knows prior knowledge of the object shape, then a better suited fitting parametric function can be used instead of a general polynomial function. Additionally, due to the fact that the PSO technique gets the parameters of the fitting function, it can be used to interpolate sub-pixel values and to increase the original phase resolution or interpolate where fringes do not exist or are not valid. A drawback is the selection of the optimal initial PSO parameters (such as swarm size, inertial, velocity) that can increase the convergence speed.

Aknowledgements. We wish to thank to the Centro de Invcstigacion en Computación by their support to this proyect and in general to the CIC-IPN, and Centro de Investigaciones en Optica during the image recollections and tests; to CONACYT by the scolarship by the support to this research.

We also acknowledge the support of the Consejo Nacional de Ciencia y Tecnologıa de Mexico, Centro de Investigaciones en Optica, A.C. and Centro de Investigacion en Computacion del Instituto Politecnico Nacional de Mexico

References

1. F. Martín et al, “New advances in Automatic Reading of VLP's”, Proc. SPC-2000 (IASTED), Marbella, España, 2000, 126-131.
2. D. Malacara, M. Servin, Z. Malacara, Interferogram Analysis for Optical Testing, Marcel Dekker, New York, 1998.
3. D. Malacara, Optical Shop Testing, Wiley, New York, 1992.
4. K. Creath, in: E. Wolf (Ed.), Progress in Optics, vol. 26, Elsevier, Amsterdam, 1988, p. 350.
5. K. Creath, in: D. Robinson, G.T. Reid (Eds.), Interferogram Analysis, IOP Publishing, London, 1993, p. 94.
6. M. Takeda, H. Ina, S. Kobayashi, J. Opt. Soc. Am. 72 (1981) 156.
7. X. Su, W. Chen, Opt. Laser Eng. 35 (2001) 263.
8. K.H. Womack, Opt. Eng. 23 (1984) 391.
9. M. Servin, R. Rodriguez-Vera, J. Mod. Opt. 40 (1993)
10. D.C. Ghiglia, G.A. Mastin, L.A. Romero, J. Opt. Soc. Am. 4 (1987) 267.
11. X. Su, L. Xue, Opt. Eng. 40 (2001) 637.
12. M. Servin, F.J. Cuevas, D. Malacara, J.L. Marroquin, R. Rodriguez-Vera, Appl. Opt. 38 (1999) 1934.

13. M. Servin, J.L. Marroquin, F.J. Cuevas, Appl. Opt. 36 (1997) 4540.
14. J. Villa, M. Servin, Opt. Laser Eng. 31 (1999) 279.
15. J.A. Quiroga, A. Gonzalez-Cano, Appl. Opt. 39 (2000) 2931.
16. F.J. Cuevas, M. Servin, O.N. Stavroudis, R. Rodriguez-Vera, Opt. Commun. 181 (2000) 239.
17. F.J. Cuevas, M. Servin, R. Rodriguez-Vera, Opt. Commun. 163 (1999) 270.
18. M. Servin, J.L. Marroquin, F.J. Cuevas, J. Opt. Soc. Am. A 18 (2001) 689.
19. Kennedy, J., Eberhart, R.C., 1995a. Particle Swarm Optimization. Proc IEEE Int. Conf. On Neural Networks, Perth, pp. 1942-1948.
20. Kennedy, J., 1997. The particle swarm: social adaptation of knowledge. IEEE International Conference on Evolutionary Computation, April 13-16, pp. 303 – 308.
21. Kennedy, J., Spears, W.M., 1998. Matching Algorithms to Problems: An Experimental Test of the Particle Swarm and Some Genetic Algorithms on the Multimodal Problem Generator. Proceedings of the IEEE Int'l Conference on Evolutionary Computation. pp. 39-43.
22. J. Hadamard, Sur les problems aux derivees partielles et leur signification physique, Princeton University Bulletin 13, Princeton, NJ, 1902.

Multi-agent Systems and Simulation

Coordinated Multi-agent Exploration

Alfredo Toriz P., Abraham Sánchez L.[+] and Maria A. Osorio L.[•]

LIRMM - UMR55606 CNRS Montpellier, France
[+] Facultad de Ciencias de la Computación, BUAP
[•]Escuela de Ingeniería Química, BUAP
alfredot@hotmail.com, asanchez, aosorio@cs.buap.mx

Abstract. Many successful robotic systems use maps of the environment to perform their tasks. In this paper, we propose a cooperative exploration strategy for multi-agent robots. This proposal is a parallelization of the basic SRT method, the following functionalities were added to it: cooperation to increase the efficiency, coordination to avoid conflicts and communication to cooperate and to coordinate. The goal in robot exploration must be to minimize the overall exploration time, and multiple robots produce more accurate maps by merging overlapping information that helps to stabilize the sensor uncertainty and to reach the goal. We present simulation results to show the performance of the proposed technique.

1 Introduction

Although most mobile robotic systems use a single robot that only operates in its environment, a number of researchers have considered the advantages and disadvantages of the potential use of a group of robots that cooperate for the accomplishment of a required task [1], [2], [3]. Multi-agent systems (MAS), may be regarded as a group of entities called agents, interacting with one another to achieve their individual as well as collective goals. The research domain of multi-agent robot systems can be divided into subdomains according to the task given to the robot group [4]. At present well-studied subdomains are motion planning, formation forming, region-sweeping and combinations of the foregoing. We focused this paper in the region-sweeping task. In the region-sweeping task, one can consider two activities.

In the first activity, a group of robots receives the order to explore/map an unknown region. The goal is to obtain a detailed topography of the desired area. In most exploration approaches, the boundary between know and unknown territory (the frontier) is used in order to maximize the information gain. In [5], the robots merge the acquired information in a global grid-map of the environment, from which the frontier is extracted and used to plan the individual robot motions. The approach presented in [3] proposed to negotiate robot targets by optimizing a utility function which takes into account the information gain of a particular region. The utility of a particular frontier region from a viewpoint of relative robot localization and the accuracy of map merging were considered in

© G. Sidorov, M. Aldape, M. Martínez, S. Torres. (Eds.)
Advances in Computer Science and Engineering.
Research in Computing Science 42, 2009, pp. 189-200

Received 25/02/09
Accepted 27/04/09
Final version 04/05/09

[6]. The incremental deployment algorithm considers that the robots approach the frontier while they retain visual contact with each other [7]. A multi-robot architecture proposed in [8] guide the exploration by a market economy, whereas [9] proposes a centralized approach which uses a frontier-based search and a bidding protocol assign frontier targets to the robots.

Closely related to the exploring/mapping activity, the second one is called complete coverage, where the robots have to move over all of the free surface in the space [10], [11]. Generating maps is one of the fundamental tasks of mobile robots. Many successful robotic systems use maps of the environment to perform their tasks. The questions of how to represent environments and how to acquire models using this representation therefore is an active research area, see [12] for an excellent overview.

This paper presents a strategy to explore an unknown environment by multi-agent robots. The strategy is a parallelization of the SRT (Sensor-based Random Tree) method, which was presented in [13]. The extension of the SRT method to multi-agent robots is called, the Multi-SRT method. A decentralized cooperation mechanism and two coordination mechanisms are introduced to improve the exploration efficiency and to avoid conflicts. The basic steps of the exploration approach are presented in Section II. Simulation results in different environments are discussed in Section III. Finally, conclusion and future work are detailed in Section IV.

2 Cooperative Exploration

MAS may be comprised of homogeneous or heterogeneous agents, it is considered as crucial technology for the effective exploitation of the increasing availability of diverse of heterogeneous and distributed on-line information sources. MAS is a framework for building large, complex and robust distributed information processing systems which exploit the efficiencies of organized behavior. Teamwork and communication are two important processes within multi-agent robots designed to act in a coherent and coordinated manner [2], [4].

An extensive amount of research has been carried out in the areas of localization, mapping and exploration for single autonomous robots, but only fairly recently has this been applied to multi-robot teams [12]. In addition, nearly all of this research has taken an existing algorithm developed for single-robot mapping, localization, or exploration, and extended it to multiple robots, as opposed to developing a new algorithm that is fundamentally distributed [1], [5], [14]. An interesting exception is some of the work in multi-robot localization, which takes advantage of multiple robots to improve positioning accuracy beyond what is possible with single robots [15]. As in the case with single-robot approaches, the research into the multi-robot version can be described using familiar categories and using either range sensors (such as sonar or laser) or vision sensors. While the single-robot version of this problem is fairly well understood, much work remains to be done on the multi-robot version. For example, one question is about the effectiveness of multi-robot teams over single-robot versions, and to what extent adding additional robots brings diminishing returns. This issue has begun to be studied, but much remains to be determined for the variety of approaches available for localization, mapping, and exploration.

The design of the cooperative exploration strategy proceeds from the parallelization of the basic SRT method, each robot builds one or more partial maps of the environment, organized in a collection of SRTs [16]. Each node of an SRT represents a configuration q which was visited by at least one robot, together with the associated local safe region (LSR). An arc between two nodes represents a collision-free path. The tree is incrementally built by extending the structure in the most promising direction via a biased random mechanism. The presence of other robots in the vicinity is taken into account at this stage in order to maximize the information gain and guarantee collision avoidance.

Consider a population of n identical robots. Each robot is equipped with a ring of range finder sensor or a laser range finder, the sensory system provides the local safe region $S(q)$. The robots move in a planar workspace, i.e., $\mathbf{R}^2$ or a connected subset of it; the assumption of planar workspace is not restrictive, 3D worlds are admissible as long as the sensory system allow the reconstruction of a planar LSR for planning the robot motion [17]. Each robot is a polygon[1] or another shape subject to non-holomic constraints. The robot also knows its configuration q, one can eliminate this assumption by incorporating a localization module in the method. The robots know its ID number and each robot can broadcast within a communication range R_c the information stored in its memory (or relevant portions of it) at any time. The robot ID number is included in the heading of any transmission. The robot is always open for receiving communication from other robots inside R_c.

The exploration algorithm for each robot is shown in Figure 1. First, the procedure BUILD_SRT is executed, i.e., each robot builds its own SRT, $\mathcal{T}$ is rooted at its starting configuration q_{init}. This procedure terminates when the robot can not further expand $\mathcal{T}$. Later, the robot executes the SUPPORT_OTHERS procedure, this action contributes to the expansion of the SRTs that have been built by others robots. When this procedure finishes, the robot returns to the root of its own tree and finishes its exploration.

```
BUILD Multi-SRT($q_{init}$)
$\mathcal{T}.init(q_{init})$
BUILD_SRT($q_{init}.\mathcal{T}$);
SUPPORT_OTHERS($q_{init}$);
```

Fig. 1. The Multi-SRT algorithm.

[1] Polygonal models make it possible to efficiently compute geometric properties, such as areas and visibility regions.

The procedure BUILD_SRT is shown in Figure 2. In each iteration of the BUILD_SRT, the robot uses all available information (partially collected by itself and partially gained through the communication with other robots) to identify the group of engaged robots (GER), i.e. the other robots in the team with which cooperation and coordination are adequate. This is achieved by the construction of the first group of pre-engaged robots (GPR), or robots that are candidates to be members of the GER, and are synchronized with them (BUILD_AND_WAIT_GPR). Then, the robot collects data through its sensory systems, it builds the current LSR (PERCEIVE) and updates its own tree $\mathcal{T}$. The current GER can now be built (BUILD_GER). At this point the robot processes its local frontier (the portion of its current LSR limit leads to areas that are still unexplored) on the basis of $\mathcal{T}$ as well as any other tree $\mathcal{T}_i$ gained through communication and stored in its memory (LOCAL_FRONTIER).

If the local frontier is not empty, the robot generates a random configuration contained in the current LSR and headed towards the local frontier, if not, the target configuration is fixed to the node father with a backward movement (PLANNER). If the GER is composed only by the same robot, the robot moves directly to its target. Otherwise, the paths advanced by the robot in the GER are checked for mutual collisions, and classified in feasible and unfeasible paths (CHECK_FEASIBILITY). If the subset $\mathcal{G}_u$ of robots with unfeasible paths is vacuum, a coordination stage takes place, perhaps, confirming or modifying the current target of the robot (COORDINATE). In particular, the motion of the robot can be banned by simply readjusting the target to the current configuration. Then, the function MOVE_TO transfers the robot to the target (when this is different from q_{act}). The loop is repeated until the condition in the output line 15 is verified: the robot is unable to expand the tree $\mathcal{T}$ (no local frontiers remaining) and therefore it has to move back to the root of its SRT. For more details of the most important stages of the BUILD_SRT procedure, one can see [16].

If the subset $\mathcal{G}_u$ of robots with unfeasible paths is not vacuum, the coordination function is invoked. The first step is to choose a master robot within $\mathcal{G}$. This can be complemented in many ways through a deterministic procedure known by all the robots, for example, the robot with the highest ID number can be elected. Two cases are possible then:

1) If the robot is the master, it invokes an ORGANIZE function, whose task is to rearrange the vector $\mathcal{Q}_g$ that contains the targets of the robots in the GER and obtain a feasible collective motion. Here, the change may mean whatever, simply accepting or readjusting the target of a robot to the current configuration (i.e., authorizing/prohibiting the motion) or adding a third option, for example, changing to a new target. We have devised two versions of the function, ORGANIZE1 and ORGANIZE2.

2) If the robot is not the master, it enters in a waiting phase, which ends with the receipt of a specified signal from the master.

```
**BUILD_SRT**$(q_{init}, \mathcal{T})$
$q_{act} = q_{init}$;
do
  BUILD_AND_WAIT_GPR();
  $\mathcal{S}(q_{act}) \leftarrow$ PERCEIVE$(q_{act})$;
  ADD$(\mathcal{T}, (q_{act}, \mathcal{S}(q_{act})))$;
  $\mathcal{G} \leftarrow$BUILD_GER();
  $\mathcal{F}(q_{act}) \leftarrow$ LOCAL_FRONTIER$(q_{act}, \mathcal{S}(q_{act}), \mathcal{T}, \bigcup \mathcal{T}_i)$;
  $q_{target} \leftarrow$ PLANNER$(q_{act}, \mathcal{F}(q_{act}), q_{init})$;
  if $q_{target} \neq NULL$
    if $|\mathcal{G}| > 1$
      $(\mathcal{G}_f, \mathcal{G}_u) \leftarrow$ CHECK_FEASIBILITY$(\mathcal{G})$;
      if $\mathcal{G}_u \neq \emptyset$
        $q_{target} \leftarrow$ COORDINATE$(\mathcal{G}_f, \mathcal{G}_u)$;
    $q_{act} \leftarrow$ MOVE_TO$(q_{target})$;
while $q_{target} \neq NULL$
```

Fig. 2. The BUILD_SRT procedure.

The final operation is to retrieve and return the robot's (possibly modified) own target from $\mathcal{Q}_g$.

ORGANIZE1 (organization via arbitration) implements a simple mechanism for arbitration in $\mathcal{G}$. In particular, all the robots contained in the feasible subset $\mathcal{G}_f$ are allowed to move (their target configurations are left unchanged). The robots that are in the unfeasible subset $\mathcal{G}_u$, are not allowed to move (their target configuration is initialized to the current configuration) with the exception of one whose motion is authorized (this strategy is guaranteed to avoid conflicts).

The selection of the robot authorized in $\mathcal{G}_u$ can be done on the basis of several criteria. The one we implemented chooses randomly one of the robots whose local frontier is empty: any of the robots whose target is their parent node (i.e., robots running BUILD_SRT and robots advised to backtrack by the planner) or robots that are moving along the trees initiated by other robots with the goal of helping them in the expansion (robots that are running the SUPPORT_OTHERS procedure and are still in the stage of transfer). This strategy is motivated by the fact that if the motion was unauthorized, such robots will have to wait until their path is clear, because they cannot change their goal (compared to the robots whose local frontier is not empty and the planner can propose a different destination). A non-ethical approach would be to randomly choose between robots in $\mathcal{G}_u$ using a probability proportional to the length of its local frontier.

ORGANIZE2 (organization through replanning) tries to modify the targets of the robots in $\mathcal{G}$, in order to maximize the number of simultaneous feasible moves. This can be done by formalizing the problem as follows. Consider the set of target configurations $\mathcal{Q}_{\mathcal{G}}$ associated with the GER $\mathcal{G}$. Two target configurations in $\mathcal{Q}_{\mathcal{G}}$ are compatible if they can be reached by the corresponding robots with paths that are not intercepted. Let G be the compatibility graph associated with $\{\mathcal{G}, \mathcal{Q}_{\mathcal{G}}\}$ and defined as the indirect graph whose nodes represent the robots in $\mathcal{G}$ and whose arcs connect pairs of nodes with compatible targets. A maximum clique in G is a full subgraph of G with maximum cardinality, corresponding to a maximum subset of robots with compatible targets. The ideal objective of the ORGANIZE2 is to modify the set of target configurations $\mathcal{Q}_{\mathcal{G}}$ to maximize the cardinality of the maximum clique associated with the constraint that the target of each robot is either accepted, changed into another configuration directed to the robot's local frontier (if this is not empty) or to the robot's current configuration (the move is not authorized). This is a very complex problem whose solution requires the computation of maximum cliques as a subproblem. To find a satisfactory solution in a given amount of time, we have adopted a randomized search technique conducted by the master as a sequential game with complete information.

```
COORDINATE($\mathcal{G}_f, \mathcal{G}_u$)
master_id ← MASTER_ELECTION($\mathcal{G}$);
if my_id = master_id
    $\mathcal{Q}_{\mathcal{G}}$ ← ORGANIZE($\mathcal{G}_f, \mathcal{G}_u$);
else
    WAIT;
return $\mathcal{Q}_{\mathcal{G}}$(my_id);
```

Fig. 3. The coordination function.

The procedure SUPPORT_OTHERS can be divided into two major phases, which are repeated over and over again. In the first phase, the robot picks another robot to support it in his exploration, or, more precisely, another tree that helps it to expand (there may be more than one robot acting on a single tree). In the second phase, the selected tree is reached and the robot tries to expand it, tying subtrees constructed by the procedure BUILD_SRT. The main cycle is repeated until the robot has received confirmation that all the other robots have completed their exploration.

Research into exploration methods has usually addressed the problem in isolation, rather than together with the SLAM problem. Frontier-based exploration is one of the simplest techniques and has been widely tested. Using this exploration technique, the robot moves towards boundaries or frontiers between known and unknown parts of the environment. Frontier-based exploration is relatively easy to integrate into most existing mobile robots architectures, although its performance in realistic long term robot experiments with uncertainty is untested.

3 Simulation Results

In order to illustrate the behavior of the Multi-SRT exploration approach, we present two strategies, the Multi-SRT-Radial (see [18] for more details.) and the Multi-SRT-Star. The strategies were implemented in Visual C++ V. 6.0, taking advantage of the MSL library's[2] structure and its graphical interface that facilitates to select the algorithms, to visualize the working environment and to animate the obtained path. The library GPC developed by Alan Murta was used to simulate the sensor's perception systems[3]. GPC is a C library implementation of a polygon clipping algorithm. In the simulation process, the robot along with the sensor's system move in a 2D world, where the obstacles are static; the only moving object is the robot. The robot's geometric description, the workspace and the obstacles are described with polygons.

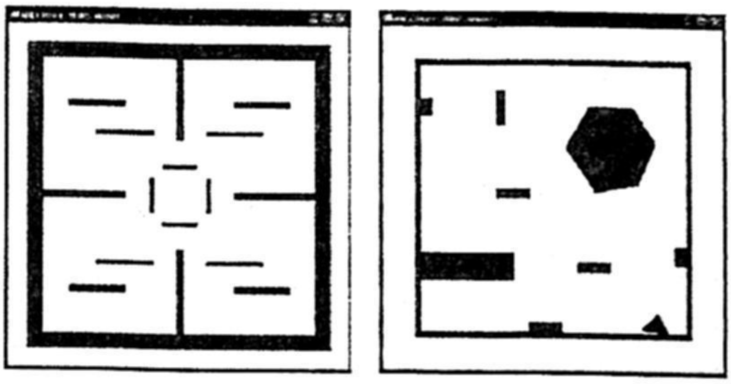


Fig. 4. Environments used for the tests of the Multi-SRT.

The tests were performed on an Intel © Pentium D processor-based PC running at 2.80 GHz with 1 GB RAM. One can consider two possible initial deployments of the robots. In the first, the robots are initially scattered in the environment; and in the second, the exploration is started with the robots grouped in a cluster. Since the Multi-SRT approach is randomized, the numerical results were averaged over 20 simulation runs. Environment coverage is not reported because it was complete in all our simulations. Figure 4 illustrates the environments used for the simulation part. The first is a square region with a garden-like layout, where each area can be reached from different access points. The second is also a square, it contains many obstacles of different shapes. These first results showed are obtained by using the Multi-SRT-Radial strategy. The performance

[2] http://msl.cs.uiuc.edu/msl/

[3] http://www.cs.man.ac.uk/~toby/alan/software/

of the method is evaluated in terms of exploration time (the time required by the last robot of the team to return home). The polygonal representation facilitates the use of the GPC library for the perception algorithm's simulation. If S is the zone that the sensor can perceive in absence of obstacles and SR the perceived zone, the SR area is obtained using the difference operation of GPC between S and the polygons that represent the obstacles.

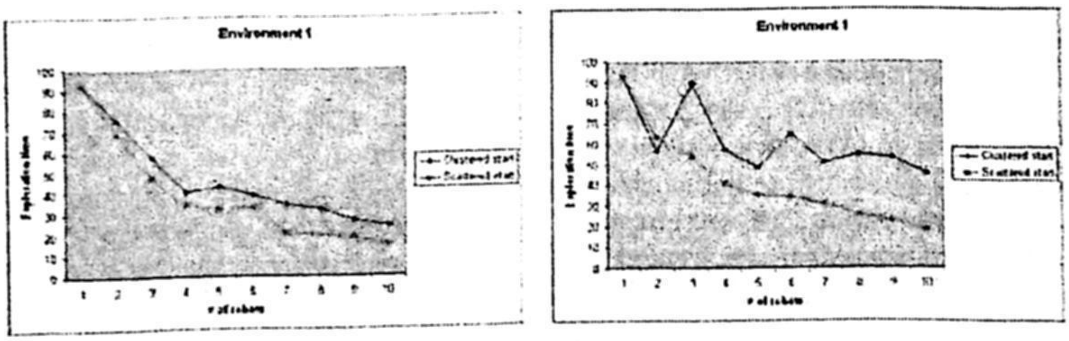


Fig. 5. Environment 1 exploration with scattered and clustered start. To the left with unlimited communication range and in the right with limited communication range.

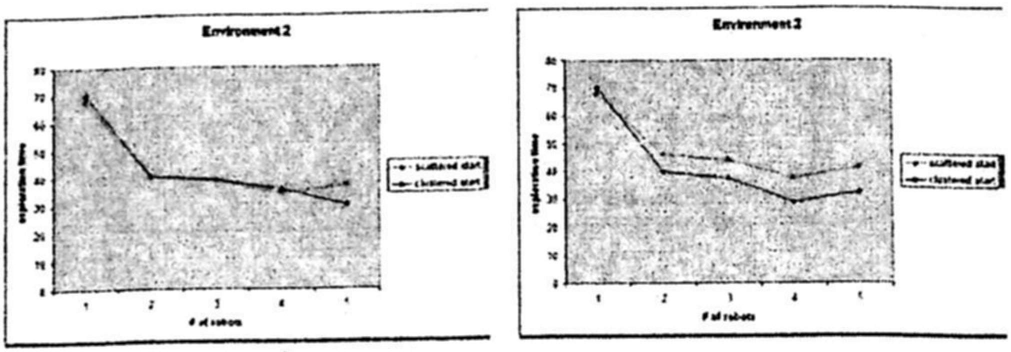


Fig. 6. Environment 2 exploration with scattered and clustered start. To the left with unlimited communication range and in the right with limited communication range.

Exploration time for teams of different cardinality are shown in Figures 5 and 6, both in the case of limited and unlimited communication range. In theory, when the number of robots increases, the exploration time would quickly have to decrease. This affirmation is fulfilled in the case of the scattered start; note however that, in the case of the clustered start, there are examples where this affirmation is not verified. We consider that an increment of the number of evenly deployed robots corresponds to a decrement of the individual areas they must cover (see Figure 7, for the unlimited communication range case). In the case of a limited communication range, when the robots are far apart at the start, they can exchange very little information during the exploration process. The total travelled distance increases with the number of robots because more robots try to support the others in their expansion. We used ORGANIZE2 for coordination (the performance of ORGANIZE1 is similar). As the number of robots increases, communication chains are formed and the total distance decreases. Due to the corresponding importance of the coordination phase, the waiting time[4] in both coordination strategies grows with the number of robots. We did not report the corresponding results to the exploration using the Multi-SRT-Star strategy because in many cases the exploration of the environment is not completed.

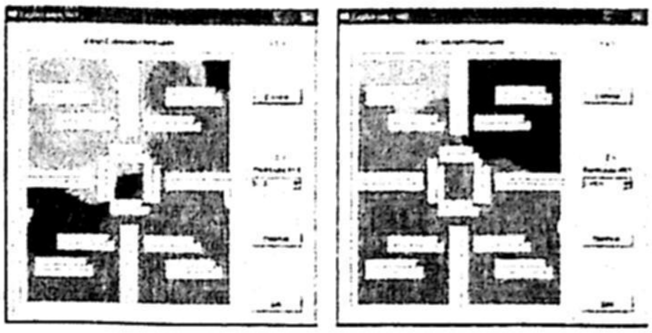


Fig. 7. The explored regions with clustered and scattered starts with a team of 4 robots.

Figures 8 and 9 show the Multi-SRT and the explored region for the environment 1 with a team of 5 and 10 robots in the case of unlimited communication range. We can see the difference when the robots are evenly distributed at the start or are clustered. At the end, the environment has been completely explored and the SRTs have been built. In these figures, we can observe that each robot built its own SRT and when one of them finished, this entered the support phase.

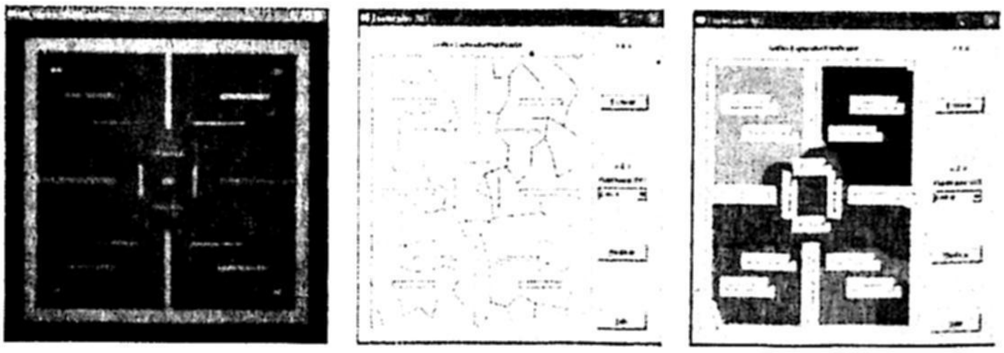


Fig. 8. Environment 1 with 5 robots. The Multi-SRT and explored regions with scattered start. Time = 54.348 secs with 141 nodes.

[4] The average percentage of the exploration time that a robot spends waiting due to coordination.

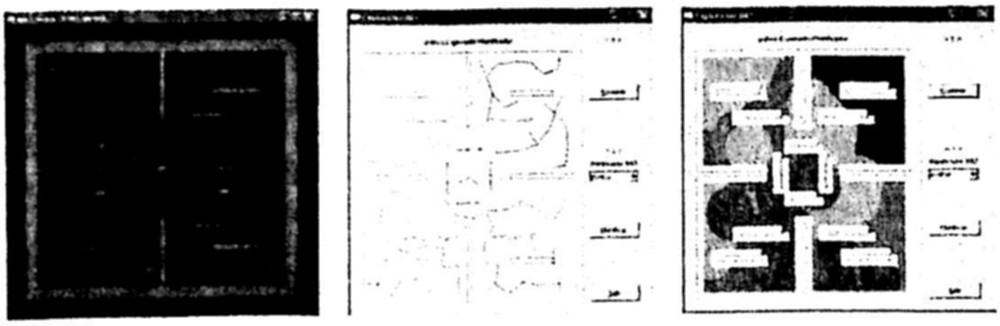


Fig. 9. Environment 1 with 10 robots. The Multi-SRT and explored regions with scattered start. Time = 26.658 secs with 32 nodes.

If we compare the exploration times in the figures 8 and 9, we can affirm that to greater number of robots, the exploration time decreases. If we also observe the placed figure in the center of both figures 8 and 9, one can note that the trees are united, this indicates that the support phase took place. Single robots can not produce accurate maps like multi-robots. The only advantage of using a single robot is the minimization of the repeated coverage.

However, even though repeated coverage among the robots decreases the mission's efficiency, some amount of repeated coverage is a desirable situation for better efficiency. Additionally, this better efficiency can be achieved by coordination among robots.

3.1 Discussion

To solve a multi-agent task, either centralized or decentralized (distributed) approaches can be used. A centralized model uses a powerful agent to plan and schedule the subtasks for every agent. This control agent has a global knowledge concerning the environment and the problems. It can deliberately plan for better performance. However, for tasks with NP complexity, the centralized approach is impractical. Furthermore, the control agent must be powerful enough to achieve satisfactory performance. High design complexity, high cost and low reliability are the other drawbacks of this approach.

On the other hand, a distributed approach decreases design complexity and cost, while increasing the reliability. Agents are autonomous and equal. An agent plans for itself and communicates with the others in order to accomplish the global task. Since every agent interacts directly with the environment, it is reactive. However, each agent has only local knowledge of the task and the environment. Hence, it cannot make the best decision of the global task alone. Furthermore, negotiation and social cooperation rules for conflict resolution are required to coordinate among them.

One distributed approach is to let each agent work alone. Whenever an agent cannot achieve its goal by itself, it requests help from the others. Moreover, every agent always offers help when it can. This is *help-based cooperation.* It is simple and effective in achieving the overall task. Its problems include: too many helpers for an agent is a waste; too many agent requiring help will cause a deadlock; some agents obtain help more often than the other agents (unbalanced load); and local knowledge limits the system performance.

Another approach is to let agents together. For a given task, agents coordinate a global plan for performing the task by exchanging their local knowledge. Its is *coordination-based cooperation.* With enough information, a task can be optimally achieved. Moreover, each agent can optimize its decisions. The problems of this approach include overhead due to coordination, complexity in choosing optimal decisions, and the amount of storage for saving the exchanged information.

The local coordination procedure implemented in our work guarantees that the collective motion of the robots is feasible from the collision viewpoint. The approach does not need a central supervision. The selection of exploration actions by each robot is spontaneous and it is possible on the basis of the available information.

4 Conclusions and Future Work

The use of multi-robot system brings in general many advantages. In exploration, it aims at significantly reducing the time required to complete the task. Possible applications include surface inspection, mine sweeping, surveillance, search and rescue missions and planetary operations. The aim of this research work is to develop a robust exploration technique for single and multi-robot system in an unknown indoor environment. For the case where multiple robots are employed, the system must be scalable, decentralized and tolerant to temporary lost of communications between some robots.

We have presented an interesting approach for cooperative exploration based on the SRT method. The Multi-SRT considers two decentralized mechanisms of cooperation at different levels. The first simply consists in making an appropriate definition of the local frontier that allows each robot to plan its motion towards the areas apparently unexplored for the rest of the team. The second allows a robot that has finished with its individual exploration phase, to support others robots in their exploration task. Additionally, we compared Multi-SRT-Radial strategy with Multi-SRT-Star strategy, the results obtained with the radial perception strategy are more interesting.

SRT is an interesting method for single/multiple robot indoor exploration and mapping. The method combines local frontier-based exploration technique and global graph-based representation of the environment to produce a robust autonomous exploration strategy.

In this proposal. we assumed the robots have good localization and a common frame of reference, but not necessarily the same start position. We are currently working in an integrated exploration strategy for one single robot (the continuous localization procedure is based on natural features of the safe region). The problem of simultaneously localizing a group of mobile robots is still open. We can also consider an extension of the Multi-SRT exploration method, where the robots constantly maintain a distributed network structure.

References

1. Y. Cao, A. Fukunaga and A. Kahng, "Cooperative mobile robotics: Antecedents and directions", *Autonomous Robots*, Vol. 4, (1-23) 1997
2. G. Dudek, M. Jenkin, E. Milios and D. Wilkes, "A taxonomy for multi-agent robotics", *Autonomous Robots*, Vol. 3, (375-397) 1996
3. W. Burgard, M. Moors, and F. Schneider, "Collaborative exploration of unknown environments with teams of mobile robots", *Plan-Based Control of Robotic Agents, LNCS*, Vol. 2466, (2002)
4. Ota J, "Multi-agent robot systems as distributed autonomous systems", *Advanced Engineering Informatics*, Vol. 20, (2006) 59-70
5. B. Yamauchi, " Decentralized coordination for multirobot exploration", *Robotics and Autonomous Systems*, Vol. 29, (1999) 111-118
6. J. Ko, B. Stewart, D. Fox, K. Konolige and B. Limketkai, "A practical, decision-theoretic approach to multi-robot mapping and exploration", *IEEE Int. Conf. on Intelligent Robots and systems*, (2003) 3232-3238
7. A. Howard, M. Mataric and S. Sukthane, "An incremental deployment algorithm for mobile robot teams", *IEEE Int. Conf. on Intelligent Robots and Systems*, (2002) 2849-2854
8. R. Zlot, A. Stenz, M. Dias and S. Thayer, "Multi-robot exploration controlled by a market economy", *IEEE Int. Conf. on Robotics and Automation*, (2002) 3016-3023
9. R. Simmons, D. Apfelbaum, W. Burgard, D. Fox, M. Moors, S. Thrun and H. Younes, "Coordination for multi-robot exploration and mapping", *17th Conf. of the American Association for Artificial Intelligence*, (2000) 852-858
10. E. U, Acar, H. Choset, A. A. Rizzi, P. N. Atkat and D. Hull, "Morse decompositions for coverage tasks", *International Journal of Robotics Research*, Vol. 21, (2002) 331-344
11. M. A. Batalin and G. S. Sukhatme, "Coverage, exploration and deployment by a mobile robot and communication network", *Proc. International Workshop on Information Processing in Sensor Networks*, (2003) 376-391
12. S. Thrun, W. Burgard and D. Fox, "Probabilistic robotics", *MIT Press*, (2005)
13. G. Oriolo, M. Vendittelli, L. Freda and G. Troso, "The SRT method: Randomized strategies for exploration", *IEEE Int. Conf. on Robotics and Automation*, (2004) 4688-4694
14. W. Burgard, M. Moors, C. Stachniss and F. Schneider, "Coordinated multi-robot exploration", *IEEE Transactions on Robotics*, Vol. 21, No. 3, (2005) 376-378
15. S. I. Roumeliotis and G. A. Bekey., "Distributed multirobot localization", *IEEE Transactions on Robotics and Automation*, Vol. 18, No. 5, (2002) 781-795
16. A. Toriz Palacios, "Estrategias probabilisticas para la exploración cooperativa de robots móviles", *Master Thesis*, FCC-BUAP (in spanish), 2007
17. I. Kamon, E. Rimon and E. Rivlin "Range-sensor-based navigation in three-dimensional polyhedral environments", *The International Journal of Robotics Research*, Vol. 20, No. 1, (2001) 6-25
18. J. Espinoza L., A. Sánchez L. and M. Osorio L., "Exploring unknown environments with mobile robots using SRT_Radial", *IEEE Int. Conf. on Intelligent Robots and Systems*, (2007) 2089-2094

A Specialist Agent in Marketing Strategies for Quotes Establishment

Miriam Salcedo, Darnes Vilariño, Fabiola López,
Josefa Somodevilla and Mireya Tovar

Faculty of Computer Science, BUAP- México.

Abstract. The design and implementation of a specialist agent on market strategies are proposed under the rules of TAC-CAT agents International Contest. The main goal of this specialist agent is to create a market and then, to establish a set of policies in order to control agents' behavior who want to trade with him. One of the most important policies to obtain a profitable market is the quota policy, therefore a detailed explanation of the quota policy' strategies are provided. Finally, the results from tests on each strategy are shown.

Keywords: Electronic market, agents, TAC CAT, marketing strategies, economics.

1 Introduction

Along the time, the human being has had the necessity to negotiate diverse goods and services to satisfy its basic needs. Buyers and sellers are people who want to negotiate, they have to find a meeting point where they can negotiate, and this point is named market. Little by little well established big markets and informal small markets were appearing. Small markets being studied by Microeconomics usually trade with specific goods.

The concept of electronic market comes with the birth of the Internet. Electronic markets widely facilitate the negotiations between buyers and sellers groups. Agents are used to carry out these negotiations on Internet, which are entities capable of realizing autonomous tasks in an efficient way.

TAC CAT Market Design Agents International Competition [1] came out in 2007 as a TAC SCM [2] branch due to the exponential growing in electronic markets. In this tournament, the goal is that contenders make a specialist agent capable to create a market where buyers and sellers can negotiate under a set of rules that control agent's behavior, bids and transactions at the market.

This work shows design aspects and details of implementation of a specialist agent in market strategies based on TAC CAT contest. Policy of quotes receives special attention, since is one of the most important to guarantee a profitable market. Microeconomic theory' strategies are used to establish a quota policy, for which some test and results are provided.

Section 2 describes CAT game dynamics in a general way. In section 3, the design of the specialist agent in market strategies is discussed. A detailed explanation of the strategies used for the implementation of quotas policy is given in section 4. Section 5

© G. Sidorov, M. Aldape, M. Martínez, S. Torres. (Eds.)
Advances in Computer Science and Engineering.
Research in Computing Science 42, 2009, pp. 201-212

Received 09/03/09
Accepted 27/04/09
Final version 07/05/09

presents implementation of this policy. In Section 6 results of several tests are shown, and finally, the conclusions and ongoing work are presented.

2 TAC CAT: The Game

The CAT tournament is based on client–server model; Server performs the game on a JCAT platform [3]. The behavior of server and clients is regulated by CATP protocol [4]. CAT clients are represented on the platform as independent agents and they communicate among them through the server. CAT clients are buyers and sellers named negotiating agents, and specialist agents are the ones that represent the market.

Buyer and seller agents are provided by organizers of the contest. Each of them is provided with two strategies: a) market selection strategy and b) trading strategy [3]. Both strategies continuously change during the game and specialists do not know about them. Each negotiating agent has a set of private values (for trading goods) and a limited budget, both of them unknown as well.

Each contender designs and implements a specialist agent to create and control his own market, establishing the following policies:

Accepting policy. This policy judges whether a shout made by a trader should be accepted in the market.

Charging policy. This policy determines the quotes issued by markets which will be charged by the specialists (Figure 1(b)). These quotes are:

- Registry quote
- Information quote
- Shout quote
- Transaction quote
- Profit quote

Pricing policy. This policy determines the transaction price for matched ask-bid pairs.

Clearing policy. This policy determines how and when to clear the market, that is, how to match accepted shouts (matching function), and when to perform transactions over already matched shouts.

Specialists must register at CAT server before the beginning of the game, in order to establish a communication channel with the rest of agents in the game.

A CAT game lasts a certain number of virtual days. A day last a certain number of rounds and a round last a number of milliseconds. These data is announced by the CAT server to all of the clients before the game begins.

Before the game start, the Server freely offers the following information:

- Number of game days (NTD)
- Number of rounds per day(NR)
- Milliseconds per round
- Number of traders in the game (NTN)
- Number of specialists in the game (NTE)

At the beginning of the day, the specialist performs following activities:

- Establish policies
- Publish quotes

Along a trading day, the specialist must:

- Register traders
- Inform (to server or other traders)
- Analyze shouts
- Match shouts
- Execute transactions
- Charge quotes

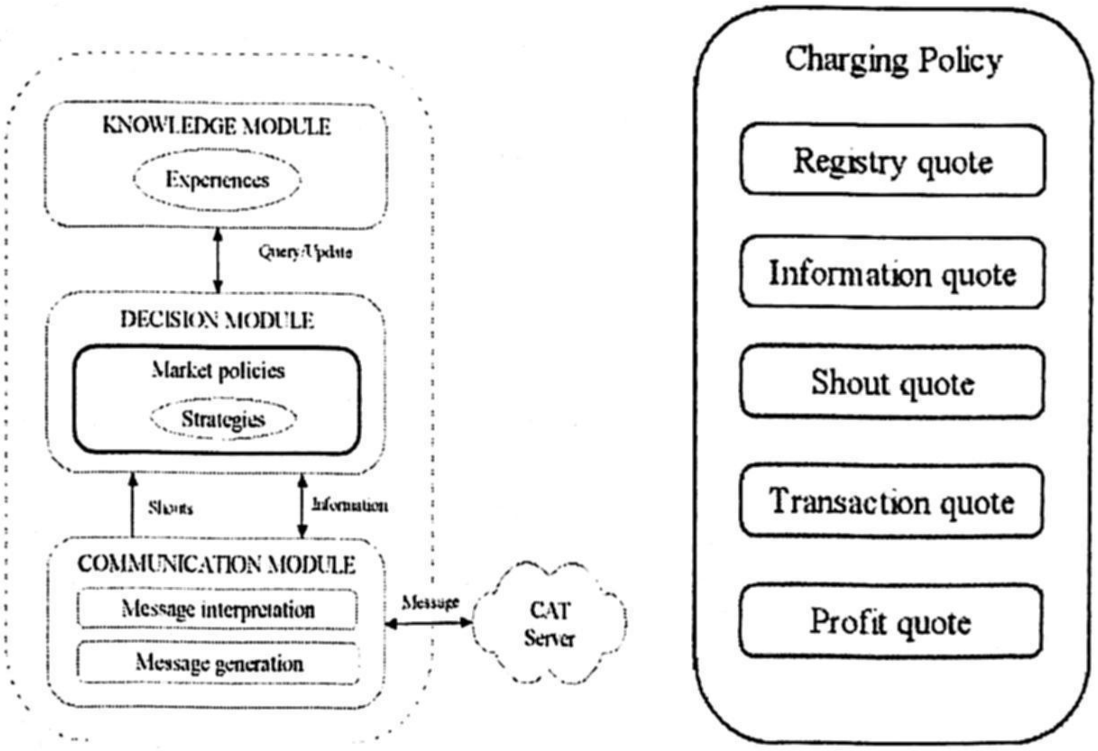
At the end of the day, the specialist must obtain the following information to be sent to the server:

- Number of registered traders in its market (CN)
- Number of matching achieved in its market (NE)
- Number of transactions realized in its market (NT)
- Profit

When game ends some duties have to be performed, firstly each agent's score, such as CN, NE, NT, are summed out and a winner is declared according to the highest scoring.

3 Design: Architecture and Strategies

The architecture of the proposed specialist agent (named Tianuani) in market strategies is shown in Figure 1.



(a) (b)

Figure 1. (a) Tianuani agent's architecture. This architecture comprises three main modules: knowledge, decision and communication [5]. (b) The charging policy. Quotes charged by the specialist.

Tianuani splits a trading day in three steps or states:

- Traders Capture State (TCS). Specialist begins the day in this state; this is the time when the traders pick its markets up. The goal of this phase is to get as much traders as possible.
- Profit Maximization State (PMS). After capturing a certain number of traders, specialist changes to EMU state, in which it tries to increase profit as much as possible.
- Transaction Maximization State (TMS). Once all transactions between the matched shouts with higher profits have done, then specialist changes to EMT, then it tries to match the higher number of shouts as possible no matter how much profit could be obtained.

Tianuani uses different strategies to establish these policies. Some of these strategies have already been tested in real stocks markets with human actors, and some others only represent theoretical models proposed by micro-economics researchers.

Based on which state the specialist is, it applies one of the next policies:

Accepting Policy. Use a combination of these strategies: *Self-beating accepting (AS)* [6] and *Equilibrium beating accepting (AE)* [7]. The paper [11] shows the design and implementation of this policy.

Charging Policy. Use a combination of these strategies: *Bait-and-switch charging (GB)*, *Charge-cutting charging (GC)* [6] and *Learn-or-lure-fast charging (GL)* [8]. This police will be explained in more detail in the next section.

Clearing Policy. Clearing is done after each round is finished. For shout matching the following strategies were used: *Equilibrium matching (ME)* [9, 10] and *Max-volume matching (MV)* [6].

Pricing Policy. Use *Discriminatory k-pricing (PD)* [6] and *N-pricing (PN) strategies* [7]. The paper [11] shows the design and implementation of this policy.

4 Tianuani's Charging Policy

It is necessary to establish a right quota value, because that'd be charged in order to get a profitable market. This is not only for increasing the utilities but also for increasing the number of traders negotiating in it. In particular, trader agents provided by TAC CAT, posses market selection strategies oriented to the quotes analysis. In the Microeconomics literature, diverse strategies for this policy can be found, some of them are mentioned below:

- *Fixed charging (GF).* This strategy establishes quotes in a specific fixed level [7].
- *Bait-and-switch charging (GB).* In this strategy, the specialist modify its quotes when it capture a certain number of traders, and then it increases them

slowly in order to increases its profit. The quotes only get down whether the number of traders is below a certain threshold [7].

- *Charge-cutting charging (GC).* This strategy establishes quotes based on the lowest charges imposed by the markets in previous days. *GC* is based on the fact of traders prefer markets with the lowest quotes [7].
- *Learn-or-lure-fast charging (GL).* This strategy adapts quotes based on objectives, following the schema used by ZIP trade strategy [7].

Based on the above strategies, a combination of GC, GB and GL strategies is proposed to be used by the specialist according to its current state.

At the beginning of a trading day, specialist is in TCS state, establishing quotes, the starting quota represents the lowest value from the day before. Once a certain number of traders are captured by the specialist, it increases quotes inside of a predefine interval.

After leaving TCS state, the specialist increases slowly quotes based on defined intervals at PMS state. The goal here is, maximize as possible, profit from operations carried out in PMS state. These intervals delimit quotes increasing.

When specialist is in TMS state, quotes can be diminished in order to motivate traders to increase number of bids. The goal, in TMS state, is to maximize the number of transactions when the clearing policy is applied.

5 Strategies Implementation for Quote Policy

This policy determines the different quotes charged by specialists. A description and implementation of the algorithms for the quote charges are provided for each one in this section.

5.1 Registry Quote (CR)

Specialist charges a quote or fee when traders want to register in his market. The quote is a fixed fee previously established by the specialist. Three intervals based on the number of registered traders (C_N) are defined to control increment/decrement of this quote. The number of desirable traders (N_D) is given in expression (1).

$$N_D=|N_{TN}*0.70|/N_{TE}. \tag{1}$$

Where:

N_{TN}: Total number of traders in the game
N_{TE}: Total number of specialists in the game

Intervals are defined as follows:

Small register interval (*irp*): $C_N \in [0,0.25*N_D]$
Medium register interval (*irm*): $C_N \in [0.25*N_D, 0.5*N_D]$
Large register interval (*irg*): $C_N \in [0.5*N_D, N_D]$

Where 0.25 and 0.5 represents the percent of desirable traders

5.2 Information Quote (CI)

Traders and specialists agents can request information to any other specialist participating in the contest. A trader has to subscribe and pay an information fee in order to get information from a specialist (number of matchings, number of traders, number of transactions). The maximum number of subscribers N_{TS} *is defined in expression (2).*

$$N_{TS}=(N_{TN}-C_N)+(N_{TE}-1) \qquad (2)$$

In the same way as for registry quote, three intervals based on the percent of subscribers (C_S) managed by the market are defined.

Small register interval (*isp*):	$C_S \in[0,0.25*N_{TS}]$
Medium register interval (*ism*):	$C_S \in[0.25*N_{TS}, 0.5*N_{TS}]$
Large register interval (*isg*):	$C_S \in[0.5*N_{TS}, N_{TS}]$

5.3 Shout Quote (CO)

When a specialist accepts a shout from a trader, a shout fee is charged to this trader. The number of shouts received during the day up to the moment (N_O), is calculated by the expression (3).

$$N_O=(N_E*2)+L_b.size+ L_a.size. \qquad (3)$$

Where:

N_E Number of matching shouts
L_b: List of unmatched bids
L_a: List of unmatched asks
$L_b.size$: Size of L_b
$L_a.size$: Size of L_a

The minimum number of shouts (N_{MO}) is defined by the expression (4)

$$N_{MO}=|C_N*0.25|. \qquad (4)$$

Number of shouts to be incremented (N_{IO}), by expression (5).

$$N_{IO}=|C_N*0.85|. \qquad (5)$$

5.4 Profit Quote (CB)

Once a transaction is done, specialist charges a profit's percentage obtained by traders working in this transaction.

Buyer agent's profit in a transaction *i* (pr_{bi}) is calculated by expression (6).

$$pr_{bi}=|v_{bi} - P_T| \quad (6)$$

Where:

v_{bi}: Bid price on transaction *i*
P_T: Transaction price established by specialist

The seller profit in a transaction *i* (pr_{si}) is calculated by expression (7).

$$pr_{si}=| P_T - v_{si} | \quad (7)$$

Where:

V_{si}: Ask price on transaction *i*

The number of desirable matching (N_{ED}) is defined by expression (8).

$$N_{ED}=|C_N*0.85/2|. \quad (8)$$

Also, three variables are defined, based on the percent of desirable matching, for the profit quote increment

$N_{EB}=0.25$
$N_{EM}=0.5$
$N_{EA}=0.75$

5.5 Transaction Quote (CT)

When a bid is matched with an ask, Specialist charges this quote CT. The fee is calculated by expression (9).

$$CT=|CB*0.1| \quad (9)$$

5.6 Algorithms

Before a round starts method *Analize_Quotes()* is executed , establishing intervals for each quote.

Variables:
C_D←Days counter along the game
C_R←Round counters during the day
R_E←Number of rounds in TCS, calculated by expression (10).

$$R_E=|N_R*0.25| \tag{10}$$

Alg. 1. Operations for each round [11].

```
While (¡round_end)
    Register_trader()
    Register_subscriber()
    Shout_Analysis()
    Inform()
$C_R$=$C_R$+1
If (State=PMS) then
    Match_PMS()
Else if (State=TMS) then
    Match_TMS()
If {(State=TCS)&&[( $C_N$>= $N_D$)|| ( $C_R$>= $R_E$)]} then
    If ($C_D$=0) then
          State=TMS
    Else
          State=PMS
Analize_quotes()
```

Alg. 2. : Analize_quotes().

```
Analize_Quotes()
// Registry quote
    If ($C_N \in irp$) then $C_R$=0
    If ($C_N \in irm$) then $C_R$= $C_R$
    If ($C_N \in irg$) then $C_R$= $C_R$+0.1
    If ($C_N$ >$N_D$) then $C_R$= $C_R$+0.5
//Information quote
    If ($C_S \in isp$) then CI=0
    If ($C_S \in ism$) then CI=CI+0.5
    If ($C_S \in isg$) then CI=CI+1
//Shout quote
    If (State=PMS) then
          If ($N_O$ <=$N_{MO}$) then CO= 0
          If ($N_O$ >=$N_{IO}$) then CO=CO+ 0.1
    If (State=TMS) then CO=0
//Profit quote
    If ($N_E$ <$N_{ED}$) then CB= $N_{EB}$
    If ($N_E$ =$N_{ED}$) then CB= $N_{EM}$
    Else CB= $N_{EA}$
```

Algorithms 3 to 7 for charging each of the quotes are shown below.

Alg. 3. Charge for registry quote.

```
Charge_cr (trader)
    Send_message(trader,CR)
    Receive_message()
    Profit=Profit+CR
```

Alg. 4. Charge for Information quote

```
Charge_ci (subscriber)
    Send_message (subscriber,CI)
    Receive_message()
    Profit=Profit+CI
```

Alg. 5. Charge for shout quote.

```
Charge_co(shout)
    Send_message(shout,CO)
    Receive_message()
    Profit=Profit+CO
```

Alg. 6. Charge for Transaction and profit quotes.

```
Charge_ct_cb (bid, ask, $P_T$)
    $v_{bi}$= bid.price
    $v_{si}$= ask.price
    $pr_{bi}=| v_{bi} - P_T|$
    $pr_{si}=| P_T - v_{si}|$
    bid_profit_fee = $pr_{bi}$*CB
    Cobrar_ct(bid, bid_profit_fee)
    Send_message(bid, bid_profit_fee)
    Receive_message()
    ask_profit_fee = $pr_{si}$*CB
    Charge_ct(ask, ask_profit_fee)
    Send_message(ask, ask_profit_fee)
    Receive_message()
    Profit=Profit+( bid_profit_fee + ask_profit_fee)
    $N_T$= $N_T$+1
```

Alg. 7. Charge for Transaction quote

```
Charge_ct (shout, profit_fee)
    Transaction_fee=| profit_fee *0.1|
    Send_message(shout, transaction_fee)
    Receive_message()
    Profit=Profit+ transaction_fee
```

In the next section experimental results using shout accepting and prices policies like is proposed in [11] and quote policies proposed in this paper are shown.

6 Experiments and Results

All the experiments were carried out and tested on JCAT platform using a local server [3].

6.1 Experimental Setup

Each experiment considered four specialists M1, M2, M3 (dummies) and Tianuani, the total number of traders are 100 (50 sellers and 50 buyers). The elapsed time was 100 days of game. Each day has 10 rounds and each round lasted 500 milliseconds. The graphics in section 6.2 represent the impact on the market of the proposed charging policy.

6.2 Results

Figure 2 shows profit obtained for each specialist, when they apply its quote policy.

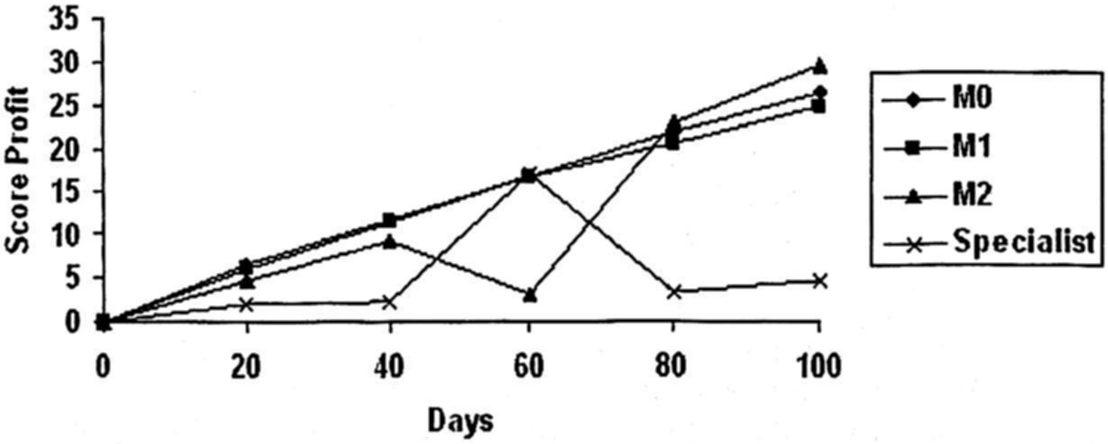


Fig. 2. Total Profit

Figure 2 shows that in the 20 first days of competition, the specialist remains more time in TCS state, therefore its profit keep constant and low. Between days 20 to 60 the specialist keeps itself more time in PMS state and its profit increases reaching the same profit of the other specialists (M0, M1 and M2). In the last days, Tianuani is more time in TMS state, it lost profit, since it had to maximize the number of matching by decrementing its quotes.

In Figure 3 transactions scores of each specialist is shown. It can be noted Tianuani from the 40 day beats the rest of the agents. This improved behavior is due to the decreasing of the quotes and then a greater number of shouts were captured.

In Figure 4, the trader's attraction for each specialist, in terms of the total percentage of registered traders in the game, is shown.

It can be noted Tianuani's behavior beats the remainder agents. From day 0 to 20 Tianuani keeps capturing traders in an increasing way and from day 40 until 100 it performs steady, outperforming in general the dummies'. Because of the quotes are decreased by Tianuani, a greater number of traders showed interest in to trade with it.

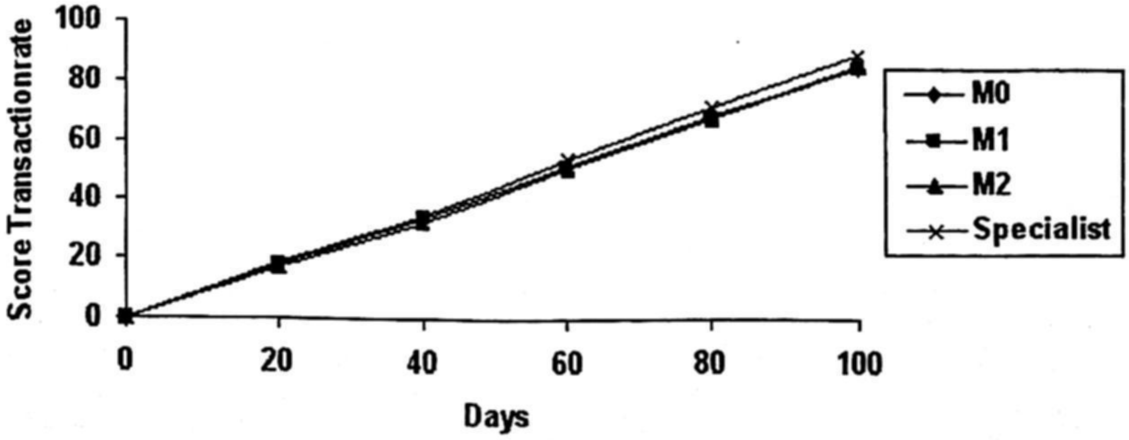


Fig. 3. Scores on transactions.

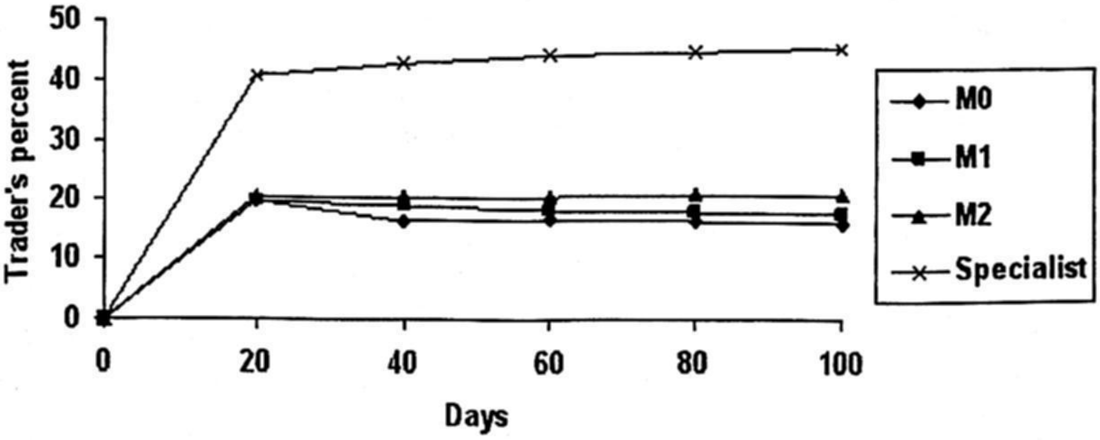


Fig. 4. Traders registered by specialist

7 Conclusions and Ongoing Work

Design and implementation of strategies were developed in order to establish charging policy. Tianuani obtains better number of transactions respect to the rest of the specialists (dummies). Considering performance based on how much traders Tianuani can attract, ours performed better than the reported results in [11], using both accepting and price policies at the same time.

Tianuani goes through diverse development steps. Initially, improvements for accepting and pricing policies were implemented. Results at this moment report a better trader's capture respect to the other specialists but the number of transactions was under the rest. Second improvement step consisted on to improve charging policy which positively impacted in a greater number of transactions and registered traders respect previous reports. Actually, an efficient clearing policy is under test given more capabilities to interact when all policies can be applied to a specific market situation.

Even though Tianuani enhanced its overall performance by improving its policies' strategies, it is still been updated. TAC CAT World Competition is coming soon and Tianuani will be playing and also train it with specialist agents from different countries.

References

1. Market Based Control, distributed resource allocation in complex computacional systems. CAT Tournament. Available : http://www.marketbasedcontrol.com/blog/index.php?page_id=5
2. TAC Trading Agent Competition. Homepage, http://www.sics.se/tac/page.php?id=1
3. Jinzhong Niu, Peter McBurney. CAT Document 002. JCAT: TAC/CAT Competition Platform. Version 1.04. University of Liverpool and Brooklyn Collage. June, 2008.
4. Jinzhong Niu. CAT Document 001 TAC Market Design; Communication Protocol Specification, Versión 1.19, University of Liverpool and Brooklyn College, June 12, 2008.
5. Salcedo M., Ramírez J.C., López F. Vilariño D., Tovar M. Development of an specialized agent on the TAC CAT Competition. 6th. National Conference on Computer Science 2008.
6. J. Niu, k. Cai, E. Gerding, P. McBurney, S. Parson. 2008. Characterizing Effective Auction Machanisms: Insights from the 2007 TAC Market Design Competition. Preecedings of the Seventh International Conference on Autonomous Agents and Multi-Agents Systems (AAMAS 2008). Estoril, Portugal. May 2008.
7. J. Niu, K. Cai, S. Parsons, and E. Sklar. Reducing price fluctuation in continuous double auctions through pricing policy and shout improvement rule. In Proceedings of the Fifth International Joint Conference on Autonomous Agents and Multiagent Systems, Hakodate, Japan, 2006.
8. D. Cliff and J. Bruten. Minimal-intelligence agents for bargaining behaviours in market-based environments. Technical report, Hewlett-Packard Research Laboratories, Bristol, England, 1997.
9. K. A. McCabe, S. J. Rassenti, and V. L. Smith. Designing a uniform price double auction. In Friedman and Rust [3], chapter 11, pages 307–332.
10. Wurman, P.; Wellman, M.; and Walsh, W. 1998. Flexible double auctions for electronic commerce: Theory and implementation. Decision Support Systems 24(1): 17-27.
11. Salcedo M., Vilariño D. et al. Accepting and Pricing Policies for a specialist agent in market strategies. Sending to the ESSAS 2009.

Some Experiments on a Geometrical Approach to Build Maps in Indoor Environment*

Luis M. Valentin-Coronado, Victor Ayala-Ramirez and Raul E. Sanchez-Yanez

Universidad de Guanajuato, DICIS, Guanajuato, México
luismvc@laviria.org, {ayalav, sanchezy}@salamanca.ugto.mx

Abstract. In this paper, we propose some experiments performed to model flat indoor environment using a mobile robot. Our approach uses straight lines as the geometric primitives to describe walls and relevant objects of the scenario under exploration. We use sensor data acquired by a laser range finder. We analyze several cases that could arise in the interpretation of sensor information and its importance for the building task.

1 Introduction

Sensor-based exploration enables a mobile robot to explore an unknown environment and build the map of that environment. That is the reason why techniques for robot map building has been a highly active research area in robotics. Robot map building is based on acquiring a set of data by the sensors and then integrate them into a representation of the environment. The mapping problem is generally regarded as one of the most important problems in the pursuit of building truly autonomous mobile robot [1]. There are basically three types of maps. The Occupancy Grid Maps, first introduced by Moravec and Elfes [2], have been widely used [3], [4], [5]. The occupancy grid maps represent the environment as a two dimensional array of cells, each of which indicates the probability of being occupied. The occupancy grid maps are considerably easy to construct and to maintain even in a large scale environments [6], [7]. Since the intrinsic geometry of a grid corresponds directly to the geometry of the environment, the robot's position can be determined by its position and orientation. On the other hand grid-based approach suffer from their space and time complexity. This is because the resolution of a grid must be fine enough to capture every important detail of the world.

Topological maps are an abstract and compact representation of the environment that captures key places and their connectivity for localisation and navigation [8]. Topological maps represent the environment as a list of significant places, called nodes, connected by arcs. The topological approach in contrast with the grid-based, determines the position of the robot relative to the model based on landmarks. Topological approach often have difficulty determining if two places

* Partial funding from project *Herramientas Mecatrónicas para la Implementación de Entornos Virtuales*, Concyteg GTO-2005- C04-18605.

© G. Sidorov, M. Aldape, M. Martínez, S. Torres. (Eds.)
Advances in Computer Science and Engineering.
Research in Computing Science 42, 2009, pp. 213-223

Received 16/03/09
Accepted 27/04/09
Final version 07/05/09

that look alike are the same or not. The advantage for such a representation is its compactness and its potentiality. Examples of topological approaches include the work by Mataric [9] and Gasca-Martínez [10].
Geometric feature based map are build from measurements acquired by a sensor at different instants of time, such as a laser range finder, thus generating a perceptual model defined by line segments. In this paper we propose a geometric map based on line segments as a way to reduce the size of data structure storage. We can then use it in efficient path planning and localization process.
Paper organization is as follows: In Section 2 we pose the problem addressed by this paper and the main elements of our approach. We show in Section 3 the test we have carried out to model on indoor environment and the discussion of the results we have obtained. Paper is finalized by stating our main findings and the work to be done (Section 4).

2 Problem Formulation

The general problem we want to solve is to let a mobile robot explore an unknown environment using its laser range finder and build a map of the environment. Consider an indoor environment as shown in Figure 1, if we place the robot in a position (x, y, θ), we can get the representation of the environment sensed by the robot, based on a feature extraction algorithm. We propose to model such an environment using a geometric map composed of line segments. With this kind of representation, a model of the environment will consist of a sequence of lines. Each line is described by six parameters: its slope-intercept equation in global coordinates and the (x, y) coordinates of the beginning and the end points of the known line segment. In Figure 2, we can see the different processes needed for this approach.

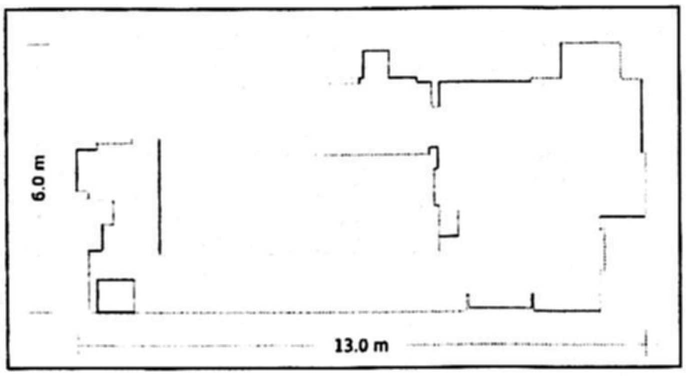


Fig. 1. Geometric map of an environment.

Robot is assumed to be at an initial unknown position (x, y) and with an also unknown orientation θ. Data acquisition is performed by using a laser range

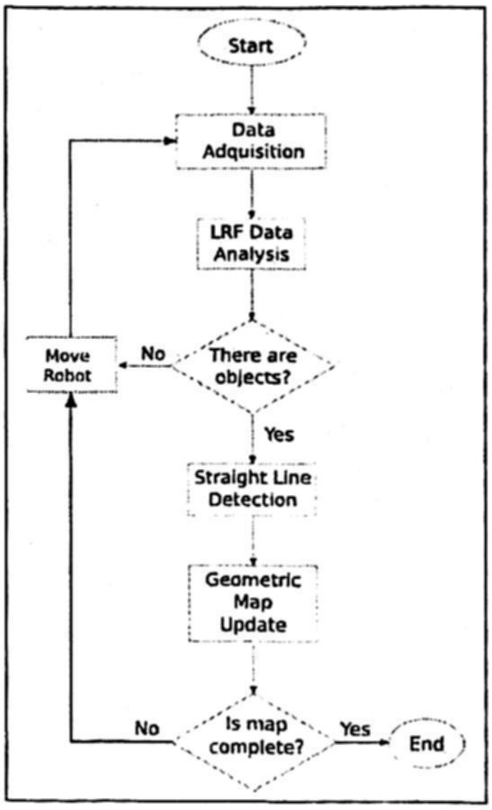


Fig. 2. Information flow through the proposed system.

finder. The acquired laser range finder data is analysed to know if there is relevant objects in the neighborhood of the robot. It is well known that the laser range finder detects objects robustly when they are in a range lower than 10 meters. So we analyze the readings of the laser range finder to detect if there is any object inside this area defined by this range and the current orientation of the robot.

If the robot does not detect any interesting object to start building a map, it changes its configuration to inspect another region of its surrounding environment. Firstly, our strategy is to cover all possible orientations of the robot in its current position. If no object is detected when this operation has been completed, the robot moves forward in the last orientation that has been inspected. Alternatively , we could use free space criteria to decide in which direction the robot has to move forward. This motion is executed until a given time interval is elapsed or until an object comes into the region of the robot, whatever occurs first. If no object has been detected at the end of the forward motion, the motion command is re-issued. Rotation of the robot is only executed in the initialization step, this helps the robot to find near walls not detected in their current orientation.

When the robot detect relevant objects, a straight line detection module tries to identify the walls in the environment under exploration. To do so, we detect the sensor data clusters originated by all the walls in the detection range of the

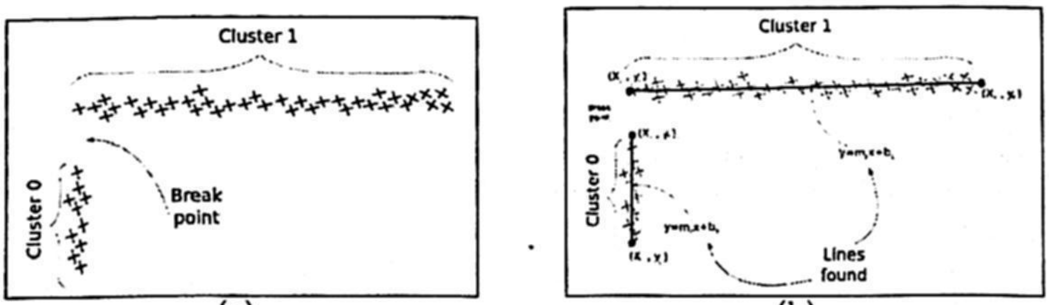
robot. They are then analyzed to fit a straight line model in global coordinates to represent the wall in the map representation.
A geometric map update module incorporates new line models in the current map. The information fusion step considers the motion of the robot in the global coordinate system and it tries to refine the wall description by analyzing line overlapping, parameter similarity, etc.

2.1 Clusters Detection

The clustering process refer to the process of classifying data by groups based on the calculation of a minimum distance that must exist between two consecutive points to consider that they must belong to different clusters. In this work we use an adaptive clustering method proposed by Borges *et al.* [11].
The clustering process is a very important step because we detect the number of groups without break points (see Figure 3(a)), based on a threshold distance parameter, which allows us to determine the different lines accurately.

2.2 Wall Modelling

Once we have all the cluster from the data, we use the *Iterative End Point Fit (IEPF)* algorithm [12] in order to calculate the existing lines associated to the walls of the environment. This step let us to compute the line model only using points that actually belong to a single wall. As a result, we obtain the six parameters of the line that best fits the data cluster (see Figure 3(b)).



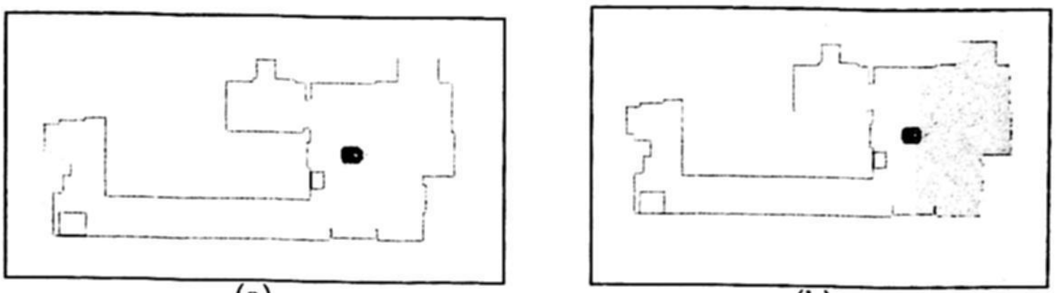
(a) (b)

Fig. 3. Clusters and breakpoints (a). Line found (b).

3 Experiments and Results

All the experiments were conducted both on the XidooBot robot, a LRF-equipped P3AT robotic platform and in the robot simulator provided by its manufacturer (see Figure 4). We have evaluated our qualitative approach by letting the robot to explore the environment in Figure 1. From a starting position (x, y, θ) (see Figure 5(a)) the robot acquired range information using its laser range finder (shown in Figure 5(b)).

Fig. 4. XidooBot.



(a)

(b)

Fig. 5. Starting position (x, y, θ) (a) and the measurement acquisition (b).

The actual data readings are shown in Figure 7. The laser sensor of our robotics platform acquires 181 equally spacial angular readings for angles from −90° to 90° with respect to the instantaneous robot orientation (see Figure 6 for definition of the robot-centered coordinate system).

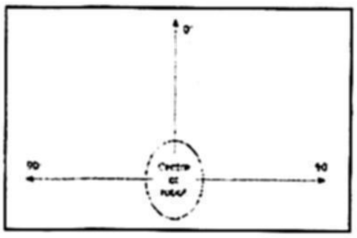


Fig. 6. Robot-centered coordinate system.

Figure 8 shows the decomposition that the clustering algorithm obtains for sensor data in Figure 7. As it is shown, three clusters are identified. Clusters identification depends in the threshold parameter of the Borges *et al.* [11] algorithm.

For each cluster, we apply the *IEPF* algorithm to estimate the parameters of all

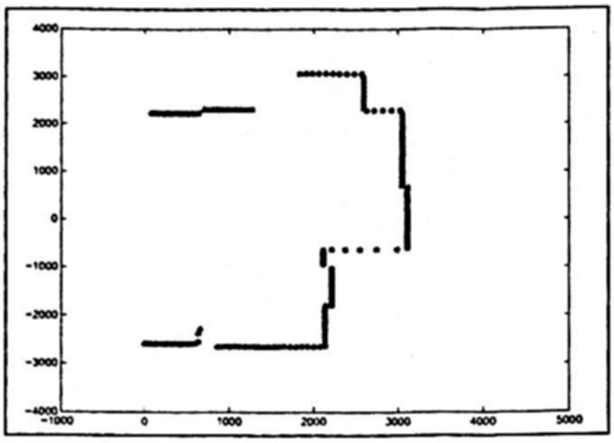
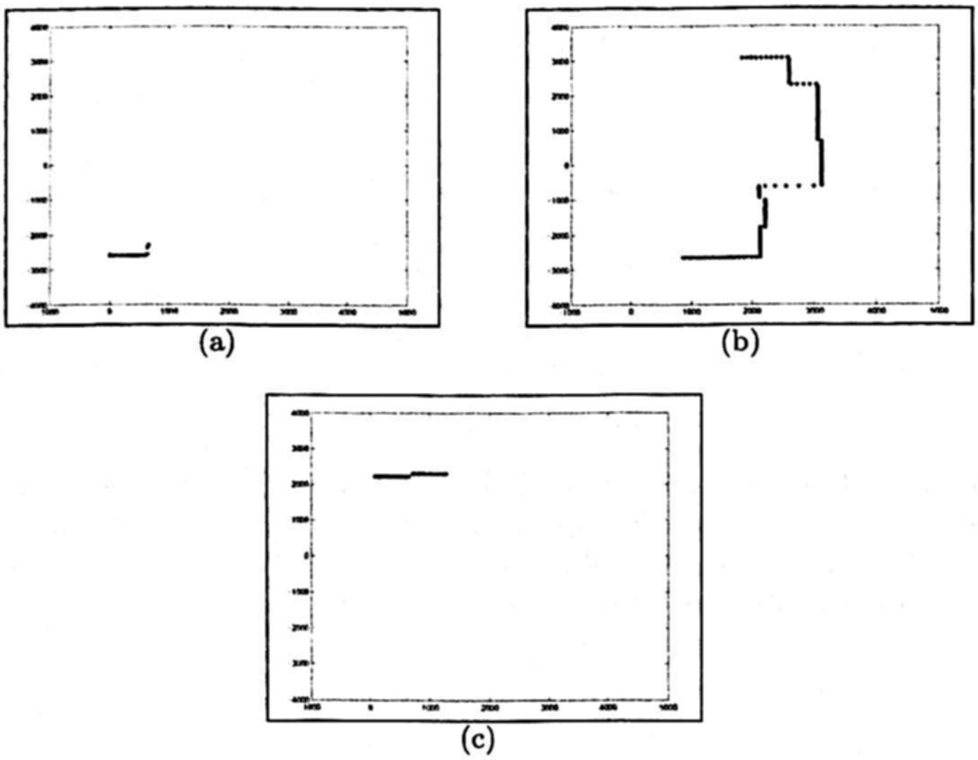


Fig. 7. Data acquired by the robot.



(a)

(b)

(c)

Fig. 8. Data clusters identified for data in Figure 7. Data cluster 1 (a). Data cluseter 2 (b). Data cluseter 3 (c).

the straight line present in the cluster under analysis.

In Figure 9 we represent the lines obtained when the above procedure has been applied. An overlay of the actual sensor data and the wall models is shown for comparative purposes in Figure 10.

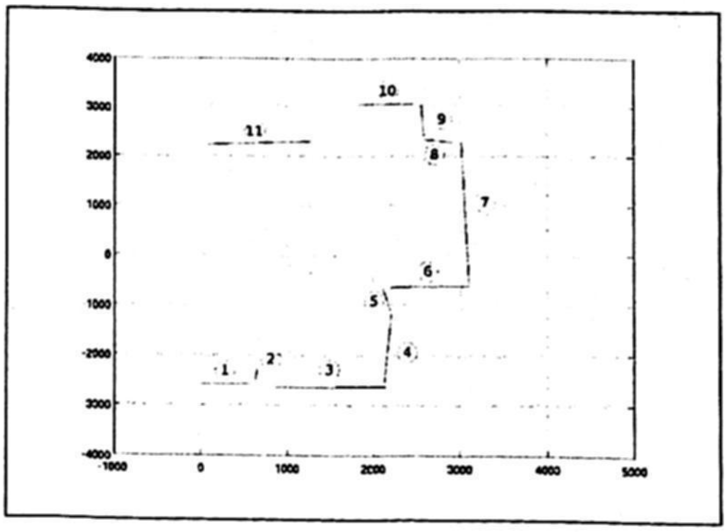


Fig. 9. Line model obtained from sensor data in Figure 7.

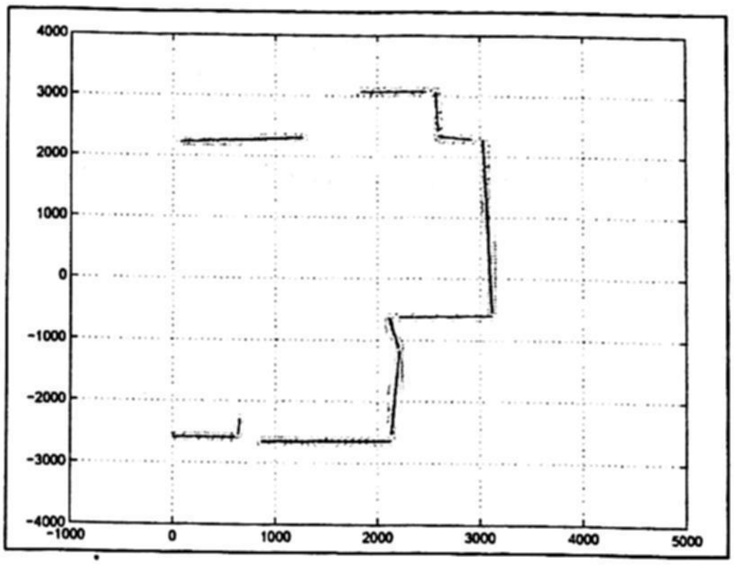
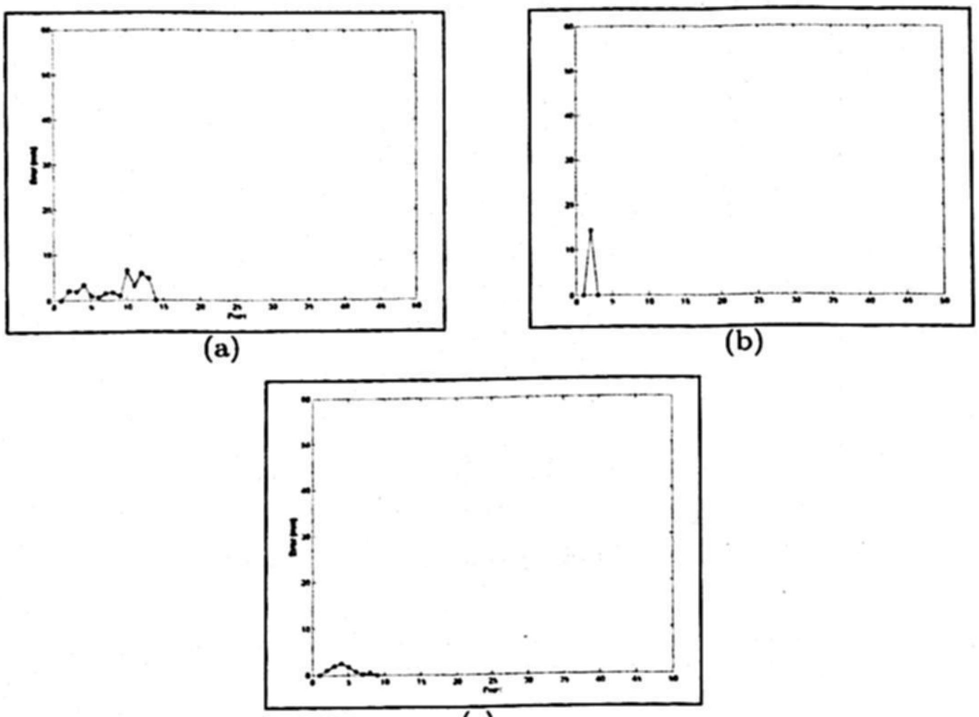


Fig. 10. Comparison between actual sensor data and the scenario model obtained.

3.1 Results Analysis

For each line detected we computed the modeling error as the euclidean distance between each point detected to the associated line (see Equation 1). From this information we have identified different behaviors of the error function that could arise when modelling on environment. Lines 1,2 and 10 present a similar behavior with regard to the modeling error function. They present a small error magnitude explained because they are short line segments. Figure 11 depicts this behavior.

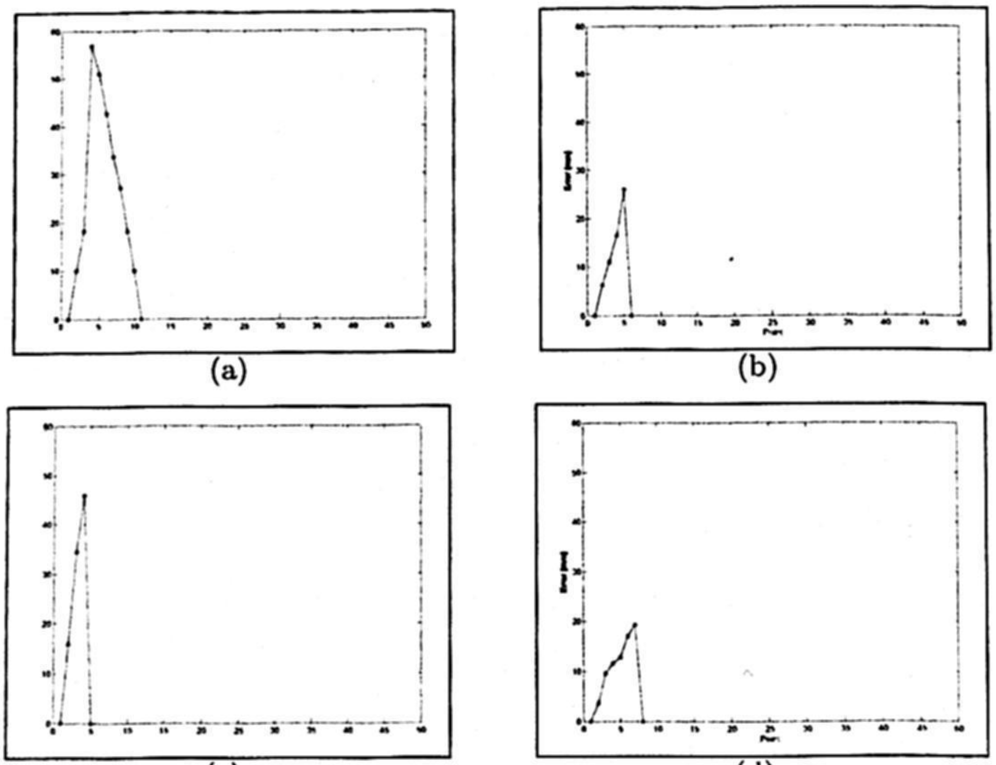


(a)

(b)

(c)

Fig. 11. Lines with a small modeling error. Line 1 error (a). Line 2 error (b). Line 10 error (c).



(a)

(b)

(c)

(d)

Fig. 12. Lines where end has not been detected accurately. Line 5 error (a). Line 6 error (b). Line 8 error (c). Line 9 error (d).

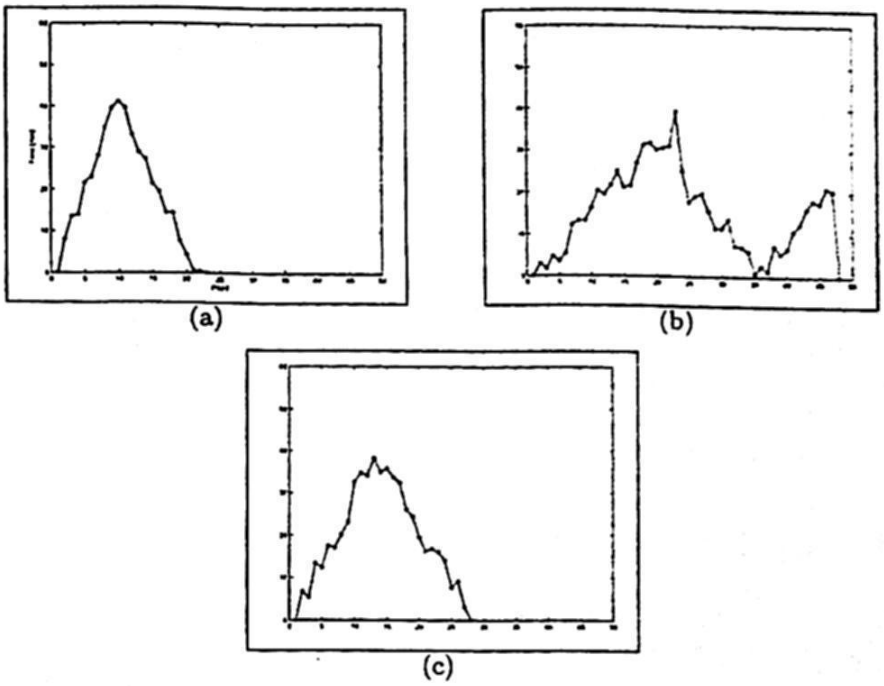


Fig. 13. Lines with an error trough all the line segment. Line 4 error (a). Line 7 error (b). Line 11 error (c).

$$Error_i = \frac{|m_j x_i + (-1)y_i + b_j|}{\sqrt{m_j^2 + 1}} \tag{1}$$

with:

x_i, y_i being the data coordinates associated to the j-th line
m_j, b_j being the slope-intercept parameters to the j-th line

A second group of lines (lines 5,6,8, and 9) present a modeling error significantly larger in one end of the line segment than in the other. This behavior is caused by a bias in the detection of the end where the larger error occurs. We show this in Figure 12.

Figure 13 depicts a third behavior. In the group of lines including lines 4,7 and 11, there is an error through all the line segment. That is originated because two wall were merged into a single line model. Even if this seems to be an inconvenient of our method, we can reduce the problem by increasing the *IEPF* algorithm sensitivity.

Figure 14 shows line 3 modeling error. This behavior is different from the other three groups. Error is not significant but it is present all along the line segment. In fact, it is the longest segment that has been modeled using a single straight line. That is the behavior that we could expect if we explore a typical scenario in indoor robotics. For all the above cases, error modeling for the maximal modeling

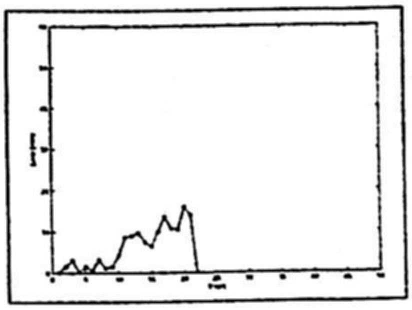


Fig. 14. Expected typical behavior modeling-error in indor environments (line 3).

deviation is 56.86 mm.

4 Conclusion and Perspectives

We have presented a geometrical approach to map building of flat indoor environment. Our method is based in the interpretation of laser sensor reading and its modeling by six-parameters straight lines model. We have detailed the elements and the algorithms used for this task. We have also analyzed our results and we have grouped the behavior of the line composing the model or our scenario. The approach works well as we have shown qualitatively in our paper. Future work will included an analysis of the modeling error when the robot executes its navigation in larger environments.

References

1. Thrun, S.: Robotic mapping: A survey. In Lakemeyer, G., Nebel, B., eds.: Exploring Artificial Intelligence in the New Millenium. Morgan Kaufmann (2002)
2. Moravec, H., Elfes, A.E.: High resolution maps from wide angle sonar. In: Proceedings of the 1985 IEEE International Conference on Robotics and Automation. (1985) 116 – 121
3. Brunskill, E., Roy, N.: SLAM using incremental probabilistic pca and dimensionality reduction. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Barcelona, Spain (2005)
4. Zalma, E., Candela, G., Jaime, G., Thrun, S.: Concurrent mapping and localization for mobile robots with segmented local maps. In: Proceedings of the IEEE International Conference on Intelligent robots and Systems (IROS). (2002) 546 – 551
5. Yamauchi, B., Schultz, A., Adams, W.: Mobile robot exploration and map-building with continuous localization. In: In Proceedings of the 1998 IEEE/RSJ International Conference on Robotics and Automation. (1998) 3175 – 3720
6. Buhmann, J., Burgard, W., Cremers, A., Fox, D., Hofmann, T., Schneider, F., Strikos, J., Thrun, S.: The mobile robot Rhino (1995)
7. Thrun, S., Buecken, A.: Learning maps for indoor mobile robot navigation. Technical Report CMU-CS-96-121, Computer Science Department, Pittsburgh, PA (1996)

8. Werner, F., Gretton, C., Maire, F., Sitte, J.: Induction of topological environment maps from sequences of visual places. In: Proceedings of the IEEE International Conference on Intelligent Robots and Systems (IROS), Nice, France (2008) 22 – 26
9. Mataric, M.: A distributed model for mobile robot environment-learning and navigation. Technical Report TR-1228, MIT (1990)
10. Gasca-Martínez: Navegación topológica para robots autónomos en escenarios virtuales. Master thesis, Universidad de Guanajuato (2008)
11. Borges, G.A., Aldon, M.J.: Line extraction in 2D range images for mobile robotics. J. Intell. Robotics Syst. **40** (2004) 267–297
12. Duda, R.O., Hart, P.E. In: Pattern Classification and Scene Analysis. John Wiley and Sons (1973) 98–105

GEMPA: Graphic Environment for Motion Planning Algorithms

Antonio Benitez and Alejandro Mugarte

Department of Informatics Engineering
Universidad Politécnica de Puebla
Tercer Carril del Ejido "Serrano"
San Mateo Cuanalá s/n. 72820 México
antonio.benitezruiz@gmail.com
http://geocities.com/antonio.benitezruiz

Abstract. This paper describes a graphic tool to simulate virtual environments involved in three dimensional space. GEMPA (Graphic Environment for Motion Planning Algorithms) was developed to be used to simulates the behavior of motion planning algorithms. Therefore, this tool includes the object's representation using different kind of files; the animation of free collision path for a motion planning problem, and graphic user interface to navigate through the virtual environment. Important elements to give more realism to the environments as illumination and textures are considered. Initially, this tool can be used as a didactics software for computer graphics and animation courses showing how transformation on three dimensional space can be used to animate the object movement, but the main application is the simulation of motion planning algorithms.

Key words: 3-D graphic environment, simulation and visualization tool.

1 Introduction

Computer graphics has advanced to a point where generating images of striking realism and complexity has become almost commonplace. However, making objects move convincingly within these pictures remains difficult, particularly as object models grow increasingly complex. The specification and control of motion for computer animation has emerged as one of the principal areas of research within the computer graphics community.

Computer graphics has grown phenomenally in recent decades, progressing from simple 2-D graphics to complex, high-quality, three-dimensional environments. In entertainment, computer graphics is used extensively in movies and computer games. Animated movies are increasingly being made entirely with computers. Even nonanimated movies depend heavily on computer graphics to develop special effects. The capabilities of computer graphics in personal computers and home game consoles have now improved to the extent that low-cost systems are able to display millions of polygons per second.

© G. Sidorov, M. Aldape, M. Martinez, S. Torres. (Eds.)
Advances in Computer Science and Engineering.
Research in Computing Science 42, 2009, pp. 225-236

Received 07/03/09
Accepted 27/04/09
Final version 05/05/09

There are also significant uses of computer graphics in nonentertainment applications. For example, virtual reality systems are often used in training. Computer graphics is an indispensable tool for scientific visualization and for computer-aided design (CAD).We need good methods for displaying large data sets comprehensibly and for showing the results of large-scale scientific simulations. Over the last few years, many different systems have been developed to represent and simulate scenarios with different kind of objects in virtual environments.

The representation of different environments in such a system is used for a widely researched area, where many different types of problems are addressed, related to animation, interaction, and motion planning algorithms to name a few research topics. Although there is a variety of systems available with many different features, we are still a long way from a completely integrated system that is adaptable for many types of applications.

This motivates us to create and build a visualization tool for planners capable of using physics-based models to generate realistic-looking motions. The main objective is to have a solid platform to create and develop algorithms for motion planning methods that can be launched into a digital environment. The developed of these tools allows to modify or to adapt the visualization tool for different kind of problems.

This paper is organized as follows: First we present the support architecture for GEMPA including different kind of files used to recover information about the objects inside the environment(Section 2). Then we review the graphic user interface (GUI) to navigate trough the virtual graphic environment, Section 3. After we describe how GEMPA is used to simulate motion planning algorithms and to build simple environments in Section 4, we continue in Section 5 to present interesting results for 3-D objects transformations and simulation of motion planning algorithms. We conclude the paper and describe work in progress in Section 6.

2 GEMPA Architecture

GEMPA architecture is supported by necessary elements to represent objects, geometric transformation tools and visualization controls. These elements are integrated to reach initial goals of visualization and animation applied to motion planning problems.

2.1 Related Work

Important graphics tools has been developed to visualize and understand motion plannig problems. As an example, the description of two visualization tools will be described; VIZMO++ and MPK.

VIZMO++. VIZMO++ (A visualization/authoring tool for motion planning) is a 3D visualization/authoring tool for files provided/generated by OBPRM motion planning library [2]. VIZMO++, was developed for visualizing and editing motion planning environments, problem instances, and their solutions. The tool offers a nice and easy to use graphical user interface (GUI) that allows you to display workspace environments, roadmap, path, and start/goal positions. It enables users to interact with and edit the environment. Application where VIZMO++ can be used are: User-Guided Path Planning ; Particle Transport Seismic Ray Tracing and Campus Navigator [1].

MPK - Motion Planning Kit. Motion Planning Kit (MPK) is a C++ library and toolkit for developing single- and multi-robot motion planners [3]. It includes SBL, a fast single-query probabilistic roadmap path planner [5] .

MPK can handle arbitrary kinematic tree structures and an arbitrary number of robots and obstacles at the same time. New robots with any combination of prismatic and revolute joints can be defined and added without recompiling. Some importants features of MPK are: C++-library and workbench for motion planning; Allows arbitrary number of robots and obstacles; New robots can be added without recompiling; Arbitrary kinematics for 'hardwired' robots; Efficient and exact dynamic collision checker.

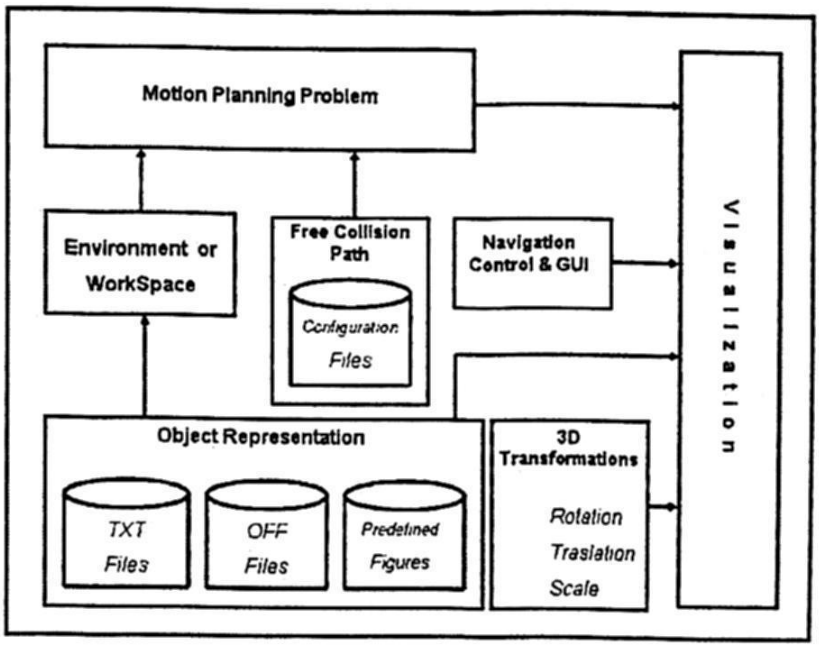


Fig. 1. Several modules conform the initial GEMPA architecture which offer interesting functionalities; visualization 3-D environments as well as animation of motion planning algorithms.

2.2 Recovering Objects Representation

People focus to solve problems using computer graphics, virtual reality and simulation of motion planning techniques used to recover information related to objects inside the environment through file which can storage information about triangle meshes. Hence, several objects can be placed on different positions and orientations to simulate a three-dimensional environment. There exist different formats to represent objects in three-dimensional spaces (3-D), however, two conventions used for many tools to represent triangle meshes are the most popular; objects based on *off - files* and objects based on *txt - files*. In motion planning community there exist benchmarks represented through this kind of files. GEMPA is able to load the triangle meshes used to represent objects from *txt* or *off - files*.

On the other hand, GEMPA allows the user to built news environments using predefined figures as spheres, cones, cubes, etc. This figures are chosen from a menu and the user can place them using translation, rotation and scale transformations. Once the user has built the environment, it can be saved as a *txt - file*. Besides this files can be load from now on.

Each module on GEMPA architecture is presented in Figure 1. There, we can see that initially, the main goal is the visualization of 3-D environments and the animation of motion planning algorithms. In the case of visualization of 3-D environments, information is recovered form files and the user can navigate through the envorinment using mouse and keyboard controls. In the second case, the animation of motion planning algorithms, GEMPA needs information about a problem. This problem is described by two elements; the first one is called workspace, where obstacles (objects) and robot representation and configuration (position and orientation) is recovered from files; the second, a set of free collision configuration conform a path, this will be used to animate the robot movement from initial to goal configuration. Both goals are supported on 3-D geometric transformations.

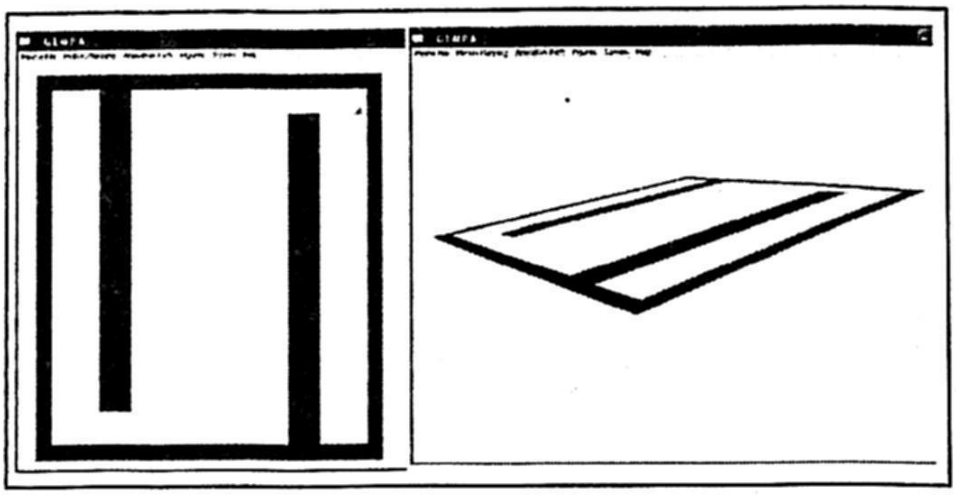


Fig. 2. Two different views of two-dimensional environment since the X-Y plane are painted.

3 GUI and Navigation Tools

Graphic user interface (GUI) is an important element for every computer system. This GUI become essential when application are used to simulate virtual environments, where navigation tools help the user to perceive realism and immersion characteristics. Initially, GEMPA has incorporated two modes to paint an object; wire mode and solid mode. Next, Lambert illumination is implemented to produce more realism, and finally transparency effects are used to visualize the objects.

Along the GUI, camera movements are added to facilitate the navigation inside the environment to display views from different locations.

3.1 2-D and 3-D Visualization

GEMPA was developed using 3-D transformations (translation, rotation and scale), that means that, if we want to display objects in 2-D, the user only have to place zero where *z - coordinate* or *gamma-angle* is required. Two dimension environment are painted on the *XY* plane. In Figure 2, two different views of the same environment are displayed on the *XY* plane.

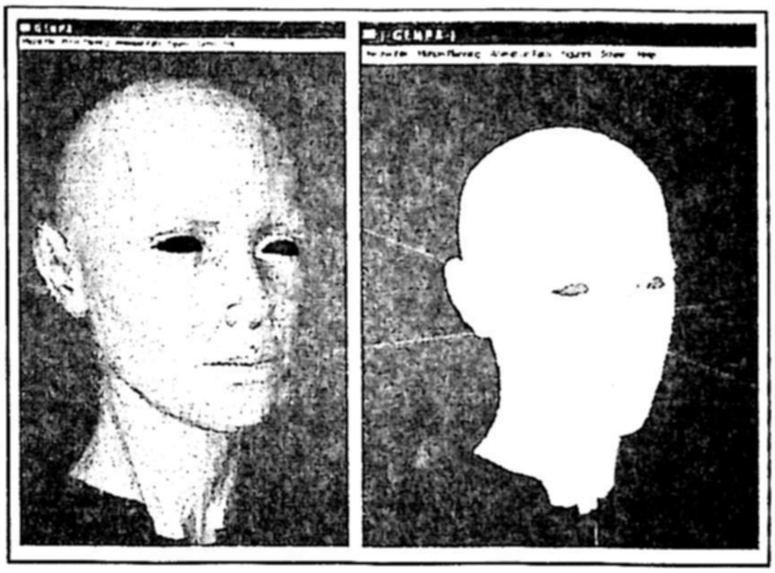


Fig. 3. In the left side, an object is painted using Lambert illumination , in the right side, transparency effect is applied on the object. Both features are used to give more realism the environment.

3.2 Navigation Controls and Realism Characteristics

This controls allow the user navigate through the environment. GEMPA support this navigation using keys and mouse movements. Hence, the tool include:

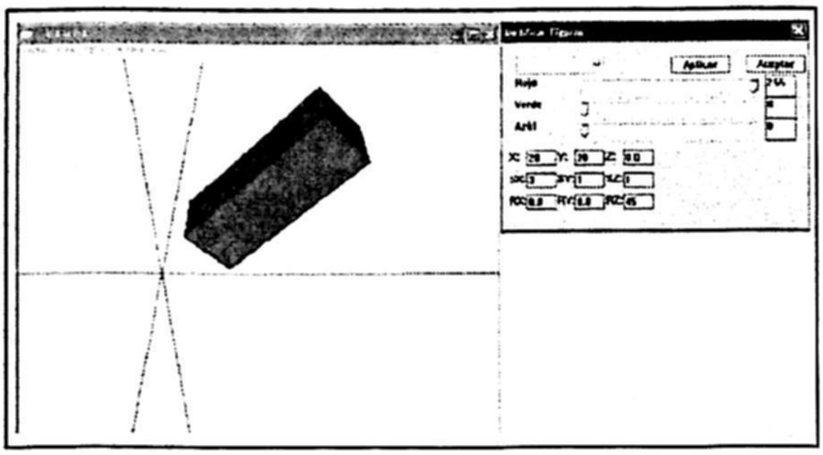
Camera movement on X-axis, Y-axis and Z-axis (zoom), and rotation around X,Y and Z axis. Both kind of controls are available since keyboard and mouse.

On the other hand, to give more realism the environment displayed by GEMPA, Lambert illumination has bee included along with transparency effect on the objects; colors for every predefined figure can be chosen. Besides, the tool is able to paint the axis to give the user information about the position where the observed is placed.

In Figure 3 we can see two samples where Lambert illumination and transparency effect are used to paint each object respectively.

4 Environments Construction and Motion Planning Algorithms

Even though GEMPA is able to read information about complex 3-D environments used to simulate motion planning algorithms [4, 7, 6], this tool besides allows the user to create or built his own environments.



X: 2.0 Y: 2.0 Z: 0.0
SX: 3 SY: 1 SZ: 1
RX: 0.0 RY: 0.0 RZ: 45

Fig. 4. The GUI includes a dialog box where position, orientation and color can be specified to create a figure.

4.1 Building Environments

GEMPA include the capability to incorporate objects with predefined figures as cubes, cones, spheres and torus. These figures are placed at the origin of a reference frame and the tool contemplate functionalities to apply geometric transformations to scale, move, or rotate the object, this can be seen in Figure 4, where a dialog box is provided to allow the user to change values for the position and orientation where every object will be placed inside the workspace. Therefore, an environment can be conformed by different objects distributes through the workspace, an example of such distribution is presented in Figure

5. It is important to say that, although figures used to build the environment seems simples, the user can construct more complex elements using repeatedly these objects.

Once an environment has been built, a function to save the environment as a *txt - files* is available. The file generated contains information about each object into the workspace and its position and orientation (*configuration*). This way, the user can reload the environment to be reused.

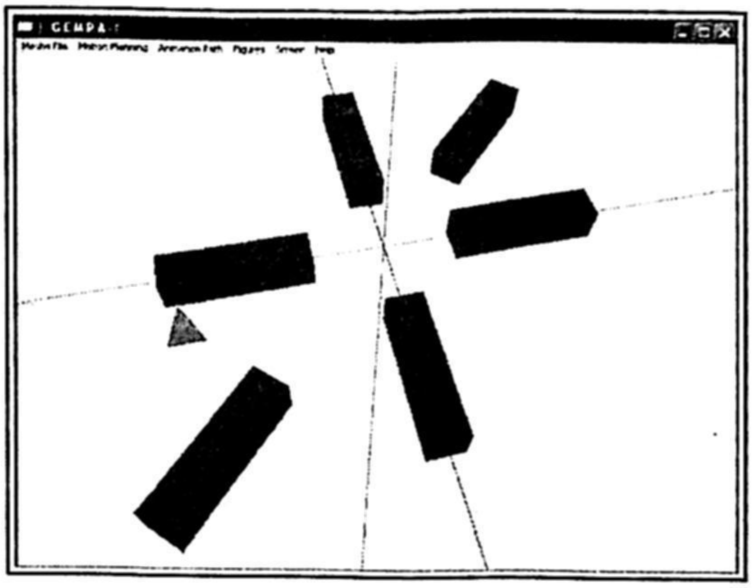


Fig. 5. The environment is built applying successive 3-D transformations on every figure to distribute them on the workspace.

4.2 Simulation of Motion Planning Problems

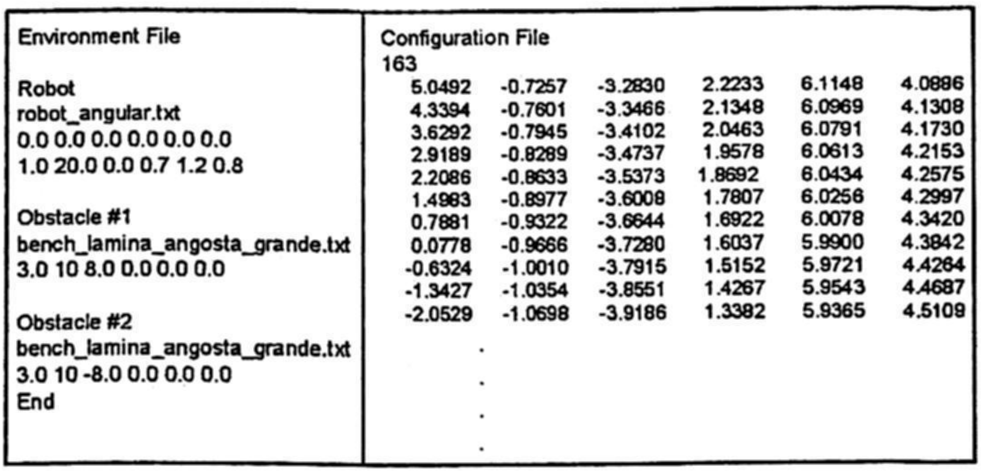
Motion planning methods are used in robotics to solve motion problems. Motion planning algorithms called probabilistic roadmap methods (PRM) are used to find a free collision path between a initial configuration and a goal configuration. There are many applications of motion planning algorithms. For this work, only PRM for free flying objects are considered as an initial application of GEMPA. Taking into account this assumption, the workspace is conformed by a set of obstacles (objects) distributed on the environment, these objects has movement restrictions, that mean that, the obstacles can not change their position inside the environment.

In addition, an object that can move through the workspace is added to the environment and is called robot. The robot can move through the workspace using the free collision path to move from the initial configuration to the goal configuration.

For PRM for free flying objects, only a robot can be defined and the workspace can include any obstacles as the problem need.

4.3 Simulating Free Collision Paths

PRMs generate as result a free collision path between init and goal configurations (if and only if the algorithm is able to find it). This path is a set of n configuration, where each configuration is represented as a six-tuple of parameters, which the three first parameters represent the position an the three last values represent the orientation. This free collision path is saved as a *configuration - file*, and the number of configurations will determinate how acute will be the animation. On the other hand, GEMPA include the capability to recover from an *environment - file* information about the position and orientation for each object inside a workspace including the robot configuration. Hence, GEMPA can draw each element to simulate the workspace associated. Therefore, initially GEMPA can recover information about the workspace, an example of this file can be see in Figure 6 (left side), *environment - file* include initial and goal configuration for the robot, beside to include x,y,z parameter for position and α, β, γ parameters for orientation for every objects inside the wokspace. Along whit this *environment file*, a *configuration - file* can also be loaded to generate the corresponding animation of the free collision path. This *configuration - file* has the form presented in Figure 6 (right side). This file is conformed by n six-tuples $(x, x, y, \alpha, \beta, \gamma)$ to represent each configuration included in the free collision path.



```
Environment File

Robot
robot_angular.txt
0.0 0.0 0.0 0.0 0.0 0.0
1.0 20.0 0.0 0.7 1.2 0.8

Obstacle #1
bench_lamina_angosta_grande.txt
3.0 10 8.0 0.0 0.0 0.0 0.0

Obstacle #2
bench_lamina_angosta_grande.txt
3.0 10 -8.0 0.0 0.0 0.0 0.0
End
```

```
Configuration File
163
   5.0492   -0.7257   -3.2830   2.2233   6.1148   4.0886
   4.3394   -0.7601   -3.3466   2.1348   6.0969   4.1308
   3.6292   -0.7945   -3.4102   2.0463   6.0791   4.1730
   2.9189   -0.8289   -3.4737   1.9578   6.0613   4.2153
   2.2086   -0.8633   -3.5373   1.8692   6.0434   4.2575
   1.4983   -0.8977   -3.6008   1.7807   6.0256   4.2997
   0.7881   -0.9322   -3.6644   1.6922   6.0078   4.3420
   0.0778   -0.9666   -3.7280   1.6037   5.9900   4.3842
  -0.6324   -1.0010   -3.7915   1.5152   5.9721   4.4264
  -1.3427   -1.0354   -3.8551   1.4267   5.9543   4.4687
  -2.0529   -1.0698   -3.9186   1.3382   5.9365   4.5109
               .
               .
               .
```

Fig. 6. In the left side, an example of environment file is showed, in the right side an example of configuration file is presented.

Once GEMPA has recovered information about workspace and collision free path, the tool allows the user to display the animation on three different modes.

Mode 1: Animation painting all configurations.
This modality paint every configuration from the free collision path. Each configuration is painted without clear the environment, that is, although visualization seems confused, all configurations are painted at the same time. An example of

this mode is presented in Figure 7 (left-side).

Mode 2: Animation painting configurations using a step control.
In these option, each configuration is painted after than previous configuration has been deleted from the environment, avoiding two configuration can be overlapped during the animation. Besides, under this molality, each configuration is painted manually when de user request it.

Mode 3: Animation using automatic step.
In this case, an automatic animation is displayed, showing one configuration at time and using an automatic step. This mode is the most used when the user need to generate a video of the simulation. An example of this mode can be seen in Figure 7 (right-side).

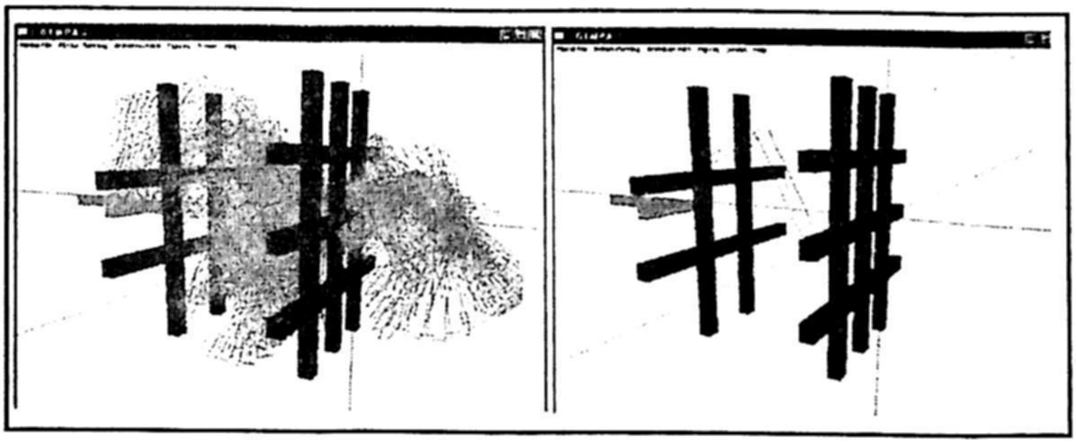


Fig. 7. Animation painting all configurations (left side), and animation using automatic step (right side) are displayed.

5 Applications

The proposed tool can be used to simulate and/or to implement later practical systems in different areas of computer science such as graphics, computational geometry and robotics. Four cases will be described as initial GEMPA applications.

Simple Figures and Geometric Transformations.
Initially, GEMPA has been an effective one in classroom teaching. It not only cuts down, significantly, on the instructor's time and effort but also motivates senior/graduate students to pursue work in this specific area of research. Important results has been reach to explain computer graphics topics, where geometric transformation are used to move, rotate or scale an object into a tree dimen-

sional environment. An example of this application can bee seen in Figure 4.

Simulation OFF Environments.
Important benchmark in motion planning are defined using *off - files*, hence the importance of managing and include this kind of files for objects representation. It is important to say that *off - files* not only allow represent a triangle meshes, this files also can use any polygon to represent an object. GEMPA provides the capability of building environments using these different polygons.
Figure 8 shows a combination of TXT and OFF files objects to build a three dimensional environment. **Simulation Motion Planning Algorithms.**

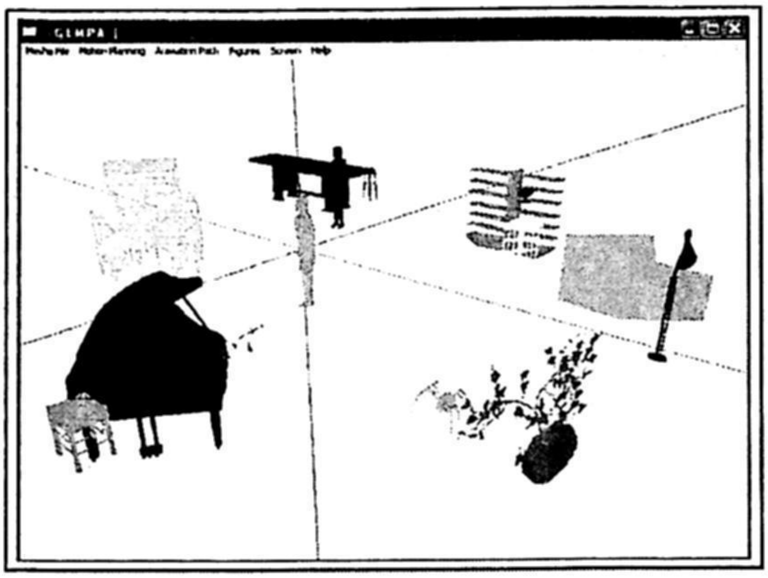


Fig. 8. Graphic environment built using OFF and TXT - Meshes Files.

A basic problem of PRM were described in section 4.2. Actually, GEMPA only can display animation for free flying objects problems defined on two-dimensional and three-dimensional spaces. Nevertheless, we are working to integrate different PRMs to GEMPA. In Figure 9 the free collision path to solve the problem is painted using the animation painting all configurations mode. In similar way, a 3-D sample of motion plannig problem is presented in Figure 10.

6 Conclusions and Work in Progress

GEMPA is a 3-D visualization tool for simulation environments and problems of PRMs (Probabilistic Roadmap Methods). This initial version of GEMPA was developed for visualizing and editing motion planning environments and animate collision free paths. GEMPA offers a nice and easy to use graphical user interface (GUI) that allows you to display workspace environments, path, and

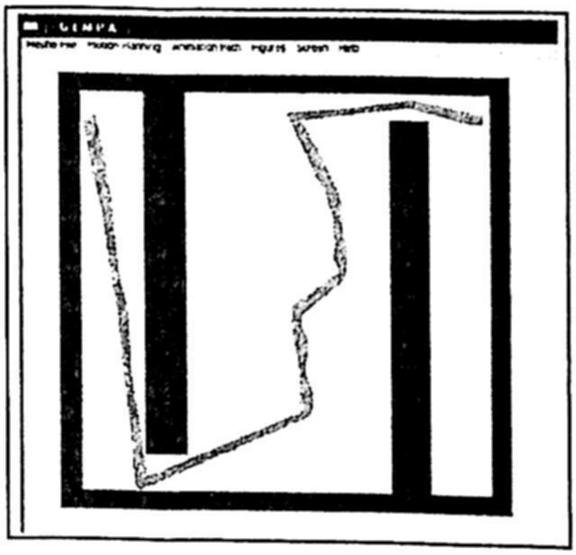


Fig. 9. A complete free collision path is painted as an example of two-dimensional motion planning problem.

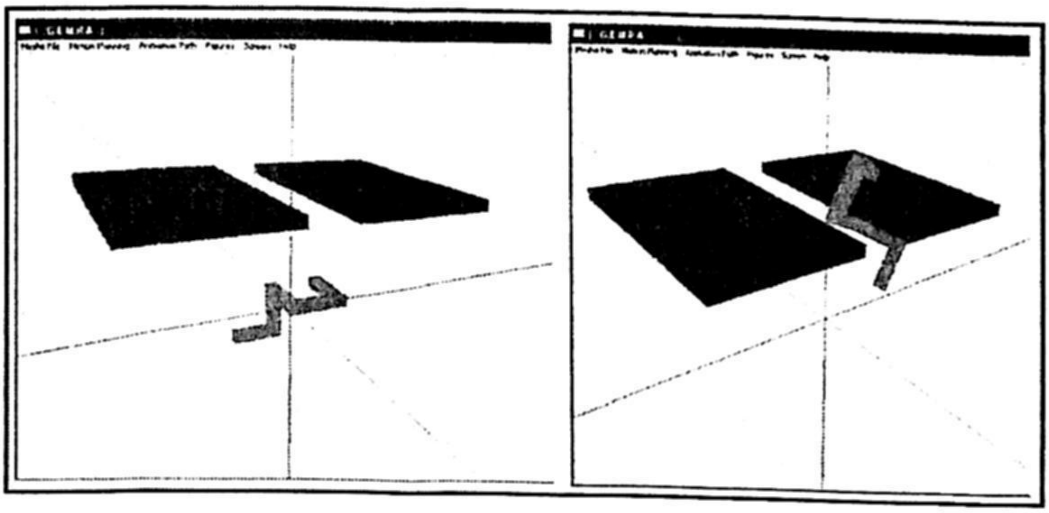


Fig. 10. 3-D Motion planning problem is displayed, where the collision free path is painted as an animation mode (one configuration at time).

start/goal positions. It enables users to interact with the environment. For example, it allows users manipulate objects configurations (or obstacles) and robot configurations, save new environments to be able to work on them later.

There many interesting a useful functionalities to be added to GEMPA. Now, this project is working on developed the follows goals.

Integration of PRMs to GEMPA. This will allows the user to generate solutions for motion planning problems selecting between at least two different techniques. Besides, GEMPA will be able to display the roadmap generated by the selected method. This functionality will enable users to run new queries.

Hence our tool will provides a convenient interface to select planners and set their parameters.

Simulation of Collision Detection Algorithms. Collision detection is a fundamental problem in robotics, computer animation, physically-based modeling, molecular modeling and computer-simulated environments. A realistic simulation system, which couples geometric modeling and physical prototyping, can provide a useful toolset for applications in robotics, CAD/CAM design, molecular modeling, manufacturing design simulations, etc. Due that motion planning algorithms used collision detection routines to solve the problems, GEMPA is developing a module capable to integrate a show the performance of collision detection algorithms.

Visualization and control for Kinematic Chains. Important application of robotics are related with articulated robots, these can be controlled using kinematic chains, applying forward and inverse kinematic to calculate them movements. Therefore, further GEMPA version is going to contemplate functionalities to manage this kind of robots, particularly forward and inverse kinematics for humanoids representation.

References

1. Jyh-Ming Lien Aime Vargas E. and Nancy M. Amato. Vizmo++: a visualization, authoring, and educational tool for motion planning. In *RIn Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, pages 727–732, 2005.
2. L. Dale C. Jones and N. Amato, B. Bayazit and D. Vallejo. Obprm: An obstacle-based prm for 3d workspaces. In L. Kavraki In P. Agarwal and M. Mason, editors, *Robotics: The Algorithmic Perspective*, pages 156–168. AK Peters, 1997.
3. M. Saha F. Schwarzer and J.-C. Latombe. Adaptive dynamic collision checking for single and multiple articulated robots in complex environments. In *IEEE Tr. on Robotics and Automation*, 2005.
4. L. Kavraki and J.C. Latombe. Randomized preprocessing of configuration space for path planning. In *IEEE Int. Conf. Robot. and Autom*, pages 2138–2139. xxx, 1994.
5. J.C. Latombe. *Robot Motion Planning.* Kluwer,Boston, MA, 1991.
6. Steven M. LaValle and James J. Kuffner. Rapidly-exploring random trees: Progress and prospects. In *International Workshop on Algorithmic Foundations of Robotics (WAFR).*, page xxxx, 2000.
7. R. Motwani L.E. Kavraki, J.C. Latombe and P. Raghavan. Randomized preprocessing of configuration space for path planning. In *Proc. ACM Symp. on Theory of Computing.*, pages 353–362, 1995.

Virtual Human Locomotion Synthesis: A Survey

Xochitl Hernández V., Abraham Sánchez L. and David Nuñez R.

Facultad de Ciencias de la Computación, BUAP
14 Sur esq. San Claudio, CP 72570
Puebla, Pue., México
hevexo3@hotmail.com, asanchez@cs.buap.mx, dnunezr@yahoo.com.mx

Abstract. The ideal animation engine is the one that exactly matches the human motions, with its subtle and unpredictable changes and variations. Virtual characters need to navigate in order to interact with their environment. The autonomy is ensured by locomotion controllers to accomplish navigation tasks. This paper surveys the set of techniques developed in Computer Graphics for virtual human locomotion synthesis. We review the recent advances in this field by presenting some of them in our multipurpose tool developed for research and teaching.

1 Introduction

The synthesis of realistic human motion is a challenging research problem with broad applications in movies, special effects, cartoons, virtual environments and games. Due to the quality and realism of the result, the use of motion captured data has become a popular and effective means of animating human figures. Since it is an inherently off-line process, there has been great interest in developing algorithms that are suitable for interactive applications [1].

Traditionally, human animation has been separated into facial animation and body animation, mainly because of lower level considerations. Facial animation results primarily from deformations of the face. Controlling body motion generally involves animating a *skeleton*, a connected set of segments corresponding to limbs and joints. Using geometric techniques, animators control the skeleton locally and define it in terms of coordinates, angles, velocities, or accelerations. The simplest approach is motion capture.

We introduce a multipurpose tool for the animation of virtual characters (MAVIC) as an available simulation platform for research and teaching. MAVIC is an extensible simulation framework and rapid prototyping environment for computer animation.

MAVIC offers solutions to the problems of flexible reuse of controllers, actuators and interactive participation in 3D animation. In addition, MAVIC provides basic modeling capabilities and support for kinematic animation such as motion capture and keyframing. In this first version, the tool includes some elements for the support of physical simulation of virtual humans.

The paper is organized as follows. Section II presents the state of the art in human virtual animation. To create realistic animation, Section III describes the

© G. Sidorov, M. Aldape, M. Martinez, S. Torres. (Eds.)
Advances in Computer Science and Engineering.
Research in Computing Science 42, 2009, pp. 237-248

Received 14/03/09
Accepted 27/04/09
Final version 05/05/09

two-stages locomotion planning as an effective new approach in this direction. Section IV presents the various modules integrated in our system from locomotion planning to the dynamic control. Finally, we give some concluding remarks and future work in Section V.

2 State of the Art in Character Animation

Existing motion synthesis techniques can be divided into three research directions: hand-driven, model-driven and data-driven methods [2], [3], [4], [5], [6], [7], [8].

The hand-driven methods are the oldest and simplest ways for animation creation. An animator determines manually postures ("keyframes") of an articulated character by defining positions and orientations on its joints at specific animation times ("key-times"). The final motion results in a smooth interpolation between the corresponding key-frames.

Model-driven methods explicitly describe how movements are generated, and use computation to propagate changes in high-level parameters to changes in low-level parameters. Therefore, these methods typically have a small number of high-level parameters which can be modified to change the generated motion. We can distinguish two classes of model-driven techniques, either based on kinematics or physics.

Motion capture technique offers an alternative source of highly realistic movement sequences used for a variety of data-driven motion synthesis methods. The motion acquisition systems allow the estimation of the joint positions and/or orientations of a real performer. These parameters are then adapted to a virtual character in order to result in an animation which reproduces the original motion accurately.

All presented methods, classified by hand-driven, model-driven and data-driven techniques have advantages and disadvantages. Illustrated in Figure 1, the different approaches are summarized according to two criteria: the motion realism and the motion control freedom provided to an animator. Hand-driven techniques allow the creation of realistic animations with a very important freedom for the animator to control the motion, constrained neither by visual realism nor physical accuracy. Actually, most of the 3D movies and games use keyframe animation sequences. However, this very flexible method has a price. It requires incredible investment of time and only skilled and talented designers produce realistic human motions. Model-driven approaches generate motions with less motion control freedom and realism. However, kinematic methods concentrate the control on a small set of high-level parameters, allowing an easier and more intuitive motion parameterization. In addition, the motion is generally created on-the-fly, but suffers from too robotic and jerky movements. The other part of model-driven methods, namely those based on physics, produce more realistic results but need lot of computational time (over several minutes). Moreover, for a valid resulting motion, physical accuracy does not imply visual realism. For example, a grasping movement can be simulated with various physically correct alternatives some of which may appear oddly and unnatural. Apart from these drawbacks, the motion control is driven by many parameters, difficult to directly interpret. Hence, data-driven methods have been developed.

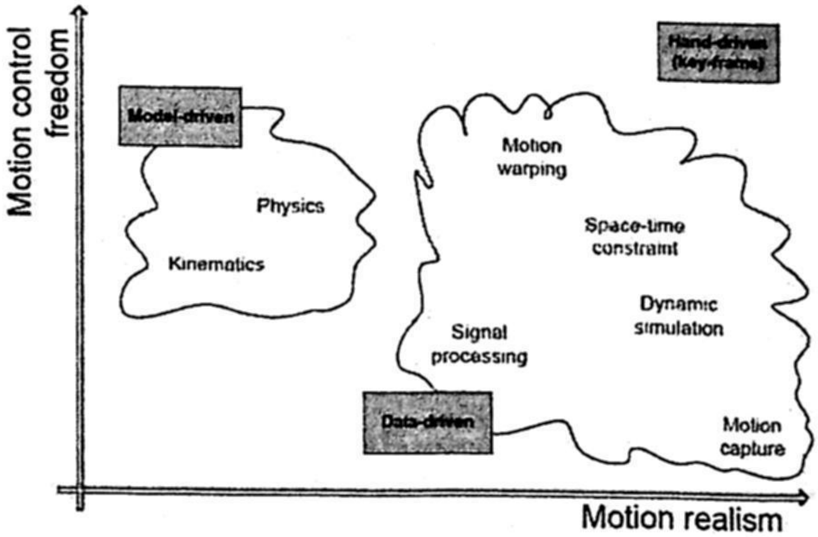


Fig. 1. Summary of motion synthesis approaches according to the motion realism versus motion control freedom for animators.

2.1 Discussion

Since the advent of motion capture systems which return high realistic motion without any control, the number of data-driven methods has exploded. Generally, these approaches produce motions with high realism thanks to the captured input data. The methods based on signal processing offer a weak control on the animation. In fact, the correspondences between motion parameters and frequencies are difficult to establish and the induced filtering can lead to less impressive results as the original motions. Motion warping techniques give lot of control freedom to the animators as key-frames are modified by hand. However, the motion realism is not ensured on the entire final sequences. Results can be improved with space-time constraint methods which consider a motion as a whole continuous sequence. In contrast, the constraint definitions have to be as intuitive as possible, reducing also the animator control. Finally, dynamic simulation allows to modify and enhance an original motion capture data by adding physical accuracy. The main advantage is that simple input data like rough keyframes are sufficient in order to produce convincing results. However, besides their expensive computational cost, these methods are limited to high dynamic motions. Nevertheless, the data acquisition is one of the stumbling blocks for data-driven methods. In fact, motion capture systems are expensive and need lots of pre-processing work to clean the recorded motions before using them.

Moreover, while data-driven synthesis could be applied to any creature whose movements can be captured, in practice it is primarily applicable to humans. In contrast, most of the hand- and model-driven methods can be applied to a wider range of body, with varying topology and joint structure.

3 The Two Stages Locomotion Planner

Motion planning increases the autonomy of virtual characters. Generally, given initial conditions and a goal, a path planner method provides a path to follow. Concerning the special case of locomotion tasks, the planner provides the path taking into account the obstacle avoidance problem. Additionally, this path is transformed into a trajectory which gives the adequate locomotion parameters at each time step. We have been inspired by an interesting proposal that raises the solution to the locomotion problem in two stages [9]. Authors apply PRM (Probabilistic Roadmap Methods) to get a collision-free 2D path in a 3D environment. This path represented by Bézier curves generates a trajectory encapsulating linear and angular speed variations. Finally a walking animation is generated according to those variations.

The two steps of the approach previously mentioned are the following: 1) Bounding the character's geometry by a cylinder allows motion planning for navigation to be reduced to planning collision-free trajectories for this cylinder in 3D. 2) Simply computing a collision-free path in the environment is not enough to produce realistic animation. A motion controller is used to animate the actor along the planned trajectory. We detail the two stages of the approach.

3.1 Path Planner

The RRT approach, introduced in [11], has become the most popular single-query motion planner in the last years. RRT-based algorithms were first developed for non-holonomic and kinodynamic planning problems. RRT-based algorithms combine a construction phase with a connection phase. Until now we do not have knowledge if some author uses this method for the planning process. In fact, in our implementation we can use the unidirectional planners or the bidirectional ones.

3.2 Motion Controller

While motion capture provides eye-believable motion, recorded sequences are fixed. The objectives of a motion controller are to readapt captured sequences to fit a given scenario while preserving their believability. A motion controller computes automatically time-parameterized trajectories driving all the degrees of freedom of a character (or any mechanical system), given the input state that defines a goal to reach.

We address the problem via a geometrical formulation in the two dimensional control space, inspired by Pettré et al [9]. The key idea is to transform each motion capture of a given database into a single point lying in a two dimensional velocity space (linear and angular velocities). These points are then structured into a Delaunay triangulation allowing efficient queries for point location and nearest neighbor computations. The control scheme is based on blending operator working from the motion capture library.

Motion capture blending and extraction of postures from the synthesized locomotion cycles is the two last stage of the controller process. The blending involves interpolating the angular trajectories of the selected motion captures according to their respective weight. The interpolator manipulates the motion frequency spectra previously computed for each motion sample. A complete locomotion cycle is synthesized with characteristics corresponding to the user's directives. Finally, postures are extracted from the new cycle to produce animation (see Figure 2).

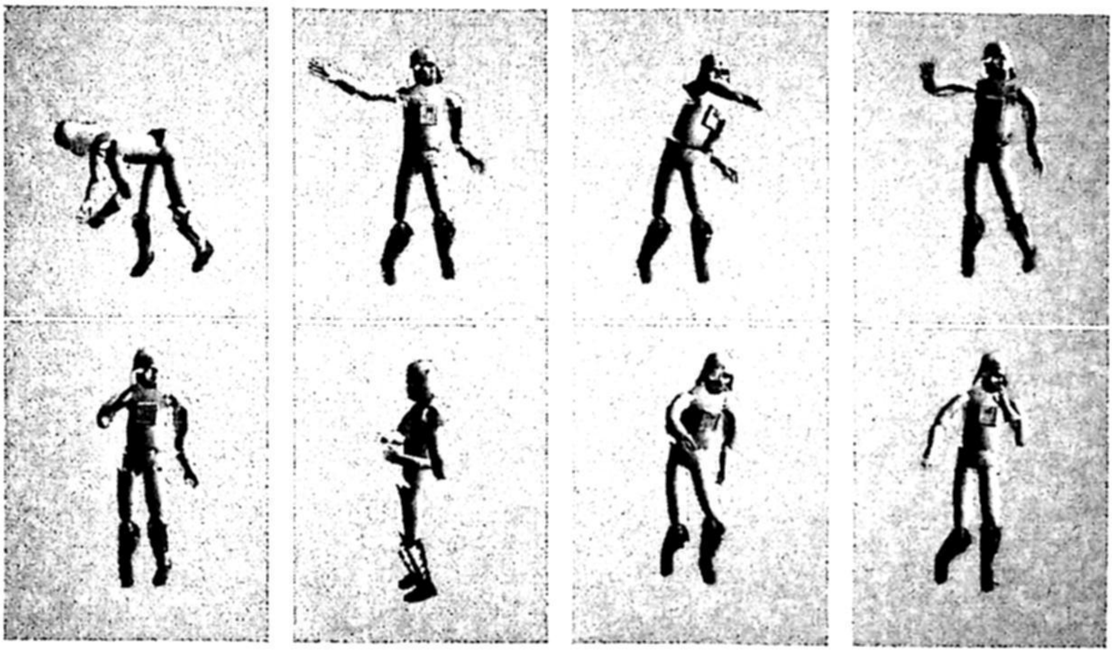


Fig. 2. Examples of different postures obtained with the proposed system. These motions can be characterized by emotional expressiveness or control behaviors.

4 System Design

MAVIC's design focus is on the creation of a modular and open animation system. In some cases, compromises had to be made to find a balance between competing requirements, like speed and data abstraction. Other times, seemingly contrasting goals like tight integration and modularity had elegant solutions that allowed both to be fulfilled. MAVIC was implemented on an Intel © Pentium IV processor-based PC running at 2.6 GHz with 2 GB RAM, using C# and OpenGL. An important feature of our system is the use of ODE as a physics engine, and in addition as collision checker. ODE is an open source, high performance library for simulating rigid body dynamics. It has advanced joint types and integrated collision detection with friction.

Animating very complex model such as virtual humans is usually done by extracting a simpler representation of the model, a skeleton, that is an articulated figure made of rigid links connected by hinges. The number of joints of the model and the degrees of freedom (dofs) depend largely on the desired reality or quality. In this work, we use a relatively simple model with 52 dofs. Motion capture data depend on the model [1], the structure of the virtual human is modeled in two levels. Pelvis and legs are used for the locomotion, all the 18 dofs are said to be active dofs. The 34 other ones are said to be reactive dofs, they deal with the control of the arms and the spine (see Figure 3).

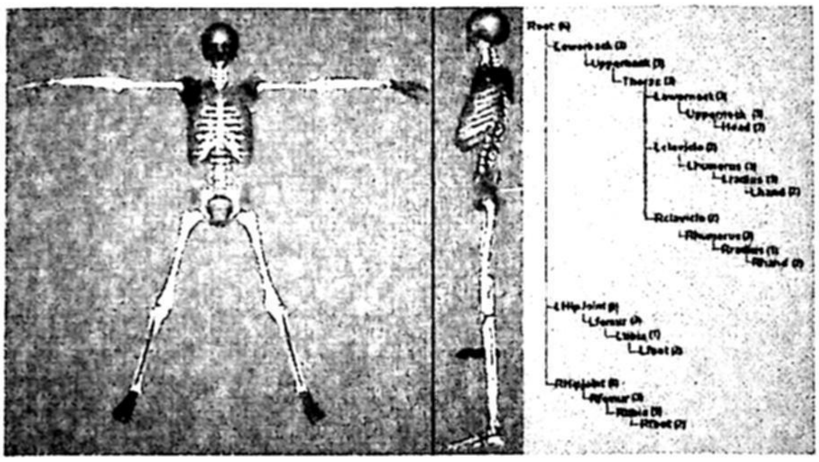


Fig. 3. Functional model for human locomotion.

Human figure walking or running is essentially a quasi-nonholonomic system, since the typical turning radius is usually subject to some minimum value depending upon the velocity [2]. For ease of control, we abstract a human-like character as a particle with an orientation. The particle is constrained to move on a floor. The orientation of the particle is aligned with its velocity. Such a particle is often called as a vehicle. Humans tend to only walk forward, not backward or sideways (no direction reversals during the path following) [10].

The path following phase is modeled as one involving an oriented disc smoothly tracking a geometric path in the plane. The disc center corresponds to the projected point at the base of the character's geometry, and the orientation of the disc corresponds to the character's forward-facing direction (see Figure 4). Since the linear velocity of the disc is constrained to always lie along the forward-facing

[1] The motion library is provided by the CMU Graphics Lab Motion Capture Database.

[2] Of course, a person can turn in place, this will only happen at the beginning or end of a path, not in the middle of a path.

direction, the character can walk or run forward. While turning is modeled by considering the disc's rotational velocity about its center.

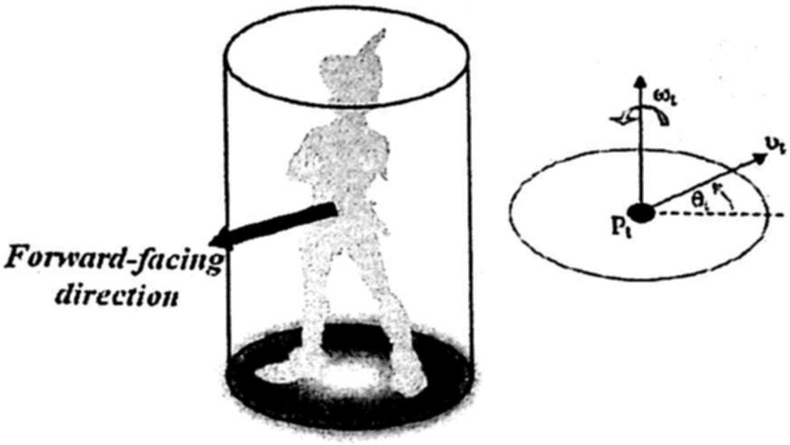


Fig. 4. The character's geometry is bounded by an appropriate cylinder. The center of the disc is the projection of the origin of the root joint of the character onto the walking surface.

A discrete time simulation of the following state variables is used: P_t position (x_t, y_t) of the disc center, θ_t orientation, v_t linear speed along the direction of θ_t, ω_t angular speed about P_t.

The tuple $(P_t, \theta_t, v_t, \omega_t)$ represents the simulated state of the character at time t. At each time step, any combination of the following two controls may be applied: a_t linear acceleration along the direction of θ_t, α_t angular acceleration about P_t. These controls model the four basic fundamental actions for the character (speed up, slow down, turn left, turn right). Speeding up and slowing down are represented by positive and negative values of a_t respectively. Positive values of α_t correspond to left turns, while negative values correspond to right turns.

Once the controls a_t and α_t have been specified, the state variables are integrated forward discretely by the time step Δt. In these experiments, simple fixed-step Euler integration can be used. For this kind of integration, the state propagation equations are approximated by:

$$\begin{aligned}
x_{t+\Delta t} &= x_t + (v_t \cos\theta_t)\Delta t \\
y_{t+\Delta t} &= y_t + (v_t \sin\theta_t)\Delta t \\
\theta_{t+\Delta t} &= \theta_t + \omega_t \Delta t \\
v_{t+\Delta t} &= v_t + a_t \Delta t \\
\omega_{t+\Delta t} &= \omega_t + \alpha_t \Delta t
\end{aligned}$$

The simulation proceeds in this fashion iteratively. As long as the values of the controls are reasonable relative to the size of the time step Δt, the motion will be smooth and continuous. The method to compute the two controls is based on proportional derivative control. Given the current state of the system, a desired state is calculated. The controls are then computed to move the system towards the desired state.

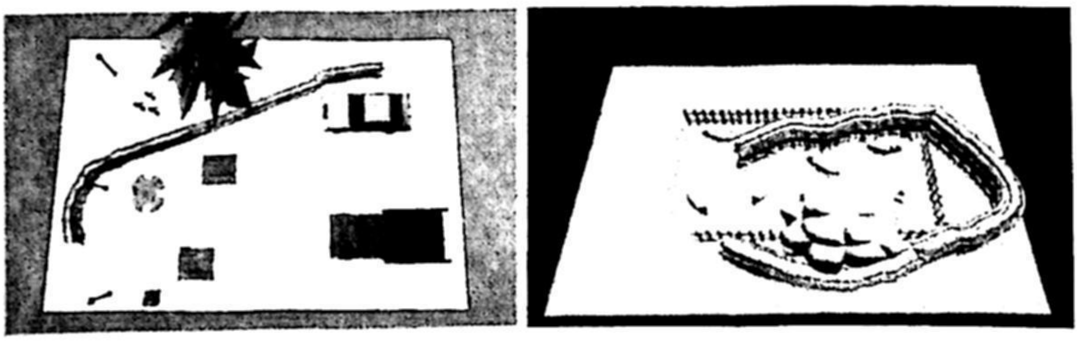


Fig. 5. A complete locomotion planning in two complex environments.

To deform a motion with precise goals, the first solution consists in modifying the body joint orientation in order to get a new posture. In [7], the authors introduce a variant of displacement mapping called motion warping. The animator interactively defines a set of keyframes inducing a set of constraints. These are used to derive a smooth deformation preserving the fine structure of the original motion. However, it is difficult to ensure the geometric constraint enforcement between keyframes. In addition, motion warping methods are purely geometric techniques and operate on each dof independently, without understanding the motion structure. They are not well suited for adjustments requiring coordinate movements, such as grasping actions with modification of object location. In such cases, not only the joint hand is affected, but also other joints like the elbow or shoulder.

In order to alter coherently multiple dofs of an original motion and over a continuous time period, constraint-based techniques can be applied. Discussed and classified in [12], these methods provide effective tools to interactively manipulate a motion clip by changing some important movement properties. Space-time constraints were first introduced to the graphics community by Witkin and Kass [5]. In our system, we have implemented both strategies to avoid collisions of the upper bodies of the virtual character, see Figure 6.

Given an appropriate model of the environment and a simulator, physics-based animation enforces realistic motion. The cost of this realism is the loss of fine-grained control over objects in the scene. An animator can no longer simply specify an object's trajectory or velocities over time, since those values may not be physically achievable. For example, the center of mass of an object in flight, in the absence of any external forces except gravity, is constrained to follow a parabolic path.

While fine-grained control is necessary to achieve certain effects, an animator might prefer to let specific objects behave autonomously. This is particularly

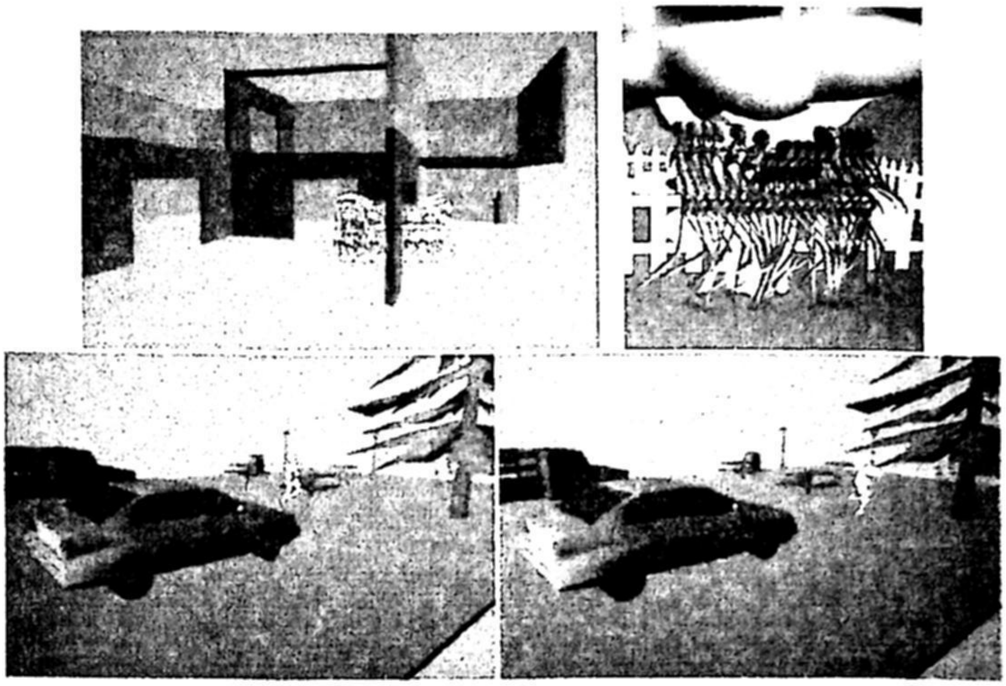


Fig. 6. Motion warping vs. constraint-based techniques.

useful in a scene where many objects are interacting in a complex manner. Rather than tediously specifying the trajectories of all the objects, the animator would often prefer to focus on control of a few foreground characters, allowing the background characters to evolve in a natural manner.

Control of animated objects in a physics-based environment ultimately involves the specification over time of actuations consisting of forces and torques. Given these, the equations of motion determine resulting accelerations, which the simulator then integrates to produce velocities and positions to update the state vector.

Synthesizing the control required to produce a desired motion is a difficult problem. Consider the control that humans use to produce a walking motion. While much progress has been made in a general understanding of the principles of biped locomotion, there has been little progress made toward a controller for physics-based human locomotion that can be applied to general gaits and terrains. Yet, walking is second nature to us; conscious control is rarely required for navigating smooth terrains. In fact, the human body senses and interprets many cues from the environment, using these to orchestrate muscle activity in a manner that puts most sophisticated robotic control systems to shame.

A general solution of the control problem produces a mapping from the state of a system to actions that will produce an appropriate path to a desired goal state. In physics-based animation, the state is described in terms of the dofs of the system and their derivatives. The space spanned by the dofs and their derivatives is called the state-space. The problem can be thought of as a motion

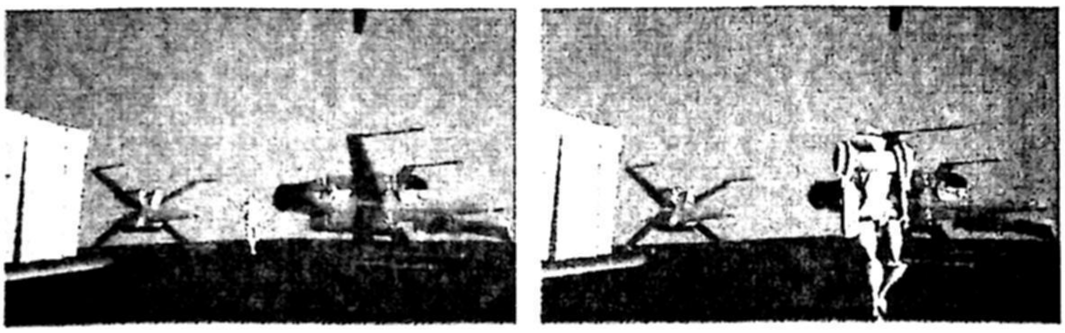


Fig. 7. With the motion controls implemented it is possible in real-time to change of the walking task to the running task.

planning problem through state-space as seen in Figure 8, which shows a two-dimensional space. A major obstacle to solving this problem is dealing with the size of the space, which grows exponentially with the number of dimensions. This is often referred to as the curse of dimensionality. For high-dimensional spaces, it can be difficult or altogether impossible to compute a path between points in the state-space, even when they are close together. These difficulties arise because of the inherent constraint of a physical environment such as inertia or gravity. A path between two points in state-space does not exist when no actuations exist that can control the system to go from one state to the other.

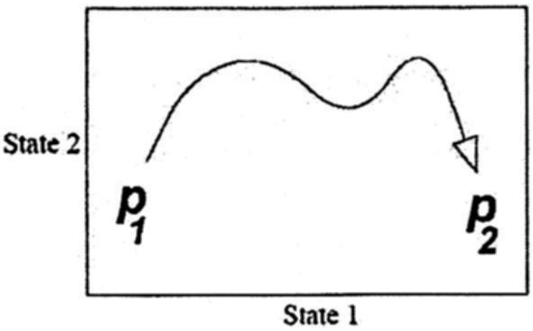


Fig. 8. A path through state-space, p_1 and p_2 are initial and goal states.

In order to make the approach practical, LaValle and Kuffner [13] integrated it within the framework of probabilistic algorithms, with the RRT approach. Here, the author's idea is to propagate the state of the robot into the future by choosing random bang-bang controls. From the state p_1, an exploration tree is thus built by applying random control inputs u over a constant time interval δt. The search is finished when p_2 is approximated with the predefined accuracy ϵ. A tradeoff has to be found between the accuracy with which the goal is reached and the time it takes to find a solution. Although the RRT approach does not produce a time-optimal solution it provides, in reasonable time, a solution for real-sized problems.

Our approach for the dynamic motion controller is the following. The method controls a character based on input motion specified by a user, and environmental physical input in a physically based character animation system. In the system, the angular acceleration of character's joints are controlled so as to track the user-input motion. Dynamic simulation then generates the resulting animation. When environmental physical inputs is applied to the character, the dynamic motion control computes the angular joint accelerations in order to produce dynamically changing motion in response to the physical input.

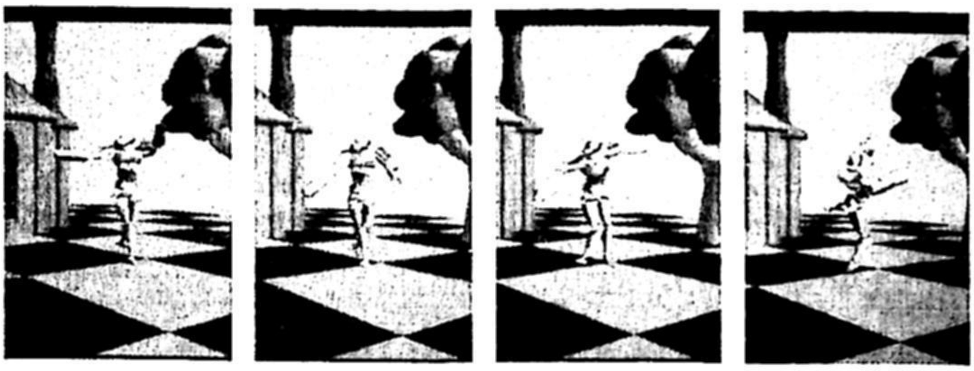


Fig. 9. Dancing animation with environmental physical input.

5 Conclusions and Future Work

Synthesizing realistic animation motions remains one of the great challenges in computer graphics. It seems that obeying physical laws is an important criterion of plausibility of motion. An interesting method to produce such physically motions is to animate characters from captured motion data that are inherently valid. These motions are adapted to different representations of the character to various environments or to additional kinematical constraints. The kinematic and kinetic adaptations (by interpolation, edition, retargeting or blending) may introduce physical inaccuracies in virtual human animation. It is thus necessary to be careful when such methods are used. Whenever the modifications introduce visually apparent errors in the physics of a motion, dynamics improvements may be added as a post-process or may correct the adaptation algorithm. We can for example, use the approach proposed by Safonova and Hodgins [14] to analyze the correctness of some physical properties.

With MAVIC, we have built an open, extensible animation system that engages the animator to interactively direct a 3D animation. The plug-in architecture allows a large library of diverse actuators and controllers to be implemented and integrated using a standard object-oriented interface We use MAVIC to build sophisticated, interactive 3D environments. We see the MAVIC platform as the basis for research into computer graphics in the areas of character animation, physically-based motion and control. An interesting aspect in the locomotion planning is the coordination of multiple virtual humans. Our work group have developed an interesting approach for the coordinated motion of virtual characters by combining centralized and decoupled planning methods. In the future, we will integrate to MAVIC this interesting solution.

References

1. N. M. Thalmann and D. Thalmann (Editors), "Handbook of virtual humans", *John Wiley & Sons, Ltd*, (2004)
2. F. Multon, L. France, M. P. Cani and G. Debunne, "Computer animation of human walking: A survey", *Journal of Visualization and Computer Animation*, Vol. 10, (1999) 39 -54
3. N. I. Badler, J. D. Korein et al., "Positioning and animating figures in a task-oriented environment", *The Visual Computer*, Vol. 1, No. 4, (1985) 212-220
4. A. Watt and M. Watt, "Advanced animation and rendering techniques: Theory and practice", *ACM Press*, (1992)
5. A. Witkin and K. Kass, "Space-time constraints", *Proc of the ACM SIGGRAPH*, (1988)
6. A. Bruderlin and L. Williams, "Motion signal processing", *Proc of the ACM SIGGRAPH*, (1995)
7. A. Witkin and Z. Popovic, "Motion warping", *Proc of the ACM SIGGRAPH*, (1995)
8. Z. Popovic and A. Witkin, "Physically-based motion transformation", *Proc of the ACM SIGGRAPH*, (1999)
9. J. Pettré, J. P Laumond and T. Siméon, "A 2-stages locomotion planner for digital actors", *Eurographics, Symposium on Computer Animation*, (2003)
10. J. J. Kuffner, "Goal-directed navigation for animated characters using real-time path planning and control", *Proc. of the CAPTECH*, (1998) 171-186
11. S. LaValle and J. J. Kuffner, " Rapidly-exploring random trees: Progress and prospects", *Algorithmic and Computational Robotics: New Directions, A K Peters*, (2001) 293-308
12. M. Gleicher, "Comparing constraint-based motion editing methods", *Graphical Models*, Vol. 63, No. 2, (2001) 107-134
13. S. LaValle and J. J. Kuffner, "Randomized kinodynamic planning", *International Journal of Robotics Research*, Vol. 20, No. 5, (2001) 378-400
14. A. Safonova and J. K. Hodgins, "Analyzing the physical correctness of interpolated human motion", *Eurographics Symposium on Computer Animation*, (2005) 171-180

Computer Networks

An Approach of Content Management based on Structured Peer-to-Peer Networks

Luis Enrique Colmenares Guillén and Angel Omar Mendoza Rojas

Facultad de Ciencias de la Computación FCC de la
Benemérita Universidad Autónoma de Puebla BUAP
Apdo. Postal J-32, Ciudad Universitaria, Puebla, México.
lecolme@cs.buap.mx, aomendoza87@gmail.com

Abstract. In this work, first evaluates routing Dht and the integration with metadata in query; in the future during the second phase would evaluate the proposal in dynamic scenes considering churn-rate and during the third phase would implement the application in platforms of development for P2P networks, that offer some applications in management of contents such as: collaborative, distributed and scalable. The main contribution of this first stage is the improvement obtained by the Distributed Hash Table (DHT) when adding metadata in the query. Semantic Routing reduces the number of hops averages to locate an object. This routing aid Dht-bamboo is used to improve the search of documents, because it obtains the interest of participants that could have on the documents. The metadata give us expressiveness Dht. In Dht-bamboo, the metadata consumes more bandwidth in the communication. The number of metadata within query increases the cost of communication between all the participants of the collaborative application.

1 Introduction

The three architectures more used of content delivery systems are: Wide World Web (WWW), Content Delivery Network (CDN) and Peer to Peer (P2P). The WWW and CDN are client-server architectures. In the WWW, the objects are static and small and the requests are usually frequent. The excess of requests can create unavailability of servers. A solution to this problem would be to have a great number of servers and high resources of storage, like for example, the CDN [1] that contains many objects of audio and video. One disadvantage of CDN is that represent greater costs in the storage and communication. Another solution is using the P2P networks because they are decentralized, distributed, and autonomous.

There are three considerations that are important for the management content:

1. The growth of non structured data. The structured data is kept in a data warehouse; data-marts and/or applications from databases. The non structured data are: Audio, video, text without structure, such as the body of an email or a document made by a text processor, spreadsheets or presentations.
2. The management of content in the network edge. The document management and the content management use metadata to classify and to search

Received 18/03/09
Accepted 27/04/09
Final version 06/05/09

information. The document management includes the information lifecycle, from its creation to the visualization by the final users. The management of contents is oriented to the reusability of the content.

3. The collaboration of Internet inside and between users from universities is frequent and useful in nowadays. The collaborative applications such as BSCW [3] and MOODLE [4] are client-server architectures in the WWW, and all the contents are in the server; the participants are dependent of the robustness of the server in order to have the active contents. Nevertheless, the disadvantages of these centralized applications are that they represent a single point of fault. Groove [5] is a system P2P that provides collaborative tools like chat, forums of discussion, interchange of files and calendars.

Now, the prototype makes the basic functions of a system of content management and the creation and search of the content with mechanisms of caching and routing. This work is organized as follows: in section 2, the description of the requirements of the system, in section 3, the architecture of the management of content for a P2P network based on collaboration. In section 4: a case of study, the used platform and the evaluation, and in the section 5, gives some conclusions of the contributions.

2 Architecture Requirements

In this section outlined the requirements that differentiate our system from the existing content management systems. The requirements in order to improve the delivery content are the following: The data types to share, the use of open standards and the management of content. Insert new: control of the content, indexes and searches P2P, content-based on interest and management of content off-line.

2.1 Non Structured Data

The characteristics of the data or contents are: they have small size, they are not structured and they change frequently, although, the contents might have greater size, and greater use of the bandwidth and storage. A discussion about data not structured is in [2].

2.2 Open Standards and Communication

The use of structured P2P networks based on Distributed Hash Table (Dht) as OpenDht [8] and Bamboo [9] is event-driven and has a single-threaded programming style [9]. Many programmers can be familiarized with this style of programming when using Graphics User Interface (GUI) libraries such as Java Swing or the Gimp toolkit (GTK++). When code is event-driven written it is easy to understand and it can generate high concurrence. The configuration file of bamboo is similar to XML files. Each node of the network Bamboo has a file with these characteristics.

2.3 Overlay peer-to-peer Networks

The overlay networks are divided into three different topologies of peer-to-peer: unstructured, structured and hierarchical. In the peer-to-peer unstructured they have a centralized manager index; the disadvantage is that they represent a single point of failure. All peers are equal and peers do not have restrictions imposed by the topology. In the unstructured P2P networks [7], the distributed searches are flooding, and have the disadvantage of saturating the network of messages, increasing some times, their delivery time. In the cases, where the documents are rare to locate, the search time could end before finding the document. A characteristic that contributes to the success of the unstructured P2P networks, is that they support versatility in query or partial-match [10], that is to say, within query the objects are described by their types and properties (for example: Title, composer, interpreter).

In hierarchical networks, the peers are divided into two groups: The Hubs (called SuperPeers or SuperNodes), which are dedicated servers with large memory and computational power and the leaf-peers responsible to store the contents (because the central server does not store contents). The Hub server recognizes the resources that are for sharing and are available.

Finally, the structured networks have a regular topology, circles, torus or cubes. In addition they use distributed hash tables (Dht). Within Dht, the mechanism search (lookup) provides a model with deterministic searches that hide to the user: request of routing, costs of churn-rate, load balance and availability.

Some works, that have used efficiently Dht in the networks overlay, are [11], [12], [13], [9], and [14].

Two main disadvantages of the P2P structured systems based on Dht are identified:

1. The requests are randomly sent towards peers without associating the content, that is to say, the use of keywords lacks of flexibility in the mechanism of lookup and routing.
2. The indices in the hash tables present transitivity, because continuously peers are connected and disconnected of the network, affecting the robustness of the system.

2.4 Content Classification and Control

A Content Management System (CMS) should be able to manage content coming from two main sources: participants and repositories. Moreover, the system should access a database, remote web server, or XML files.

Metadata offer a good way to classify any piece of information and this makes it possible to search for content in a very simple and efficient way.

2.5 Indices and Searches P2P

In the search methods, it is fundamental to have an appropriate mechanism of indices and an optimization of queries that are propagated by the indices. The engineering of the indices is the heart of the search methods of P2P. Now, the applications use indices distributed in the P2P network. The known example is Gnutella [7] that has semantic indices. In many applications, the keywords are associated with a document, an administrative domain or a key in a database.

The indices in Dht are semantic-free with data-centric references [15]. These references are based on the name of the data, without interest in the location. The semantic indices capture the relation of the objects. While, that Dht semantic-free guarantees that they can find a key, even if it there is no relation with the content.

The query is a keyword that represents content or a document. The keywords or metadata have been studied in size and number. For example, the queries that contain, keywords in the Web, demonstrated in an analysis that the communication costs are increased as the keywords grow, indicating that they do not scale well [16]. Nevertheless, the indexed in keywords is possible in structured P2P networks [17], because the size of the metadata is usually single-key. Some Works [18] and [19], include metadata in the request of routing, improving the precision of the search.

2.6 Content-based Interest

The content-based interest is a characteristic has been linked by affinities of social networks that currently remain a challenge to solve [26].

2.7 Management of Content Off-line

Many implementations like Bamboo, Kelips [20], or [21] use an algorithm broadcast as they can be epidemic, rumor, among others. For example: the nodes that are affected in their tables of Dht when a node is connected or disconnected, sends a message (a rumor), to their near neighbors like their distant neighbors, propagating in the entire network the rumor and updating the corresponding indices within Dht.

At this point, should remark that the Architecture is still in a prototyping phase, and only some of the requirements above have been addressed:

1) The non structured data is displayed as indices; these indices associate a document creating a document list.
2) Routing was added so that it determines the preferences of the objects and allows that the requests are based on interests. A caching mechanism with a LRU policy was designed and the size of caching is static.
3) Open standards for communication between nodes are used. A distribution of Zipf for requests of documents is used. One Dht of a structured P2P network is installed in a computers cluster.
4) Requirements 2.4 and 2.7 have not been implemented yet.

3 An Approach of Management of Content for a Network P2P based on Collaboration

Dht-Bambo allows the communication between the nodes, the high level architecture of open-Dht is used, Fig. 1. The nodes OpenDht act like gateway, each node executes the code of OpenDht allowing the communication via RPC to any user who belongs to the network. The users (student or teacher) are nodes that can be situated outside of the infrastructure of Open-Dht. The nodes Open-Dht participate in routing and storage of Dht.

The three phases of our architecture that provide management of content in the context of structured P2P based in Dht are: 1. Communication begins with caching indexed 2. Searches with keyword in routing, make simultaneous requests for the location of content. 3. Replication with a broadcast algorithm to update content.

3.1 Communication

In this phase, the system initializes the collaborative architecture. We identified two roles that can be defined in the same PC. The role of participant is in the physical network and the role of participant-Dht, which belongs to the Open-Dht. Two zones are created when you enter the collaborative architecture: zone of replication of the participants and caching indexed of the participants-Dht. In Fig. 2, the indexed of a participant-Dht is the zone of caching; this indexing is a local index where the documents ordered by popularity are located. For example, a metadata "a" is associated with the node in the form of IP 192.168.0.1 within query, the request that makes the node open-Dht, this is composed of metadata and IP address and is called metadata-node. This metadata-node is in the participants who previously are asking for theme "a". The access to the network of contents is simple, by means of script in Perl all the participants-Dht are created. The participants have the infrastructure to communicate with other participants via RPC in Internet. The phase of communication allows during the disconnection the information of the active session to be stored. When the participant connects again, automatically the metadata-nodes update all the participants who belong to their shared areas.

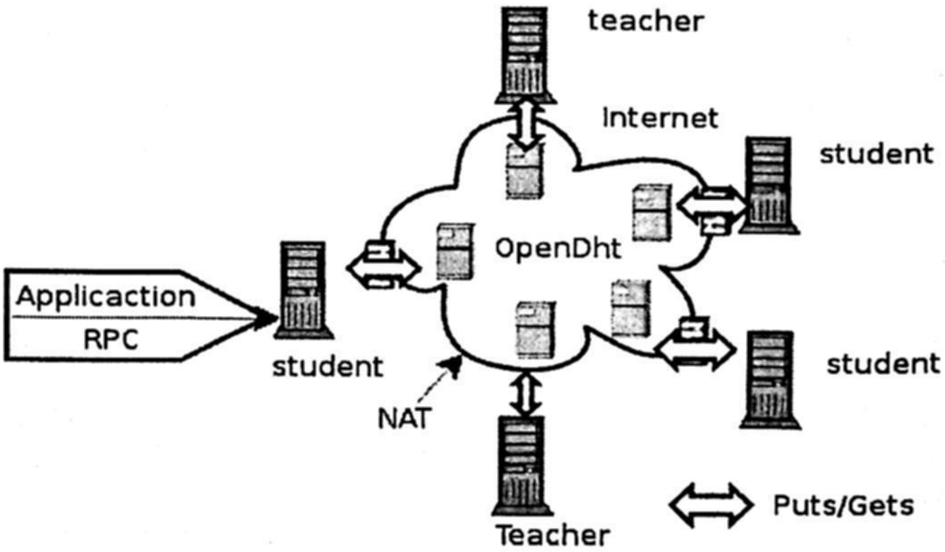


Fig. 1. Collaborative Architecture with OpenDht.

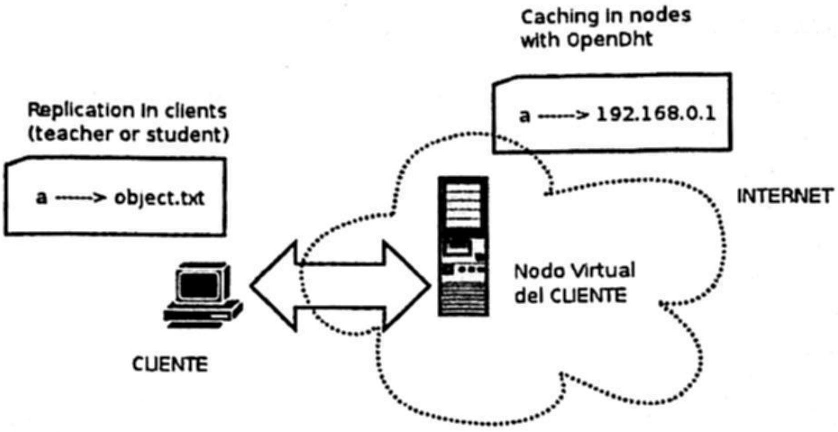


Fig. 2. Communication of clients and OpenDht nodes

3.2 Searches

In this phase, there are two mechanisms to make precise searches, first provides additional routing and second within query, there are metadata that represent the keyword for a location based on the content of the Theme. The first mechanism is called search of the participant-Dht, the participant uses different routings, and in the second mechanism the participants locate contents in their shared areas.

3.2.1 Search of the Participant-Dht

There are two routings in our architecture: routing of Dht-bamboo and routing semantic based on caching, see Fig. 3. For Dht Routing (DR), use routing of bamboo, is one Dht based on Pastry [11]. Dht-bamboo [9] assigns to each node, a key. The main mechanism is lookup; this has a greedy protocol that comes progressively near to the key in each hop. Each node maintains a set of neighbors who are used to send contents. These neighbors are: a) Near neighbors, which are numerically near and b) distant neighbors, that guarantees to choose some of them, for a network of *n* nodes, has a maximum of *O(log n)* hops and the number of neighbors does not exceed *O(log n)*. In bamboo, the near neighbors are called leaf set and the distant neighbors routing table.

A new class of bamboo is created, Semantic Routing (SR) works with the Dht. The mechanism of lookup, now, has three ways: leaf-set, routing table, and SR.

Info-semantic is defined, as a pair of data (metadata, node). These data, represent the metadata = key and value = IP node. This space of memory that uses info-semantic is caching. The key is the theme and the value is the node that has the theme.

At each request, the node, keeps in its zone from caching the metadata-node. The participants, who have interest in the document, keep the connection metadata-node. And in next requests, the content is given in a smaller number of hops. The participant, who makes for the first time a request of an object, uses the DR. Info-semantic is limited in each node, that is, the node can be full and use the DR. If a node makes a request and the object that is requested is not found in caching, also the DR is used. The substitution policy is made in order to eliminate at least the most recently used (LRU), and updating caching in each participant node.

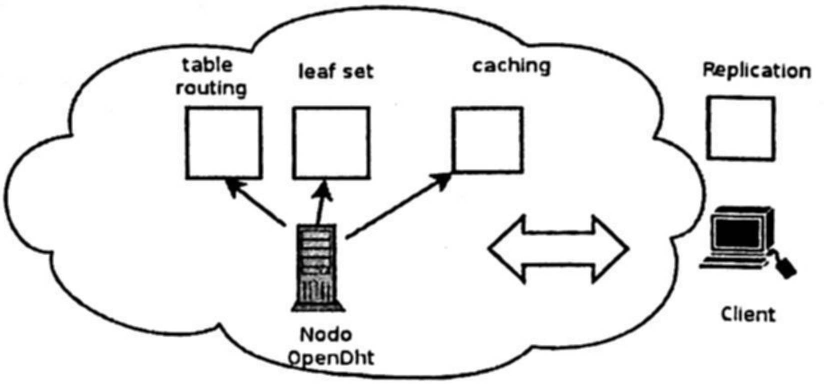


Fig. 3. Routings

3.2.2 Search in the Participant

The mechanism search in the participant is similar to the search in a system of files or a page Web. It is a continuation of the search in the participant-Dht. That is to say, when the way in routing was chosen, it traverses through different participants-Dht. The participant-Dht who has the content is chosen, so the content in its shared areas is located and success in the search is obtained. The Dublin Core schema [25] provides a consolidated and nearly standardized way for documents classification.

3.3 Replication

The works of replication like Beehive [22] or Symmetrical Replication [23] that need O(1) messages when the nodes enter/leave the network, do not relate the content from node, that is to say, they relate the content to numerical approaches like Dht. Bamboo supports an epidemic algorithm during the update of indices in the tables of the nodes when churn-rate exists. When extending this algorithm, and updating the new zone of caching, the new information is available in the nodes.

4 Case Study: The Magazine Collaborative

The motivation of this work is to have a collaborative magazine in the network overlay structured P2P, where the participants are students or professors. These participants are from different universities from all the country that can be located in physical or administrative limits, with security mechanisms: firewall, router, NAT, among others.

The work of professors and students is the creation of content where they can collaborate and share contents. This magazine is scalable, distributed and decentralized. The storage of the contents is directly proportional to the use of the themes and the degree of collaboration. The collaborative magazine cannot be accessed by certain documents because the participants do not have sufficient bandwidth for the communication of messages or when a document is highly acceded.

When a participant has decided to provide the content to another participant, these users have a relation. Regularly, those relations are created in run time and can change on the time. Each participant has a shared zone, where the files are associated by interest.

The users have interest in the contents. This allows creating content based on interests in a zone of caching. The zone of caching has a fixed size. When added caching to the users, improved response time, although to have a new difficulty, the update of caching that is solved including a policy of Least Recently Used (LRU) caching.

The users have permissions to make changes to contents. In addition, the system maintains coherence and consistency of the contents, even though the user is off-line. In the storage, Dht supports the primitive, *put* (key, value) and *get* (key) within routing. The idea is extended this algorithm and measure the impact that it generates to have more content in the node, in static and dynamic scenes.

4.1 Platform

The platform to measure our architecture this made in Bamboo-DHT [9] and the language scripting is Perl; the code is within a cluster of 16 nodes. The nodes of cluster have dual microprocessors AMD Opteron 64, 1.4 GHz, connected by Ethernet Gigabit, RAM memory of 2 GB a hard disk SCSI of 36 GB and OS fedora Linux Core 6x86-64. The simulations are execution-driven. All the file logs are stored in each participant of the network overlay P2P in run time.

4.2 Evaluation

For the evaluation three modules or classes in Bamboo were added. The first module, is caching, that assigns cache in each participant-Dht and allows the node to choose between two routings. The second module makes simultaneous requests in parallel. This allows making requests of queries from different participant nodes at the same time. The third module is the replication of content, and this in phase of elaboration. This will consist basically of the extension of the epidemic algorithm of Dht-bamboo to update the contents in caching of each node.

In the evaluation the following parameters are considered in the tables:

1.Requests based on the law of Zipf (simultaneously from different participant nodes).

2. Numbers of nodes with info-semantic: Use a local index and create this zone. Info-semantic is the zone of caching.

3. Limit of the Caching in the node: The zone of caching is limited; it uses a policy LRU and the objects from most popular to least popular.

4. Structured Metadata: The filtration of the searches of contents is made with metadata that represent a theme. The queries allow a unique key within routing of Dht. This association key-metadata within routing allows precise searches.

4.2.1 Static Scenes

The evaluations are made both routing SR and DR. In Table 1, shows the parameter that was used for static scenes with routing Dht-bamboo. The number of tests that were made was approximately 100 by each scene, although from test 45 the graphical one shows a uniform behavior with respect to the growth of the average of the number of hops.

The Table 2 shows the parameter that was used for static scenes with SR. A metadata in query of Dht is used; in addition caching that was made vary in the size of each node or participant. The number of objects is important, if are a small number of requests, 5, 10 or 20 requests. The objects will vary of 100, 250, 500, 750, and 1000 with duration of the object of one week (7*24*3600) or a day (24*3600) guaranteeing that the objects always are in the network bamboo. The communications between nodes of the cluster are stable in approximately 85% of the nodes. This guarantees 85 % of objects. In the number of requests: 10000, 12500, 15000, 20000, observed that the limits of the average are reduced. The size of caching used are: 5, 10, 15, 20. The use of caching represents use more the Dht in SR. In scenes 1, 2, 3 of table 2, using small caching, would more use of Dht-bamboo. In experiments 4, 5, 6 used greater caching, would use less the Dht-bamboo that is reflected in a smaller number of hops, as is in Fig. 5.

Table 1. Routing Dht-bamboo parameters.

Scenes: Dht-bamboo	No.Nodes (parameter 2)	No. Objects	No.Requests (parameter 1)	Without *caching*, without metadata
1	500	100	10000	-------------
2	500	250	12500	-------------
3	500	500	15000	-------------
4	500	750	15000	-------------
5	500	1000	20000	-------------

Table 2. Routing Semantic parameters.

Scenes: info-semantic (parameter 4)	Nodes (parameter 2)	Objects	Requests (parameter 1)	Size of caching (parameter 3)
1	500	100	10000	5
2	500	250	12500	5
3	500	500	15000	5
4	500	750	15000	10
5	500	1000	20000	15
6	500	1000	20000	20

In Fig. 6, so as in Fig. 4 and Fig. 5, the minimum, 3er quartile and the average of the number of hops by searching each object were found. The graph of Fig. 6 shows the following: The five more popular objects of the average, the total of objects represent 95% of the total requests. The requests are sent from 500 nodes of a total of 500 available nodes. Each node sends *n* requests of *m* available object. Each experiment of 500 nodes uses 500**n* requests. In *x*-axis, the objects are ordered from the most popular to the least popular, indicated by the named percentage and with their labels from left to right. In the *y*-axis, the number of total hops is shown.

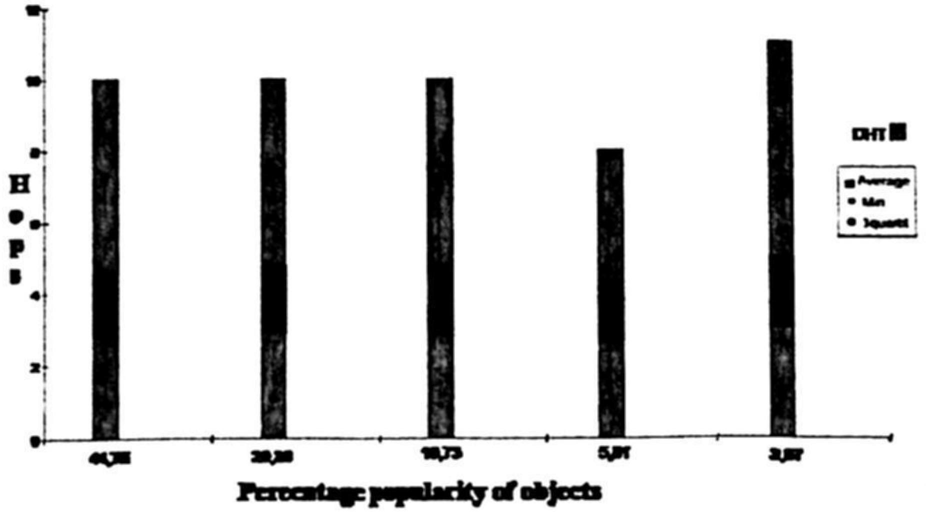


Fig. 4. Dht-bamboo Routing

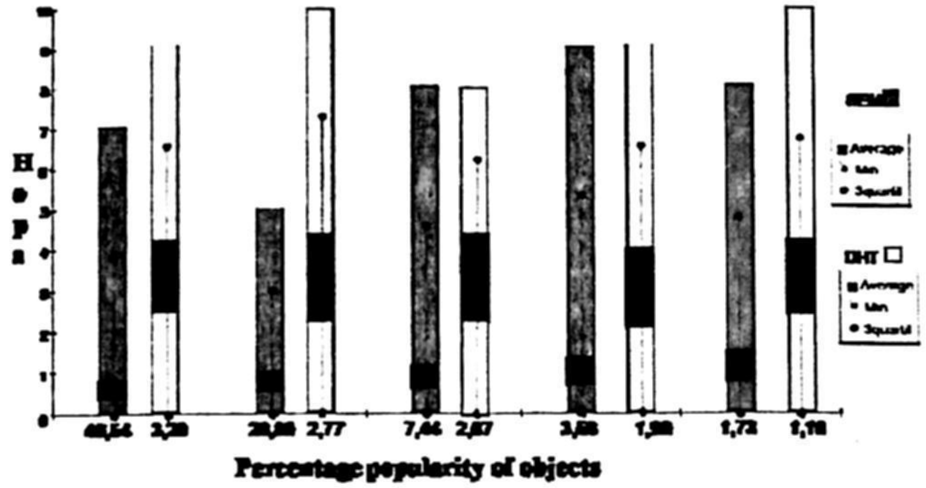


Fig. 5. Semantic Routing

In Fig. 6, there are five most popular objects with three bars, respectively. The two first bars of each object represent semantic routing proposed and it is a combination of both routings. The last bar of each object represents routing of Dht-Bamboo without alterations. The bars are divided in quartiles; the maximum number of hops for a same object is the high part of the bar. The third quartile represents the superior part of the line that is within the bar, the average or second quartile represents the high part of the black picture and the first quartile is the minimum hop that was obtained by requests of a same object that represents the final part of the line. The average of hops (superior limit of the black box) diminishes when the object is more popular. We observed in the graph that improvement in the number of hops by routing formed by DR and SR are the two initial bars. Routing of DR represents the third bar of each popular object; the average is over the two bars.

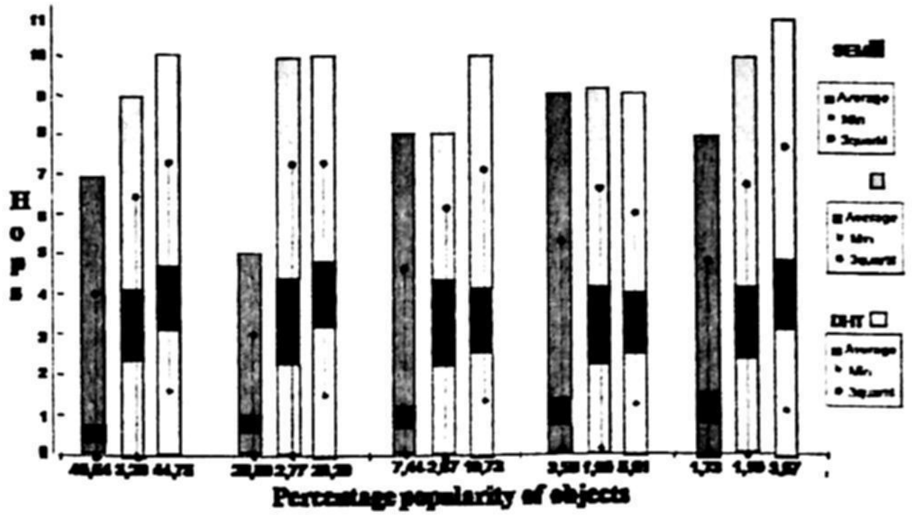


Fig. 6. Comparison of Routings

5 Conclusions

The main contribution of this work is to improve the routing of the Dht-Bamboo, with semantic routing that reduces the number of hops average. The culmination of this work will have three contributions, now had fulfilled the two first: 1. a mechanism of content delivery that allows possible routings to reach the content, when adding semantic routing in Dht-bamboo, obtained a reduction in the number of hops averages to find an object and 2. A search mechanism using semantic information, giving expressivity to Dht, single-key was used and it does not affect the costs of the communication between nodes of Dht, and 3. The last contribution will end with the class of bamboo that extends the broadcast algorithm, in addition, the evaluation in dynamic scenes, with parameters likes churn-rate and limitation of resources.

References

1. Akamai Technologies, Inc. [Online] http://www.akamai.com/. (2009).
2. Robert Blumberg, Shaku Atre. The Problem with Unstructured Data. DM Review Magazine. (2003).
3. BSCW (Basic Support for Cooperative Work) http://bscw.fit.fraunhofer.de/ (2009).
4. Modular Object-Oriented Dynamic Learning Environment (Moodle) http://moodle.org (2009).
5. The Microsoft Office Groove. http://office.microsoft.com/es-es/groove/default.aspx/ (2009).
6. "The emergence of Distributed Content Management and Peer-to- Peer Content Networks", by Gardner Consulting, (2001).
7. T. Kleinberg and R. Manfredi, http://www.gnutella.com, (2009).

8. S. Rhea, B. Godfrey, B. Karp, J. Kubiatowicz, S. Ratnasamy, S. Shenker, I. Stoica, and H. Yu. OpenDHT: a Public DHT Service and its Uses. SIGCOMM' 05, Philadelphia, Pennsylvania, USA, august 21-26, (2005).
9. The Bamboo Distributed Hash Table. A Robust, Open-Source DHT. http://www.bamboo-dht.org y http://opendht.org (2009).
10. E. Cohen, A. Fiat, and H. Kaplan. A Case for Associative Peer to Peer Overlays. ACM SIGCOMM Computer communication review, vol. 33, issue 1 pp 95-100. ISSN: 0146-4833, (2003).
11. Rowstron, A., Druschel, P.: Pastry: Scalable, Distributed Object Location and Routing for Large-scale peer-to-peer Systems. In IFIP/ACM Int. Conf. On Distr. Systems Platforms (middleware) (2001) 329-350.
12. Stoica, I., Morris, R. Karger, D., Kaashoek, M. F., Balakrishnan, H.: Chord: a Scalable Peer-to-Peer Lookup Service for Internet Applications, In Conf. On applications, Technologies, Architectures, and protocols for comp. Communications (2001) 149-160.
13. Zhao, B. Y., Kubiatowicz, J. D., Joseph, A. D.: Tapestry: An Infrastructure for Fault-Tolerant Wide Area Location and Routing. TR UCB/CSD-01-1141, UC Berkeley (2001).
14. Ratnasamy, S., Francis, P., Handley, M., Karp, R., Shenker, and S.: A Scalable Content Addressable Network. In: ACM SIGCOMM 2001. (2001) 161-172.
15. S. Shenker. The Data-Centric Revolution in Networking. Proceedings of the 29th VLDB Conference, Berlin, Germany, (2003).
16. J. Li, B. T. Loo, J. M. Hellerstein, M. F. Kaashoek, D. Karger, and R. Morris. On the Feasibility of Peer-to-Peer Web Indexing and Search. In Proc. IPTPS'03, (2003).
17. A. T. Clements, D. R. K. Ports, and D. R. Karger. Arpeggio: Metadata Searching and Content Sharing with Chord. In Proc. IPTPS' 05, (2005).
18. P. Triantafillou and I. Aekaterinidis. Content-based Publish-Subscribe over Structured P2P Networks. DEBS' 04, (2004).
19. Y. Choi and D. Park. Mirinae: A peer-to-peer Overlay Network for Large-scale Content-Based Publish-Subscribe Systems. In proc. of ACM conference NOSSDAV' 05, Washington, USA, (2005).
20. I. Gupta, K. Birman, P. Linga, A. Demers, R. van Renesse. Kelips: Building an Efficient and Stable P2P DHT through Increased Memory and Background Overhead. Proceedings of the 2nd International Workshop on Peer-to-Peer Systems, IPTPS' 03,(2003).
21. S. El-Ansary, L. Onama, P. Brand, and S. Haridi. Efficient Broadcast in Structured P2P Networks. In Proc. of the 2nd. Int. workshop on Peer-to-Peer Systems, IPTPS' 03, (2003).
22. V. Ramasubramanian, E. Gün, Beehive: O (1) Lookup Performance for Power-law Query . Distributions in Peer-to-Peer Overlays. Proceedings of the 1st conference on Symposium on Networked Systems Design and Implementation, (2004).
23. Ghodsi A., Alima L., and Haridi S. Symmetric Replication for Structured Peer-to-Peer Systems. In the 3rd Int. Workshop on Databases, Information Systems and Peer-to-peer computing, Trondheim, Norway (2005).
24. Adamic, L.A. "Zipf, Power-laws and Pareto– a ranking tutorial", http://www.hpl.hp.com/research/idl/papers/ranking/ranking.html (2009).
25. http://dublincore.org/, (2009).
26. H. Garcia-Molina, "Web Information Management Past, Present and Future". In the Proceedings of the International Conference on Web Search and Web Data Mining, Palo Alto, California, USA, 2008, pp. 1-10.

NSIS Signaling Protocol Simulator for Heterogeneous Quality of Service Inter-domain Negotiation

María del C. Vargas G., Ernesto E. Quiroz M. and Miguel A. Álvarez C.

Centro de Investigación y Desarrollo de Tecnología Digital del IPN
Av. del Parque No. 1310, Mesa de Otay Tijuana, Baja California, México CP-22510
Tels.: (664) 623-13-44; Fax: (6664) 623-13-88,
correo-e: {eequiroz, vargas, malvarez}@citedi.mx,

Abstract. This paper presents an IP domain Simulator whose local Quality of Service (QoS) provision rules follow ITU-T specifications. The Simulator inter-operates with a multi-domain, non QoS-homogeneous environment. By means of the NSIS signaling protocol, the local IP domain negotiates QoS agreements with "foreign" domains, effects performance parameters exchange, QoS adaptation in the local domain, route resource reservation, issues the data transmission initiation command and performs session termination. Since the adaptation policies from external to local QoS specifications is left open by the NSIS documentation, the Simulator sets the framework to interchangeably test and compare diverse QoS-adaptation approaches.

1 Introduction

With no regard to what type of application is to be transported by the Internet Protocol (IP), it will only provide the best effort service, which is insufficient for most of today's multimedia applications. In order that proper Quality of Service is provisioned, it has been necessary to add higher level protocols. In the context of a single IP domain, IntServ (Integrated Services), and DiffServ (Differentiated Services) have been widely used. When two or more domains are comprised, consistent end-to-end QoS has to be provided. With this objective, a host of protocols have been substantiated, among them: RSVP [1], YESSIR [2], Boomerang [3], BGRP and others [4, 5]. The diverse and at the same time stringent performance demands of the ever growing multimedia applications have outrun these solutions. The Next Steps in Signalling (NSIS) group of the Internet Engineering Task Force (IETF) is the most ambitious and relevant effort for inter-domain QoS provisioning at the time. NSIS started in 2002 releasing the NSIS Threats draft [6]. IP domains working under NSIS maintain their local QoS models (UMTS, DiffServ, etc.), while NSIS provides procedures and formats for QoS information exchange between domains, strategies to homogenize dissimilar QoS classes, IP and routing support and interaction [7].

NSIS has two layer architecture (Fig. 1). The upper layer (NSLP: NSIS Signalling Level Protocol) defines fields and message formats, and messages interchange [8]; while the second layer (NTLP: NSIS Transport Level Protocol) interworks with IP and lower layer protocols to deliver user data and NSLP messages [8].

© G. Sidorov, M. Aldape, M. Martinez, S. Torres. (Eds.)
Advances in Computer Science and Engineering.
Research in Computing Science 42, 2009, pp. 263-273

Received 09/03/09
Accepted 27/04/09
Final version 07/05/09

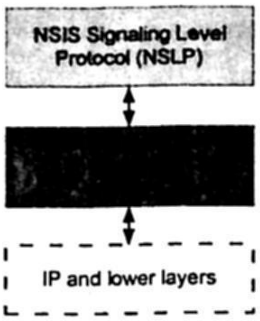


Fig. 1. NSIS layer architecture.

In this paper, a Simulator implementing the NSL Protocol in a local domain governed by ITU-T QoS specifications [9] is presented.

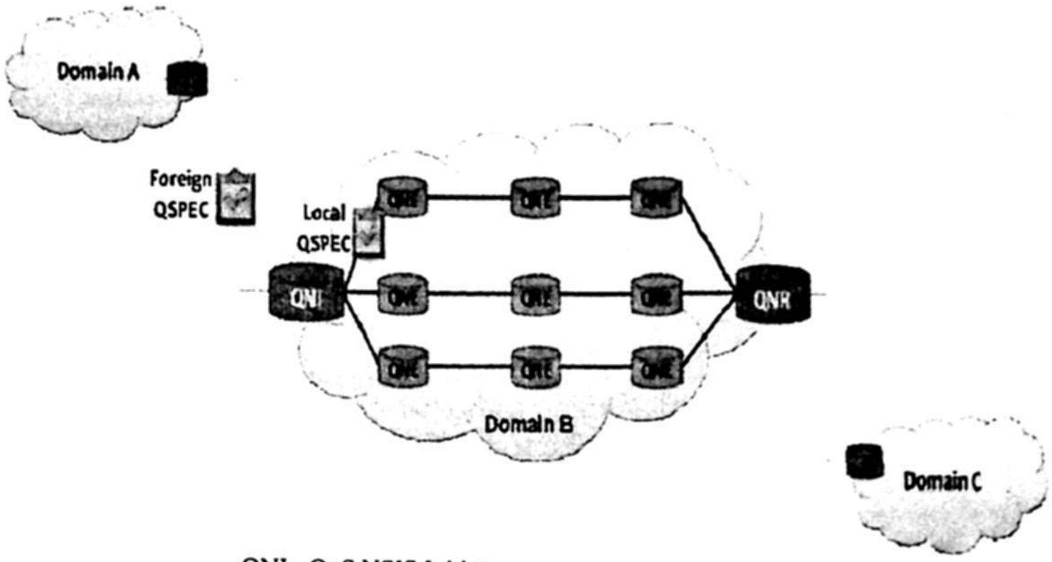
The rest of the paper is organized as follows. Section 2 explains the NSLP, including packet control information, network entities definition, types of messages and their utilization. Section 3 presents the NSLP Simulator development, NSIS entities implementation, QoS class/subclass definitions and signalling procedures for session establishment. Section 4 shows Simulator test scenarios and discusses on the outcomes of the QoS level assignments and their validity. Finally some conclusions on the Simulator features and future work are stated.

2 NSIS Signaling Level Protocol

NSLP establishes and maintains signalling among network entities of a session path, in order to reserve resources for the transmission of a data stream [10]. It interprets foreign QoS specifications (QSPECs) and adapts them to local QSPECs (Fig. 2).

2.1 QoS Specification

NSLP organizes in a QSPEC all QoS parameters, formats, and processing rules of the domain from which the session establishment comes. The adaptation procedure of a foreign QSPEC to a local QSPEC is not specified by the NSLP, and therefore is open to interpretation and/or definition by the Domain Administrator. Our implementation is explained in section 3. The local QSPEC is composed of four objects (see Fig. 3). (a) Desired QoS. Describes the best fit of a local-to-an-external QSPEC. (b) Available QoS. Is the available resource of an entity equal or nearest to the QoS level requested an entity can assign. (c) Minimum QoS. The minimum QoS level an Entity can offer to a session. (d) Reserved QoS. Resources reserved by a specific entity for a session [11]. Listed in the QSPEC objects, the QoS parameters are the QoS descriptors that specify the expected performance of a session. Restrictions parameters {Delay (ms), Delay variation (ms), Packet loss rate (packets/sec) and Packet error rate (%)} are preferred for NSLP performance description.



QNI: QoS NSIS Initiator
QNE: QoS NSIS Entity
QNR:QoS NSIS Receptor

Fig 2. Foreign and local QSPECs.

QoSDesired	QoSAvailable	QoSMinimum	QoSReserved
Parameter1	Parameter 1	Parameter 1	Parameter 1
:	:	:	:
:	:	:	:
Parameter n	Parameter n	Parameter n	Parameter n

Fig. 3. QSPEC structure.

2.2 NSLP Entities

Each node (router) in an NSIS domain is called an entity. According to their position in the network topology there are three types.

- QoS NSIS Initiator (QNI): Edge router capable of receiving and processing incoming sessions. Starting with a foreign QSPEC whose values can be expressed in numeric or descriptive form. If the values are descriptive, the QNI translates them to a numeric equivalent. The numeric foreign QSPEC is the reference to compare and find the most fitted local parameters. The QNI can now go through its resource table, to see if it has the required or similar QoS available. If it does, QoS class is reserved and diminished from the resource table.
- QoS NSIS Entity (QNE): Operates in the intermediate routers of an NSIS domain. Receives the local QSPEC and seeks through its resource table for the QoS requested. If available it is reserved, discharged from the resource table, and loaded in a message to the next entity in the path.
- QoS NSIS Receptor (QNR): Edge router in the receiving end of a domain. On receiving a QSPEC, the QNR performs the same tasks described for the QNE. Additionally builds a RESPONSE message for the QNI, which collects the reservation

status of all entities participating in the session [7], [12]. If the QNR is not the final destination of the session, it will pass on the original foreign QSPEC to the next domain (Fig. 4).

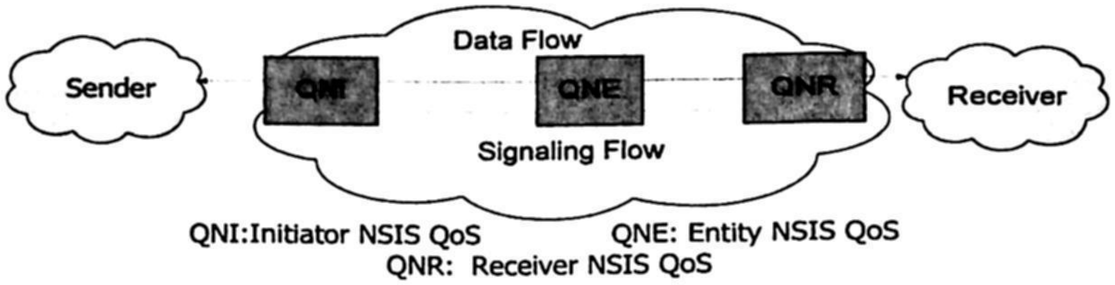


Fig. 4. Shows the entities that compose a domain NSIS.

2.2 NSLP Messages

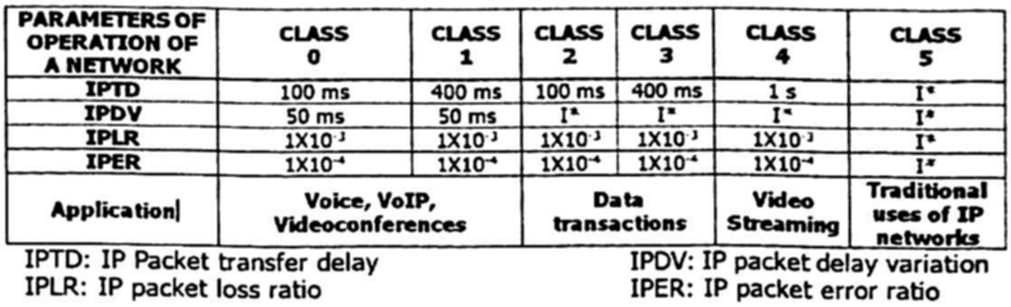
QSPECs are carried within various types of messages throughout the NSIS domain [11]. There are three main messages:

- RESERVE: Travels from QNI to QNR. Signals the entities to manipulate its reservation states (creation, update and modification).
- QUERY: Goes from QNR to QNI. Collects information about the available resources in the entities of a path, which allows the QNI to decide on whether to accept or reject a session request.
- RESPONSE: Its direction is from QNR to QNI. Provides information in reply to previous messages like RESERVE or QUERY. Advices the QNI that a path reservation has been successful, and delivers the reservation values assigned by each entity in the path. Is used in reply to a QUERY when a reservation error has appeared.

3 NSLP Simulator

The Simulator replicates the *modus operandi* of the NSL Protocol in an IP network. The local domain's QoS classes are those of the Y.1540 and Y.1541 (ITU-T) [9], which together define packet delivery performance parameters (Table 1). In order that the QoS ITU-T framework be adapted to the NSIS objects, a modification to the ITU-T table was introduced, consisting of a subdivision of each class into three subclasses (Premium, Regular and Basic), to correspond respectively with desired, available and minimum QoS.

Table 1. ITU-T QoS specification.



PARAMETERS OF OPERATION OF A NETWORK	CLASS 0	CLASS 1	CLASS 2	CLASS 3	CLASS 4	CLASS 5
IPTD	100 ms	400 ms	100 ms	400 ms	1 s	I*
IPDV	50 ms	50 ms	I*	I*	I*	I*
IPLR	$1X10^{-3}$	$1X10^{-3}$	$1X10^{-3}$	$1X10^{-3}$	$1X10^{-3}$	I*
IPER	$1X10^{-4}$	$1X10^{-4}$	$1X10^{-4}$	$1X10^{-4}$	$1X10^{-4}$	I*
Application|	Voice, VoIP, Videoconferences		Data transactions		Video Streaming	Traditional uses of IP networks

IPTD: IP Packet transfer delay IPDV: IP packet delay variation
IPLR: IP packet loss ratio IPER: IP packet error ratio

3.1 Simulator´s Operation

The control flow chart of Fig. 5 depicts all functions needed to be performed by the Simulator for a session to be completed, initiating with the request arrival until the user-data transmission ceases and the session is terminated. Since the diagram intends to be self-explanatory, additional clarification is given only to the QNI´s processes, which require more "intelligence" then the rest of the entities.

On the arrival of a foreign QSPEC, the QNI determines if there are descriptive parameters. All descriptive values are fitted to a corresponding numeric. Once all values are numeric the QNI performs the QoS level adaptation, to obtain a QSPEC with local values from an external QSPEC that follows a different value reference. This is an unspecified issue in NSIS, bequeathed to the network Administrator. We implemented a straightforward comparison to match the foreign QSPEC values with those of our modified ITU-T QoS specification, first by individual parameter, then by group with majority of hits. This way a class and subclass is assigned as a best fit to the external QSPEC. Other approaches can be tried in this stage [13].

The QNI then inquires in its resource table, to look for the QoS level required. If available, the QoS level is reserved and diminished from the resource table. Then a predefined route to the QNR is designated for the RESERVE message. Finally the role of the edge router as QNI or QNR is defined.

3.2 Programming Tools

The Simulator was developed in Visual Basic, using the Visual Studio 2005 platform. The rationales for this are as follows:

- Visual Studio supports several programming languages, so it would ease the workload in case a migration might be needed.
- Object oriented architecture. One of the main functions of the Simulator is to load QoS objects, which can be performed straightforward with Visual Basic. [14].
- A future version might be done and transported to XML (Extensible Markup Language), which is a de facto standard for networks device management [15].
- Visual Studio allows working in console mode, which could facilitate a future NSLP learning environment.

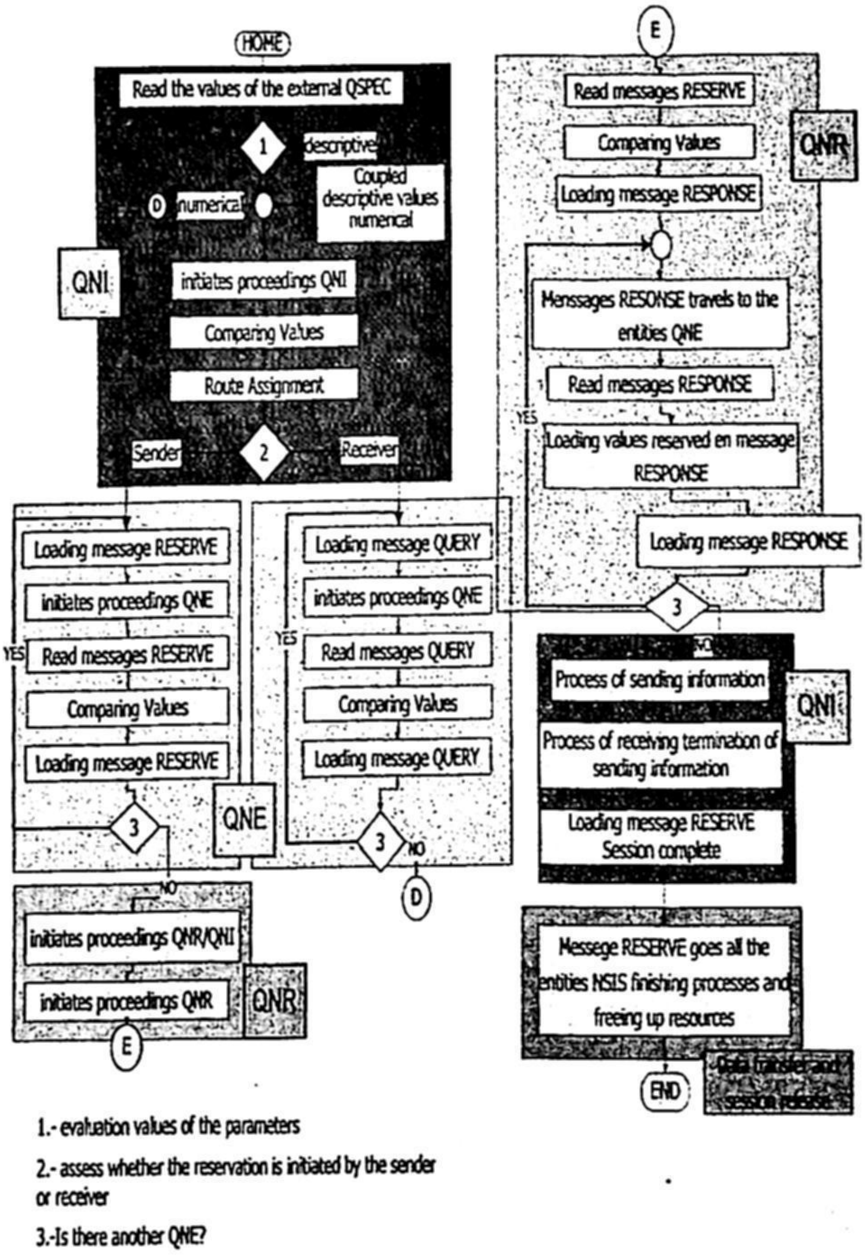

HOME
Read the values of the external QSPEC
1
descriptive
D
numerical
Coupled descriptive values numerical
QNI
initiates proceedings QNI
Comparing Values
Route Assignment
Sender
2
Receiver
Loading message RESERVE
initiates proceedings QNE
YES
Read messages RESERVE
Comparing Values
Loading message RESERVE
3
QNE
NO
initiates proceedings QNR/QNI
initiates proceedings QNR
QNR
E
Loading message QUERY
initiates proceedings QNE
YES
Read messages QUERY
Comparing Values
Loading message QUERY
3
NO
D
E
Read messages RESERVE
Comparing Values
QNR
Loading message RESPONSE
Menssages RESONSE travels to the entities QNE
Read messages RESPONSE
YES
Loading values reserved en message RESPONSE
Loading message RESPONSE
3
NO
Process of sending information
QNI
Process of receiving termination of sending information
Loading message RESERVE Session complete
Messege RESERVE goes all the entities NSIS finishing processes and freeing up resources
END
Data transfer and session release
1.- evaluation values of the parameters
2.- assess whether the reservation is initiated by the sender or receiver
3.-Is there another QNE?

Fig 5. Diagram shows the operation of the simulator based on NSLP.

4 Simulation Scenarios and Results

Over 150 simulations were made, each one representing a session. For the sake of comparison, two of them are selected to highlight different conditions of the network entities, which lead to a differentiated processing for each one. In the first case, the

QoS required is maximum, and the QoS assigned ends up degraded one level. For the second case, the Desired QoS is regular, and is also assigned a lesser level..

The main premises for the Simulator operation are:

- Topology. A simple network topology was adopted (Fig. 2, Domain B). It provides three alternative routes with three QNEs each, enough to test all NTLP functions.
- Foreign QSPECs. Twenty QSPECs were formulated (Table 2). The QSPECs are fed in sequential order, and upon getting to the 19th the process is repeated iteratively. These QSPECs cover the whole spectrum of characteristics a Domain could encounter, mainly: (a) Parameters are descriptive and numeric. (b) Signaling initiated by the QNI and QNR. (c) Range of values. The two instances to be analyzed are QSPEC 3 and 19.
- Entity´s Resource Table. Each entity contains a table to manage a limited pool of resources. Each class has 10 premium instances, 10 regular and 10 basic. This list is updated according to the incoming demands, and the sessions being finished.

Table 2. List of external QSPEC-sessions fed to the Simulator.

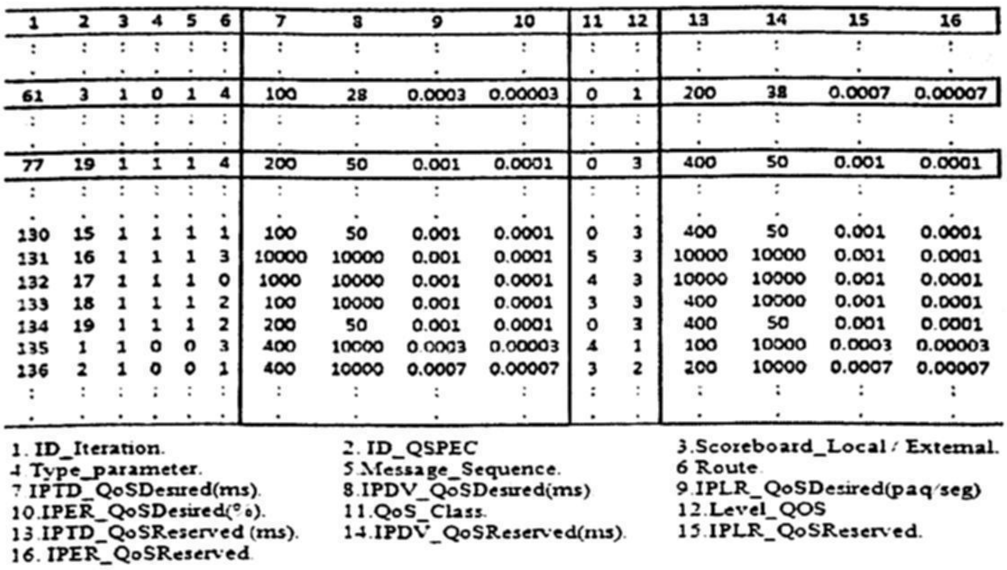
No. QSPEC	Message Sequence	QoS Class	Packet Transfer Delay (ms)	Packet Delay Variation (ms)	Packet Loss Ratio (paq/s)	Packet Error Ratio (%)
0	0	5	2000	300	0.0002	0.00002
1	0	4	350	100	0.0003	0.00003
2	0	3	300	70	0.0007	0.00007
3	0	0	58	30	0.0002	0.00002
4	0	5	3000	500	0.0005	0.00005
5	0	4	LOW	LOW	LOW	LOW
6	0	3	HIGH	HIGH	HIGH	HIGH
7	0	0	MEDIUM	MEDIUM	MEDIUM	MEDIUM
8	0	5	LOW	LOW	LOW	LOW
9	0	4	HIGH	HIGH	HIGH	HIGH
10	1	3	250	98	0.0003	0.00003
11	1	0	300	47	0.0007	0.00007
12	1	5	2000	300	0.0002	0.00002
13	1	4	350	100	0.0003	0.00003
14	1	3	300	70	0.0007	0.00007
15	1	0	HIGH	HIGH	HIGH	HIGH
16	1	5	MEDIUM	MEDIUM	MEDIUM	MEDIUM
17	1	4	LOW	LOW	LOW	LOW
18	1	3	HIGH	HIGH	HIGH	HIGH
19	1	0	MEDIUM	MEDIUM	MEDIUM	MEDIUM

4.1 Test Outcomes Analysis

Table 3 records the QNI sessions information in 16 items. Columns 7-10 carry the Desired QoS values. Columns 13-16 hold the actual resources the QNI had available to match the request.

Sessions 61 and 77 are processed by the QNI on request of external QSPECs 3 and 19 respectively. As can be seen, in session 61 the values of Desired and Reserved objects belong to the same QoS level.

Table 3. Data excerpt from the QNI log.

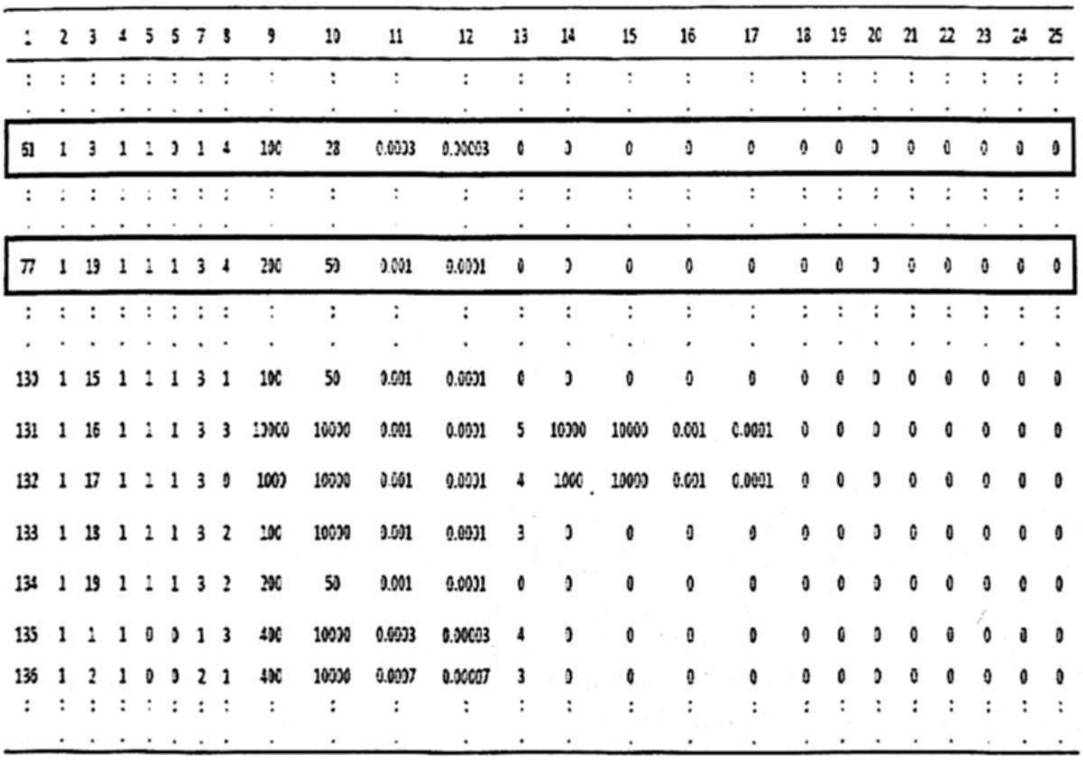


1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
61	3	1	0	1	4	100	28	0.0003	0.00003	0	1	200	38	0.0007	0.00007
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
77	19	1	1	1	4	200	50	0.001	0.0001	0	3	400	50	0.001	0.0001
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
130	15	1	1	1	1	100	50	0.001	0.0001	0	3	400	50	0.001	0.0001
131	16	1	1	1	3	10000	10000	0.001	0.0001	5	3	10000	10000	0.001	0.0001
132	17	1	1	1	0	1000	10000	0.001	0.0001	4	3	10000	10000	0.001	0.0001
133	18	1	1	1	2	100	10000	0.001	0.0001	3	3	400	10000	0.001	0.0001
134	19	1	1	1	2	200	50	0.001	0.0001	0	3	400	50	0.001	0.0001
135	1	1	0	0	3	400	10000	0.0003	0.00003	4	1	100	10000	0.0003	0.00003
136	2	1	0	0	1	400	10000	0.0007	0.00007	3	2	200	10000	0.0007	0.00007
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

1. ID_Iteration. 2. ID_QSPEC 3.Scoreboard_Local / External.
4 Type_parameter. 5.Message_Sequence. 6 Route.
7 IPTD_QoSDesired(ms). 8.IPDV_QoSDesired(ms) 9.IPLR_QoSDesired(paq/seg)
10.IPER_QoSDesired(%). 11.QoS_Class. 12.Level_QOS
13.IPTD_QoSReserved (ms). 14.IPDV_QoSReserved(ms). 15.IPLR_QoSReserved.
16. IPER_QoSReserved.

Columns 7-10 of session 77 display the descriptive-to-numeric conversion of the Desired QoS, while the Available QoS assigned values stated in columns 13-16 are of a lesser level. The QNI loads the Desired QoS values in the RESERVE message (Table 4), and sends it to the next QNE in the route. Entities QNE 4, QNE5 QNE6 launch the process of comparing, adapting, and assigning QoS values in relation to the RESERVE message bore values (see Table 4).

Table 4. RESERVE message registry.



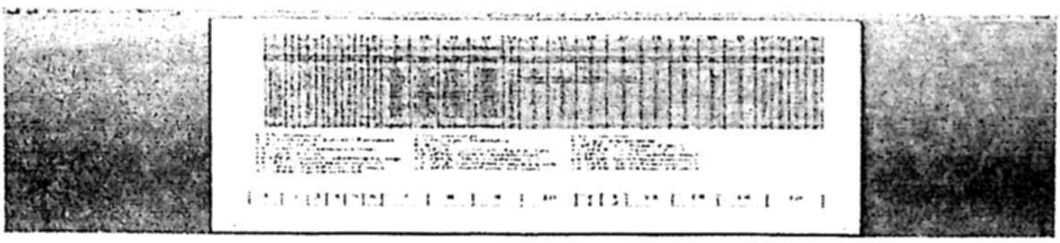
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
61	1	3	1	1	0	1	4	100	28	0.0003	0.00003	0	0	0	0	0	0	0	0	0	0	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
77	1	19	1	1	1	3	4	200	50	0.001	0.0001	0	0	0	0	0	0	0	0	0	0	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
130	1	15	1	1	1	3	1	100	50	0.001	0.0001	0	0	0	0	0	0	0	0	0	0	0	0	0
131	1	16	1	1	1	3	3	10000	10000	0.001	0.0001	5	10000	10000	0.001	0.0001	0	0	0	0	0	0	0	0
132	1	17	1	1	1	3	0	1000	10000	0.001	0.0001	4	1000	10000	0.001	0.0001	0	0	0	0	0	0	0	0
133	1	18	1	1	1	3	2	100	10000	0.001	0.0001	3	0	0	0	0	0	0	0	0	0	0	0	0
134	1	19	1	1	1	3	2	200	50	0.001	0.0001	0	0	0	0	0	0	0	0	0	0	0	0	0
135	1	1	1	0	0	1	3	400	10000	0.0003	0.00003	4	0	0	0	0	0	0	0	0	0	0	0	0
136	1	2	1	0	0	2	1	400	10000	0.0007	0.00007	3	0	0	0	0	0	0	0	0	0	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Upon the RESERVE message arrival, the QNR repeats the reservations operations in like manner as the QNEs. Now the QNR ready a RESPONSE message to inform the QNI that the reservation process was successfully completed (see Table 5). The RESPONSE message gathers the Desired QoS and Reserved QoS objects of all the QNEs in its trajectory to deliver to the QNI.

The rest of the messages, and data generated during the initiation, follow-up, and tear-down of a session can also be observed.

Table 5. RESPONSE message registry.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
61	1	3	1	1	0	1	4	100	38	0.0003	0.00003	0	0	0	0	0	0	0	0	0	200	38	0.0007	0.00007
:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
77	2	19	1	1	1	3	4	200	50	0.001	0.0001	0	0	0	0	0	0	0	0	0	200	50	0.001	0.0001
:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
130	2	15	1	1	1	1	1	100	50	0.001	0.0001	0	0	0	0	0	0	0	0	0	100	28	0.0003	0.00003
131	2	16	1	1	1	2	3	10000	10000	0.001	0.0001	5	10000	10000	0.001	0.0001	0	0	0	0	10000	10000	0.001	0.0001
132	2	17	1	1	1	1	0	1000	10000	0.001	0.0001	4	1000	10000	0.001	0.0001	0	0	0	0	100	10000	0.0003	0.00003
133	2	18	1	1	1	1	2	100	10000	0.001	0.0001	3	0	0	0	0	0	0	0	0	100	10000	0.0003	0.00003
134	2	19	1	1	1	2	2	200	50	0.001	0.0001	0	0	0	0	0	0	0	0	0	200	38	0.0007	0.00007
135	2	1	1	0	0	1	3	400	10000	0.0003	0.00003	4	0	0	0	0	0	0	0	0	0	0	0	0
136	2	2	1	0	0	2	1	400	10000	0.0007	0.00007	3	0	0	0	0	0	0	0	0	0	0	0	0
:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.



5 Conclusions

The Simulator presented follows NSIS signaling layer protocol specifications. Its prominent operative characteristics are: (a) Parameters value fitting mechanism. (b) Assign numeric equivalents to descriptive, or a combination of descriptive-numeric values. (c) NSIS entities construct and issue messages with parametric values derived from the entity's particular state.

Simulator's QoS policies follow ITU-T (Y.1540 and Y.1541) specifications. An adaptation was introduced to the QoS Classes table to allow for the use of the three NSIS QoS containers (desired, available and minimum).

The tests carried on the Simulator yielded a heuristic validation of its operation. Sessions 61 and 77 having external QSPEC with numerical and descriptive values and also different QoS level requirements were analyzed, to contrast the differences in processing. In both instances the QoS levels assigned were of inferior qualification, in accordance to the available resources at the Entities at the time of the local QSPEC arrival.

What appears to be a weakness in the reservation strategy of NSIS was detected. It comes out when an NSIS entity assigns a lower QoS level then requested, then this QoS is transported as a request to following entities, and depending on resources, the process can repeat, lowering furthermore the QoS level. This will be a drawback in large chains.

Continuation of this work is envisioned using the Simulator as a framework to test known and novel QoS-adaptation strategies, subject not specified by NSIS, but which constitutes the core of the heterogeneous to equivalent QoS-level conversion. Implementations of other QoS models (UMTS, IntServ, etc.) are also envisaged.

Acknowledgments. This work was financially supported by SIP-IPN grant 2009-0995 and COFAA-IPN Exclusivity Scholarship.

References

1. R. Braden, L .Zhang, S. Berson, S. Herzog, S. Jamin, "Resource ReSerVation Protocol (RSVP) Version 1 Functional Specification"; RFC2205, Septiembre 1997.
2. P. Pan, H. Schulzrinne, "YESSIR: A Simple Reservation Mechanism for the Internet", Computer Communication Review; Vol. 29; No. 2; April 1999.
3. J. Bergkvist, I. Cselenyi, D.Ahlard, "Boomerang-A Simple Resource Reservation Framework for IP" Internet-Draft; May 2001.
4. R. Sofia, R. Guerin, P. Veiga, "A Study of Over-reservation for Inter-Domain Control Aggregation Protocols", proceedings of ICNP, November 2002.
5. Bless,R; "Dynamic Aggregation of Reservations for Internet Services"; Proceedings of the Tenth International Conference on Telecommunication Systems- Modeling and Analysis (ICTSM 10); Vol. 1; pp. 26-38; October 2002.
6. H.Tschofenig, "NSIS Threats", draft-tschofenig-nsis-threats-01.txt, July 2002.
7. M. Marchese. QoS over Heterogeneous Networks. Italy: John Wiley & Sons, 2007.
8. R. Hancock, G. Karagiannis, J. Loughney, S. Van den Bosch, "Next Steps in Signaling (NSIS): Framework", RFC 4080; June 2005.
9. R. Ramachandran. L.Tujiao. "ITU-T Standards for IP-Based Networks", School of Mathematical, Statistics and Computer Science, Victoria University of Wellington, Wellington, New Zealand, 18 May 2006.
10. J. Ash. C. Dvorak. A. Morton. P. Tarapore, Y. El Mghazli, y S. Van den Bosch. "NSIS Network Service Layer Protocol QoS Signaling Proof-of-Concept". Internet Draft-concept-01, February 2004.
11. G.Ash, A.Bader, C. Kappler, D. Oran "QoS NSLP QSPEC Template", Internet-Draft-nsis-qspec18, april 2007.
12. N.Sanz. et al. "Movilidad en Redes Heterogéneas", Telefónica Investigación y Desarrollo, Universidad de Cantabria y Universidad Carlos III de Madrid, submitted for publication.

13. Gaxiola P.C.G., Quiroz M.E.E., Lepe A.O.I., "QoS Assignment with GSS for Incoming IP Sessions in UMTS", Journal Research in Computing Science Vol. 23, 2006. Mexico. (ISSN 1870-4069).
14. General Information of Visual Studio, Microsoft. November 2008. [On line] available at: http://msdn.microsoft.com/es-mx/vstudio/products/default.aspx
15. Network Management System: Best Practices White Paper 15114". CISCO. November 2008. [On line] available at:
http://www.cisco.com/application/pdf/paws/15114/NMS_bestpractic.pdf

Performance of a Realnetworks Video Streaming System over an Internet Infrastructure in Mexico

Mireya S. García-Vázquez[1], Alejandro A. Ramírez-Acosta[2]
and Miguel Colores-Vargas[1]

[1] Instituto Politécnico Nacional-CITEDI, Av. Del Parque No.1310, Tijuana BC,
colores, mgarciav@citedi.mx
[2] Dpto. R&D, PILIMTEC, Châteaugiron, Francia.
alramirez10@yahoo.fr

Abstract. Nowadays, streaming technology has become the most efficient way of distributing video over the Internet. The application of this technology requires a system of video diffusion, where the video quality received by users-clients depends largely on the server and the codec used. This paper presents the experimental evaluation of the RealNetworks video streaming system over an Internet infrastructure in Mexico. Each element that defines the diffusion architecture of video streaming is studied and its performance evaluated, particularly, the study is focuses in the characterization and analysis of the video streaming server performance. On demand and live are two types of streaming video analyzed in this work. The server performance is obtained based on the perception of quality video received by clients, the subjective metric used is the mean opinion square MOS. The results obtained provided information about optimal diffusion architecture of streaming video in Mexico.

Keywords: Streaming, networking, video, realnetworks.

1 Introduction

Currently, the video streaming systems have been implemented in many parts of the world and applied on a wide variety of applications. The penetration of this technology in the market has increased due to the advantages offered with respect to other means of diffusion and their interactivity. However, although this technology has taken an important place in some countries, in Mexico, the infrastructure and the bandwidth offered by Internet providers have not reached a level that gives support to the streaming systems in an efficient way. According to the OECD (Organization for Economic Co-Operation and Development) statistics [1], Mexico has the lowest rank places in Internet speed.

In order to obtain the best performance of an IPTV system (Internet Protocol TV) in Mexico under his present conditions of the infrastructure of telecommunications, it is important to make an analysis that takes into account the different parameters that influence the implementation of this system. Although there are several works in

© G. Sidorov, M. Aldape, M. Martinez, S. Torres. (Eds.)
Advances in Computer Science and Engineering.
Research in Computing Science 42, 2009, pp. 275-283

Received 15/03/09
Accepted 27/04/09
Final version 06/05/09

Latin America that evaluate some aspects of a streaming diffusion system [2-3], there is a lack of analysis that considers the different aspects as an integral system.

Of the three dominant commercial streaming media products (Microsoft Windows Streaming Media, RealNetworks RealSystems, and Apple QuickTime), RealNetworks [4] produces the most popular streaming media clients (Real Player) and servers (Helix Server) in the world. The Helix server supports multiple video diffusion formats and is available for the most operating systems. Therefore, the present work is focus on RealNetworks system solution. The aim is to experimentally measure the behavior of system under various Mexican network conditions with particular attention in the characterization and analysis of the video streaming server performance.

In this article the section 2 explain the video streaming technology; the section 3 describes the methodology for the experimental tests and the used metrics; section 4 analyzes the data obtained from experiments. Finally, section 5 summarizes our conclusions.

2 Video Streaming Technology

The streaming term represents a bundle of technologies that enable the PC or set top box for IPTV to deliver media files in real-time, with no download wait over the internet [5-6]. The content is read while it is stored in a video buffer. The streaming is base on a client-server model that allows multimedia data stream should arrive and play out continuously without interruption. The general principle of this set of technologies is that the audio/video content is coded according to a predefined format and bit-rate. Then, this coded content is sent via Internet. The audio/video bit-stream is fragmented to a series of network packets, which are sent out to the user, via Internet protocol (IP) [6]. The client can access to the video content via media player. This is an application that, while it memorizes a video or audio segment (~ 6 to 10 seconds of content), it display this information and so on. The figure 1, show the main stages involved on the streaming transmission process [5,6-7]. The streaming server can store content and/or deliver it to the clients. It can stream two types of diffusion [5,7]: *Streaming on demand* and *Streaming live*.

In general, streaming technologies support the latest digital media standards based on MPEG-4 AVC/H.264 [8-9].

3 Evaluation Methodology

The performance of a video streaming diffusion system depends largely on the performance of each element that composes the system, mostly the video codecs, network infrastructure and streaming servers used.

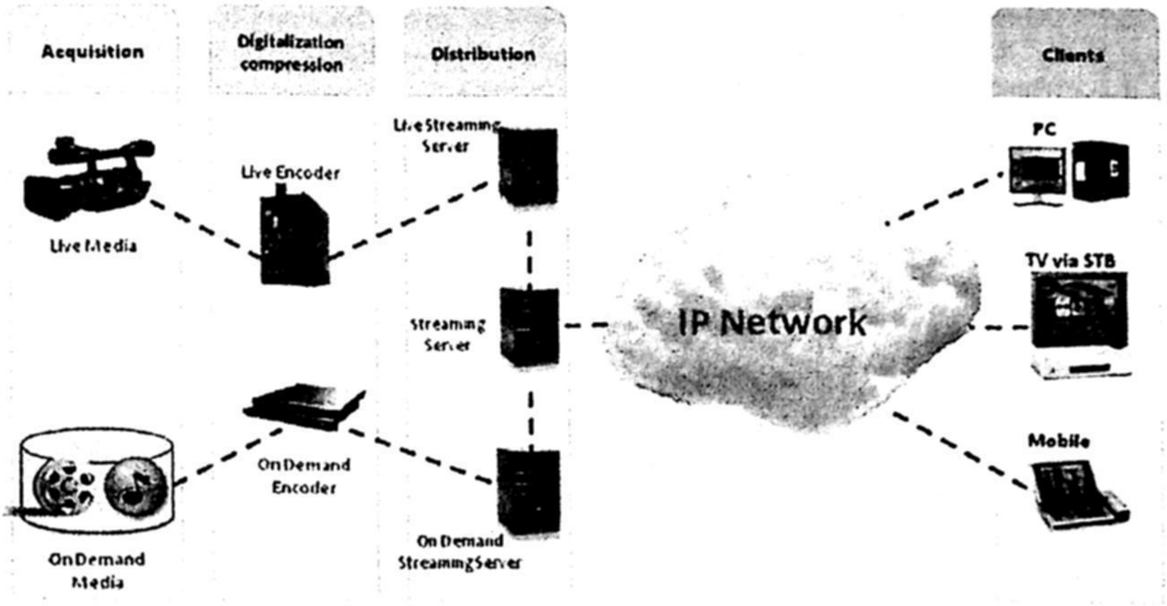


Fig. 1. Video Streaming Architecture.

3.1 Test Sequence Library

To evaluate codecs and streaming server we collect test material established and made available by the technical community. The test library contains original video sequences in ITU-R 601 format [10], raw video YUV 4:2:0 format. (see tables 1 and 2).

Table 1. Test sequences for QCIF resolution.

Resolution: QCIF (176x144 pixels). Application: video conference and video phone		
Sequence	Frames per second	Frame Numbers
Hall	30	300
News	30	300
Foreman	30	300

These sequences are widely used within of the video compression community due to their variation content of motion and texture [7].

Table 2. Test sequences for CIF resolution.

Resolution: CIF (352x288 pixels). Application: video web		
Sequence	Frames per second	Frame Numbers
Mobile	30	300
News	30	300
Stefan	30	300

3.2 Evaluation Metrics

3.2.1 Metric for Video Codecs

With the aim to compare codecs, we choose the PSNR (Peak Signal Noise Ratio) to measure the video quality and degradation of encoding sequences. PSNR is one of the most widespread objectives metric used to evaluate video quality on video codecs [7]. PSNR measures the error between a reconstructed image and the original one.

3.2.2 Metric for Streaming Server

The evaluation of the streaming servers, unlike the evaluation codecs, is based on the video quality received by clients under different restricting conditions of bandwidth. The measure of quality in the video was subjectively measured, that is, based on human perception. ITU [11] has proposed *Mean Opinion Score* (MOS) as a measure of perceived video quality. The participants are then asked to score the perceived quality of the shown media content from 1 ("worst") to 5 ("best"). The mean of the scores (MOS) provides a quantitative indicator of the perceived quality [11]. In this work we suggest an alternative evaluation for measuring performance server, analyzing strictly technical details that arise during testing of diffusion video streaming such as latency time to start, handling buffers, latency time in jumping, handling the bandwidth and the video quality [12]. The latter is valued according to the characteristics and defects presented in the pictures that make up the videos and test is conducted toward a possible mapping the value MOS. The following describes the mapping.

- MOS = 1, constant freezing on the image.
- MOS = 2, blocks effect, the image freezes and blurring.
- MOS = 3, constant video with small defects.
- MOS = 4, few robotic movements and blocks effect during abrupt changes of scene.
- MOS = 5, clients don't see video defects.

3.3 Platform Configuration

The number of supported streams, the protocol used for the application, the supported video formats and the compatibility with multiple operating systems are the most important features of the streaming servers. The streaming server Helix [4], evaluated in this study has the advantages over other servers: support of multiple video diffusion formats, version to unlimited number of clients and is available for the most operating systems known. To evaluate its performance, several tests were achieved on Local Area Network (LAN) architecture. The platform testing topology used for the experimental evaluation is shown in figure 2.

- **Video Streaming on demand.** When a video is requested to the server, it searches in its hard drive the requested file. The file has been previously encoded. Once located, it is encoded and packed in a suitable format. Then it is send to the client as streaming packets through the IP network.

- **Video Streaming Live.** The video is captured in raw format and sent it to the encoding tool in real-time. Subsequently, the encoded video is sent to the server. Then it sends to the client the file as streaming packets through the IP network. The client can access the content through the IP address set by the server, only during the event.

In both types of diffusion, the bandwidth in the reception side has been limited by software. The purpose of this is to simulate various network connectivity conditions of a public network (Internet) in Mexico. They were used application protocols (MMS, RTSP and HTTP). The server and clients were implemented on a two types of computers. The characteristics are shown in the table 3.

Table 3. Characteristics of equipment used.

Operating System: Microsoft Windows XP Pro V 5.1.2600 Service Pack 2
RAM Memory: 2.048,00 MB
Features of the computer for the streaming server
Processor: Dual x86 CORE™ 2 Authentic AMD ~2812 MHz
Features of the computers for the clients
Processor: Dual x86 Intel® CORE™2 Genuine Intel ~2400 MHz

3.4 Coding Process

The test sequences are coded at different bitrates (20, 56, 64, 128, 150, 256, 350, 500, 750, 1000, 2000 bps), QCIF/CIF resolutions [7] and three video streaming formats supported for Helix server: Real Media (.rm), Windows Media (.wmv) and Quick Time (.mov). It is crucial identify and use the codec that represent the best video quality with the fewest bits. For the video diffusion with the Helix servers usually used the following codecs:

- RealNetworks [7], Real Video 10 – This codec is suspected to be based on H.264 standard.
- Microsoft [7], Windows Media Series 9 Series 9.00.00.2980 – Codec, VC-1 standard, SMPTE 421M [13].
- Apple [7], Quick Time 7 Pro – Codec, H.264 standard.

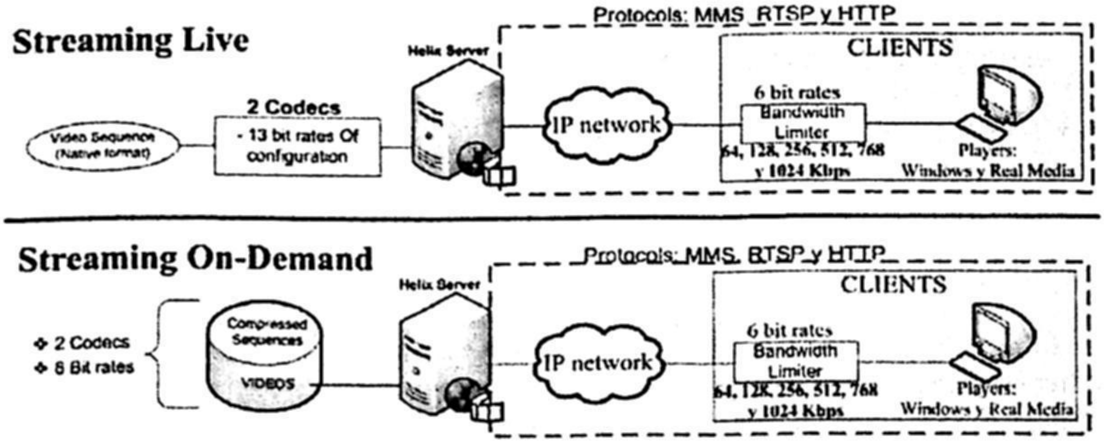

Streaming Live
Protocols: MMS RTSP y HTTP
Helix Server
CLIENTS
Video Sequence (Native format)
2 Codecs
- 13 bit rates Of configuration
IP network
6 bit rates
Bandwidth Limiter
64, 128, 256, 512, 768 y 1024 Kbps
Players: Windows y Real Media
Streaming On-Demand
Protocols: MMS, RTSP y HTTP
Helix Server
CLIENTS
✣ 2 Codecs
✣ 6 Bit rates
Compressed Sequences
VIDEOS
IP network
6 bit rates
Bandwidth Limiter
64, 128, 256, 512, 768 y 1024 Kbps
Players: Windows y Real Media

Fig. 2. Schematic of the platform used for testing.

4 Experimental Results

4.1 Streaming Server

In each video received by clients, it was measured subjectively the video quality to determinate the best platform over an Internet infrastructure in Mexico.

4.1.1 Evaluation for Streaming on Demand

The following points are the results of tests on streaming clients (Windows Media Player and Real Media Player).

Referring to the quality of the video: -For low bit rates (until 256 Kbps), the WMV format (MOS=3), showed the highest perceptive video quality. The RM format showed defects on the image (MOS=2). -For high bit rates (after 2000 Kbps), the sequences received as RM format had no defects (MOS=5). The diffusion tests with WMV format showed some defects (MOS=4).

As regards the bandwidth management: it was observed that clients Real Media maintain a dynamic bit rate value during transmission depending of the available bandwidth, i.e. send packages in advance when the bandwidth is allowed. In addition, when the videos are encoded at a bit rate higher than available bandwidth, the server slows down the speed of client-server connection to a value of less than 20 Kbps. With respect to the Windows client, it maintains the connection speed client-server close to the narrow bit rate of the request video, even where it is not possible to deploy any image (MOS = 1).

The video requests in RM format have a start time of almost zero. However, requests for video format WMV generate latency higher than eight seconds.

4.1.2 Evaluation for Streaming Live

The tests results for this type of diffusion were the following (Windows Media and Real Media).

Referring to the quality of the video: Similarity with the diffusion test of video streaming on demand, for video transmitted to low bit rates (until 256 Kbps), the WMV format presented the best subjective quality.

As regards the bandwidth management: There are differences in bandwidth management and planning when using different video formats. The RM format has values close to the dynamic bit rate encoding. By contrast, the bit rate is maintained at a stable value with WMV format.

Through evaluation tests, it was noted that the bit rate used to encode the videos, it must maintain a margin to be lower that available bandwidth, to obtain an acceptable visual quality for clients (MOS ≥ 3). The table 4 shows the maximum bit rate encoding used in diffusion tests to maintain fluently video for different conditions of bandwidth between client-server.

Table 4. Limited bandwidth.

Available bandwidth between server-client is: (kbps)	Optimal bit rate to encode and transmit is: (kbps)
100	80
128	100
150	120
256	220
350	310
500	460
750	700
900	835
1024	910

5 Conclusions

The aim of this paper was to present and to discuss the performance of a video streaming architecture based on Helix server and three compatible commercial proprietary codecs used in the video streaming technology over an Internet infrastructure in Mexico.

The tests realized with the different combinations from coders, players, protocols (MMS, RTSP, HTTP), and bandwidths among others parameters, gave as a result a hybrid architecture. This one is the one that better adapts to the infrastructure of Internet in Mexico (see figure 3), to offer a service of correct video streaming.

The tests of this architecture also confirmed the results of the codec evaluations [7]: for live broadcasting, the Windows Media Codec 9 Series had the best performance in both quality and compression. However, for diffusion on demand, Real Video codec 10 had the best performance.

This hybrid architecture presented an extraordinary performance in tests for both types of distribution of video streaming.

For live diffusion, the best solution consists of an encoder Windows Media Series 9. It sends the encoded content to Helix server. Then the server sends to the client the file as streaming packets through the IP network using MMS application protocol. The content may be received and reproduced by any client Windows Media.

For diffusion on demand, it is required to encode files with Real Producer RealVideo10 codec. It sends the encoded content to Helix server. Then the server sends to the client the file as streaming packets through the IP network using RTSP application protocol. The content may be received and reproduced by any client Real Media.

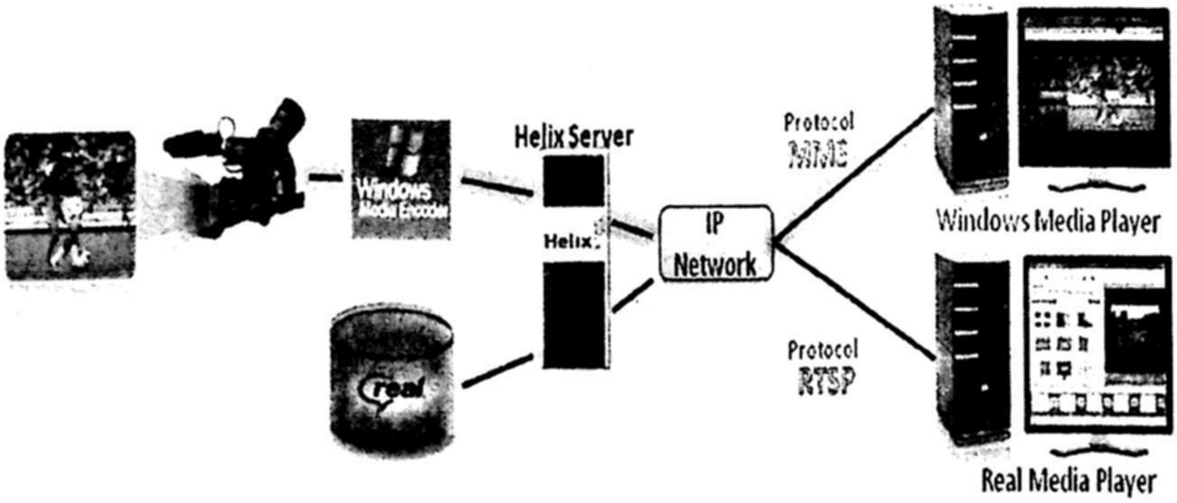


Fig. 3. Proposal architecture.

Acknowledgment. This work was support by IPN-SIP2008-09 project and PILIMTEC project.

References

1. Robert D, Atkinson, Daniel K. Correa and Julie A. Hedlund, "ITIF Broadband Rankings 2008", The Information Technology and Innovation Foundation, 2008.
2. C. González, "Desarrollo de un sistema de Videoteca Digital a través de Streaming", III Jornadas de Bibliotecas Universitarias. Arequipa, Perú, 2007.
3. M. Venegas, A. Yañez and A. González, "Transmisión de Video de alta Calidad a través de Redes IP Utilizando herramientas de Código Abierto". SENACITEL, XI Congreso Internacional de Telecomunicaciones, Chile, 2006.
4. Real Networks (Media servers), http://www.realnetworks.com/products/media_delivery.html Accessed march 2009.
5. David Austerberry, "Technology of Video and Audio Streaming". United States of America: Elsevier Science & Technology Books, 2004.
6. Wes D. Simpson, "Video over IP: A Practical Guide to Technology and Applications". Elsevier Focal Press Computers 2006, ISBN-13978-0-240-80557-3.

7. Mireya S. García, Alejandro A. Ramírez and Juan M. Colores, "MPEG-4 AVC/H.264 and VC-1 Codecs Comparison used in IPTV Video Streaming Technology", CERMA'08, Cuernavaca, Morelos, México. Sept. 30-Oct. 3, 2008. IEEE Computer Society ISBN: 978-0-7695-3320-9.
8. R. Koenen, "Overview of the MPEG-4 standard", International Organization for Standardization, 2002. http://www.chiariglione.org/mpeg/standards.htm, Accessed march 2009.
9. Iain E. G. Richarson, "H.264 and MPEG-4 Video Compression, Video Coding for the Next Generation Multimedia", John Wiley & Sons, 2003. ISBN-97-8047-0-869604.
10. ITU-R, "Encoding parameters of digital television for studios", Recommendation BT. 601-4, 1994.
11. ITU-R Rec. BT. 500, "Methodology for the subjective assessment of the quality of television pictures", Geneva, 2002.
12. Juan M. Colores, "Estudio comparativo de sistemas de diffusion de video afluente". Master Thesis. IPN-CITEDI, Tijuana, México, Sept. 2008.
13. SMPTE 421M, "VC-1 Compressed Video Bitstream Format and Decoding Process".

Digital Signal Processing

Two Reconstruction Algorithms of Non Gaussian Processes on the Output of a Polynomial Converter

V. Kazakov and Y. F. Olvera

Department of Telecommunications, SEPI-ESIME, IPN, México D.F.
vkazakov41@hotmail.com, yairfom@hotmail.com

Abstract. Two extrapolation algorithms are investigated for reconstruction procedures of non Gaussian processes. We investigate the process on the output of polynomial converter driven by Gaussian process. The optimal reconstruction algorithm is investigated on the basis of the conditional mean rule with help of cumulant functions. In this case the error reconstruction function depends on the given samples. Another algorithm is non optimal, because the reconstruction operation is realized by using of a covariance function of the output process only. The case with the polynomial of the third order is investigated in detail.

1 Introduction

The problem of reconstruction of a signal that passes through determinate samples has been investigated since XIX century. The classical Sampling Theorem is usually associated with the names of Whittaker, Kotelnikov and Shannon (or WKS theorem), and it has been proved for deterministic functions with the limited spectrum. This classical theorem has been generalized on stochastic stationary processes by A. Balakrishnan [1]. Following Balakrishnan´s theorem [1] all types of random stationary processes with a limited power spectrum can be reconstructed without error by the unique reconstruction function sin x / x when the number of samples is equal to infinity. But some important characteristics like the Probability Density Function (pdf), limit number of samples and high order moments are not mentioned in this theorem. In fact, there are some publications where recommendations of Balakrishnan´s theorem are applied for the cases with limited number of samples and with an arbitrary pdf of processes.

In the present paper we analyze the statistical description of Sampling-Reconstruction Procedure (SRP) of non Gaussian process. This process is formed on the output of a converter with a non linear polynomial characteristic. The process in the input is Gaussian Markovian. We take an extrapolation case (only one sample).

The first algorithm is optimal. This algorithm is based on the conditional mean rule [2]. The evaluation of a reconstructed realization is formed by the conditional mean. This evaluation provides a minimum mean square error which

© G. Sidorov, M. Aldape, M. Martinez, S. Torres. (Eds.)
Advances in Computer Science and Engineering.
Research in Computing Science 42, 2009, pp. 287-295

Received 01/04/09
Accepted 27/04/09
Final version 07/05/09

is described by the conditional variance. Using the method suggested in [3], we obtain the conditional mean and the conditional variance of the output non Gaussian process on the bases of conditional cumulant functions [4].

The second algorithm is non optimal, because the reconstruction operation is based on the covariance function of the output process only. In other words, the extrapolation procedure of non Gaussian process is formed like the extrapolation function of Gaussian process does.

We analyze two mentioned variants for the polynomial characteristic of the third order in detail. The conclusion is: it is necessary to take into account the pdf of sampled process in the statistical SRP description of random processes.

2 Features of the Process on the Output of a Polynomial Converter

In a polynomial case, in order to find the right methodology that describes the properties of the non-linear converter, we assume the non-linearity expressed by the formula:

$$\eta(t) = g[x(t)] = a_o + a_1 x(t) + a_2 x^2(t) + ... + a_n x^n(t) \qquad (1)$$

where $a_i (i = 0, 1, 2, ..., n)$ are constants.

We choose three variants transfer functions of third order:

$$\eta(t) = g(\xi) = \xi^3 \qquad (2)$$

$$\eta(t) = g(\xi) = \xi^3 + 2\xi \qquad (3)$$

$$\eta(t) = g(\xi) = 1.6\xi^3 + 5.1\xi \qquad (4)$$

they are shown in the Fig. 1.

The expressions (2) – (4) have the unique inverse functions $\xi(t) = h(\eta(t))$.

We suppose that the process $\xi(t)$ is Gaussian Markov with the covariance function:

$$k(\tau) = \sigma^2 \exp(-\alpha \mid \tau \mid) \qquad (5)$$

and the mathematical expectation is zero.

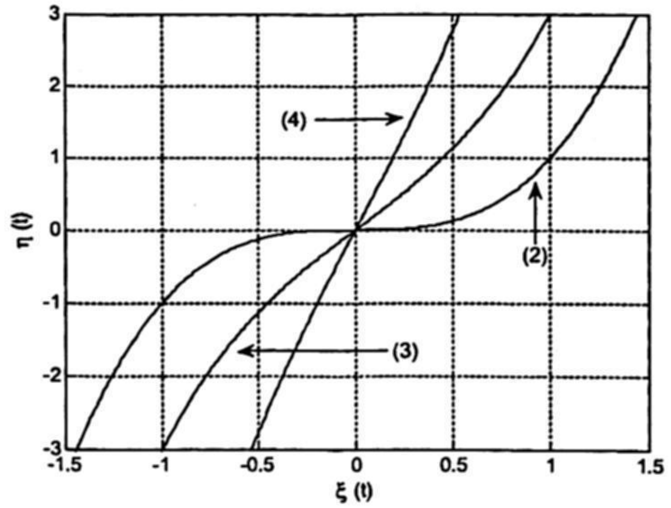


Fig. 1. Polynomial Non-linearity

We put $\sigma^2 = 1$. Firstly we determinate the unconditional characteristics of the process $\eta(t)$.For getting an one-dimensional pdf for the transfer functions, we use the next methodology [3]:

$$w(\eta) = w(h(\eta))\frac{1}{\mid \frac{d\eta}{d\xi} \mid} = \frac{1}{\sqrt{2\pi}} \exp\{-\frac{1}{2}\xi^2\}\frac{1}{\mid \frac{d\eta}{d\xi} \mid} \tag{6}$$

The graphs of these functions are presented in Fig. 2. Knowing pdf (6), we can calculate all required moments of the output process. It is clear that the mathematical expectation $\langle \eta(t) = 0 \rangle$ for all types of non linearity, formula (7). The variance calculations give the values: 15, 31 and 113.35 for the cases (2), (3) and (4) respectively, formula (8).

$$m_n = \langle \eta \rangle = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\{-\frac{1}{2}\xi^2\} \cdot g(\xi) \cdot d\xi \tag{7}$$

$$\sigma_\eta^2 = \langle \eta^2 \rangle = \langle \eta \rangle^2 = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\{-\frac{1}{2}\xi^2\} \cdot [g(\xi)]^2 \cdot d\xi \tag{8}$$

Rewriting (1) for two time moments t and t+τ, multiplying both expressions and applying the average operation, we can find the covariance function of the output process, restricting by the order 3, by the following formula [4]:

$$K_\eta[\tau] = \nu_1^2 K_\xi[\tau] + \frac{\nu_2^2}{2!} K_\xi^2[\tau] + \frac{\nu_3^2}{3!} K_\xi^3[\tau] \tag{9}$$

where the coefficients ν_n are determined by [4]:

$$\nu_n(m_\xi, \sigma_\xi^2) = n! a_n \tag{10}$$

Although the extrapolation algorithm only depends on the sample $\eta(T_n)$. The output covariance function is showed in Fig. 3.

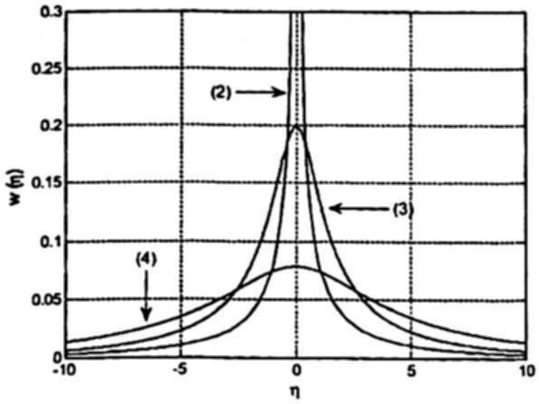


Fig. 2. Probability Density Function

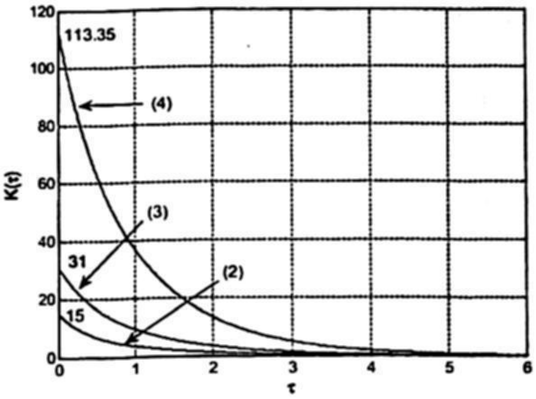


Fig. 3. Covariance output function

In Fig. 2 we can see the dispersion of the possible values. The covariance function grows according the dispersion on the fpd.

3 The Optimal Extrapolation Reconstruction Algorithm

Let us fix a set of samples Ξ =$[\xi_1, \xi_2, ..., \xi_n]$. Owing to the unique inverse functions $\xi(t) = h(\eta(t))$ we find the corresponding set η =$[\eta_1, \eta_2, ..., \eta_n]$. Then, we can apply the statistical conditional average operation to both parts of the expression (1). In result we have the expression for the reconstruction function:

$$\tilde{m}_1^\eta(t) = a_0 + a_1\tilde{m}_1^\xi(t) + ... + a_n\tilde{m}_n^\xi(t) \tag{11}$$

where $\tilde{m}_1^\eta(t)$ is the conditional mathematical expectation of the output process, $\tilde{m}_i^\eta(t)(i = 1, ..., n)$ are the conditional moment functions of the order i of the input process.

Let us calculate the square of both parts of the expression (1) and fulfill the conditional average operation, this yield:

$$\tilde{m}_2^\eta(t) = a_0^2 + a_1^2\tilde{m}_2^\xi(t) + ... + a_n^2\tilde{m}_{2n}^\xi(t) + 2a_0a_1\tilde{m}_1^\xi(t) + ... + a_{n-1}a_n\tilde{m}_{2n-1}^\xi(t) \tag{12}$$

Knowing (11) and (12) we can find the required conditional variance or the error reconstruction function:

$$\tilde{\sigma}_\eta^2(t) = \tilde{m}_2^\eta(t) - [\tilde{m}_1^\eta(t)]^2 \tag{13}$$

The equations (11) – (13) show that the reconstruction function and the error reconstruction function require the high order conditional moments on the input. In order to determine the output moment functions we can apply some connection expressions between moments and cumulants [4,5]. The Gaussian pdf is described by the two first cumulants. Then, the relations between the conditional moments $\tilde{m}_1^\xi(t)$ on the input and the conditional first $\tilde{k}_1$ and second cumulants $\tilde{k}_2$ are:

$$\begin{aligned}
\tilde{m}_1^\xi(t) &= \tilde{k}_1 \\
\tilde{m}_2^\xi(t) &= \tilde{k}_2 + \tilde{k}_1^2 \\
\tilde{m}_3^\xi(t) &= 3\tilde{k}_2\tilde{k}_1 + \tilde{k}_1^3 \\
\tilde{m}_4^\xi(t) &= 3\tilde{k}_2^2 + 6\tilde{k}_2\tilde{k}_1^2 + \tilde{k}_1^4 \\
\tilde{m}_5^\xi(t) &= 15\tilde{k}_2^2\tilde{k}_1 + 10\tilde{k}_2\tilde{k}_1^3 + \tilde{k}_1^5 \\
\tilde{m}_6^\xi(t) &= 15\tilde{k}_2^3 + 45\tilde{k}_2^2\tilde{k}_1^2 + 15\tilde{k}_2\tilde{k}_1^4 + \tilde{k}_1^5
\end{aligned} \tag{14}$$

The conditional mathematic expectation $\tilde{k}_1$ and the conditional variance $\tilde{k}_2$ are expressed by the formulas [6]:

$$\tilde{k}_1 = \tilde{m}^\xi(t) = m^\xi + \sum_{i=1}^{N}\sum_{j=1}^{N} K_\xi(t - T_i)a_{ij}[\xi(T_j) - m^\xi(T_j)] \tag{15}$$

$$\tilde{k}_2 = \tilde{\sigma}_\xi^2(t) = \sigma_\xi^2 - \sum_{i=1}^{N}\sum_{j=1}^{N} K_\xi(t - T_i)a_{ij}K_\xi(T_j - t) \tag{16}$$

a_{ij} is an element of the inverse covariance matrix:

$$\mid a_{ij} \mid = \mid K_\xi(T_i, T_i) \mid^{-1} \tag{17}$$

From (11) - (16) one can see that the conditional variance $\tilde{\sigma}_\eta^2(t)$ depends on the values of samples, but in a Gaussian case it does not.

We consider a particular case when the set of samples has only one term . Let us choose the following number and values of input $\xi(T)$ and output $\eta(T)$ samples, on Table 1.

Table 1. Input and Output samples

	$\xi(T)$	$\eta(T) = \xi^3$	$\eta(T) = \xi^3 + 2\xi$	$\eta(T) = 1.6\xi^3 + 5.1\xi$
a)	0.4	0.064	0.864	2.142
b)	0.7	0.343	1.743	4.119
c)	1	1	3	6.7
d)	1.3	2.197	4.797	10.15
e)	1.5	3.375	6.375	13.05

The results of the reconstruction function are presented in Fig. 4.a, 5.a and 6.a for (2), (3) and (4) respectively, we can observe that they have a nonlinear performance. The error reconstruction function is presented in Fig. 4.b, 5.b and 6.b for (2), (3) and (4) respectively, all those curves converge at the value of their variance, they characterize the minimum error for the optimal reconstruction algorithm of the extrapolation type, because of the output process is not Gaussian, the error reconstruction depends on the sample value.

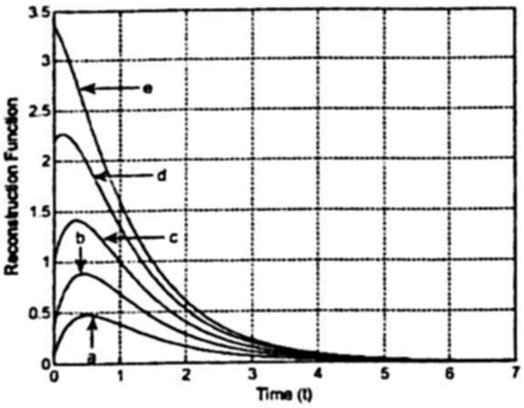


Fig. 4.a. Optimal Reconstruction Function for $\eta(t) = \xi^3$

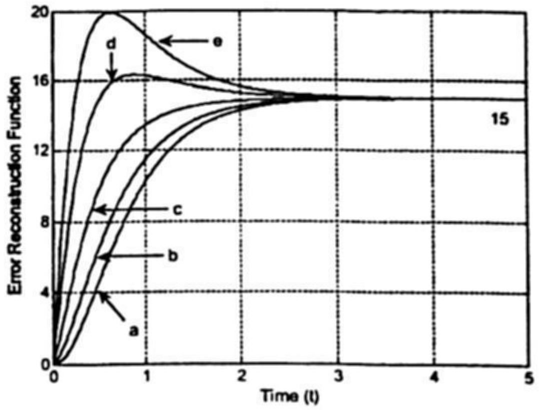


Fig. 4.b. Optimal Error Reconstruction Function for $\eta(t) = \xi^3$

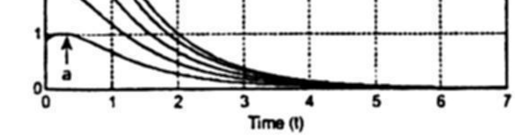


Fig. 5.a. Optimal Reconstruction Function for $\eta(t) = \xi^3 + 2\xi$

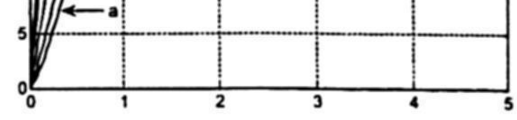


Fig. 5.b. Optimal Error Reconstruction Function for $\eta(t) = \xi^3 + 2\xi$

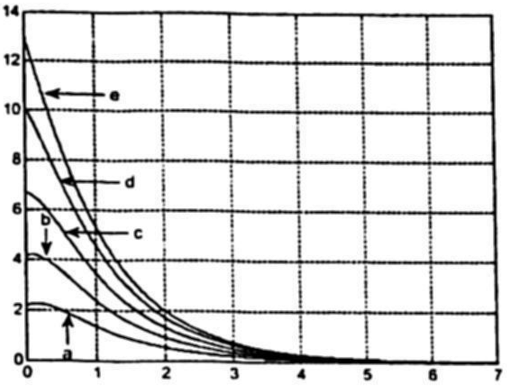


Fig. 6.a. Optimal Reconstruction Function for $\eta(t) = 1.6\xi^3 + 5.1\xi$

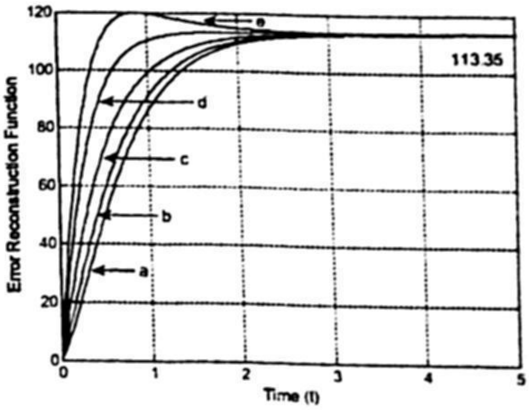


Fig. 6.b. Optimal Error Reconstruction Function for $\eta(t) = 1.6\xi^3 + 5.1\xi$

All those groups of curves have the clear physically significances, and explain the sense of the non Gaussian distribution of the output process.

4 Non Optimal Extrapolation Reconstruction Algorithm

The non optimal algorithm is based on the knowledge of the covariance output function only. We just need the Gaussian approach to describe the reconstruction, so we must to apply the next formulas:

$$\tilde{m}^{\eta}(t) = m^{\eta} + \sum_{i=1}^{N}\sum_{j=1}^{N} K_{\eta}(t-T_i)a_{ij}[\eta(T_j) - m^{\eta}(T_j)] \tag{18}$$

$$\tilde{\sigma}_{\eta}^2(t) = \sigma_{\eta}^2 - \sum_{i=1}^{N}\sum_{j=1}^{N} K_{\eta}(t-T_i)a_{ij}K_{\eta}(T_j - t) \tag{19}$$

when N=1, we have:

$$\tilde{m}^{\eta}(t) = m^{\eta} + \frac{K_{\eta}(t-T_1)}{\sigma_{\eta}^2}[\eta(T_1) - m^{\eta}(T_1)] = \frac{K_{\eta}(t-T_1)}{\sigma_{\eta}^2}[\eta(T_1)] \tag{20}$$

$$\tilde{\sigma}_{\eta}^2(t) = \sigma_{\eta}^2[1 - R_{\eta}^2(t-T_1)] \tag{21}$$

where $R_{\eta}(\tau) = \frac{K_{\eta}(\tau)}{\sigma_{\eta}^2}$

Substituting (5) – (10) on the last formulas, we obtain the reconstruction function that is a group of linear curves; and the error reconstruction function which is only one curve that converge to their variance, as in optimal algorithm.

There is just one curve for each transfer function because this method doesn't take the samples for finding the error reconstruction. The graphics for the reconstruction function are in the Fig. 7, they are the same for (2), (3) and (4) using a) $\xi(t) = \eta(t) = 0.4$, b) $\xi(t) = \eta(t) = 0.7$, c) $\xi(t) = \eta(t) = 1$, d) $\xi(t) = \eta(t) = 1.3$, e) $\xi(t) = \eta(t) = 1.5$; this is because of the lineal method. The graphics of the error reconstruction function are showed in the Fig. 8 for (2), (3) and (4).

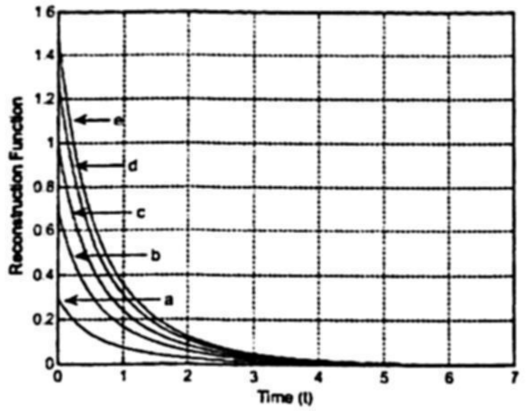


Fig. 7. Non Optimal Reconstruction Function

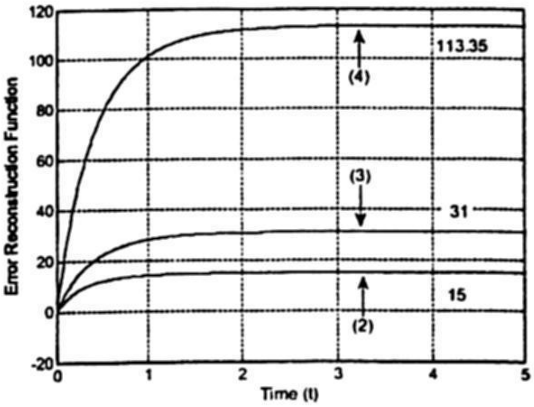


Fig. 8. Non Optimal Error Reconstruction Function

The non optimal reconstruction is characterized by one error curve (Fig. 8). This fact is connected with the Gaussian approximation of the output process. The real situation is another; the non Gaussian pdf of the output process determines absolutely another behavior of the error reconstruction curves. The error reconstruction function depend on the samples, the Fig 4.b, 5.b and 6.b show that the big samples $(\xi(T_1 > 1))$ have bigger error reconstruction function. It means that it is necessary to take into account the pdf of sampled processes.

5 Conclusions

Two different reconstruction algorithms are investigated for output processes of non linear polynomial converters driven by Gaussian Markovian process. Both principal characteristics (reconstruction function and error reconstruction function) are obtained. Comparison of these algorithms shows that it is necessary to take into account pdf of sampled process. The non optimal algorithm gives non correct results. And, the optimal case is an easier methodology to use.

References

1. A. Balakrishnan: A note on the Sampling Principle for Continuous Signals. IRE Trans. On IT, IT-3, (1957).
2. P. E. Pfeiffer: Probability for Applications. Springer Verlag, (1990).
3. V. A. Kazakov, Regeneration of samples of random processes following non linear inertialess convertions . Telecommunication and Radio engineering, Vol. 43, pp. 94-96, No. 10, (1988).
4. A. N. Malakhov: Cumulant analysis of non Gaussian processes and their transformations. Moscow, "Sov. Radio", (1978).
5. A. Stuart, K. Ord: Kendall´s advanced theory of statistics. VI edition, vol. I, Distribution Theory, Arnold Edition, London (1994).
6. R. L. Stratonovich: Topics in the Theory of Random Noise. New York, Gordon and Breach, (1963).

What is the Required Series Length for Correct Self-similarity Analysis?

Julio C. Ramírez Pacheco[1,2] and Deni Torres Román[1]

[1] CINVESTAV Unidad Guadalajara, Av Científica 1145, 45010
Col. el Bajío, Zapopán, Jalisco, México
cramirez@gdl.cinvestav.mx, dtorres@gdl.cinvestav.mx
[2] Universidad del Caribe, SM 78, Manzana 1, lote 1, 77528, Cancún, QRoo, México
jramirez@ucaribe.edu.mx

Abstract. It is well-known that *self-similar* and *Long-memory* signals appear in many fields of science. LAN, VBR sources, WWW traces, wireless traffic, etc. are among the ones with this behaviour in computer networking. An important question in these applications is how long a measured trace should be to obtain reliable estimates of the *Hurst*-index. This paper addresses this question by first providing a thorough study of estimators for short series based on the behaviour of bias, σ, $\sqrt{\mathrm{MSE}}$ and convergence when using *Gaussian H-sssi* signals. Results show that *Whittle*-type estimators behave the best when estimating H for short signals. Based on the results, empirically derived minimum trace length for the estimators is proposed. Finally for testing the results, the application of estimators to real traces is accomplished. Inmediate applications from this can be found in the real-time estimation of the *Hurst*-index which is useful in agent-based control of QoS parameters.

1 Introduction

Self-similar stochastic processes are the ones which present scale-invariant statistical behaviour [27] [12] [5]. These processes are widely applied as models for different types of phenomena in a wide range of fields of science [23] [21] [22] [14] [24]. In the computer networking area, these processes are used for the modelling of aggregate LAN, VBR video, wireless and WWW traffic among others [17] [6] [8] [20] [19]. In these studies, traffic was measured and then analyzed in order to find whether it fits the *self-similar* model or not. The traces used in these studies consisted of hundreds of thousands of points in the LAN case and nearly 100000 for each VBR video trace. Obtaining such lengths usually implies a long measurement time. For off-line study, the above lengths are acceptable while for applications of real-time administration of QoS metrics based on accurate *Hurst*-index estimation, the above are unacceptable. The paper first studies the behaviour of estimators to short time series and then adresses the problem of obtaining the minimum length required for accurate real-time estimations of the *Hurst*-index. The requirement is thus obtaining high accuracy with minimum length. The accuracy shoud be comparable with that of long series, where accuracy in this case, is based on metrics such as standard deviation,

© G. Sidorov, M. Aldape, M. Martinez, S. Torres. (Eds.)
Advances in Computer Science and Engineering.
Research in Computing Science 42, 2009, pp. 297-307

Received 08/03/09
Accepted 27/04/09
Final version 06/05/09

bias and $\sqrt{\text{MSE}}$. Convergence analysis is also a useful tool for accomplishing the above. Thus, the paper addresses the following issue: given specifications of bias, variance or MSE. what the series length N should be?, i.e.. suppose stochastic process Ψ posseses *Hurst*-index H, find the minimum length N_{min} such that for every realization(of length N_{min}) the estimated *Hurst*-indexes $\hat{H}$ are similar to that of H. For accomplishing the above, the paper is organized as follows. Section 2 briefly summarizes *self-similar* stochastic processes and estimation methods. Section 3 provides description of the methodology used for finding the minimum value N while section 4 presents a detailed study of the behaviour of estimators to short time series, the problem of finding the minimum length for the estimators and the application of these results to real LAN traces. Finally section 5 concludes the paper.

2 Self-*similar* Signals and Estimation of *H*

Self-similar processes are the ones whose distributional properties are invariant to dilations in time and suitable compression of amplitude. Let $Z = \{Z_t\}_{t \in I}$, where $I = \mathbb{R}$ or $\mathbb{R}_+$, be a real-valued stochastic process, it is said that Z is ***self-similar iff*** there exist an $H \in \mathbb{R}$ such that for any $a \in \mathbb{R}_+$ the following holds $\{Z_{at}\}_{t \in I} \overset{d}{=} \{a^H Z_t\}_{t \in I}$, where $\overset{d}{=}$ is in the distributional sense. Usually, the interest is in *H-ss* processes with stationary increments for which the above holds with $H > 0$. The above definition is called the strict one. A relaxed version of the above is obtained by the second-order definition one which requires invariance on second-order statistical properties under scaling. Formally let Z_t be a continuous-time stochastic process, it is said that Z_t is second-order *self-similar* if $EZ_t = a^{-H} EZ_{at}$ and $R_{zz}(t, s) = a^{-2H} R_{zz}(at, as)$. Computer networking requires discrete-time models, then discrete versions of the above are needed. Let $X = \{X_t, \in \mathbb{Z}\}$ be a discrete-time process, possibly obtained by sampling a continuous time random signal, it is said that X is strictly ***self-similar iff*** there exists an $H \in (0, 1)$ such that for any $m \geq 1$ $X \overset{d}{=} m^{1-H} \Gamma_m(\{X\})$. $\Gamma_m(.)$ is the block aggregation process which receives as input a length N time series and outputs an length N/m time series. A relaxed version of strict discrete *self-similarity* is given by second-order *self-similarity* in the exact sense. Let X be a discrete-time stochastic process, it is said that X is exact second-order *self-similar* if its correlation coefficient satisfies

$$\rho(k) = \frac{1}{2}\{(k+1)^{2H} - 2k^{2H} + (k-1)^{2H}\}. \tag{1}$$

A process having a correlation function of the above form also satisfies $\text{Var}(X) = m^{2-2H}\text{Var}(\Gamma_m(\{X\}))$ and $\text{Cov}(\Gamma_m(X_t), \Gamma_m(X_{t+k})) = m^{2-2H}\text{Cov}(X_t, X_{t+k})$. In computer networking an even more relaxed version of (1) is used. A process X is said to be asymptotically second-order *self-similar* if the correlation function of $\Gamma_m(\{X\})$ as $m \to \infty$ is equal to that of an exact second-order *self-similar* stochastic process in discrete-time i.e., equal to (1). If in the exact or asymptotic definition of discrete *self-similarity* we let $k \to \infty$, then, $\rho(k) = ck^{2-2H}$

which implies *long-range* dependency. It means an asymptotic or exact *self-similar* process is *long-range* dependent provided $H \in (0,1)$ and $k \to \infty$. The greater the parameter H the smoother the process is and the slower the decay to zero of the autocorrelations. Several methods of estimation of the parameter H have been proposed, the methods can be classified as time-domain, frequency-domain and time-scale methods. Among the time-domain methods it is found the R/S statistic [18] [13], variance-time plot(dispersional analysis), variance of residuals(DFA), absolute moment, MAVAR, Higuchi's method, scaled window variance [7], Whittle, etc [26] [25]. GPH, Periodogram and other modified periodograms methods are found in the frequency-domain class which in turn take advantage of the power law behaviour of the *self-similar* processes near the origin. Time-scale methods include wavelet based estimators such as Abry-Veitch estimator [2] [3] [1] [4] [28]. Software tools for *self-similarity* analysis are also important since they collect a number of estimators and methodologies for improving the anaysis of *self-similarity*. In this context, studies have demostrated that `fARMA` and `SelQoS` are the most robust and accurate ones while a widely used one, `Selfis` [16] [15], is inaccurate and not robust. This paper, makes use of the R package `fARMA` for obtaining the results of estimations. The choice of `fARMA` is based on the excellent programming capabilities of the `S` language.

3 Methodology

We first provide a detailed study of the behaviour of estimators to short series and then propose minimum lengths for the estimators. The behaviour of estimators is studied by applying a given estimator to N time series and then obtaining BIAS, σ and $\sqrt{\text{MSE}}$. Comparison of the behaviour of the estimators subject to these statistics and for varying length aids in finding the minimum length. Also a convergent analysis shows the evolution of estimators in time and will be useful in this paper. Next we describe these steps in more detail. In order to apply the estimators, N time series should be obtained. Synthetic signals with known H are obtained by the simulation of Gaussian *H-sssi(fGn)* series using the Davies and Harte method [9]. The considered lengths for the traces were $N = \{2^i, i = 6, 7, \ldots, 16\}$ and for each length 100 traces with *Hurst*-index $H \in \{0.5, 0.6, 0.7, 0.8, 0.9, 0.99\}$ were generated. Thus, a total of 6600 'exact' fractal signals were generated. For each set of estimations of a particular H, the following statistics are computed BIAS $= H_0 - \bar{X}$, where H_0 is the nominal value, standard deviation σ and $\sqrt{\text{MSE}} = N^{-1} \quad {}_{i=1}^{N} (x_i - H_0)^2$. Based on the results of BIAS, σ and $\sqrt{\text{MSE}}$ a minimum length is proposed. The length N_{min} is obtained from accurate estimations(BIAS ~ 0.03,$\sigma \sim 0.015$). In addition, the classification of estimations based on these values(that of BIAS and σ) is also proposed. *High accuracy* estimations are obtained when BIAS ≤ 0.03 and $\sigma \leq 0.015$, *acceptable* estimations when BIAS $\in (0.03, 0.05)$ and $\sigma \leq 0.02$ and *biased* estimations when BIAS > 0.1. Once the minimum length is obtained for *fGn*-type series, the application of these results is performed to long synthetic and real traces. For such series Z of length M, $M >> N_{min}$, the following is studied: let $t_0, t_1, \ldots, t_k$ be a sequence of points in the x-axis, where $t_{i+1} > t_i$ and $(t_{i+1} - t_i) < N_{min}$, to each block of Z of length N_{min}, $\{Z_j\}_{j=t_i}^{t_i+N_{min}-1}$, apply a *Hurst* estimation method $\Theta_{t_i}^{N_{min}}(.)$ to these blocks. Repeat until $t_k + N_{min} > M$ for any k. A plot of t_i versus $\Theta_{t_i}^{N_{min}}(.)$ should result in a signal with little variation(the variation should conform σ) if N_{min} is correctly set. The proposed length N_{min} is related to convergence analysis of a series, which is also studied. Convergence of any estimator is obtained by first partitioning the original series Z into blocks of size $m << M$ to obtain $Z = \{\Psi_1^m, \Psi_2^m, \ldots, \Psi_i^m\}$, where $\Psi_i^m = \{Z_{(i-1)m}, Z_{(i-m)m+1}, \ldots, Z_{im}\}$. Next apply a Hurst methodology $\Theta(.)$ to $\cup_{j=1}^{N/m}\{\Psi_i^m\}$, $j = 1, 2, \ldots, N/m$ to obtain $\Theta_1^m, \Theta_1^{2m}, \ldots, \Theta_1^{jm}$. Plot Θ_i^{jm} versus jm for $j = 1, 2, \ldots, N/m$ to visualize the convergent behaviour of estimator $\Theta(.)$.

4 Simulation Results

4.1 Perspective, Bias and σ Plots

Figure 1 shows a perspective plot of estimations of the *Hurst*-index for traces with $H = 0.90$ and for varying length when applying *wavelet*-based and Whittle techniques. Left plot corresponds to wavelet technique while right plot to Whittle. Note from the left plot that wavelet-based techniques experience high bias and variability when estimating the *Hurst*-index for short time series. The length of the traces for these highly variable estimations is in the order of $N < 2^{12}$. When $N \in (2^{11}, 2^{12})$, the estimations show aceptable results and only when $N > 2^{13}$, the estimations are highly accurate. Similar behaviour is obtained for traces with *Hurst*-index value different that $H = 0.90$. The same kind of plot when using *Whittle*-type estimations is shown in right plot of the figure for traces

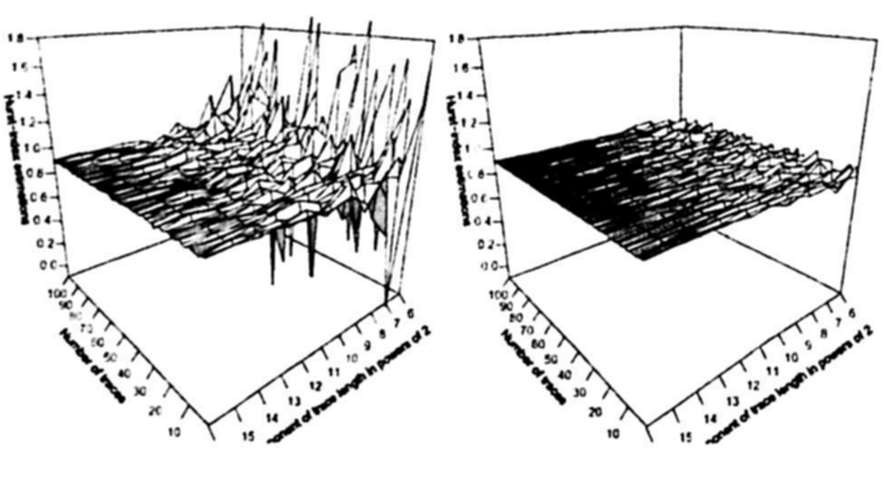


Fig. 1. Wavelet and Whittle Estimations for 100 fGn traces with $H = 0.90$, left plot corresponds to Wavelet estimations and the right to Whittle-type estimations.

with $H = 0.90$. As can be seen from the figure, in general, Whittle-type estimations present high accuracy and low variability for short *long-range dependent* time series. For values of H below 0.90, the estimated *Hurst*-index present high variability when $N < 2^{10}$ and when $H \geq 0.90$, generally, the estimations are accurate for $N > 2^9$. Comparing the results of *Whittle*-type estimators with that of *wavelet*-based techniques is seen that the *Whittle* ones are more robust to short series in the context of exact simulated *fGn* traces. Figure 2 shows a perspective plot of estimations of the *Hurst*-index for traces with $H = 0.90$ and for varying length when applying periodogram and R/S Statistic method. Recall that periodogram method is based on the behaviour of *long-memory* series' PSD near origin. As seen in the left plot of the figure, the periodogram method behaves similarly to wavelet method for short series, i.e., with high bias and variability. A similar behaviour is obtained when using different values of the *Hurst*-index H. The right plot of figure also shows the same kind of plot as the Periodogram for R/S algorithm. Note from the plot that R/S method present high variablity no matter what the length of the trace is. The variablity diminishes when the lengths of the traces are longer but it is difficult to establish a minimum length for accurate estimations of H. In contrast to the other methods, the surface in the R/S perspective plot is always rough. The results then imply that additional analysis methods should be applied to R/S statistic method in order to deal with this biased behaviour. Finally, figure 3 shows the perspective plots for variance-type method. Variance-type method presents a similar behaviour to those of wavelet and periodogram methods. Unless wavelet and periodogram, the range of accurate estimations for variance-type in the *fGn* context is different. For accurate estimations, the length of the trace should be at least $N \geq 2^{15}$.

Figure 4 shows the bias experienced by every method studied for $H = 0.90$.

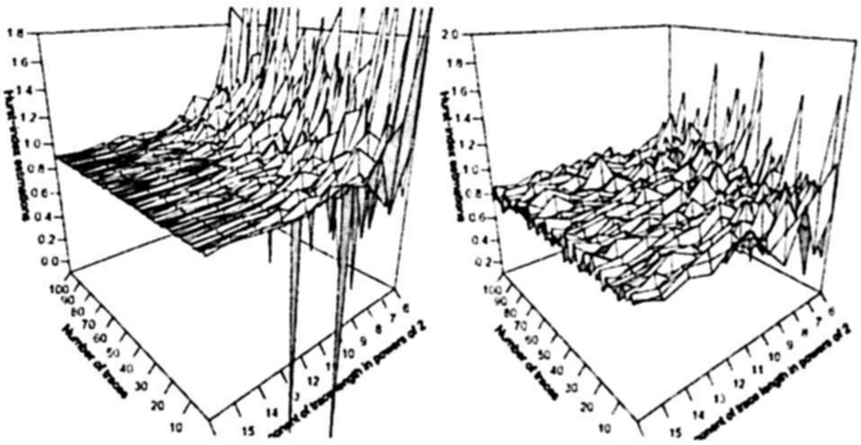


Fig. 2. Periodogram and R/S Statistic estimations for 100 fGn traces with $H = 0.90$, left plot corresponds to Periodogram estimations and the right to R/S Statistic.

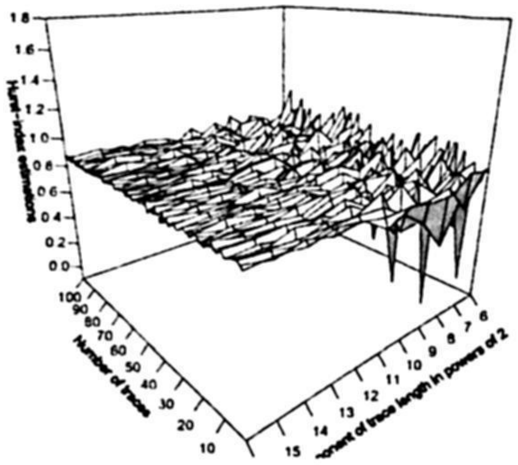


Fig. 3. Variance estimations for 100 fGn traces

Note from the figure that Whittle and Wavelet methods are the methods whose behaviour for short series is better than the others. Unless Wavelet, Whittle method behaves with less irregularity for short series and for $N \geq 2^8$, the biases are not significant. For wavelet methods, the bias is irregular for short series and stabilizes on $N \geq 2^{12}$. The other methods show irregular behaviour and high bias and unless $N \geq 2^{16}$, the estimations are not aceptable. R/S statistic method bias seems not to have stabilizing behaviour while periodogram and variance seems to stabilize for high N. Figure 5 illustrates the standard deviations for traces with $H = 0.90$ and varying length N. Note that Whittle-type method

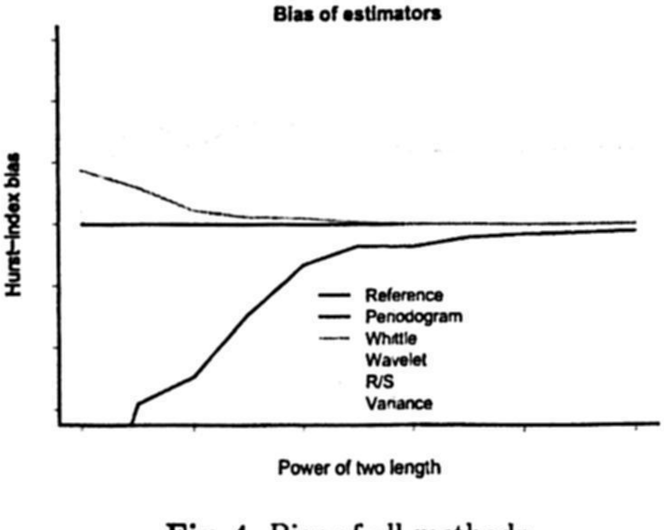


Fig. 4. Bias of all methods

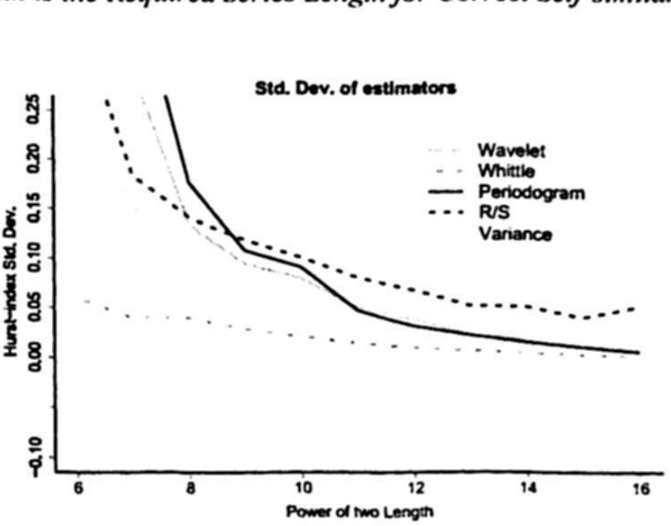


Fig. 5. Standard deviation of all methods

is the estimator which presents less variability and for $N \geq 2^8 = 256$ points, the estimations are accurate enough. Periodogram and Wavelet methods are the ones that follow in accuracy and the length required is the same when variability is taken into account. However, bias and standard deviation together indicate that the best estimator for short time series is Whittle method. Whittle method presents high accuracy when $N \geq 2^8$, Wavelet methods presents accuracy when $N \geq 2^{13}$, Periodogram acceptable estimations when $N \geq 2^{15}$ and Variance and R/S statistic method, biased estimations when $N \in (2^6, 2^{16})$.

4.2 Convergence of Estimators

For each series of length $M = 2^{16}$, a variation of the above defined convergent analysis is performed. The analysis is first applied to the first $\tau_0 = 64$ points of the series X_j, i.e., a *Hurst* estimation method $\Theta(.)$ is applied to the first τ_0 points of X_j. Then, we repeatedly apply $\Theta(.)$ to the next $\tau_0 + i\tau_u$ points of X_j, where $\tau_u = 200$ and $i = 1, 2, \ldots, k$ such that $\tau_0 + k\tau_u \leq M$. This analysis is done to 100 *fGn* series, thus obtaining the convergent behaviour of each. Once the convergent analysis is performed for each of the studied traces, the mean convergence analysis is performed. it means the mean for the 100 estimations of τ_0 and so on. The mean convergence plot $\Theta(.)$ versus $\tau_i, i = 0, 1, 2, \ldots, k$ is now an indicator of how well the estimators convergence to the theoretical *Hurst*-index value. This mean convergent behaviour was applied to the studied estimators in this paper. Figure 6 shows the mean convergent behaviour of the R/S-statistic method and variance-type method. Note from figure that R/S statistic stabilizes quickly but is biased by 0.05, thus it is difficult to propose a minimum length for this method. Unlike R/S statistic, variance-type method converges to the theoretical value($H = 0.9$ in this case) as the length increases. From the figure and the bias and σ plots it is inferred that the required length for this method should be at least 2^{16} points. Figure 7 shows the same kind of mean convergent plots for the other methods studied. Note that *Whittle*-type method stabilizes quickly

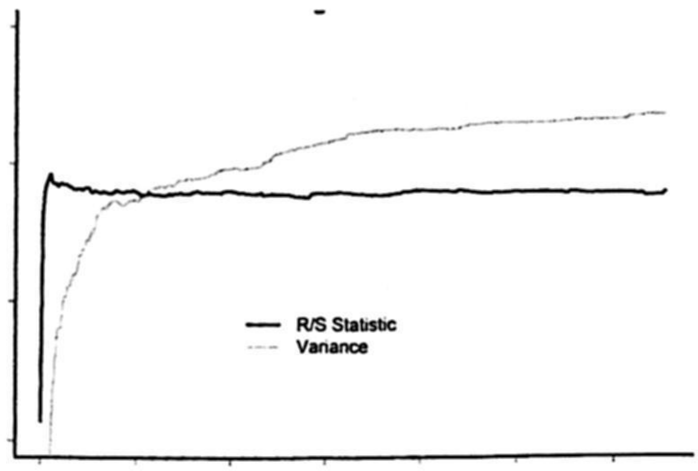


Fig. 6. Mean convergent behaviour

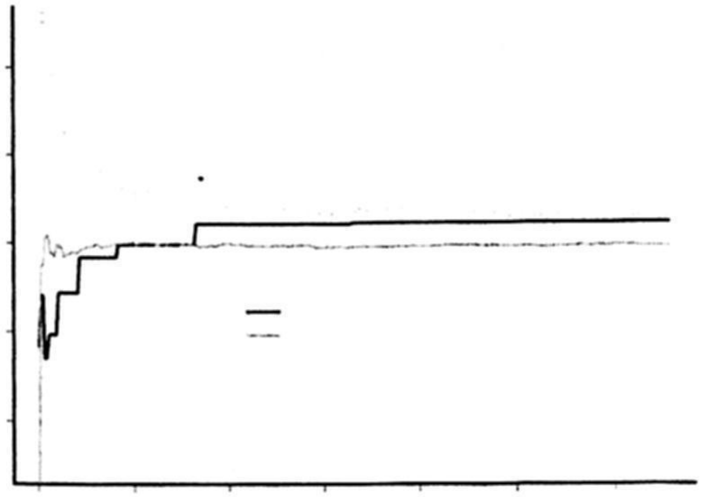


Fig. 7. Mean convergent behaviour

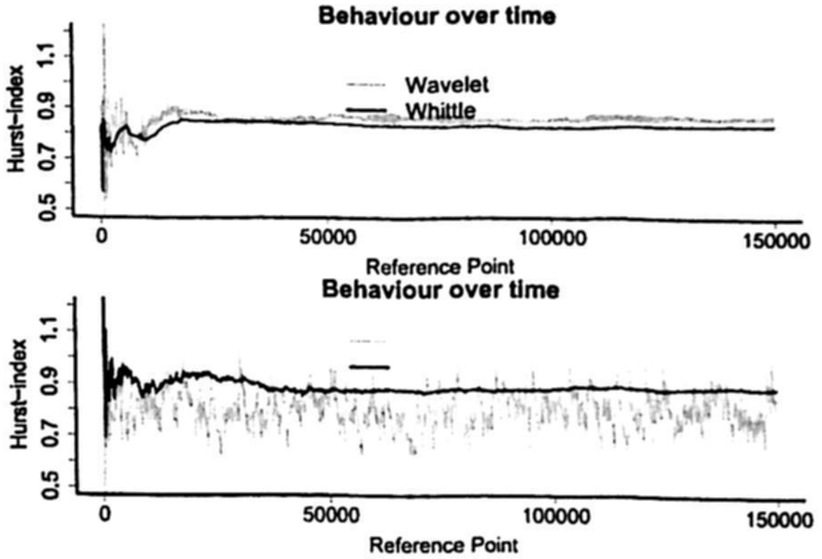


Fig. 8. LAN trace behaviour in time

with very low bias. The behaviour is similar to that of R/S statistic but unlike R/S statistic, the estimations in *Whittle*-type are not biased. The behaviour of Periodogram and wavelet method is similar. Periodogram, according to definition in section 3 is negatively biased, while wavelet is in principle positively biased, but in the long term it becomes negatively biased. From the above figure it is inferred that for accurate estimation of the *Hurst*-index at least 512 points are needed for the *Whittle*-type method. For the wavelet method based on the figures it is inferred that at least 2^{12} points are needed for accurate estimations. Periodogram method, on the other hand, needs at least 2^{13} points for accurate estimations.

4.3 Application to Real Traces

The application of the above results to a real trace is now done. The trace selected is Bellcore's Ethernet trace measured in August 1989. The trace represents one hour of traffic on a LAN. The analysis performed is the same explained in section 3, let $1 = t_0, t_1, \dots, t_k$ be a sequence of points in the x-axis, where $t_{i+1} > t_i$ and $\left(t_{i+1} - t_i\right) = 256$, to each block of trace X_j of length 256, $\{X_j\}_{j=t_i}^{t_i+256}$, apply a *Hurst* estimation method $\Theta(.)$ and finally construct the graph of the behaviour of the estimator $\Theta(.)$. Figure 8 shows the results of this analysis. Note from figure that periodogram overestimates and R/S statistic shows respectively irregular behaviour. Whittle-type estimator follows the H value reported in [17] and wavelet based method follow the reported value with high variability. Among the possible application areas of the current results are physiological time series where short time series are obtained [10] [11], administration of QoS parameters in real time, where a short measured trace is required in order to make performance decisions and in every discipline where time series length is a considerable factor affecting the performance of a system.

5 Conclusions and Future Work

This paper presented a study of the behaviour of estimators under short series in the context of *fGn* traces. Based on the study of the behaviour of thousands of *fGn* time series under bias, σ, convergent analysis and behaviour under time for a given length N_{min}, supossed to be the minimum length we arrive at the folllowing conclusions. The *Whittle*-type method behaves the best for short and long time series presenting both minimum bias and variability. The wavelet and periodogram method behave well when the time series is medium length. Variance and R/S statistic method behave with high bias and are not suitable for short-length measurements. The minimum length for accurate estimation of the *Hurst*-index was proposed for the estimation methods. Based largely on the above mentioned analyses the minimum length for *Whittle*-type method is at least 2^8, 2^{13} for the wavelet method, 2^{15} for the periodogram one and 2^{15} for the variance-type method. No minimum length was obtained for the R/S statistic method due to the high bias and variability in lengths of 2^{16}. The testing of these results in real Ethernet traces was also done. In the future we expect to study the same behaviour for f-*ARIMA* time series and also to study the effects of nonstationarities and trends on estimating the *Hurst*-index for short time series.

Acknowledgements. We finally acknowledge the reviewers for the useful comments that enriched the final version of the paper. All their comments and suggestions were taken into account in the final version of this paper. We also acknowledge CINVESTAV, Unidad Guadalajara and CONACYT for the support received.

References

1. Abry, P.: Wavelets, Spectrum Estimation and 1/f processes. Lecture Notes in Statistics **105** (1995) 15–30
2. Abry, P., Veitch D.: Wavelet Analysis of Long-Range Dependent Traffic. IEEE Transactions on Information Theory **44** (1998) 2–15
3. Abry, P., Veitch D., Flandrin, P.: Long-Range Dependence: Revisiting Aggregation with Wavelets. Journal of Time Series Analyis **19** (1998) 253–266
4. Audit, B., Bacry, E., Muzy, J., Arneodo, A.: Wavelet-Based Estimators of Scaling Behaviour. IEEE Transactions on Information Theory **48** (2002) 2938–2954
5. Beran, J.: Statistics for Long-memory Processes. New York, Chapman & Hall (1994)
6. Beran, J., Sherman, R., Taqqu, M. S., Willinger, W.: Long-range dependence in variable-bit-rate video traffic. IEEE Transactions on communications **43** (1995) 1566–1579

7. Cannon, M., Percival, D., Caccia, D., Raymond, G., Bassingthwaighte, J.: Evaluating Scaled Window Variance Methods for Estimating the Hurst Coefficient of Time Series. Physica A. **247** (1997) 606–626
8. Crovella, M., Bestavros, A.: Self-similarity in Word-Wide-Web traffic: evidence and possible causes. IEEE/ACM Transactions on Networking. **5** (1997) 835–846
9. Davies, R. B., Harte D.: Tests for Hurst Effect. Biometrika. **74** (1987) 95–101
10. Deligneres, D., Ramdani, S., Lemoine, L., Torre, K., Fortes, M., Ninot, G.: Fractal Analyses for Short Time Series: A Reassesment of Classical Methods. Journal of Mathematical Psychology. **50** (2006) 525–544
11. Eke, A., Herman, P., Bassingwaighte, J., Raymond, G., Percival, D., Cannon, M., Balla, I., Ikrenyi, C.: Physiological Time Series: Distinguishing Fractal Noises from Motions. Pflugers Archives. **439** (2000) 403–415
12. Embrechts, P., Maejima, M.: Self-Similar Processes. Princeton, Princeton University Press. (2002)
13. Giraitis, L., Kokoszka, P., Leipus, R., Teyssiere, G.: On the Power of R/S-Type Tests Under Contiguous and Semi Long-Memory Alternatives. Acta Applicandae Mathematicae. **78** (2003) 285–299
14. Hosking, J. R. M.: Modelling Persistence in Hydrological Time Series using Fractional Differencing. Water Resources Research. **20** (1984) 1898–1908
15. Karagiannis, T., Faloutsos, M.: SELFIS: A Tool for Self-similarity and Long-range Dependence Analysis. 1st Workshop on Fractals and Self-similarity in Data Mining: Issues and Approaches.(2002)
16. Karagiannis, T., Faloutsos, M., Molle, M.: A User-Friendly Self-similarity Analysis Tool. Special section on Tools and Technologies for Networking Research and Education, ACM SIGCOMM Computer Communication Review. (2003)
17. Leland, W. E., Taqqu, M.S., Willinger, W., Wilson, D. V.: On the self-similar nature of Ethernet traffic(Extended version). IEEE/ACM Transactions on Networking **2** (1994) 1–15
18. Mandelbrot, M., Wallis, J.: Robustness of R/S in Measuring Non-Cyclic Global Statistical Dependence. Water Resources Research. **5** (1969) 967–988
19. Park, K., Willinger W.: Self-similar Network Traffic and Performance Evaluation. Wiley-Interscience (2000)
20. Paxson, V., Floyd, S.: Wide-area traffic: The failure of poisson modelling. IEEE/ACM Transactions on Networking. **3** (1995) 226–244
21. Percival,.B. P.: Stochastic Models and Statistical Analysis of Clock Noise. Metrologia. **40** (2003) S289–S304
22. Podesta, J.: Self-Similar Processes in Solar Wind Data. Advances in Space Research. **41** (2008) 148–152
23. Rangarajan, G., Ding, M.: Processes with Long-Range Correlations. Berlin, Springer (2003)
24. Serletis, A., Rosengerg, A.: The Hurst Exponent in Energy Future Prices. Physica A. **380** (2007) 325–332
25. Taqqu, M. S., Teverovsky, V.: Semi-parametric graphical estimation techniques for long-memory data Athens Conference on Applied Probability and Time series analysis. **115** (1996) 420–432
26. Taqqu, M. S., Teverovsky, V., Willinger, W.: Estimators for long-range dependence: An empirical study. Fractals. **3** (1995) 785–798
27. Tsybakov, B., Georganas, N.: Self-similar Processes in Communications Networks. IEEE Transactions on Infotmation Theory. **44** (1998) 1713–1725
28. Veitch, D., Abry, P.: A Wavelet Based Joint Estimator of the Parameters of Long-Range Dependence. IEEE Transactions on Infotmation Theory. **45** (1999) 878–897

Computer Architecture and Digital Systems Design

Service-Based Access Control Using Stages for Collaborative Systems

Mario Anzures-García[1], Luz A. Sánchez-Gálvez[2],
Miguel J. Hornos[2] and Patricia Paderewski[2]

[1] Facultad de Ciencias de la Computación, Benemérita Universidad Autónoma de Puebla,
14 sur y avenida San Claudio. Ciudad Universitaria. San Manuel, 72570 Puebla, Mexico
[2] Universidad de Granada, C/ Periodista Saucedo Aranda, s/n, 18071 Granada, Spain
manzures@siu.buap.mx, luzsg@correo.ugr.es, {mhornos, patricia}@ugr.es

Abstract. The adaptation of collaborative systems includes many dimensions, some of these are: access control, concurrency control, coupling of views, and extensible architectures. In this paper, we focus on access control; for this reason, we present a set of services which are part of a SOA-based architectural model for developing groupware applications and which provide two mechanisms for security management: authentication and access control. The former controls the user access to the shared workspace, and the latter controls the access to the shared resources and the interaction among users, in order to avoid conflicts arising from cooperative and competitive activities. We specify the policy-based management of the groupwork organizational structure by means of an ontology. This allows us to define several group organizational structures and to support the groupwork dynamism, facilitating the management of the security mechanisms mentioned. We consider that in collaborative systems there are several stages, i.e. phases of collaboration. Adequate access to the shared workspace must be controlled at each stage, taking into account the roles that can participate in it. Moreover, we explain by means of an example how interactions are carried out among the services related to the access control process and how these services are sometimes adapted.

1 Introduction

Collaborative systems are focused on investigating how computer-based groupwork can improve the performance of groups of people engaged in a common task or goal [2, 3]. In this paper, we will consider mainly two key aspects of this type of system: groupwork and shared information.

We believe that, with regard to the first aspect, both static and dynamic issues have to be taken into account. The static ones are associated with the groupwork organizational structure, while the dynamic ones should support the dynamism that is inherent in the groupwork. It is very important to provide methods to adequately model these issues, using a set of elements (or concepts) and relations among them. This facilitates adaptation to the dynamic nature of the group and to the changing needs of the groupwork.

*© G. Sidorov, M. Aldape, M. Martínez, S. Torres. (Eds.)
Advances in Computer Science and Engineering.
Research in Computing Science 42, 2009, pp. 311-322*

*Received 09/03/09
Accepted 27/04/09
Final version 06/05/09*

We believe that for the second aspect above mentioned, security mechanisms should be considered. Access control to resources and activities is a key element in system security. In most systems, security is achieved through mechanisms such as: authentication, access control, data encryption, digital signature, and so forth. In the collaborative domain, special attention has been paid to authentication and access control mechanisms. The former are mechanisms that allow identification and verification of the user identity, in an attempt to protect the system from unauthorized access. The latter are mechanisms that enable the information to be protected according to security policies, by allowing access to shared resources only to authorized users.

In this paper, an ontological model specifying a policy-based approach to managing the groupwork organizational structure is supplied. This ontology-based policy establishes who authorizes users' registration, how the interaction among users is carried out, and how users' participation is defined (for example, by turns). In addition, we use a set of services that help us to provide authentication and access control to the shared workspace. The concepts related to the access control are designed as services in accordance with the ontological model of the groupwork organizational structure. In this way, the user access to the collaborative system as well as to the shared resources is facilitated.

Generally, the interaction among services is coordinated using Business Process Management (BPM), which is a top-down methodology designed to organize, manage and analyze the processes of an organization, and to undertake reengineering processes. Business processes exist as logical models that can be represented by ontologies. The ontology we propose serves as a logical model for managing the services related to access control on collaborative applications.

We consider long-term groupware applications, where sharing information takes place at various stages. A stage in a coordination model is defined as each of the collaboration moments [6]; for example, a conference management system has several stages: submission, assignment, review, and acceptance of papers. Each stage controls the roles that can participate in it, which facilitates the authentication and interaction among users in the shared workspace. None of the existent access control models for collaborative applications that we have studied takes into account this concept of stage, which facilitates the adaptation process and the access control to the shared workspace in a collaborative system.

This paper is organized as follows. Section 2 gives a brief introduction to the access control. Section 3 explains the ontological model that allows us to define several organizational structures and modify them in runtime. Section 4 describes service-based access control management and presents a real application case based on a well-known collaborative application: a conference management system. Finally, we present conclusions and outline future work.

2 Access Control

Access control models are used to decide how the available resources in the system are managed. In order to implement these models effectively and appropriately into collaborative systems, the following requirements have to be taken into account [4], in such a way that access control must:

- be able to protect any type of information or resources at different levels of granularity.
- facilitate transparent access for authorized users and rigorous exclusion of unauthorized users in a flexible manner that does not constrain groupwork.
- be expressive enough to allow high level specification of access rights, thereby managing better the increased complexity that groupware introduces.
- be dynamic, that is, it should be possible to specify and change policies at runtime.
- support delegation, revocation and management of access policies (meta access control) at runtime.
- grant access control by considering the current context of the user.

There are several access control models for collaborative environments [12], such as *Access Matrix Model* [8], *Role-Based Access Control* (RBAC) [9], *Task-Based Access Control* (TBAC) [11], *Team-Based Access Control* (TMAC) [10], *Spatial Access Control* [4], and *Context-Aware Access Control* [5]. RBAC is very effective and the most important and popular for traditional and collaborative systems, but it has several weaknesses:

- The roles in RBAC lack flexibility and responsiveness to the environment.
- RBAC supports the notion of role activation within sessions, but it does not go far enough to encompass the overall context associated with any collaborative task.
- RBAC lacks the ability to specify a fine-grained control on individual users playing certain roles and on individual object instances.
- The specification of constraints has not been discussed in the RBAC model. Constraints are an important aspect of role-based access control and a powerful mechanism for laying out higher-level organizational policy.

We take the idea of RBAC that permissions are assigned to roles rather than users. In this way, the policy has not to be changed when users modify their role within the organization.

3 Ontology of the Group Organizational Structure

We present a model that specifies the groupwork organizational structure taking into account all its static and dynamic aspects, and that allows control of the access to the shared workspace and resources. We use an ontology to model this structure. An ontology, according to Gruber, *is a formal and explicit specification of a shared conceptualization* [7]. A domain is specified using the following ontology elements: *concepts*, *relations*, *axioms* and *instances*. In our work, the conceptual modelling of the groupwork organizational structure (see Figure 1) considers the following *concepts*:

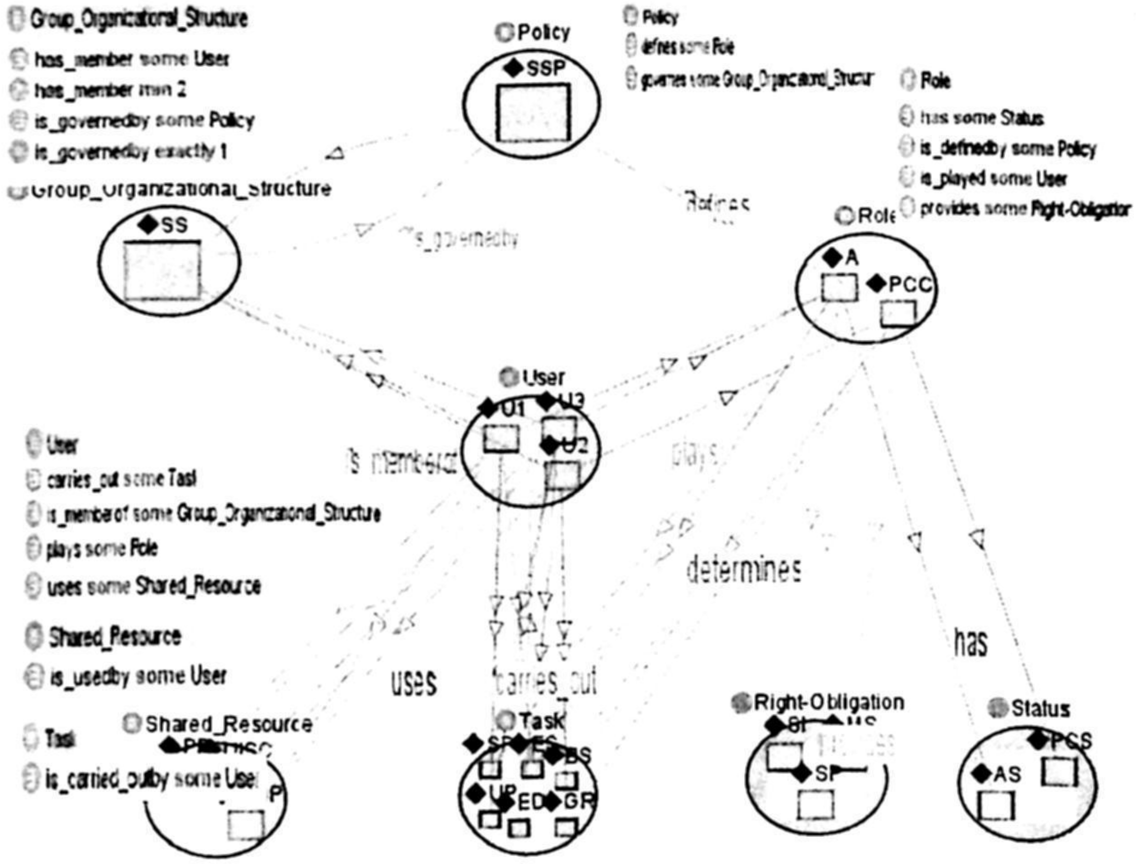


Fig. 1. Ontology of the groupwork organizational structure based on OWL [13] representing the instances for the submission stage. Figure generated by Jambalaya [15] plug-in for Protégé [14].

- *Group_Organizational_Structure,* which is governed at a given stage by a specific policy.
- *Policy,* which configures the above mentioned structure and defines a set of roles.
- *Role,* which must be played at least by one user so that the policy can operate in an appropriate way. Each role has a status and a set of rights/obligations.
- *Status,* which defines the role authority according to the relative position of this inside the organization.
- *Right-Obligation,* which constrains the user actions in the groupwork.
- *User,* which is a person or computational entity that plays one or more roles and carries out collaborative tasks.
- *Task,* which is a set of activities carried out by one or more users to achieve a common goal.
- *Shared_Resource,* which represents the resources used to carry out a task.

The concepts are associated by means of the following *relations* (see Figure 1):

- *is_governedby* (Group_Organizational_Structure, Policy), which specifies that a group organizational structure is governed by a policy at a given stage.

- *is_memberof* (User, Group_Organizational_Structure), which defines that the user is member of the group organizational structure.
- *defines* (Policy, Role), which specifies that the policy defines the roles.
- *plays* (User, Role), which indicates what roles can be played by the user.
- *has* (Role, Status), which determines that the role has a status.
- *provides* (Role, Right/Obligation), which specifies that each role provides a set of rights/obligations.
- *determines* (Role, Task), which indicates that each role determines the tasks that a user playing it can carry out.
- *uses* (User, Shared_Resource), which defines the shared resources used by the user.
- *carries_out* (User, Task), which specifies that the user carries out one or more collaborative tasks in a given stage.

Some of the main *axioms* of our ontology are (see Figure 1):

- A group organizational structure is only governed by a policy in a certain stage.
- A group organizational structure has at least two users.
- Each policy defines at least one role.
- Each role has to be played by at least one user.
- Each task has to be carried out by at least one user.

In order to define the *instances* of our ontology we consider a conference management system, which facilitates the electronic submission, assignment, review, and acceptance of papers (we refer to each of these collaboration moments as a *stage*), along with the management of the whole process. Commonly the roles that can participate are: Author (*A*), Program Committee Chair (*PCC*), and Member of the Review Committee (*MRC*). In accordance with each stage, certain roles can participate, carrying out different tasks. The boxes in Figure 1 show the *instances* corresponding with the submission stage of papers, which are:

- Group organizational structure: Submission Stage (*SS*).
- Policy: Submission Stage Policy (*SSP*).
- Role: The roles that can be played at this stage are:
  a. Author (*A*), whose *status* is *AS* and *rights/obligations* are submitting paper and information (*SP*, *SI*), and has the following tasks associated: submitting paper and information (*SPI*), editing submission (*ES*), and uploading paper (*UP*).
  b. Program Committee Chair (*PCC*), whose *status* is *PCS* and *rights/obligations* are managing the system (*MS*), and can *carry out* the following tasks: browsing submissions (*BS*), extending deadline (*ED*), and generating reports (*GR*).
- User: We define four users: *U1*, *U2*, *U3* and *U4*.
  a. *U1* and *U3* play the *A* role and they use the shared resources: their paper (*PP*) and the user interface that only allows submitting a paper (*UISP*).
  b. *U2* plays the *PCC* role, which uses the user interface without any constraint (*UIWC*).

c. *U3* and *U4* play the *MRC* role. Since this role does not participate in the submission stage, the user *U4* is not shown in Figure 1. *U3* is shown in the figure due to his/her also playing the *A* role.

4 Service-Based Access Control

The success of collaborative systems mainly depends on their capability to be reused and adapted to different and dynamic collaborative scenarios. A change in the groupwork objectives, the participants involved, the group structure, etc. of a collaborative scenario can make a previously successful collaborative system unsuitable for the new situation. The potential solution to these adaptation and reuse problems is nowadays a recognized benefit of Service-Oriented Architecture (SOA), since it is easier to reuse a service that supports the common functionality of several applications than reusing complete applications across different scenarios. The adaptation can be achieved by replacing or even only modifying one or several application services in order to change solely that part of the application that did not fit the characteristics of the new scenario. For this reason, we have proposed a SOA-based layered architecture that facilitates the development of adaptive and adaptable collaborative applications [1], along with authentication and access control to the shared environment. This proposal presents the following elements (see Figure 2):

1. **Application Layer**, which contains applications (such as a conference management system, a chat, etc.). Each collaborative application provides a user interface (web page) that the user uses to carry out a common goal with other users. This layer makes use of different services of the lower layer in order to supply users with all the necessary aspects to perform groupwork.
2. **Group Layer**, which establishes shared workspaces that are appropriate and adaptable to the needs and dynamics of the groupwork. It includes several services: Session Management, Session Management Policy, Registration, and Shared Management. We will devote a subsection to explaining in detail each of them. The interaction among these services is controlled and managed by the lower layer.
3. **Control Layer**, which contains two services: Services Control and Adaptation Control. The former, which models the business logic using the concepts and relations of the ontology proposed in Section 3, also describes the interaction flow of the services related to access control. The latter manages the adaptation process to provide adaptive or adaptable collaborative applications.

In the next subsections, devoted to explaining in detail each of the services related to access control, we will focus on the actions that allow control of interactions among users and avoid inconsistencies in the application due to cooperative and competitive activities.

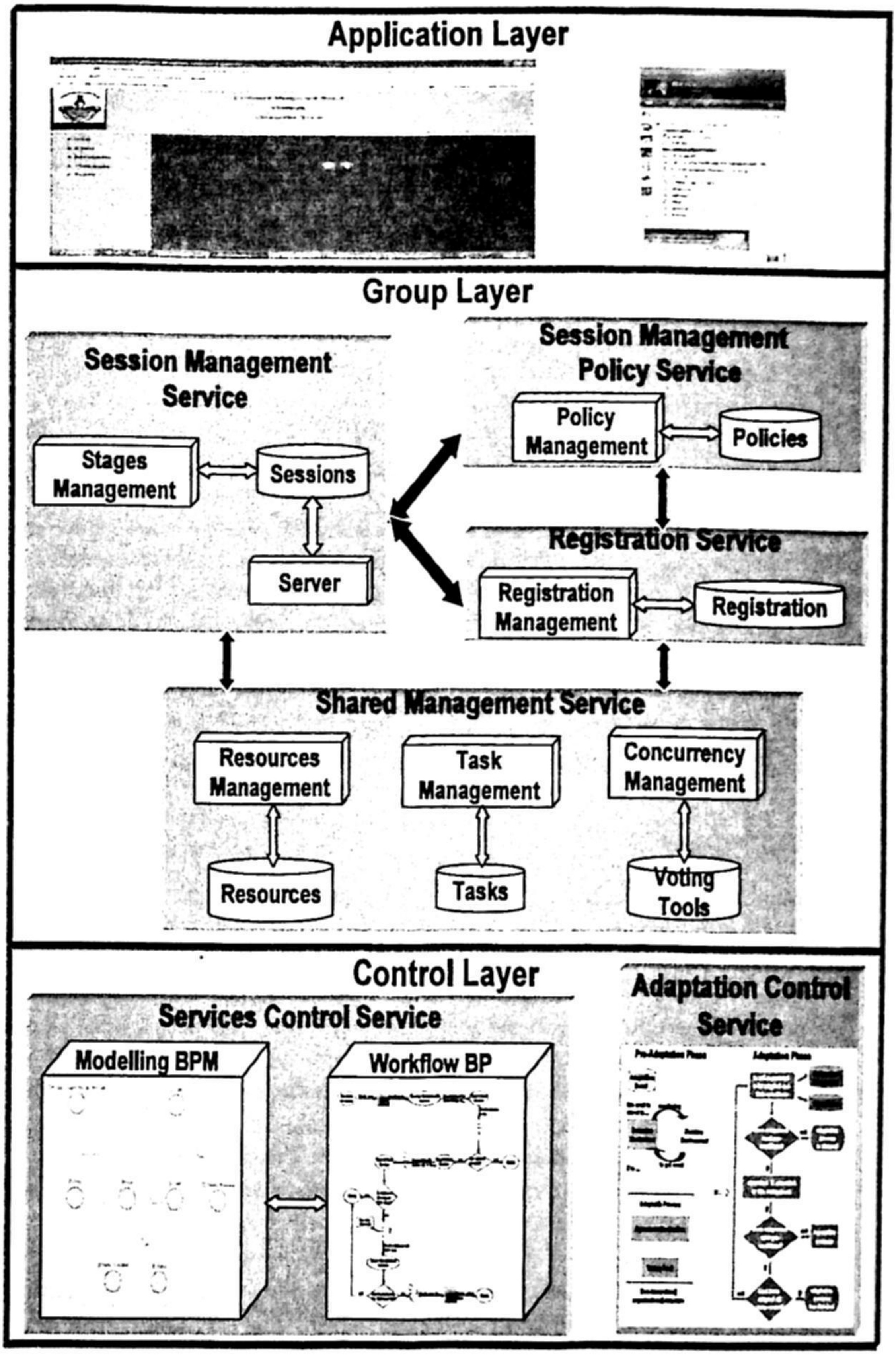


Fig. 2. General schema of the SOA-based layered architecture for collaborative systems.

4.1 Session Management Service

This service allows asynchronous and synchronous sessions to support complex and large-scale collaborative applications, and provides a mechanism that allows users to connect to sessions, as well as to join, leave, invite someone to, and excludes someone from a session. In order to facilitate the access control, this service is able to:

- Manage and control the session and the connections of users to it; in this way collaborative work is allowed.
- Store information about a user and his/her work session; with this it is possible to identify the users that are connected to each session (a user can participate in more than one session).
- Manage the collaboration moments, by establishing when a stage begins and ends. When a stage ends, it is possible to change the current policy to another more appropriate.
- Grant to the Session Management Policy Service the control of the groupwork organizational structure in order to adapt the application to the group changes and to the new needs of the groupwork.
- Send the information of the stages to the Registration Service, which associates this information with the roles that each user will play at a specific stage.
- Allow the Shared Management Service to manage all the aspects related to the shared resources in accordance with the organizational structure defined by Session Management Policy Service.
- Notify all the Group Layer services of the moment in which a user joins or leaves a session so that each service can carry out the necessary actions to avoid inconsistencies in the application.

4.2 Session Management Policy Service

This service uses the ontology presented in Section 3 to establish the groupwork organizational structure (see Figure 1) and define the access control to the shared workspace as well as to the shared resources. Consequently:

- It stores the information related to the ontology: concepts, relations, axioms and instances. This allows us to modify individually the instances of each concept without having to change the others.
- It associates roles with the current stage defined in the Session Management Service. This establishes the authorized users that can participate in a specific stage; therefore, it facilitates the access control to the shared workspace.
- It checks the constraints to be fulfilled so that the current policy always is valid; for example that each role always has to be played by at least one user.
- It sends the information about the roles that can be played at each stage to the Registration Service, in order to control access to the collaborative system.
- It transmits the tasks that must be carried out in this stage and the shared resources that can be used for these tasks to the Shared Management Service.

4.3 Registration Service

This service allows the registration of new users in an existing session through the user interface (web page) of the Application Layer so that they can participate in the groupwork. To control access to sessions, this service carries out the following operations:

- It authenticates the access to the session when the user inputs a login and a password, and stores these data, which are used by the system to corroborate that he/she is an authorized user.
- It stores the information that users submit at the moment of their registration in the system (in the case of our example, this service stores personal data and information about the paper).
- It assigns the adequate role to the registered user taking into account the valid roles at the current stage and the data submitted by him/her.
- It associates the role with its corresponding rights/obligations and its status, which determine the user behaviour in the collaborative application.
- It sends the information on roles, rights-obligations and status to the Shared Management Service so that this service can manage the interactions among users and the uses of shared resources by users.

4.4 Shared Management Service

This service enables management of the shared context, since it takes into account the session management policy in order to determine who can carry out a task and which shared resources can be used for this task. It uses the concurrency mechanism (should this be necessary) to avoid conflicts due to cooperative and competitive activities. The Shared Management Service carries out the following operations:

- It stores the tasks that must be carried out in the shared workspace.
- It connects the tasks with the roles that can perform them.
- It checks that each task is carried out by authorized users, i.e., users playing a role which allows them to perform the corresponding task.
- It stores the existing shared resources in the collaborative application.
- It associates the shared resources with the tasks to be carried out.
- It corroborates that the collaborative work is suitably carried out in accordance with the workflow of the Services Control Service.

4.5. Services Control Service

This service controls the interactions among the services related to the access control by modelling the business logic and specifying a workflow in accordance with the model. In this work, we use the ontological model specified in Section 3. The workflow establishes the interrelations of each user with the shared resources and other users as well as the different services related to the access control. Figure 3 shows the workflow for submitting a paper using the conference management system.

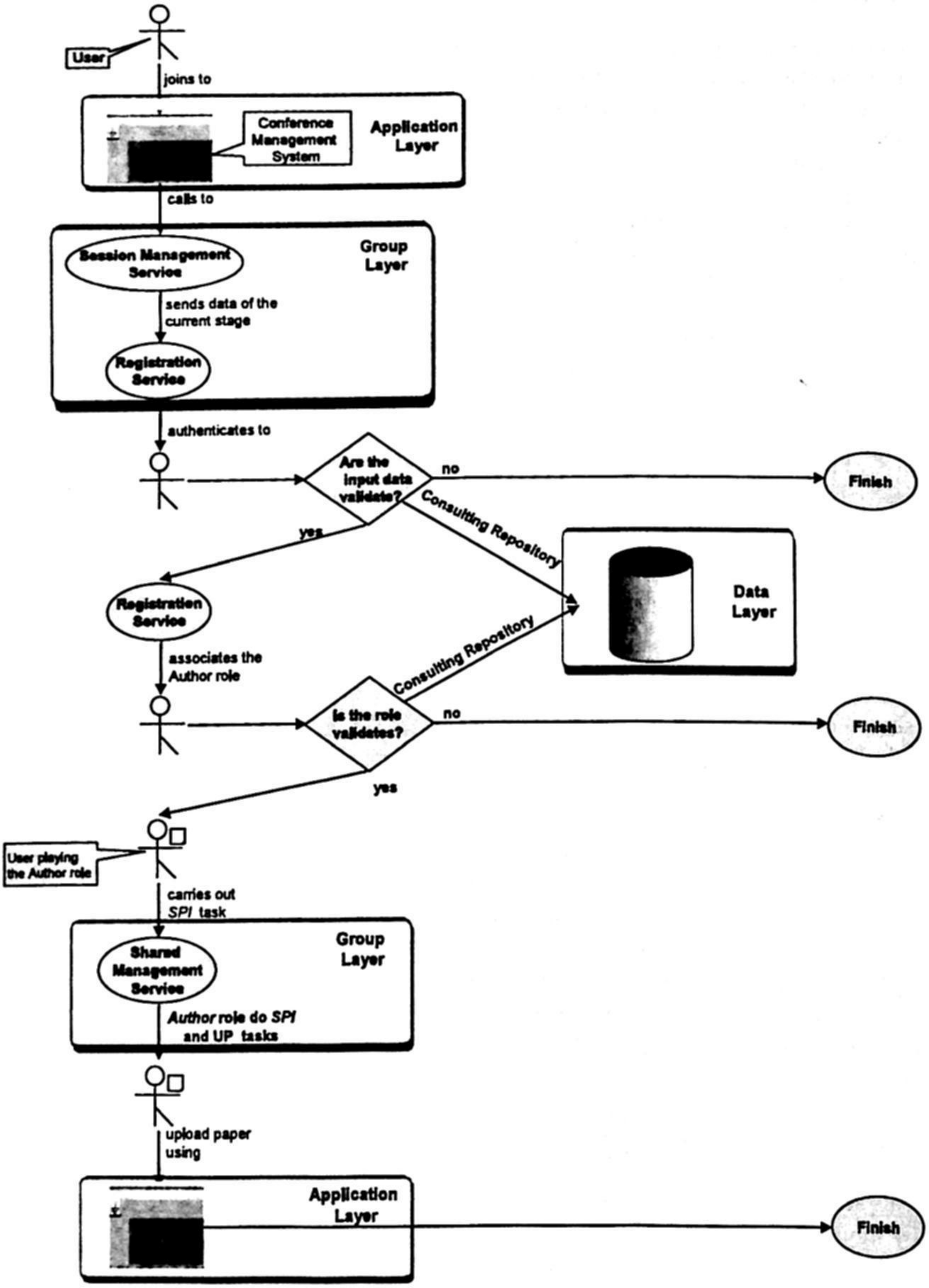


Fig. 3. Workflow for submitting a paper using a conference management system.

5 Conclusions and Future Work

In this paper, we have presented a service-based access control, which facilitates the access of users to the shared workspace. This provides all the related aspects to the security of collaborative systems, and enables groupwork with more sophisticated and appropriate access policies, as well as supporting the dynamic nature of the groupwork, in such a way that permissions, roles and constrains are part of the model; therefore, they can be changed in a just-in-time fashion. The access control uses as business logic the ontology of the groupwork organizational structure. This ontology enables adjustment of the organizational structure (for example, changing the role that a user can play in a session and/or the rights/obligations of a role), and the transformation of the model (for example, adding a new role or changing the policy established). Moreover, we have also used the concept of stage to facilitate the access control.

With regard to future work we think that it is necessary to analyze and model the access control management from the first step of the development of collaborative systems at both architectural and organizational level, in order to facilitate their dynamism and greatly decrease their final cost. In addition, we are studying how the changes in the group organizational structure affect in order to consequently adapt the access control.

References

1. Anzures-García, M., Hornos, M.J., Paderewski, P.: Architecture for Developing Adaptive and Adaptable Collaborative Applications. Lecture Notes in Computer Science (LNCS), Vol. 4758, pp. 271-274. Springer, Heidelberg (2007)
2. Beaudouin-Lafon, M., et al.: Computer Supported Cooperative Work. Trends in Software, John Wiley & Sons (1999)
3. Beaudouin-Lafon, M.: Beyond the Workstation: Media Spaces and Augmented Reality. In Proceedings of the Conference on People and Computers IX. Vol. 9, pp. 9-18 (1994)
4. Bullock, A., and Benford, S. An Access Control Framework for Multi-User Collaborative Environments. ACM GROUP (1999)
5. Covington, M., Long, W., Srinivasan, S., Dey, A., Ahamad, M., Abowd, G.D.: Securing Context-Aware Applications Using Environment Roles. In ACM Symposium on Access Control Model and Technology (2001)
6. Ellis, C., Wainer, J.: A Conceptual Model of Groupware. Proceedings of the ACM conference on CSCW, pp. 79-88 (1994)
7. Gruber, T.R.: Toward Principles for the Design of Ontologies Used for Knowledge Sharing. International Journal of Human Computer Studies. Vol. 43(5/6), pp. 907–928 (1995)
8. Lampson, B.: Protection. In Princeton Symposium on Information Science and Systems, pp. 437-443. Reprinted in ACM Operatives Systems Rev. Vol. 8-1, pp. 18-24 (1974)
9. Sandhu, R.S., Coyne, E.J., Feinstein, H.L., Youman, C.E.: Role-based Access Control Models. IEEE Computer. Vol. 29-2, pp. 38–47 (1996)
10. Thomas, R.: Team-based Access Control (TMAC). In Proceedings of 2nd ACM Workshop on Role-Based Access Control, pp. 13–19 (1997)

11. Thomas, R., and Sandhu, R.: Task-based Authorization Controls (TBAC): Models for Active and Enterprise-Oriented Authorization Management. In Database Security XI: Status and Prospects. Lin, T. Y., and Qian, X. (eds.) North-Holland (1997)
12. Tolone, W., Ahn, G., Pai, T., Hong, S.: Access Control in Collaborative Systems, ACM Computing Surveys. Vol. 37-1, pp. 29–41 (2004)
13. OWL Web Ontology Language Guide, http://www.w3.org/2004/OWL/
14. Protegé Platform, http://protege.stanford.edu/
15. Jambalaya Plug-in, http://www.thechiselgroup.org/jambalaya

Successful Development of Portable Didactic Satellite for Training and Research in Satellite Technology*

Esaú Vicente-Vivas[1], C. Álvarez-Bernal[2], E. A. Jiménez[3], Z. Carrizales[3],
C.A. Sánchez[3], J.R. Córdova[4], R. Alva[3], M.A. García[3] and G. Islas[3]

[1]Instituto de Ingeniería, UNAM,
Cd. Universitaria Coyoacan, 04510, México DF.
[2]Instituto Tecnológico de Sonora, Campus Guaymas, Sonora, México
[3]Student, FI-UNAM, Cd. Universitaria, Coyoacan, 04510, México DF.
[4]Student, Posgrado Ingeniería, UNAM,
Cd. Universitaria, Coyoacan, 04510, México DF.
evv@unam.mx, calvarez@itson.mx, factore8@hotmail.com,
zairalilian@yahoo.com.mx, carlos_ss@msn.com,
roro_send@hotmail.com, imatrion@yahoo.com.mx,
janorius@yahoo.com.mx, jcordovaa@iingen.unam.mx

Abstract. This paper depicts the global specifications and the successful results for a cost-effective system developed for training purposes in the small satellite technology field. The didactic satellite subsystems were fully designed, manufactured and tested at the Institute of Engineering UNAM. Information about intelligent subsystems for the portable didactic satellite, its operations software as well as successful results obtained in laboratory operations for the first system prototype are exposed in this paper. In addition, the planned collaborative work in the field with Instituto Tecnológico de Sonora is mentioned.

Keywords: didactic satellite, operational prototype, portable laboratory equipment, training system, cost-effective technology.

1 Introduction

Mexico has developed in the past a couple of small satellites projects, and currently is developing a couple more of them. Those projects are very important not only to bridge the enormous technological gap with industrialized countries, but also to launch a domestic satellite development and research program associated to the future Mexican Space Agency (MEXSA). The MEXSA could take us in the medium and long term to generate our own Remote Sensing and Communications satellites, according to the needs of the country. However, to conduct such a line of work, it is required the formation, qualification, and training of human resources in these technological areas.

In accordance with our experience, the development of an experimental satellite in our country requires of 4 to 8 years of work, depending on the obtained financial

* Project 52297 financed by CONACYT

© G. Sidorov, M. Aldape, M. Martínez, S. Torres. (Eds.)
Advances in Computer Science and Engineering.
Research in Computing Science 42, 2009, pp. 323-334

Received 18/03/09
Accepted 27/04/09
Final version 07/05/09

support, as well as on the magnitude of the technological and scientific objectives.

We have also detected that this type of projects allows the participation of 25 to 200 people according to the challenges and magnitude of the technological demands aimed by the satellite mission. Considering the cost of this type of projects (half million to 5 million US dollars) the ratio between the amount of participant personnel versus project cost is very low. In addition, if we consider how young Mexican University people become attracted and motivated by satellite and space projects, it is observed an extremely low efficiency in the rows of participation and training of new human resources in this field. By the way, this technology arena is related with the possibility to generate alternative solutions to national security problems from our country. That scenario took us to the development of the portable didactic satellite (PDS).

In this way, our group has successfully developed a cost-effective training system in small satellite technology employing commercial-off-the-shelf (COTS) parts as well as electronic components from automotive and services industries, Fig. 1. The system is affordable enough to be used in laboratories with the intention to offer attractive, fast, and versatile training practices and courses in satellite technology and related fields. Our goal is to use the system in High Schools, Technological Institutes and Universities, with the intention of approaching young people to the world of space applications, Science and Technology.

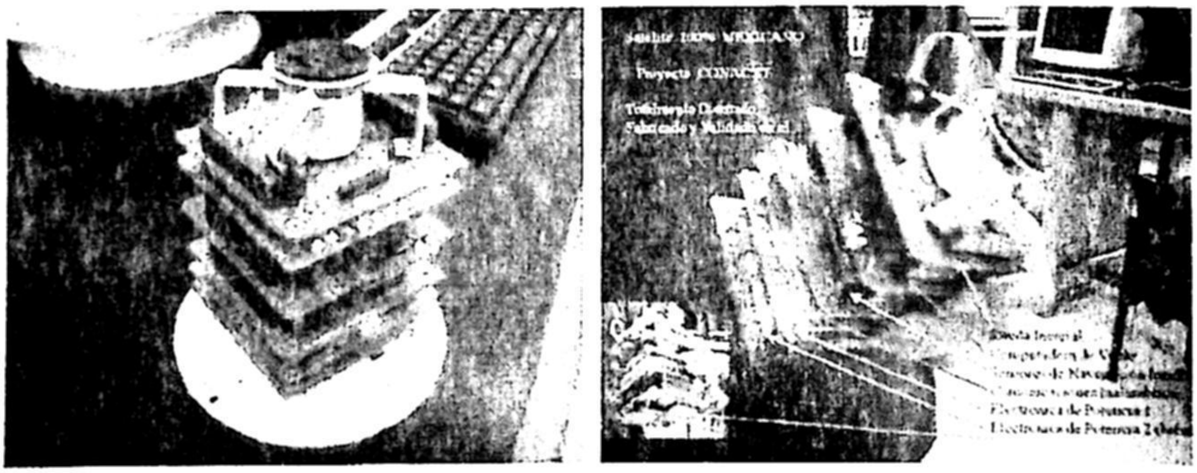


Fig. 1. Developed Mexican portable didactic satellite showing its architecture based in subsystem modularity.

To accomplish a cost-effective PDS its print circuit boards (PCB) were designed and manufactured in two layers only; the power subsystem considers the use of flexible solar panels, rechargeable batteries and an external battery charger; 1-wire sensors were preferred; and the PDS structure was chosen according with cost-effective commercial available products. In addition, several cost-effective auxiliary interfaces such as the orientation determination sensors, serial and USB expansion ports, as well as expansion cards for user defined interfaces are integrated in the PDS. In this way, the PDS is perceived as a friendly device with growing capabilities.

The PDS was projected to be useful in research laboratories to develop new solutions and modules for real satellite subsystems. In this sense, research in fields such as three axis stabilization, digital communications, satellite sensors, power systems, payload validation, flight computers, navigation autonomy, and satellite constellations will be addressed with the support of this laboratory satellite tool.

It must be mentioned that commercial availability of similar products to the PDS is rarely seen in the global market. Right now the only commercial educative satellite product detected by the authors is the Eyassat educational system developed initially by the US Air Force and later commercialized by Colorado Satellite Services, [1]. Besides, we found that very few institutions have developed their own satellite prototype to accomplish laboratory research in distributed space systems, as the case of the Israel Institute of Technology [2] and the US Naval Academy Satellite Laboratory with its "LABsat" experimental hardware, [3].

The Eyassat basic equipment starts at 8,000 dollars. However, this price is difficult to be afforded in developing countries. This is why the Mexican PDS was developed to be offered for a cost under 3,000 US dollars in order to be attractive for different schools, universities, and so on. This goal shaped the main characteristics of the PDS in order to achieve a cost-effective satellite training tool.

It is also important to highlight that we took advantage of previous experiences in space projects, [4], [5] and [6], to fast track this project.

In this way, the paper describes the PDS prototype and its successful operations in laboratory. The paper presents information for every PDS subsystem as well as software information for both the PDS and its ground station. In addition, some innovative PDS operating modes are described, among them: a reaction wheel based stabilization system and software that allows 3D visualization in a laptop by means of monitoring the PDS maneuvers performed by the user.

The PDS prototype was developed under CONACYT project 52979. Right now the satellite prototype is being employed for demonstrative purposes before federal government agencies and some Universities. The goal is to show the technological capabilities from our group to develop not only satellite training systems, but also to develop real small picosatellites and nanosatellites according to the "cubesat" standard developed in 1999 by Dr. Robert Twiggs from Stanford University in USA. In addition, the successful results obtained with the PDS are being employed to launch new and real small satellite projects in México. The new satellite projects will take advantage of experiences and technology from PDS subsystems presented in this paper. The work developed for real small satellites will be presented in future papers.

2 PDS Architecture

The satellite training system (STS) is basically formed by the PDS, operations software, and executable software for personal computers. The PDS contains operations control software to carry out, among other functions: 1) digital communications with a personal computer, that performs as the Ground Station (GS) for Telemetry acquisition and Command Shipment, 2) task delivery requested by the GS, 3) acquisition of telemetry from PDS subsystems, 4) telemetry pack up for wireless transmission, 5) protocols for GS communications, 6) real time operations with the didactic satellite, etc.

The STS has a cylindrical shape of 12 cm in diameter and 17 cm as maximum height. It also includes the ground station Software that runs in a PC and carries out the functions of telemetry acquisition and command transmission to the PDS, Fig. 1. Besides, this software allows the uploading of new operations software to the PDS by wireless means.

The PDS architecture was basically defined by its flight computer dimensions, which were fixed according with real dimensions of Picosatellites, [7], [8], [9], [10], and [11]. Therefore, the STS PCB size is 8.5cm x 8.5cm, Fig. 1. The PCB architecture allows the satellite boards to be assembled in tandem, Fig. 1. The complementary PDS subsystems are: Power, communications, sensors, inertial wheel stabilization and magnetic torquer coil stabilization.

3 Flight Computer Subsystem

The PDS has a single board flight computer, Fig. 2, which integrates lateral connectors in a bus fashion to interconnect cards in tandem. The electrical connector offers the required mechanical ties among printed circuit boards. In addition, each PCB contains screw holes at each one of the corners, which allow screwing the whole printed board array to the PDS structure.

With the bus type connectors its electrical signals are available for all the PDS cards, thus it is possible the card interconnection without caring about the order of them. Therefore, the order mentioned in the next paragraphs gives just an order to the article writing. Besides, it is important to notice out that all the bus type connectors in the cards are male type in their top part, whereas the bottom side looks like a wire-wrap connector. This allows the interconnection of cards either by the top or the bottom side.

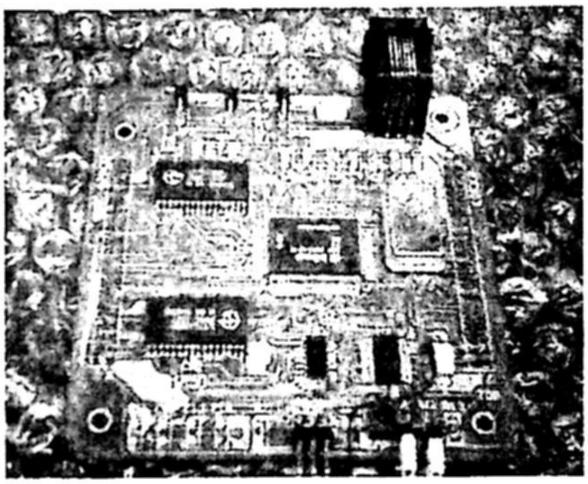


Fig. 2. PDS flight computer

The PDS single board computer is built around the 16 bit RISC SAB80C166 Siemens processor, industrial version, with extended temperature, 40 Mhz oscillator, 256 kb of RAM memory where the PDS operations program is loaded, hardware for automatic uploading of new programs to the computer and a total of 5 serial ports. The lasts support full-duplex asynchronous communication up to 625 Kbaud and half-duplex synchronous communication up to 2.5 Mbaud. The synchronous mode is employed in serial port So1 to gain access to a USB port on one side, and on the other side serial port So2 allows the communication with a 32 Mb block of Flash Memory. The Flash Memory allows to store diverse PDS telemetry data from the PDS.

On the other hand, the 100-pin SAB80C166 microcontroller internally contains important resources as follows: a watch dog timer, an interrupt controller, some 16-bit timers, 10-channel 10-bit A/D converters, two serial channels and several 16 bits I/O ports, with a total of 76 I/O lines.

The SAB80C166 allows the upload of new programs in external RAM memory, however it is required to control several electrical signals in accordance with specific feedback responses from the processor. For this purpose a small PIC microcontroller was integrated in the PCB which is also connected to the communications channel with the Ground Station. In this way, when the GS software sends the "new program download" command to the PDS, the PIC16F877 microcontroller takes over the control of the SAB processor to achieve and supervise the uploading process.

Regarding the software development for the SAB80C166 it was written in standard "C" language, programs were compiled using the BSO Tasking family of tools for the SAB80C166 microcontroller.

4 Power Subsystem

The second PDS subsystem is the intelligent power subsystem (IPS) integrated in 2 PCBs, Fig. 3. The first one contains 4 AA-sized lithium rechargeable batteries and electronics that admit the batteries to be recharged by means of solar panels or from an external battery charger. The second board integrates a PIC microcontroller, solid state switches, voltage regulators, DC/DC converters as well a latch-up protection circuitry.

In this way, the IPS constitutes a simple, small, and economic power subsystem.

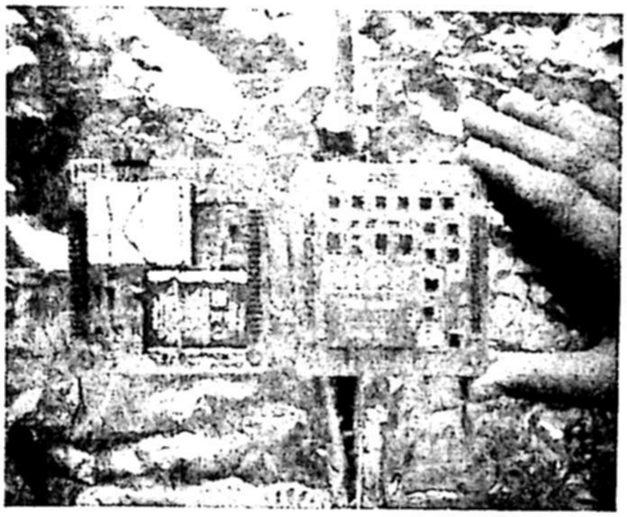


Fig. 3. PDS power subsystem.

5 Wireless Communications Subsystem

The third card constitutes the intelligent PIC based wireless communications system.

This subsystem confers a great operating versatility to the didactic satellite.

The STS communications subsystem consists of 2 wireless cards. One for the PDS and a second one for a laptop that performs as the GS, Fig. 4a and Fig. 4b.

The communications wireless card for the PDS, Fig. 4a, has the followings characteristics: it receives the radiofrequency signal through an antenna, then it passes to a filtering stage; the RF signal is taken to the RF chip CC2500 to generate serial data in SPI format; the 18F2321 PIC device controls the RF chip; the PIC device converts synchronous data into asynchronous data; finally, data are delivered to the flight computer.

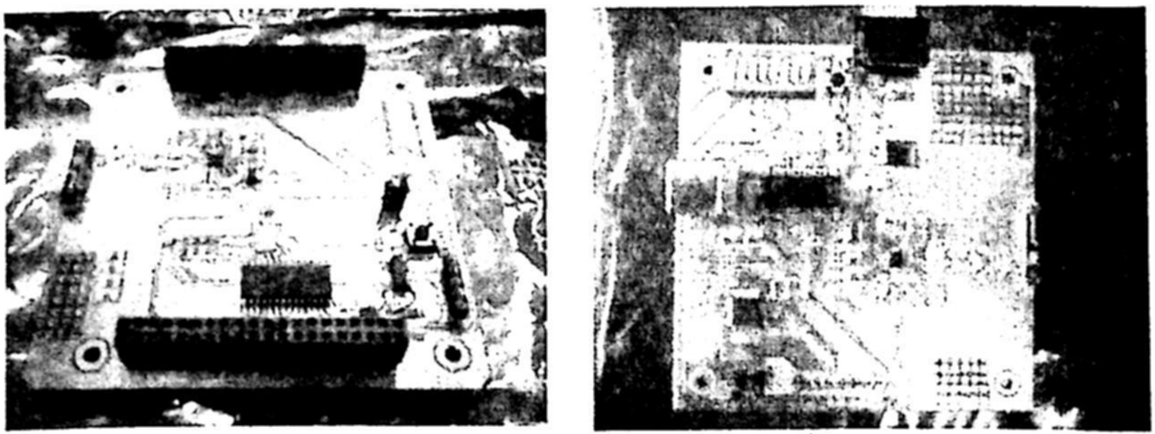

(a) (b)

Fig. 4. The STS communications subsystem, a) for PDS and, (b) for ground station.

On the other hand, the wireless communications card for GS, Fig. 4b, allows any laptop computer to interact by wireless means with PDS. Its main components are a USB interface chip that generates an asynchronous serial output signal, a PIC device to control the serial data transmission and reception among laptop and flight computer, an RF chip CC2500 for wireless transmission in the 2.4 GHz band and a surface mount antennae. In addition, the PIC is interfaced to digital microswitches to choose among 255 RF communication channels configurations to avoid data collisions when multiple RF cards are employed together. This means that up to 255 different PDS equipments can be operated together without interferences between each other.

6 Platform and Inertial Navigation Sensors

The fourth PIC based intelligent card integrates the satellite inertial navigation sensors, Fig. 5. Although, each one of the described cards has at least a pair of local platform sensors (current and temperature) this card integrates three 1-axis gyroscopes as well as three 1-axis accelerometers. In addition it contains a digital compass. The referred navigation sensors are extremely useful to provide real-time PDS navigation data which is employed at the ground station to render 3-D virtual navigation graphics. The navigation acquisition mode is requested by command, and allows the GS software to generate 3D real time animations in connection with the satellite manipulation generated by the user in the laboratory environment.

The PDS employs the 1490 digital compass from Robson Company. It is a solid-state Hall effect cheap device, 12 pin component, with cylindrical shape and requires 5 V supply. When rotated it senses the position of the four cardinal points on a compass, North, South, East and West. As well as the intermediate directions: North East, North West, South East, and South West.

It has to be noticed that most of the PDS sensors are 1-wire technology, they were selected to significantly reduce the number of traveling signals through the PDS bus. The 1-wire sensors let the microcontroller, by means of only two wires, to access an important amount of sensors. This contrasts with conventional sensors, which require two wires to connect every included sensor.

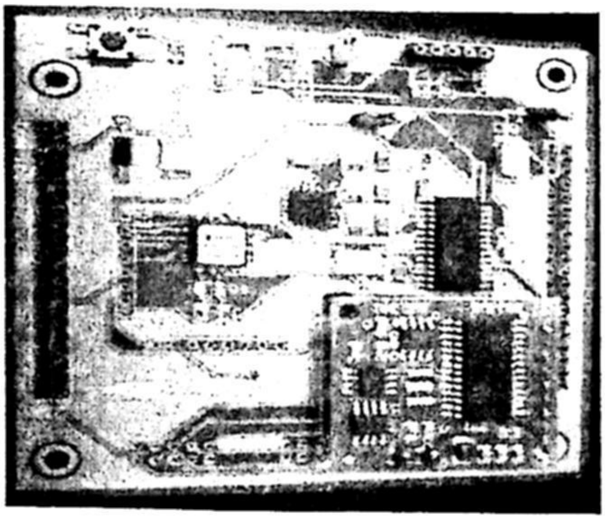


Fig. 5. PDS inertial navigation sensors.

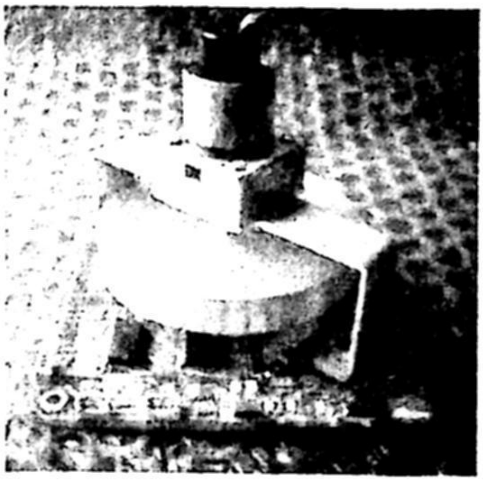


Fig. 6. PDS momentum wheel stabilization subsystem.

7 Stabilization Subsystem

The reaction/momentum wheels are attractive because they apply a torque to a single axis of a spacecraft by adding or removing energy from the reaction/momentum wheel (flywheel), causing it to react by rotating.

By maintaining flywheel rotation, called momentum, a single axis of a spacecraft can be stabilized. Consequently, several reaction/momentum wheels can be used to provide full three-axis attitude control and stability in a space vehicle, [1]. However, for automatic PDS maneuvering and automatic stabilization purposes the system includes only one reaction/momentum wheel.

For those reasons the PDS includes an intelligent satellite stabilization hardware to allow satellite maneuvering demonstration capabilities, Fig. 6. This eventually would conduct to elaborate tasks such as payload pointing towards specific targets and more important, to allow the users of the educative system to learn and understand the use of this important satellite resource.

The exposed system constitutes an intelligent stabilization module therefore it includes a dedicated microcontroller to perform both stabilizations tasks and communications with the flight computer.

In this sense, the PDS allows the emulation of a satellite stabilization system by means of a reaction/momentum wheel. It will allow the study and experimentation of PDS behavior when changes or control is applied in its system actuators. That experience and knowledge is expected to take us to the design and exploration of different small satellite stabilization control schemes.

The stabilization subsystem (SS) is made up of a flywheel driven by a DC motor and a set of six magnetic torquing coils, two different coils (coarse and fine) for each one of the ES structure axis. Three coils apply a coarse momentum while the other three provide fine momentum forces. The dedicated control is given by a PIC18F4431 microcontroller that is connected through serial port with the flight computer, this microcontroller unit receives and processes commands from the flight computer. The command and protocol software was inherited from software developed in our laboratory for a 50 Kg microsatellite mission, [7]. The chosen PIC device has enough resources and capabilities to accomplish the tasks for this subsystem. In addition, SS contains the electronic control interfaces between the microcontroller and the active stabilization actuators. They are composed by an H bridge for the motor (TA7291S from Toshiba) and six further H bridges (3 L293DD integrated circuits from ST Electronics) driven as hardware interface to control the magnetic torquing coils.

In order to carry out the motor control it was added a 16 pulses per round-trip encoder mounted in the motor along with the flywheel. This serves as feedback to the microcontroller. Besides, in case the motor could become obstructed the system has an overcurrent circuit to protect both the motor and its driving H bridge. It also has a set of LEDs that indicate the behavior of control signals applied to the stabilization actuators.

The stabilization card lodges the DC motor, an inertial mass located in the rotation axis of the motor and related electronics.

The reaction wheel is controlled both in open and closed loop configurations. When the PDS is hanged by a string from the ceiling, the last operating mode enables PDS visible and controlled movements that are quickly observed by the users.

In addition, the PDS contains a real time PDS virtual follow-on mode. This mode started by command from the ground station software forces the PDS to continuously capture information from inertial navigation sensors (3 axis gyroscope, 3 axis accelerometer and digital compass) and then transfers the acquired data by wireless means to the GS software. Then the GS software automatically and continuously draws 3D images according to the inertial navigation sensors data.

Both previous visual and interactive modes allow the user to see and easily understand the concepts of satellite supervision and satellite navigation sensors in a friendly manner.

8 PDS Operations

It is necessary to point out that under laboratory validation testing the didactic satellite is suspended in the laboratory ceiling by means of a string, as shown in Fig. 7.

Once in a suspended position, a digital command can be sent through the communications system, and then received by the PDS computer, then, as an answer action, it will actuate the motor in the rotation direction and at the RPMs indicated by the command. Under these circumstances, the didactic satellite experiences reaction forces that generate the PDS physical movements. This process allows the user to carry out several dynamic experiments for satellite stabilization control in a cost-effective fashion.

Fig. 7. The PDS under validation in laboratory.

9 PDS Structural Subsystem

The PDS structure is composed by a cheap commercially available cylindrical container manufactured with plastic, Fig. 7 and Fig. 8. When it was considered the manufacture cost of the structure, the manufacture time and the acquisition of materials, it was decided to go for the commercial cylindrical container. In addition, the last solution allows the PDS to be integrated at a lower cost through the reduction of manual labor. The chosen structure is light, cost-effective and shapes the PDS to render a satellite appearance. Once the full PDS PCBs are plugged together, the stack is fixed inside the plastic structure with screws.

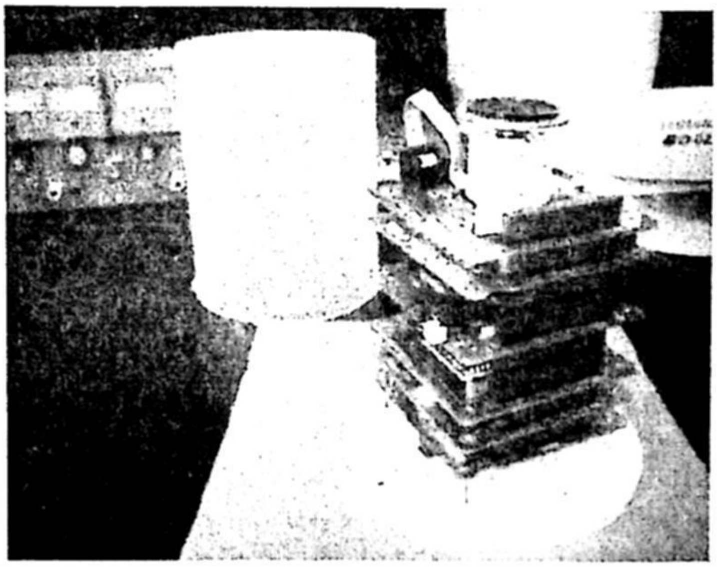


Fig. 8. PDS structural subsystem.

10 PDS Operations Software

PDS operation software is distributed into every satellite subsystem, however, the PDS flight computer allows the coordinated control required by the PDS in order to execute operating functions such as subsystem powering, telemetry gathering, stabilization tasks, wireless operations, command reception, telemetry transmission and so on.

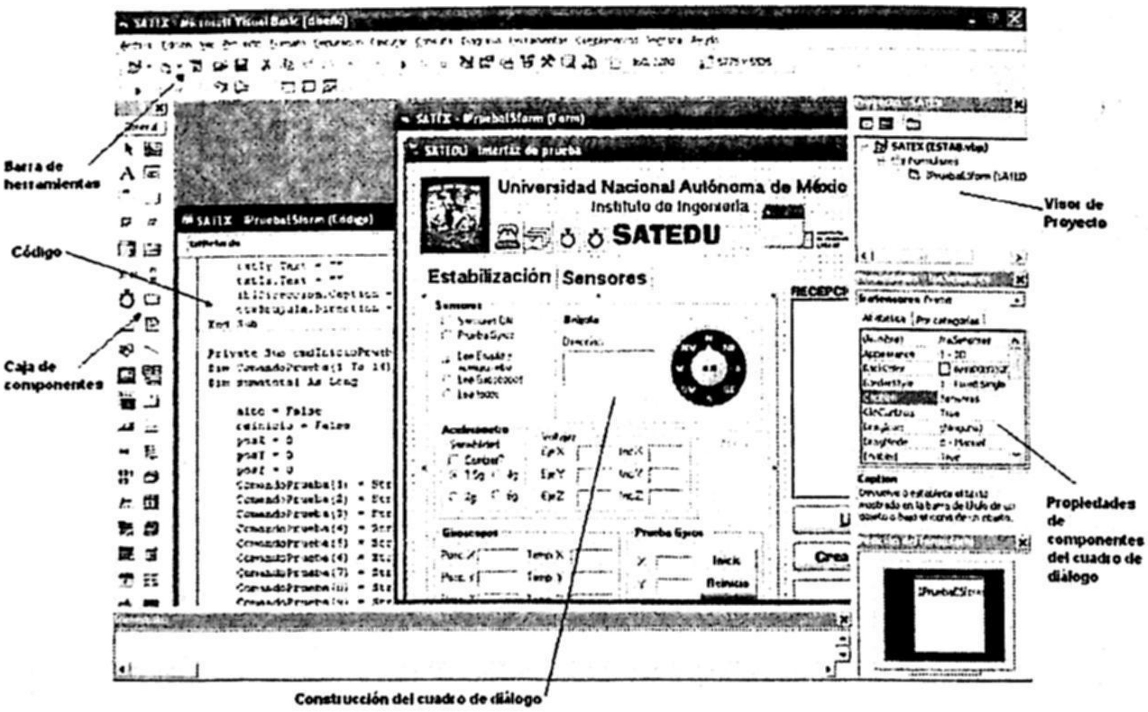
11 Ground Station Software

The GS software executes in personal computers and interacts with the PDS software to achieve both the demonstration and the training functions of the didactic satellite system, Fig. 9. In fact, great part of the friendship attributes for the whole STS system (simplicity and clarity of use) are generated by this software.

The software has several functions, such as those employed to control small satellites launched into space orbit, among them: 1) Immediate answer commands, 2) Telemetry request commands, 3) Real time telemetry requests to allow on-line PDS supervision with the help of 3D animations, 4) Satellite stabilization control commands, etc.

12 Projected Work for the Project

The whole STS software will be recorded in a compact disk (CD) which will also contain the user and the operations manuals. For these reasons, the STS final version will consist only of the PDS and a CD. This is the system we expect to share with educative institutions from our country.


Barra de herramientas
Código
Caja de componentes
Visor de Proyecto
Propiedades de componentes del cuadro de diálogo
Construcción del cuadro de diálogo
Universidad Nacional Autónoma de México
Instituto de Ingeniería
SATEDU
Estabilización
Sensores

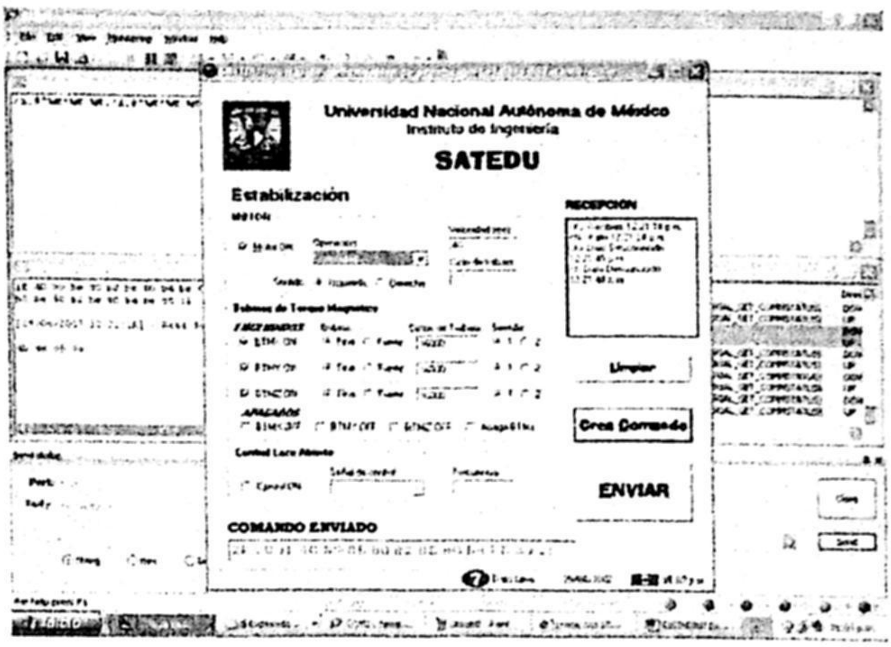


Fig. 9. Software developed for the ground station.

In addition to the system that is described in this article, a laboratory practice and other educative manuals will be elaborated later with the support of the "Instituto Tecnológico de Sonora (ITSON)" in Guaymas, México. This will allow carrying out training practices with the satellite system in a simple and friendly way, in order to teach the satellite system operation to the users. These manuals will be carefully elaborated to emphasize the clarity of concepts, as well as the progressive transference of information, besides reinforcing the learned knowledge, the questioning of treated subjects, as well as the comparison of learned concepts with the operation of commercial small satellite systems. In addition the patent acquisition as well as the possible PDS commercialization will be managed both by UNAM and ITSON Guaymas. The organization, development, and implantation of this learning and training methodology as well as future results will appear in later publications.

13 Concluding Remarks

We have presented the global architecture and operations characteristics for a portable and cost-effective system to train human resources in the field of small satellite technology. The first prototype is fully operative and its successful operations are based on 15 years of experience and work in the area. The PDS also considers the attraction of young people from our country for the satellite and the space fields. On the other hand, we expect the project will be affordable for many Institutes, Technologic Schools and Universities, and hence will represent an open gate for the new generations to participate in a field of work perceived to be far away from the academic possibilities of developing countries. In addition, the described system will attract young people to the world of Science and the Technology.

In this way, the depicted project constitutes a whole portable training system, attractive and friendly, with capabilities to be adopted as a partner to access the world of science and technology, satellite technology or towards other technological fields such as Electronics, Telecommunications and Informatics.

The system will support research in fields such as three axis stabilization, digital communications, satellite sensors, power systems, payload validation, flight computers, navigation autonomy, and satellite constellations. Those subjects will be addressed shortly with the support of this laboratory tool.

In addition the future work among UNAM and ITSON, related to the use of the PDS in laboratories from Universities, has been mentioned, along with the plans to develop efficient manuals and educative material to encourage the described satellite tool within the university environment.

References

1. Eyassat, 2007, http://eyassat.com/tiki/tiki-index.php .
2. Israel Tecch, 2007, http://www.technion.ac.il/~pgurfil/projects.html
3. USNaval, 2007, http://www.wa8lmf.net/bruninga/ea467.html .
4. Esaú Vicente-Vivas, Fabián García-Nocetti and Francisco Mendieta-Jiménez, "Automatic maintenance payload on board of a Mexican LEO microsatellite", Acta Astronautica Journal, Elsevier Science,, Volume 58, Issue 3, Pages 149-167, February 2006.
5. E. Vicente-Vivas y D.F. García Nocetti, "Computadora de Vuelo Triplex de Diseño y Manufactura Mexicana para el Microsatélite Satex", Revista Iberoamericana Información Tecnológica, ISSN 0716-8756, Vol. 17, No 1, pp. 69-76, Enero 2006, Chile.
6. E.Vicente Vivas, A.Espinoza M., C. Pineda F., J.R. Tórres F. Y A. Calvillo, "Software de Adquisición de telemetría y control de operaciones para microsatélites", 4° Conferencia Internacional en Control, Instrumentacion virtual y Sistemas Digitales "CICINDI 2002", Pachuca, Hidalgo, México, Agosto del 2002.
7. 7. Y.Tsuda et al, "University of Tokyo's CubeSat "XI" as a Student-Built Educa tional Pico- Satellite –Final Design and Operation Plan", The 23rd International Symposium of Space Technology and Science, Matsue, Japan, 2002.
8. CubeSat Origins, http://directory.eoportal.org/pres_CubeSatConcept.html .
9. The Stanford University CubeSat concept, http://cubesat.info/ .
10. CubeSats Running Projects, http://en.wikipedia.org/wiki/CubeSat#Current_running_projects .
11. H. Heidt et al, "CubeSat: A new Generation of Picosatellite for Education and Industry Low-Cost Space Experimentation ", 14th Annual/USU Conference on Small Satellites, August 2001.

Fuzzy Logic and Control

Classification and Assessment of the Water Quality using Fuzzy Inference in Shrimp Aquaculture Systems

J. J. Carbajal and L. P. Sánchez

Department of Real Time Systems and Modeling.
Centre of Computer Researches -IPN, Mexico D.F., Mexico
Phone (55) 57296000, E-mail: juancarvajal@sagitario.cic.ipn.mx

Abstract. Nowadays, there is a lack of effective tools for assessing the status of the environment in aquaculture systems, laws and standards policies does not establish measurements rules in water quality treatment. Systems based on fuzzy inference theory, have demonstrated to be useful in the treatment of environmental problems. Water quality is an important factor in faming aquaculture, and the detection of early environmental problems offers the opportunity for stopping potentially danger situations into the ponds. This study proposes a new method for evaluating the water quality in shrimp ponds based on fuzzy inference system (FIS), this method has been developed proving the importance and potentiality of the fuzzy logic theory in this area. FIS's are used to establish a relationship among the physical-chemical variables that affect the shrimp habitat. The developed model can classify the water quality status in four levels; *excellent, bad, regular* and *poor*. This work gives an alternative tool that is used in the treatment of the water management.

Keywords: Water quality, fuzzy logic, aquaculture, artificial intelligence.

1 Introduction

The water quality in oceanic research is a problem that affects daily the activities of many people that practice fishing activities. The quantity of biological data is increasing day by day and it is needed to create models of the environment conditions and to use some functional features.

The shrimp farming is an important economical activity in several countries. The farming shrimp is made in different ways; in intensive, semi intensive or extensive ponds. The shrimp production is determined by two main factors: 1) the capacity of maturing in organisms and 2) the capacity of the environment. The capacity of the environment is referred to the conditions that allow a growing and reproduction, whose would be the best in good environment conditions. The water quality into the farming ponds determines the capacity of the environment that makes influence in the life of the organism [7].

Water quality indicators have been grouped in three categories: physical, chemical and biological, each of them contains a significant number of water quality variables.

© G. Sidorov, M. Aldape, M. Martinez, S. Torres. (Eds.)
Advances in Computer Science and Engineering.
Research in Computing Science 42, 2009, pp. 337-346

Received 15/03/09
Accepted 27/04/09
Final version 08/05/09

There are variables that have more significance because they can highly affect the growing and the surviving of the organism (Table 1).

An equilibrated environment in the pond generates good growing and reproduction, a bad controlled habitat generates high stress levels in the organisms, low growing and low resistance for sickness; for example, for shrimp maturation and spawning [9], [10], [11] almost all hatcheries require availability of oceanic-quality water on a 24-h basis.

Table 1: Physical-chemical variables in a shrimp system.

High Impact	Low Impact
- Temperature	- Hydrogen Sulfide
- Oxygen dissolved	- Non Ionized Hydrogen Sulfide
- Salinity	- Nitrates
- PH	- Total inorganic Nitrogen
- Non ionized ammonia	- Silicate
	- Phosphorus
	- Chlorophyll A
	- Total suspension solids
	- Potential redox
	- Alkalinity
	- *Dioxide of Carbon*
	- Total Ammonia
	- Turbidity

2 Environmental Problems

There are a lack of methodologies that asses the water quality giving an index about its status. Some standards have been established as the NOM-001-ECOL-1996 (maximum limits of pollutants in water discharges on coastal water in Mexico) and the CE-CCA-001/89 (water quality criteria) [2], [3], [4]. However, these standards do not establish an index of the degradation status in fresh water or coastal water. International standards has been created giving a solution in this area as the ACA and NSF, however, the methodologies given by them can be only applied by fresh water bodies [5], [6]. The Canadian Council of Ministers of the Environment proposes a method that can be used in coastal water, it is based on calculate the number of failed value tests in a set of environmental variables, obtaining a water quality index (CWQI) [9]. However the ACA, NSF and CWQI indices have a number of weak points, where the lack of a reasoning process in the classification of environmental patterns is the main problem.

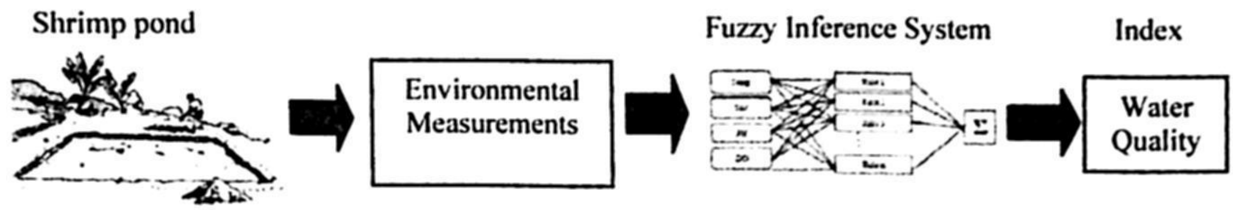


Fig. 1: The measurements of the environment will be analyzed by a model that describes the relationship between variables, obtaining a status of water quality.

3 Methods

3.1 Data Collection

Data collection was made on the shrimp farms from Sonora, Mexico. Almost 97 days were used to measure 9319 patterns; all patterns were obtained in a farming period using manual techniques and electronic devices (sensors). In the aquaculture systems, the environmental variables measured were pH, temperature, salinity and dissolved oxygen; non ionized ammonia can be estimated with pH concentrations, thus was not measured.

The environmental variables were measured with a period of 15 minutes, having a total of 96 measurements by day. The regular behaviors of each variable in one day are showed in figure 2.

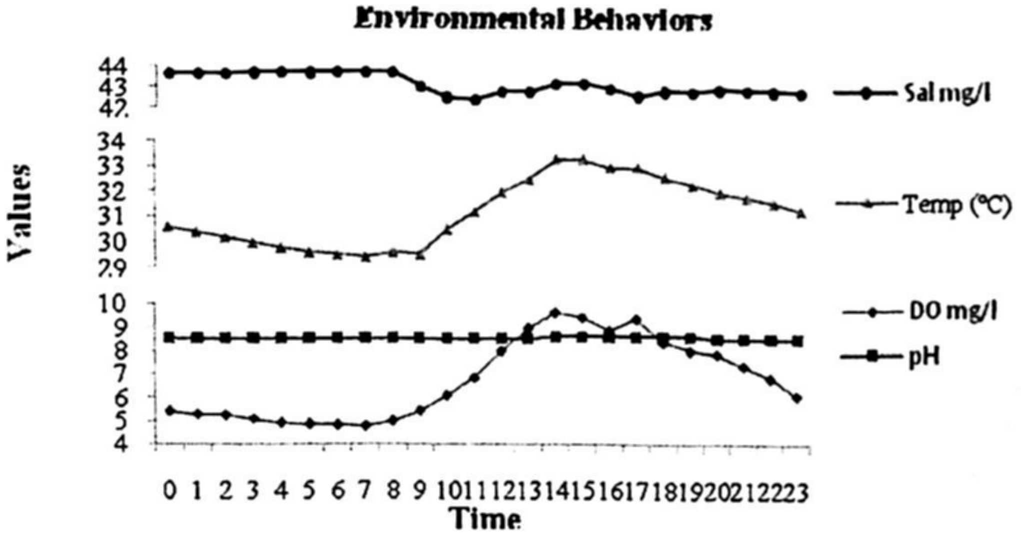


Fig. 2: Measurements one day of the physical-chemical variables in a shrimp system.

3.2 Pattern Classification

Physical-chemical variables can be classified in levels, whose represents the ecological impact in shrimp ponds, those levels are showed in Table 2.

Table 2. Ranges of classification of the environmental variables.

Variables	Hypoxia Acid	Low	Normal	High	Alkaline
Temp (°C)	------	0 - 23	23 - 30	30 - ∞	-----
Salt (mg/L)	------	0 - 15	15 - 25	25 - ∞	-----
DO (mg/L)	0 - 3	3 - 6	6 - 10	10 - ∞	-----
PH	0 - 4	4 – 6.5	6.5 - 9	9 - 10	10 - 11

In general, the environmental variables have nonlinear relations, which have been observed and proven experimentally, the equations that represent them, have been formulated, which is very hard to do. There are various equations that pretend to describe the environment, however for each local place the conditions of the environment changes and the equations created are not the indicated for that case [2], [3], [4], [5].

The water quality can be represented as a relationship between variables as follows:

$$\text{Water quality} = f(\text{Temp, Salt, DO, pH}) \tag{1}$$

where *f* is the nonlinear function, *Temp* is temperature; *Salt* is salinity and *DO* is dissolved oxygen.

The water quality status is obtained using a vector of the variables concentrations, for example in day one (Fig. 2) we have the vector v[Temp, Salt, DO, pH] = [23, 35, 3, 8], this vector is used as an input of the FIS, that calculates the water quality status (Fig 1).

3.3 Fuzzy Inference Systems (FIS)

The Fuzzy inference systems (FIS) theory was applied in this study providing a non-linear relationship between input sets (Physical-chemical variables) and output set (Water Quality Index) [10], [4]. Fuzzy inference systems (Fig. 3) uses propositions that are represented with a true or false level, while a Boolean logic proposition is only true or false [12].

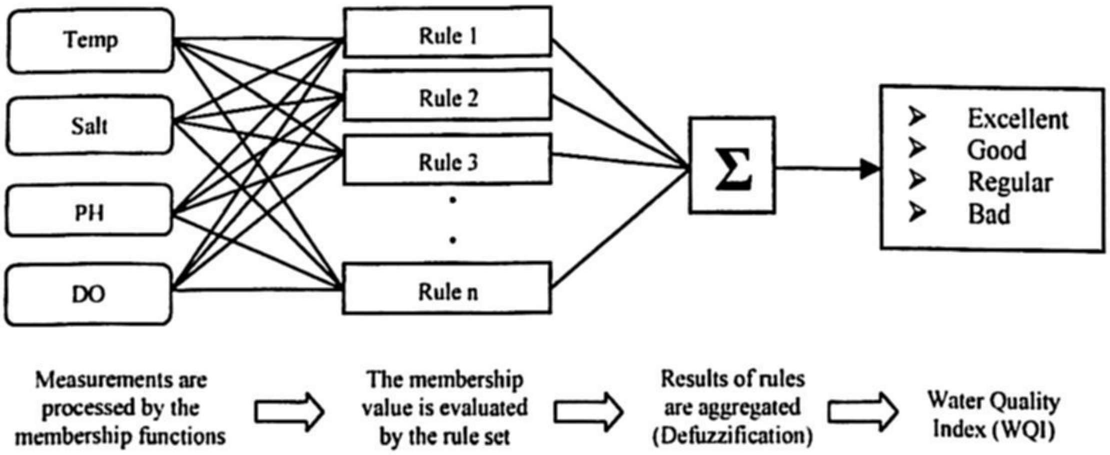

Temp
Salt
PH
DO
Rule 1
Rule 2
Rule 3
Rule n
Σ
Excellent
Good
Regular
Bad
Measurements are processed by the membership functions
The membership value is evaluated by the rule set
Results of rules are aggregated (Defuzzification)
Water Quality Index (WQI)

Fig. 3: Architecture of the Fuzzy Inference System applied to the water quality problem in shrimp farms.

The fuzzy logic operators can be used as the basis of the inference systems. Such inference methods have been studied by the expert system community. The fuzzy logic involves three important concepts: membership functions, fuzzy set operations and inference rules.

A membership function is a curve that maps an input real value in a membership value (μ) between 0 and 1. The input space is called universe of discourse (X). A fuzzy set is represented as a set of ordered pairs that assigns a membership level to each element x of the universe X.

$$A = \{x, \mu_A(x) | x \in X\} \tag{2}$$

where $\mu_A(x)$ is the membership function x in A. The election of the curve is arbitrary and it depends of the context of the problem. The operations that define de basis on fuzzy logic can be defined as:

Union (OR) $$\mu_{A\cup B}(x) = max\{\mu_A(x), \mu_B(x)\} \tag{3}$$

Intersection (AND) $$\mu_{A\cap B}(x) = min\{\mu_A(x), \mu_B(x)\} \tag{4}$$

Complement (NOT) $$\mu_{\bar{A}}(x) = 1 - \mu_A(x) \tag{5}$$

Fuzzy systems are based on knowledge rules If – Then. A fuzzy rule is a condition which some words are characterized for having a membership function. The inference process uses the fuzzy rules "IF – Then" creating a relationship between the fuzzy input set (membership values) and the output set. A fuzzy rule has the form:

If DO if low, then WQI is regular

where DO is the membership level of the measured concentration and WQI is the final membership function to analyze. There are some expressions that are frequently used by experts in water quality, that expressions will be helpful for the construction of the FIS, they structure are the follows: if dissolved oxygen is low and the salinity is high, then the water quality is regular. This kind of expressions built the fuzzy language of the FIS and it is represented as follows:

Rule 1: If Temp is normal and Salt is normal and pH is normal and DO is normal then WQI is Excellent
Rule 2: If Temp is normal and Salt is normal and pH is normal and DO is low then WQI is Good
Rule 3: If Temp is normal and Salt is High and pH is normal and DO is low then WQI is Regular

The size of the set rule depends of the number of rules that are involved in the environment; a total of 139 rules have been used in this case.

Trapezoidal membership functions define these fuzzy sets (Fig. 4), they are represented as:

$$\mu(x) = min\left\{\frac{x-a}{b-a}, 1, \frac{d-x}{d-c}\right\} \tag{6}$$

Fig. 4 shows the membership functions used in the fuzzyfication process and the output membership function.

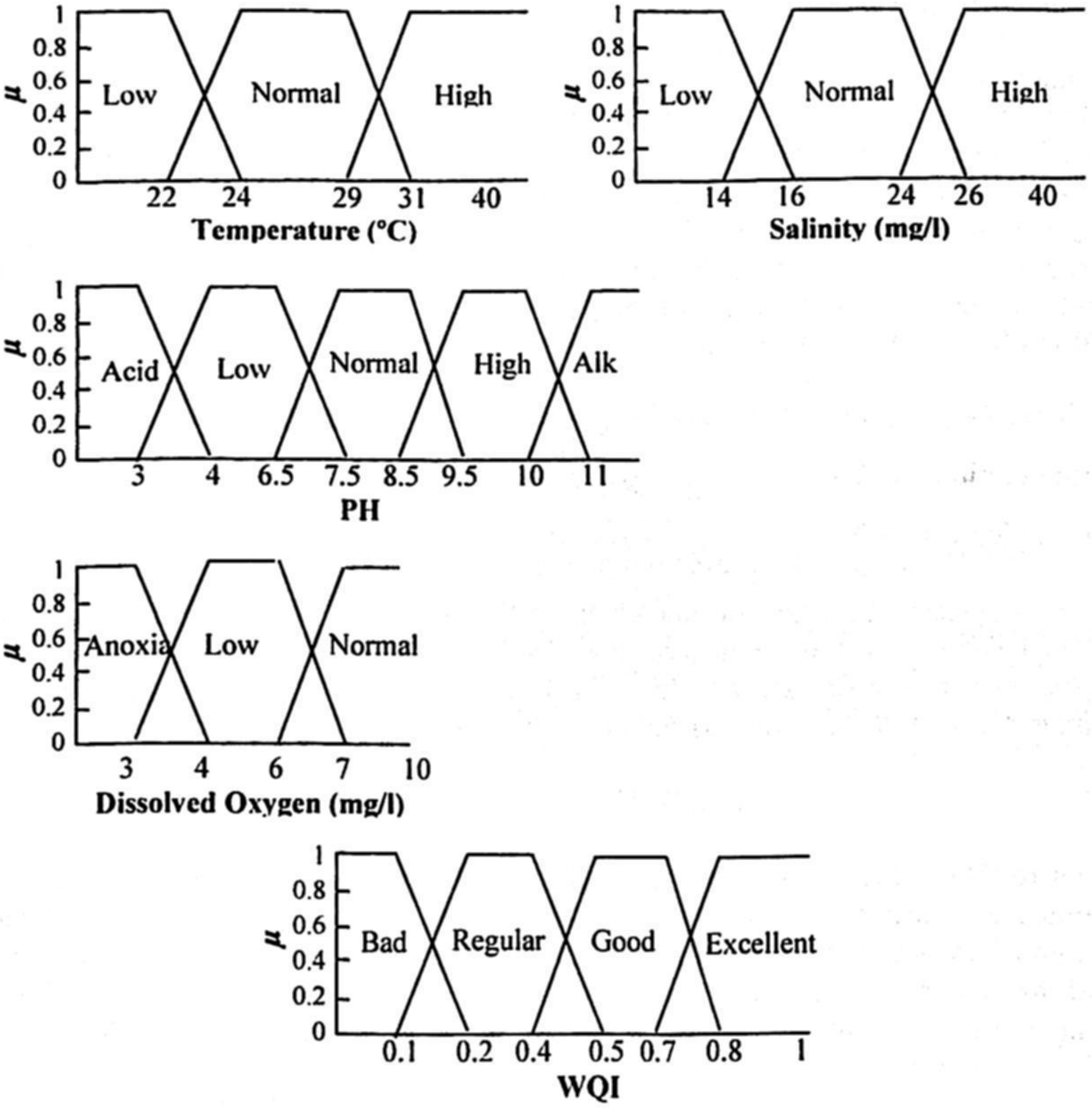


Fig. 4: Membership functions for temperature, salinity, dissolved oxygen, pH and WQI.

The output membership function is the water quality indicator (WQI). The output of the FIS is expressed in a [0, 1] range, where 0 means "poor" and 1 "excellent", the "good" and "regular" status are involved within the range, this is represented in Fig. 4.

The fuzzy outputs of the if-then rules in a real world value are joined by the aggregation process (Fig. 5) [10]. When the final membership function is obtained $\mu_{out}(x)$, the gravity center is calculated using the following equation:

$$WQI = \frac{\int x\mu_{out}(x)dx}{\int \mu_{out}(x)\,dx} \tag{7}$$

If we assume that it is necessary to evaluate the WQI in a shrimp pond, using the rule 1 and having the variables Temp, Sal, pH and DO with values 25.0, 20.0, 8.0 and 6.3 respectively, we can calculate using the membership functions proposed in the Fig. 4. For "R1" and "R2" we can calculate:

$$R_1: \mu_{out}(x) = \min\{\mu_{Temp-n}(x), \mu_{Salt-n}(x), \mu_{pH-n}(x), \mu_{DO-n}(x)\} = \min\{1, 1, 1, 0.3\} = 0.3$$

$$R_2: \mu_{out}(x) = \min\{\mu_{Temp-n}(x), \mu_{Salt-n}(x), \mu_{pH-n}(x), \mu_{DO-l}(x)\} = \min\{1, 1, 1, 0.7\} = 0.7$$

Where *n* is normal, *l* is low and μ_{out} is the membership value calculated in R1 and R2. Calculating the aggregation functions:

$$\mu_{R1}(x) = \min\{\mu_{out}(x), \mu_{excellent}(x)\} = \{0.3, \mu_{excellent}(x)\} = 0.3$$

$$\mu_{R2}(x) = \min\{\mu_{out1}(x), \mu_{good}(x)\} = \{0.7, \mu_{good}(x)\} = 0.7$$

Calculated the output membership functions (fuzzy outputs), the aggregation of these functions will generate one membership function, this is showed in Fig. 5. The WQI is evaluated using the centroid method (Eq. 7) and replacing the membership functions:

$$WQI = \frac{\int_{0.4}^{0.47}(10x-4)x\,dx + \int_{0.47}^{0.73}(0.7)x\,dx + \int_{0.73}^{0.77}(-10x+8)x\,dx + \int_{0.77}^{1}(0.3)x\,dx}{\int_{0.4}^{0.47}(10x-4)\,dx + \int_{0.47}^{0.73}(0.7)\,dx + \int_{0.73}^{0.77}(-10x+8)\,dx + \int_{0.77}^{1}(0.3)\,dx} = 0.663$$

2.4 Postprocessing

The results evaluated by the defuzzification method is one step for the assessment of the WQI, however, even the membership functions of WQI has a domain [0,1], the process cannot obtain the limit values, if the result of the output function is fully excellent, the highest value calculated is the media of the length of the output function and the final value never is obtained, this is represented using the following equations:

$$\mu_{excellent}(x) = \frac{\int x\mu_{Excellent}(x)\,dx}{\int \mu_{Excellent}(x)\,dx} = 0.873 \tag{8}$$

$$\mu_{Bad}(x) = \frac{\int x\mu_{Bad}(x)\,dx}{\int \mu_{Bad}(x)\,dx} = 0.0777 \tag{9}$$

For scaling the FIS output, the following operation fits the final value in a [0, 1] range:

$$X_{out} = \frac{X - \min(X)}{\max(X) - \min(X)}$$

Where X is the original vector and X_{out} is the linear scaled vector. The final levels of the WQI index are 1 for *excellent*, 0.66 for *good*, 0.28 for *regular* and 0 for *bad*.

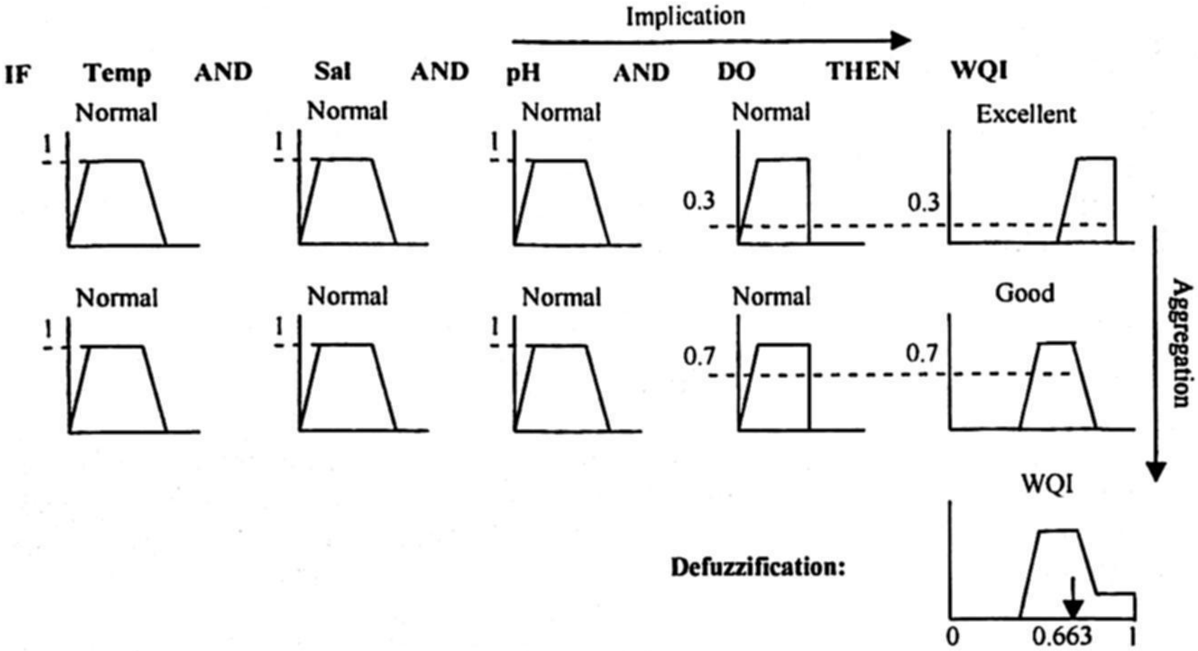


Fig. 5: Defuzzification process where the final membership functions are obtained using the aggregation of the fuzzy outputs.

4 Results

A set of measurements have been evaluated for proving the WQI. Input data extracted from shrimp farms databases have been used to assess water quality the 2007 farming period.

The validation of an index is not an easy task; although indexing processes can miss information their benefits are significant when measuring environmental impacts.

The real validation process is done when the model is proved with real values; however, the most relevant aspect to highlight here is the methodology used in this work to develop the WQI index. A comparison of the performance between deferent methodologies is proposed in this investigation. In Fig. 6, the WQI index is compared with the CCME. A comparison of the performance for the WQI and the CCME index used in environmental assessment could remark some interesting things.

The analysis of Fig. 6 shows the results of the assessment of the CCME and the WQI proposed with a data set of tree days, each day corresponding to different months of the farming period. The CCME and the WQI indices report different behaviors into the pond; the treatment of the information within the FIS influences the final score.

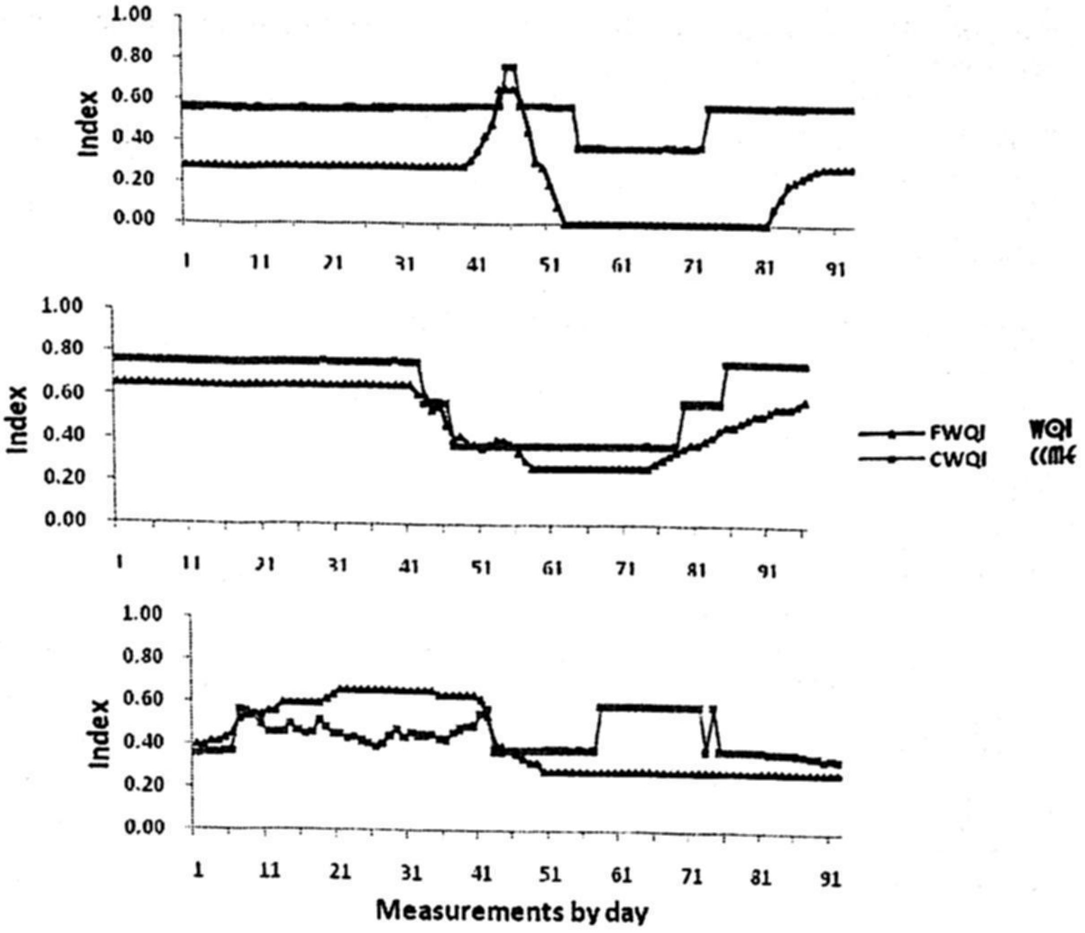


Fig. 6: Comparison of the WQI and CCME indices.

The CCME index shows values between 4.0 (regular) and 8.0 (good) in a non fuzzy environment. The WQI index shows values between 0.0 (poor) and 7.0 (good). The CWQI index is above of the FWQI index; although the final values are similar they have some differences. The main difference is because the WQI is calculafed with reasoning cases that affects the status of the water, and the especial situations are processed by the WQI, as an example we can refer the hypoxia situations with the DO

5 Conclusion

An original methodology to evaluate and classify an ecological impact in the habitat of shrimp aquaculture systems has been created. This research establishes an index that asses the status of the artificial shrimp habitat based on four levels; excellent, good, regular and poor.

In this paper, we present a robust tool for water management in shrimp ponds; the methodology is presented in the form of a fuzzy inference system and it is based on a reasoning process implemented to evaluate information data. The necessity to link fuzzy systems and reasoning methodologies to evaluate water bodies has been showed. Studies are now needed to tune the different cases that are presented in the environment.

This model can be adjusted by assessing other aquaculture marine systems that have the necessity of analyze the status of their ecosystem. Nowadays some researchers are evaluating the possibility of use this model for embedding in expert systems or for creating predicting models that allows recognizing undesirable status into the ponds.

Acknowledgements. Authors wish to thank CONACYT and the National Polytechnic Institute for supporting this work.

References

1. [ACA] Agencia Catalana del Agua (Catalonia, Spain). 2005. Available at: http://www.mediambient.gencat.net/aca/ca/inici.jsp [Accessed August 2007].
2. Brown, D.: A qualitative ecological model to support mariculture pond water quality management, Bioinformatics, Vol. 11 (1995) 595-602.
3. [CCME] Canadian Council of Ministers of the Environment (Canada). 2004. An assessment of the application and testing of the water quality index of the Canadian Council of Ministers of the Environment for selected water bodies in Atlantic Canada. National indicators and reporting office. Available at:http://www.ec.gc.ca/soer-ree/N [Accessed August 2007].
4. Hernández. J., Zirino, A., Marione, S., Canino, R., Galindo, M.: PH-density relationships in seawater. Ciencias Marinas, Vol. 29 (2003) 597-508.
5. [INE] Instituto Nacional de Ecología: La calidad del agua en los ecosistemas costeros de México. 2000.
6. Li, Y., Li, J., Wang, Q.: The effects of dissolved Oxygen Concentration and Stocking Density on Grown and Non-Specific Immunity in Chinese Shrimp, Fenneropenaeus Chinensis. Aquaculture, Vol. 256. Elservier (2006) 608-616.
7. Martínez Córdova Rafael M., "Cultivo de Camarones Pendidos. *Principios y Practicas*", Ed. AGT Editor S.A.. 1994.
8. [NSF] National Sanitation Foundation International. 2005.
9. Available at: http://www.nsf.org [Accessed August 2007].
10. Ocampo, W., Ferré, N., Domingo, J., Schuhmacher, M.: Assessing water quality in rivers with fuzzy inference systems: A case study. Environment International, Vol 2. Elservier (2006) 733-742.
11. Rodríguez, A., Antonio, J.: Aplicaciones de lógica difusa en ingeniería gráfica, XVI Congreso Internacional de Ingeniería Gráfica (2004).
12. [SEMARNAP], Secretaría de Medio Ambiente, Recursos Naturales y Pesca.: NOM-001-ECOL-1996
13. Rodríguez, A., Antonio, J.: Aplicaciones de lógica difusa en ingeniería gráfica, XVI Congreso Internacional de Ingeniería Gráfica (2004).

Intelligent Control at the Coagulation Process in a Drinking Water Treatment Plant

Hector Hernandez[1], Madaín Pérez[1], Jorge Camas[1],
Rafael Mota[1], Nicolás Juárez[1] and Marie-Véronique Le Lann[2]

[1]Instituto Tecnológico de Tuxtla, ITTG
Carretera Panam. Km. 1080, 29050-Tuxtla Gutiérrez, Chiapas, México
(hhernandezd,mperez,jcamas,rmota,njuarez)@ittg.edu.mx
http://www.ittg.edu.mx

[2]INSA, Département de Génie Electrique et Informatique
135, Avenue de Rangueil 31077 Toulouse cedex 4 France
mvlelann@laas.fr
http://www.insa-tlse.fr

Abstract. With increasing demands for high precision autonomous control over wide operating envelopes, conventional control engineering approaches are unable to adequately deal with system complexity, nonlinearities, and temporal parameter variations, and with uncertainty. Intelligent Control or self-organising/learning control is a new emerging discipline that is designed to deal with problems. The coagulation unit is a major step in the production of potable water, allowing the removal of colloidal particles and contamination sources. In order to obtain a simple model to describe the water treatment plant, a behavior model sets out, from the analysis of raw water characteristics to the entrance of the plant: (1) to develop a software sensor based on artificial neural networks for predicting on-line the amount of optimal coagulant dosage, and (2) the determination of the functional states in real time (diagnosis system based on fuzzy classifier).

1 Introduction

The water industry is facing increased pressure to produce higher quality treated water at a lower cost. The drinking water plant object of this study is the "SMAPA" (SMAPA, 2007[1]) of Tuxtla city in Mexico. The control of the plant is fundamental to maintain a good quality of service. This has motivated important efforts in the development of the methods of control and automatic monitoring in the last few years [2]. Coagulation process is one of the most important stages in surface water treatment, allowing the removal of colloidal particles [3].

This paper addresses the problem of coagulation control based on the raw water characteristics such as turbidity, temperature and pH, in a global system including the analysis and the determination of the functional states and the detection of fault. Coagulant dosing is not only one of the major control parameter in coagulation

© G. Sidorov, M. Aldape, M. Martínez, S. Torres. (Eds.)
Advances in Computer Science and Engineering.
Research in Computing Science 42, 2009, pp. 347-358

Received 09/03/09
Accepted 27/04/09
Final version 06/05/09

process, but also the major operation cost in water treatment plant. Most coagulant dosing is determined by the way of jar test. However, the jar test can only provided periodic operation information, which can not be applied to real-time control of the coagulation process, especially with a time-varying raw water quality (principally during a spring-summer runoff period of the time). Excessive coagulant overdosing leads to increased treatment costs and public health concerns, while underdosing leads to a failure to meet the water quality targets and less efficient operation of the water treatment plant.

One of the fields of machine learning more developed is the classification. This method, based on Fuzzy Logic, presents many advantages which make it well adapted to chemical or biotechnological complex processes. In the monitoring of dynamic processes, the state of operation of the system determines the situation or operating condition. The objects to be classified are the situations observed in real time. Good coagulation control is essential for water quality and economic plant operation [4], [5]. Finding a general mathematical model for biotechnological process is somehow complex. During the last decade a number de models based on artificial neuronal networks (ANN) have been developed and applied for predicting coagulant dose concentrations of water treatment process (Baba, 1990 [6]; Mirepassi, Cathers and Dhamarppa, 1995 [7]; A fuzzy neural system, coupling of ANN with fuzzy theory, has been applied to extract the control rules by learning story operational data of water treatment process [8]. Another works based on ANN's involved in the production of potable water have been carried out (Fletcher et al. 2001[9]; Baxter et al., 2002[10]; Peijin and Cox, 2004[11]).

This paper addresses the problem of automatic coagulation control based on the raw water characteristics such as turbidity, pH, temperature, etc. Some recent studies (Valentin, PhD thesis 2000 [3]; Lamrini and LeLann, 2004 [2]; and Hernandez, PhD thesis 2006 [12]) have shown the potential effectiveness of such an approach based on ANN's by means of the implementation of a neural software sensor for on-line prediction of the coagulant dosage. The innovative aspect of this work resides in the integration of various techniques in a global system including data pre-processing, automatic control of coagulation, analysis of uncertainties and the possibility of integration as entry to a system of diagnosis, which should allow the portability of the system at low cost from one site to another.

A brief description of the water treatment plant and the operation units involved in the drinking water treatment process is first provided in section 2. The LAMDA technique of classification is described in section 3. The methodology used for the design and synthesis of the Neural Software Sensor and system diagnosis is given in section 4. Finally, experimental results are presented and discussed in section 5.

2 Water Treatment Process

2.1 Overview of Water Treatment Operations

The water is the most abundant compound on the surface of the world. Water treatment involves physical, chemical and biological processes that transform raw water into drinking water which satisfies a whole of standards of quality at a reasonable price for the consumer.

The "SMAPA" water treatment plant (Tuxtla city, Mexico), which was used as an application site for this study, provides water to more than 800,000 inhabitants and has a nominal capacity to process 800 l/s of water per day. The figure 1 presents a schematic overview of the various operations necessary to treat the water, the available measurements, and the coagulant dosing point. The complete usual chain comprises the 5 great following units: pre-treatment, pre-oxidation, clarification, disinfection, and refining. The present work concerns essentially the coagulation process. Raw water is abstracted from the river "Grijalva" and pumped to the treatment works. Water treatment plants invariably include two main process units, clarification and filtration. Other units may be required depending of the quality of the water source. The coagulation process is brought about by adding a highly ionic salt (aluminum sulphate) to the water.

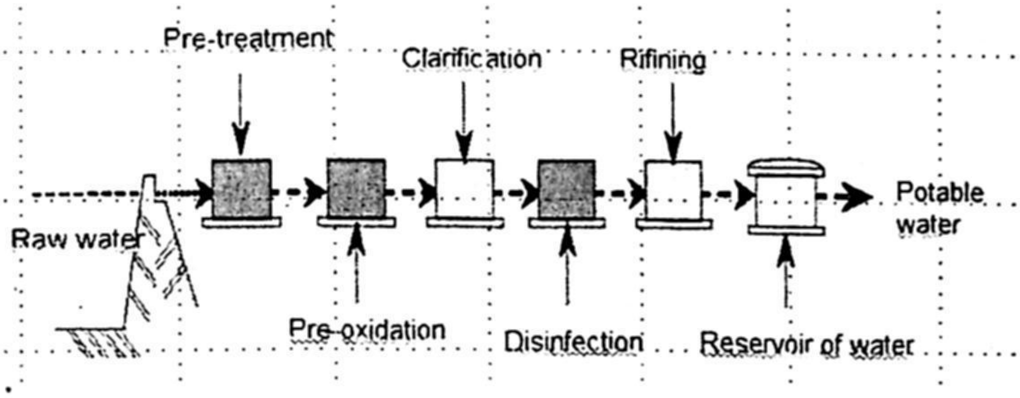


Figure 1. Potable water plant.

A bulky precipitate is formed which electrochemically attracts solids and colloidal particles. The solid precipitate is removed by allowing it to settle to the bottom of the tank and then periodically removing it as sludge. The coagulation process accounts for the removal of most of the undesirable substances from the raw water and hence tight monitoring and control of this process is essential. The next stage is filtration, where the particles passing trough the previous stages are removed. The final stages in the process are chlorination and pH adjustment. The water is then stored in a tank and ready to be transported through the water supply network.

2.2 Coagulation Control

Surface waters contain both dissolved and suspended particles. The suspended particles vary considerably in source, composition charge, particle size, shape and density. The correct design of a coagulation process and the selection of coagulants depend upon understanding the interactions between these factors. This process is one of the most important stages in surface water treatment, allowing for the removal of colloidal particles. The main difficulty is to determine the optimum quantity of chemical reagent related to raw water characteristics. Poor control leads to wastage of expensive chemicals, failure to meet the water quality targets, and reduced efficiency of sedimentation and filtration processes. In contrast, good control can reduce manpower and chemical costs and improve compliance with treated water quality targets. The traditional method of controlling coagulant dose, called the jar-test, relies heavily upon human intervention. It involves taking raw water samples and applying different quantities of coagulant to each sample. After a short period of time each sample is assessed for water quality and the dosage that produces the optimal result is used as a set point. Operators change the dose and make a new jar test if the quality of treated water changes. Disadvantages associated with such a procedure are the necessity to rely on manual intervention, and lack of adaptation to abrupt changes of water characteristics. The objective of this paper is to provide a complementary support to the jar-test allowing for the automatic determination of optimal coagulant dose from raw water characteristics, using an artificial network approach. This approach requires the availability of on-line water quality measurements at an upstream survey station.

3 LAMDA classification technique

LAMDA (Learning Algorithm for Multivariate Data Analysis) methodology is a classification technique developed by Joseph Aguilar-Martin and others in LAAS-CNRS [13]. Previous works [14], [15], [16], [17], described in detail the methodology, as well as the algorithms and functions used. We will limit ourselves in this work to present the main characteristics of the methodology and its general operation. LAMDA algorithm represents a system of classes or concepts by means of the logic connection of all marginal information available [18]. LAMDA is a fuzzy methodology of conceptual clustering and classification, in this way, the global adequacy of an object (individual) to a class is equivalent to its membership to the class, and it is calculated from the marginal adequacy of each attribute and according to the heuristic rule of Maximum Adequacy. An object is then assigned to the class which presents the greater adequacy degree. The total indistinguishability or homogeneity inside the description space from which the information is extracted is introduced by means of a special class called the Non-Informative Class (NIC), this class accepts the same adequacy degree any object of the description space, and a minimum classification threshold is therefore induced.

Let us consider a collection of objects or situations *X*, and a finite set of n qualitative or quantitative descriptors, **A**. An object is represented by a vector $\breve{x}$ and its jth component is the value taken by the jth descriptor. The information conveyed by each descriptor contributes to the membership of the element to the class by means of the Marginal Adequacy Degree (MAD). The Global Adequacy Degree (GAD) to a class is a Fuzzy Logic combination of the MAD's as shown in figure 2. Piera and Aguilar-Martin (1991) introduced the mixed connectives of lineal compensation that interpolate between a conjunctive and a disjunctive logic operator. It was shown that such interpolation is completely ordered with respect to the "exigency degree", the highest in the conjunction case (AND) and the lowest in the disjunction one.

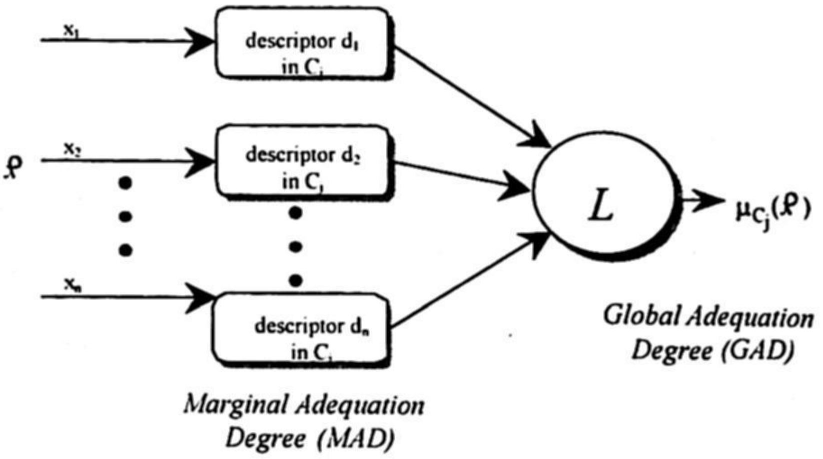

x_1
x_2
x_n
$\mathcal{X}$
descriptor d_1 in C_j
descriptor d_2 in C_j
descriptor d_n in C_j
L
$\mu_{C_j}(\mathcal{X})$
Global Adequation Degree (GAD)
Marginal Adequation Degree (MAD)

Figure 2. The Marginal and Global Adequacy Degree (MAD and GAD).

The computation for the class parameters is done independently for each component, using separately the measurement given by each sensor. Only averaging functions are needed for that algorithmic step. Each vector $\mathcal{X}$, or situation is assigned to one of the existing classes according that its adequacy GAD exceeds the NIC adequacy, otherwise this object is assigned to the NIC class in case of passive recognition. In the learning case, this vector starts the creation of a new class and contributes to its initialisation; therefore the previous knowledge of the number of classes in not needed. Waissman et al. (1998) have been proposed several MAD functions, as well as their corresponding sequential learning algorithms.

4 Methodology of Neural Software Sensor and the System Diagnosis

The coagulation process is difficult to model using traditional models. The coagulant dosage ensuring optimal treatment efficiency has been shown experimentally to be non-linearly correlated to raw water characteristics which are usually available on-line. The system developed for the prediction of the optimal coagulant dosage and the system diagnosis it was divided into two modules: (1) Determination of coagulant dosage using Artificial Neural Networks (software sensor), (2) System diagnosis that allows the process expert to obtain the classification that represents the best the process situations in order to use it later for pattern recognition.

4.1 Mode of the Prediction of Optimal Coagulation Dosage

The general method for the prediction of coagulant dose is shown in Figure 3. We analyze the following stages: pretreatment of data descriptors of the water by using ACP (principal component analysis), and the implementation of an iterative procedure to test the data for training and testing determining a confidence interval of prediction using the bootstrap technique (software sensor).

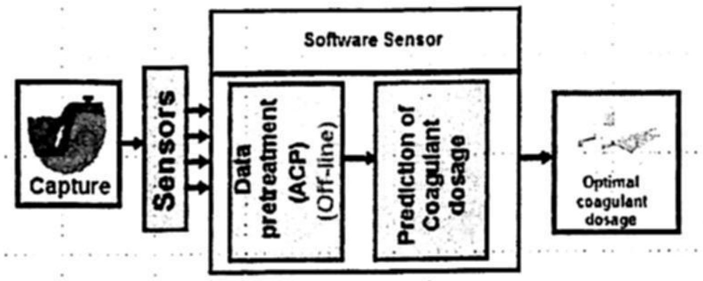

Software Sensor
Capture
Sensors
Data pretreatment (ACP) (Off-line)
Prediction of Coagulant dosage
Optimal coagulant dosage

Figure 3. Model of the prediction of optimal coagulation dosage.

4.2 Principal Components Analysis (PCA)

PCA is one of the multivariate methods of analysis and has been used widely with large multidimensional data sets. The PCA method is applied to determine the main characteristic of the variables necessary for the prediction of the optimal coagulant dosage. These characteristic variables are considered as the input variables to the neural model for which the training algorithm is performed. The PCA generates a new set of variables called principal components. Each principal component is a linear combination of the original variables. All the principal components are orthogonal to each other so there is not redundant information. The principal components as a whole form an orthogonal basis for the data space. The first principal component is a single axis in space. When each observation is projected on that axis, the resulting values form a new variable, and the variance of this variable is the maximum among all possible choices of the first axis. The second principal component is another axis in space, perpendicular to the first. Projecting the observations on this axis generates another new variable. The variance of this variable is the maximum among all possible choices of this second axis. The full set of principal components is as large as the original set of variables. But is common for the sum of the variances of the first few principal components to exceed 80% of the total variance of original data.

4.3 Artificial Neural Networks (ANN's) and Multi-Layer Perceptron (MLP)

ANN's are one of the earliest adaptive techniques in engineering and computing science. The concept of ANN was inspired by the way of the biological brain processes information. An ANN is a network of neurons or processing elements and weighted connections [19]. ANN is fundamentally a mathematical model composed by a set of nodes (artificial neurons) where information is processed. An ANN can be classified into two different categories: unsupervised or supervised learning. In the situation of an unsupervised model, the networks seek to identify features of the training patterns without external assistance. On the other hand, for the supervised learning process is necessary to use in the train input-output patterns. For each input pattern, the network generates an output pattern, which is compared with the desired output and the adaptation of the model parameters is made in relation to the observed error. One of the most studied and used ANN architecture is the Multi-Layer Perceptron (MLP). The prediction of optimal coagulant dosage from water characteristics is a non linear regression problem which can be tackled using MLP's. Consists of an input-output network, which have the neurons distributed by several layers, fully connected between adjacent layers, and where the flow of information is done in a feed-forward way. The MLP is usually trained by gradient descent methods [20], in which the error is propagated backwards through the network. Figure 4 shows a MLP with three layers: an input layer (variables of the raw water quality parameters) with n neurons, a hidden layer with H neurons and a layer with one output neuron (variable of the optimal coagulant dosing rate).

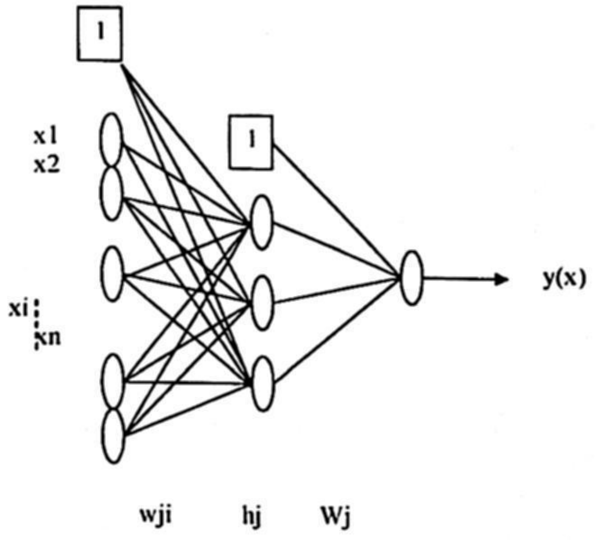


Figure 4. Multi-Layer Perceptron (MLP).

The expression of the output of MLP is given by:

$$y(x) = \sum_{j=1}^{H} W_j h_j + W_0 \qquad \text{and} \qquad h_j = \sum_{i=1}^{n} w_{ji} x_i + w_{j0}$$

Where: W_{ji} are the weights between the input layer and the hidden layer and W_j the weights between the hidden layers and the output layer.

The input nodes do not make any kind of processing and sending the input patterns to the first hidden layer is their only function. Conversely, the neurons from the other layers have the capacity of processing the received information. Each one of them performs two different operations: the weighted sum of its inputs (using the weights associated with the existing links between this neuron and the others from the previous layer), followed by a non-linear transformation (called by activation function or transfer function). The resulting output from these two actions is then sent on to the next layer. To sum up, if we have a MLP such as the one represented in figure 2 and with the same activation function, h , in all its neurons, then it can be described mathematically as:

$$h_j = \sum_{i=1}^{n} w_{ji} x_i + w_{j0}$$

The transfer function can be any function, but for most practical uses of neural networks it is important to have a continuous, completely differentiable function. Over the years many transfer functions have been proposed [21], but the most prominent ones for neural networks are linear and transfer function. In our case, we know that the relationship between the coagulant dosage with the raw water characteristics is non-linear so the choice of a sigmoid transfer function has been made. The method traditionally used to perform the training of such networks, e.g. to adjust the weighted connections, is the Backpropagation learning algorithm. The term Backpropagation refers generally to the manner in which the gradient is computed for non-linear Multi-layer networks. There are a number of variations on the basic algorithm which are based on other standard optimization techniques, such as conjugate gradient, Newton and Levenberg-Marquardt methods. Learning occurs when the network emulates the non-linear function underlying the training data set. The weighted connections are adjusted by minimizing the following criteria derived from the difference between real and neural outputs respectively t and y, as:

$$C = \frac{1}{N} \sum_{i=1}^{N} (t_i - y_i)^2$$

4.4 System Diagnosis

This part is a general procedure for monitoring the plant. It comprises the following modules: (1) Data pretreatment, (2) The method of form recognition (fuzzy classifier) to generate the class association to functional states of the process (Figure 5) and considering the stage of recognition in real time (Figure 6). At this part it is very important the active participation of the expert, both in the learning phase and the online recognition phase, which displays the current status of the water treatment plant.

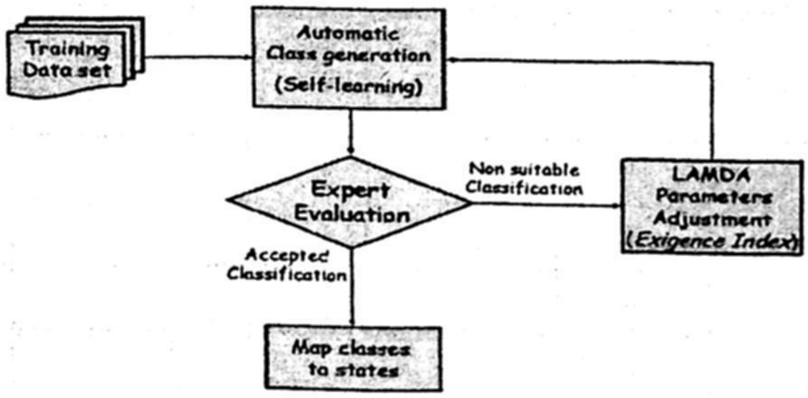


Figure 5. Association classes to functional states (analysis of historical data-non-supervised learning).

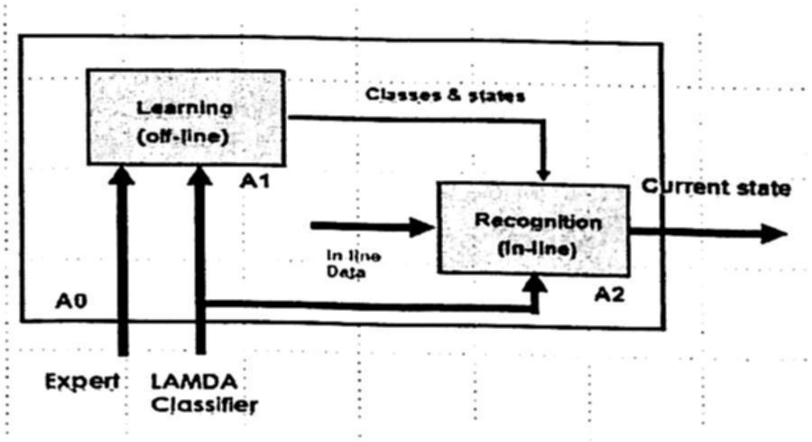


Figure 6. In-line functional states recognition (LAMDA fuzzy classifier).

5 Application the Method to the SMAPA Plant

The raw database consists of 728 measurements of 9 variables during a period of 24 months (2006-2007). Every sample underwent to different physical and chemical analysis as well as to the jar-testing in order to determine the coagulant dosage.

5.1 Prediction of Coagulant Dosage

After implementing the ACP, the number of input neurons is 4 (TUR, TC, TEMP and pH). The training was carried out over the first year (2006). The average error represented by the Matlab criterion *MSE* [22] calculated by the network is 0.085 on the training set corresponding to data of year 2007. The validation of the ANN has been performed on test data of year 2007 not included in the initial training set: the criterion *MSE* is a little higher (0.092). Figure 7 shows the prediction accuracy of the ANN model for this validation set of year 2007. The predictions given by neural network (point line) are very close to the real data.

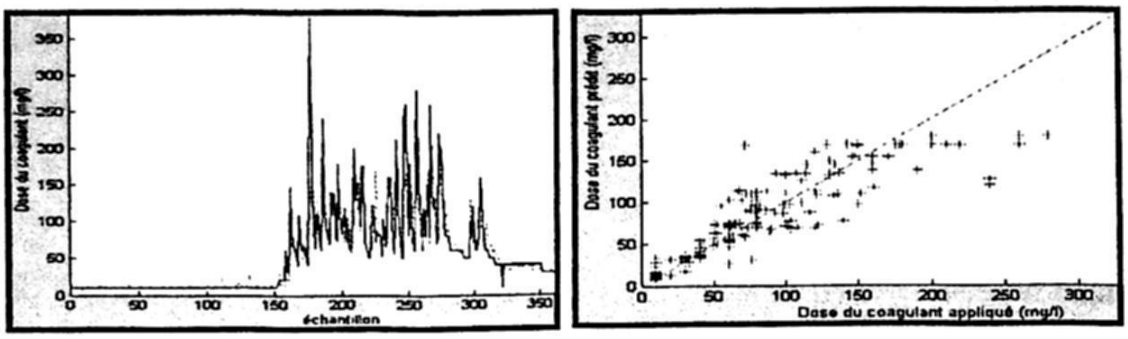


Figure 7. Left: actual (thin line) versus predicted (point line) coagulant dosage with ANN model on test data. Right: predicted versus target coagulant dose.

ANN model is compared with a multi-linear decomposition of the dose versus with the same input variables. Prediction accuracy is clearly poorer than the one of the neuronal model performance.

5.2 Results with the system diagnosis

Functional states of the water treatment process and according to the method proposed in the previous section, using the LAMDA classification algorithm as a general strategy, for obtaining the model of the plant (using historical data), as in recognition of the functional states in real time according to online measurement of the variable characteristics of the water at the entrance to the plant. Figure 8 shows the results for 5 classes. These classes are associated with 5 functional states (shown in figure) with the assistance of the expert from the drinking water treatment.

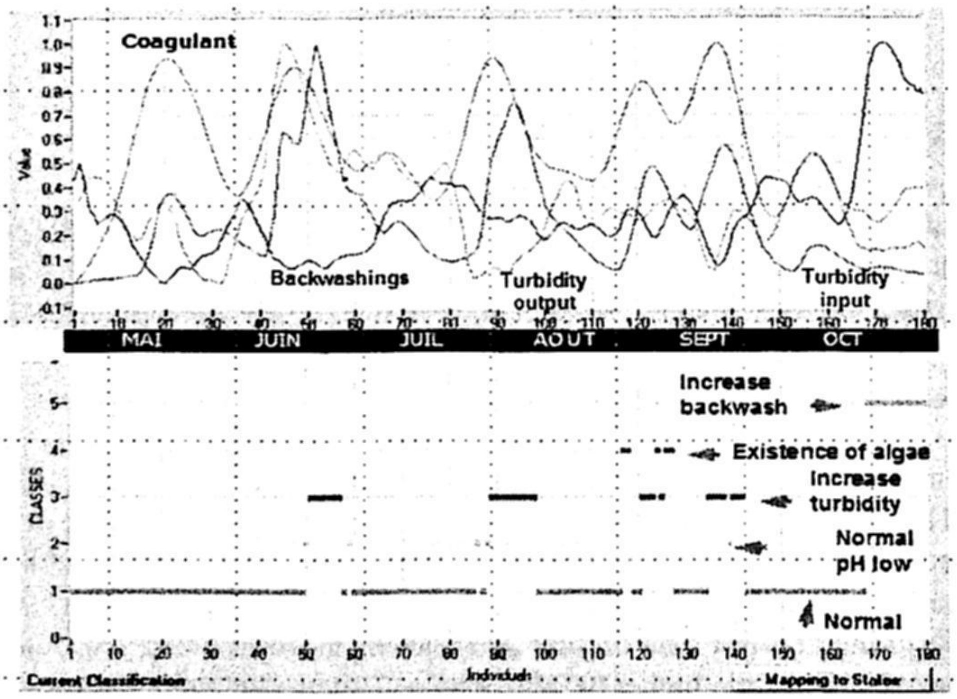


Figure 8. In-line functional states recognition (LAMDA fuzzy classifier) [23].

6 Conclusions

A neural software sensor for coagulation control was developed from Artificial Neural Networks. He supplies us on real time the coagulant dosage to inject in the coagulation unit, key stage of the process in a water treatment plant. The selection of the entries of the network was made by using the statistical technique ACP to allow eliminating the redundant information. Experimental results using historical real data have demonstrated the efficiency of this approach.

The LAMDA fuzzy classifier technique has been used to help to the interpretation of operational behaviour of a potable water treatment plant (SMAPA plant unit of Tuxtla Gutierrez, Mexico). Their classification under the unsupervised mode (i.e. with automatic class generation but controlled by the expert through the parameter α) exhibited 5 classes which could be aggregated to 5 functional states. This paper showed the realization of an intelligent monitoring system, where the reasoning task consists in the combination of data inspection and classification situations and of the expert knowledge. This strategy will be applied to the monitoring of the overall water treatment plant with measured data from different units in the plant (not only in the coagulation unit). Then, this strategy will be implemented on-line.

The water treatment units contain complex processes. Few researches so far concerned their control or diagnosis. However, front of the more and more necessity of producing a water of constant quality, the producers of potable water become made sensitive to any technique allowing to answer quickly this requirement. For that purpose, the future works aim at establishing a methodology of diagnosis based on the use of wireless sensors and the interpretation of the information obtained on all of the water treatment plant by including the value of coagulant dosage calculated by the neural software sensor.

References

1. SMAPA (2007). Sistema Municipal de Agua Potable y Alcantarillado de Tuxtla, Manual de Procedimientos. Tuxtla Gtz., Chiapas, México.
2. Lamrini, B., Benhammou, A., Le Lann, M-V.: Construction d'un capteur logiciel pour la prédiction de la dose du coagulant: application à une station de traitement d'eau potable. Laboratoire d'automatique, UCA, Marrakech-LAAS/CNRS, Toulouse, France (2004).
3. Valentin, N.: Construction d'un capteur logiciel pour le contrôle automatique du procédé de coagulation traitement d'eau potable. Thèse de doctorat, UTC/L.desEaux/CNRS (2000).
4. Lind, C.: Coagulation Control and Optimization: Part One, Public Works, Oct (1994) 56-57
5. Lind, C.: Coagulation Control and Optimization: Part Two. Public Works, Nov(1994) 32-33
6. K. Baba, I. Enbutu, M. Yoda. Explicit representation of knowledge acquired from plant historical data using neural network. Int. Joint Conf on Neural Networks. Wash, D.C., 1990.
7. A. Mirsepassi, B. Cathers, H.B. Dharmappa. Predicted of chemical dosage in water treatment plants using ANN models. IAWQ Asia-Pacific Reg.Conf. Korea, 1997, 16-561.

8. I. Enbutsu. K. Baba. N. Hara. K. Waseda. A. Nogita. Integration of multi AI paradigms for intelligent operation – fuzzy rule extraction from a neural network. Wat. Res.1998,28,11-12.
9. I. Fletcher, A. Adgar, C.S. Cox, T.J. Boheme. Neural Network applications in the water industry. The Institute of Electrical Engineers IEE pp 16/1-16/6, London, UK, 2001.
10. C.W. Baxter. S.J. Stanley. Q. Zhang. D.W. Smith. Developing artificial neural network process models of water treatment process. Eng. Sci./Rev.gen.sci.env.1(3):pp201-211, 2002.
11. W. Peijing. C. Cox. Study on the application of auto-associative neural network. IEEE ICSP'04 Proceedings, 0-7803-8406-7/04, pp 1570-1573, 2004.
12. H. Hernández De León: Supervision et diagnostic des procédés de production d'éau potable. Thèse de doctorat, INSA Toulouse. France/CNRS (2006)
13. Piera N., Desroches P., Aguilar J.: LAMDA: An incremental Conceptual Clustering System. Report 89420 LAAS-CNRS (1989)
14. Waissmann J.: Construction d'un modèle comportemental pour la supervision de procédés : Application a une station de traitement des eaux. PhD Thesis, LAAS/CNRS, Institut National Polytechnique de Toulouse, France (2000)
15. Aguado J.C.: A Mixed Qualitative-Quantitative Self-Learning Classification Tech applied to Situation Assessment in Industrial Process Control. PhD Thesis. UP Catalunya (1998)
16. Aguilar-Martin. J., López R.: The process of classification and learning the meaning of linguistic descriptors of concepts. Approx Reasoning in Decision Analysis.N.Holland (1982)
17. Waissman J., Aguilar-Martin. J., Dahhou B., Roux G.: Généralisation du degree d'adéq. marginale (DAM) de la classification LAMDA. Soc.Francophone de Classification (1998).
18. Kempowsky, T., Aguilar-Martin, J., Le Lann, M-V., Subias, A.: Learning Methodology of a supervision System using LAMDA Classification Method. LAAS/CNRS.Iberamia (2002)
19. W.S. McCulloch. W. Pitts. A logical calculs of the ideas immanent in nervous activity. Bulletin of Math. 1943, Biophysics, 5, 115-133.
20. D.E. Rumelhart, J.L. McClelland. Parallel distribution processing: exploration in the microstructure of cognition. Cambridge, MA, 1986, MIT Press, 1.
21. W. Duch. Survey of neural transfer functions. Neural Computing Surveys. 1999, 2.
22. Matlab, Neural Network Toolbox, User' Guide, Inc., (2007)
23. Kempowsky, T.: SALSA(Situation Assessment using LAMDA Classification Algorithm). User's Manual. Rapport LAAS/CNRS No. 04160 (2004).

Boundary Constraints Strategies in Differential Evolution Algorithms Applied to Optimal Control Problems

Lopez-Cruz I. L., Rojano-Aguilar A. and Salazar-Moreno R.

Postgrado en Ingeniería Agrícola y Uso Integral del Agua
Universidad Autónoma Chapingo, Chapingo, México.
ilopez@chapingo.mx

Abstract. The aim of current research is the evaluation of four strategies: clipping technique, random reinitialization, bounce-back and averaged bounce-back, to handle boundary constraints of the control inputs in solving optimal control problems using Differential Evolution (DE) algorithms are efficient in solving both multimodal, and also singular optimal control problems especially when a relatively greater number of variables have to be optimized. DE algorithms are simple and efficient evolutionary methods when are compared to other evolutionary methods, regarding the number of function evaluations to converge to a solution. Results showed that clipping and bounce-back strategies performed better on a high-multimodal and also a singular optimal control problem, than random reinitialization and averaged bounce-back strategies.

1 Introduction

During the last decade interest on the application of global optimization methods in optimal control problems has significantly increased. Evolutionary Algorithms are stochastic optimization methods that have shown several advantages as global optimization methods. They have been applied mostly to solve static optimization problems and only rarely in solving optimal control problems. It is well known that optimal control problems with singular arcs are very hard to solve by the Pontryagin minimum principle [1],[2]. Singular optimal control problems are frequently found in the optimization of bioreactors [3], [4] and likely also in other biosystems [5]. Also multimodal optimal control problems are frequently found in optimization of bioreactors [6]. Luus [6,7] has applied Iterative Dynamic Programming (IDP), which can be considered as another global optimization method, to solve multimodal and also singular control problems. Tholudur and Ramirez [8], who also used IDP, found highly oscillatory behavior of optimal control trajectories in singular optimal control problems. Therefore, they proposed two filters in order to calculate smoother optimal trajectories. Recently, Roubos *et al.* [5] suggested two smoother evolutionary operators for a Genetic Algorithm with floating-point representation of the individuals and applied this approach to calculate solutions for two fed-batch bioreactors.

Theoretical and empirical results [9] have shown that Evolutionary Algorithms (like those based in Genetic Algorithms) that use low mutation rates for mutation and high probability for crossover are not good candidates to solve optimal control problems efficiently since they may require highly number of function evaluations when

© G. Sidorov, M. Aldape, M. Martinez, S. Torres. (Eds.)
Advances in Computer Science and Engineering.
Research in Computing Science 42, 2009, pp. 359-370

Received 15/03/09
Accepted 27/04/09
Final version 06/05/09

many variables are optimized or these variables are correlated. There is a necessity of developing more efficient global optimization algorithms for solving optimal control problems, in general, and multimodal and singular optimal control problems, in particular. Lately, a new family of evolutionary algorithms named Differential Evolution (DE) has been proposed [10, 11, 12, 13] which is not only simple but also remarkably efficient compared to other Evolutionary Algorithms, Simulated Annealing and Stochastic Differential equations [10, 12]. While Differential Evolution algorithms are studied mainly on unconstrained optimization problems, the present work evaluates four strategies for boundary constraints handling: clipping technique, random reinitialization, bounce-back and averaged bounce-back on the standard differential evolution algorithm (*DE/rand/bin/1*) in solving an multimodal an also and singular arc optimal control problems. As has been shown recently [14] a so-called median filter operator considerably improved solution of singular optimal control problems by using evolutionary algorithms. Therefore, a standard differential evolution algorithm with a median filter operator was also used to evaluated the four strategies to handle boundary constraints in solving the singular arc optimal control problem.

2 The Optimal Control Problem

A continuous-time optimal control problem [15] implies to find an optimal control $u^{*}(t)$ which causes the system

$$\dot{x} = f(x((t), u(t), t)\,,\ x(t_0) = x_0\,. \tag{1}$$

to follow an admisible trajectory $x^{*}(t)$ that optimizes the performance measure given by the functional :

$$J = \phi(x(t_f), t_f) + \int_0^{tf} L(x(t), u(t), t)dt \tag{2}$$

where $x \in R^n$ denotes the states of the system and $u \in R^m$ denotes a control vector. In addition the controls are constrained $\alpha \le u(t) \le \beta$. The final time t_f is fixed. As the Hamiltonian function:

$$H(t) = \lambda^T(t) f(x(t), u(t), t) \tag{3}$$

is linear with respect to the controls, the optimal control problem becomes singular [16]. Singular optimal control problems are difficult to solve by classical methods and direct methods seem to be a promising approach. To apply a direct optimization method a parameterization of the controls is necessary, for instance piecewise constant control can be applied

$$u(t) = u(t_k)\,,\ t \in [t_k, t_{k+1})\,,\ k = 0,1,...N-1\,. \tag{4}$$

where N is the number of sub-intervals for the time interval $[t_0, t_f]$. In this way a vector of parameters $\widetilde{u} = [u_1^T, u_2^T, ..., u_{N-1}^T]$ is defined and the value that optimizes the original performance index (2) can be obtained by parameter optimization methods or solving a Non-Linear Programming (NLP) optimization problem. The numerical solution of these problems is challenging due to the non-linear and discontinuous dynamics. Likely, there is not a unique global solution. Standard gradient-based algorithms are basically local search methods, they will converge to a local solution. In order to surmount these difficulties global optimization methods must be used in order to ensure proper convergence to the global optimum.

3 Differential Evolution Algorithms

A differential evolution algorithm is as follows:

```
Generate a population ($P(0)$) of solutions.
Evaluate each solution.
g=1;
while (convergence is not reached)
      for i=1 to $\mu$
       Apply differential mutation.
       Execute differential crossover.
       Use a handling boundary constraints technique if
       necessary.
       Evaluate the new solution.
       Apply differential selection.
      end
      g=g+1;
end
```

Firstly, a population $P(0)$ of floating-point vectors $\vec{u}_i, i = 1, ..., \mu$ is generated randomly from the domain of the variables to be optimized, where $\vec{u} = [u_1, ..., u_d]$ and μ denotes the population size. Next, each vector is evaluated by calculating its associated cost function (eqn. 2), $i = 1, ..., \mu$. Notice that the evaluation of each solution implies to carry out a numerical integration of the dynamic model (1). After that, a loop begins in which the evolutionary operators: differential mutation, differential crossover and selection are applied to the population ($P(g)$), where g denotes a generation number. Differential Evolution operators are quite different than those frequently found in other evolutionary algorithms. In DE, the differential mutation operator consists of the generation of μ mutated vectors according to the equation:

$$\vec{v}_i = \vec{u}_{r_1} + F \cdot (\vec{u}_{r_2} - \vec{u}_{r_3}), \ i = 1, 2, ..., \mu \tag{5}$$

where the random indices $r_1, r_2, r_3 \in [1, 2, ..., \mu]$ are mutually different and also different from the index *i*. $F \in [0,2]$ is a real constant parameter that affects the differential variation between two vectors. Greater values of F and/or the population size (μ) tend to increase the global search capabilities of the algorithm because more areas of the search space are explored.

The crossover operator combines the previously mutated vector $\vec{v}_i = [v_{1i}, v_{2i}, ..., v_{di}]$ with a so-called target vector (a parent solution from the old population) $\vec{u}_i = [u_{1i}, u_{2i}, ..., u_{di}]$ to generate a so-called trial vector $\vec{u}'_i = [u'_{1i}, u'_{2i}, ..., u'_{di}]$ according to:

$$u'_{ji} = \begin{cases} v_{ji} & if\ (randb(j) \le CR)\ \ or\ \ j = rnbr(i) \\ u_{ji} & if\ (randb(j) > CR)\ \ and\ \ j \ne rnbr(i) \end{cases},\ j = 1,2,...,d;\ i = 1,2,...,\mu \qquad (6)$$

where $randb(j) \in [0,1]$ is the j-th evaluation of a uniform random number generator, $rnbr(i) \in 1,2,...,d$ is a randomly chosen index. $CR \in [0,1]$ is the crossover constant, a parameter that increases the diversity of the individuals in the population. Greater values of CR give rise to a child vector ($\vec{u}'_i$) more similar to the mutated vector ($\vec{v}_i$). Therefore, the speed of convergence of the algorithm is increased. As can be seen from equation (6), each member of the population plays once the role of a target vector. It is important to realize that even when $CR = 0$, equation (6) ensures that parent and child vectors differ by at least one gene (variable). The three algorithm parameters that steer the search of the algorithm, the population size (μ), the crossover constant (CR) and differential variation factor (F) remain constant during an optimization.

The selection operator compares the cost function value of the target vector $\vec{u}_i$ with that of the associated trial vector $\vec{u}'_i$, $i = 1,2,...,\mu$ and the best vector of these two becomes a member of the population for the next generation. That is,

$$\text{if } \phi(\vec{u}'_i(g)) < \phi(\vec{u}_i(g)) \text{ then } \vec{u}_i(g+1) := \vec{u}'_i(g)$$
$$\text{else } \vec{u}_i(g+1) := \vec{u}_i(g);\ i = 1,...,\mu$$

Several DE algorithms can be identified according to their type of mutation (x), number of difference vectors (y) and type of crossover (z). Commonly, the notation $DE/x/y/z$ is used to named a DE algorithm. Where x, means the way the vector to be mutated is chosen, y indicates the number of difference vectors is used, and z is the type of differential crossover implemented. For instance, the previously described algorithm is known as the $DE/rand/1/bin$ which means than the to be mutated vector is selected randomly, only one difference vector is calculated and the scheme of crossover is binomial. In general $x \in \{rand, best, current-to-rand\}$, $y \in \{1,2,...,n\}$, and $z \in \{bin, \exp\}$.

Statistics to Measure Differential Evolution Performance

To measure algorithm's convergence in addition to the mean and standard deviation from several runs the Q-measure (Q_m)[13] was used, but also the expected number of functions evaluations per success (ENES) [12] and the average (number of function) evaluations per success (AES)[12] from a number of consecutive successful trials were calculated. The Q-measure[13] is given by the ratio of a convergence measure (C) and a probability of convergence (P_C) according to equation (7):

$$Q_m = \frac{C}{P_C}. \tag{7}$$

Where the convergence measure and AES are calculated by equation (8)

$$AES = C = \frac{\sum_{j=1}^{n_s} E_j}{n_s}. \tag{8}$$

Where E_j is the number of function evaluations in the ith trial and n_s is the number of successful trials. A probability of convergence is calculated by the ration between the number of successful trial and the total number of trial given by equation (9)

$$P_C = \frac{n_s}{n_t}\% \tag{9}$$

On the other hand ENES is calculated as follows:

$$ENES = \frac{\sum_{i=1}^{n_t} E_i}{n_s}. \tag{10}$$

Boundary Constraints Handling Strategies for Optimal Control Problems

Since originally DE algorithms were designed to solve unconstrained static optimization problems, a modification is required in order to deal with constraints for the controls. A **clipping technique** has been introduced to guarantee that only feasible trial vectors are generated after the mutation and crossover operators:

$$u'_{ji}(g) = \begin{cases} \beta_j & if\ u'_{ji}(g) > \beta_j \\ \alpha_j & if\ u'_{ji}(g) < \alpha_j \end{cases} \quad j = 1,2,\ldots,d; i = 1,2,\ldots,\mu. \tag{11}$$

where α_j and β_j represent the lower and upper boundaries of the control variables, respectively. This approach has been successfully applied to optimal control problems [14], [17]. **Random reinitialization** [12] is described by equation (8) as follows:

$$u'_{ji}(g) = \alpha_j + rand_j(0,1)(\beta_j - \alpha_j)\ if\ u'_{ji}(g) < \alpha_j\ or\ u'_{ji}(g) > \beta_j \tag{12}$$

The **bounce-back** technique has been proposed recently [12] as is described in equation (9):

$$u'_{ji}(g) = \begin{cases} \overset{\rho}{u}_{r_1} + rand(0,1)(\alpha_j - \overset{\rho}{u}_{r_1})\ if\ u'_{ji}(g) < \alpha_j \\ \overset{\rho}{u}_{r_1} + rand(0,1)(\beta_j - \overset{\rho}{u}_{r_1})\ if\ u'_{ji}(g) > \beta_j \end{cases} \tag{13}$$

The averaged bounce-back strategy was suggested in [11] and [12] according to the equation (10):

$$u'_{ji}(g) = \begin{cases} (\alpha_j + \overset{\rho}{u}_{r_1})/2\ if\ u'_{ji}(g) < \alpha_j \\ (\beta_j + \overset{\rho}{u}_{r_1})/2\ if\ u'_{ji}(g) > \beta_j \end{cases} \tag{14}$$

The Smoother Operator to Singular Optimal Control Problems

A smoother operator is defined according to [8] as follows:

$$u_{j,i} = median(u_{j-F,i}, u_{j-F+1,i}, \ldots, u_{j,i}, \ldots, u_{j+F-1,i}, u_{j+F,i}) \qquad (15)$$
$$j = F+1, F+2, \ldots, N-F\,;\; i = 1,2,\ldots,\mu$$

where $F = 1,2,..$ is the filtering radius. All the boundary constraints handling strategies and the standard Differential Evolution algorithm were programmed as an m-file in the Matlab environment.

4 Multimodal Optimal Control of Bifunctional Catalyst Blend

A chemical process converting methylcyclopentane to benzene in a tubular reactor is modelled by a set of seven differential equations:

$$\dot{x}_1 = -k_1 x_1 \;. \qquad (16)$$

$$\dot{x}_2 = k_1 x_1 - (k_2 + k_3) x_2 + k_4 x_5 \,. \qquad (17)$$

$$\dot{x}_3 = k_2 x_2 \,. \qquad (18)$$

$$\dot{x}_4 = -k_6 x_4 + k_5 x_5 \qquad (19)$$

$$\dot{x}_5 = k_3 x_2 + k_6 x_4 - (k_4 + k_5 + k_8 + k_9) x_5 + k_7 x_6 + k_{10} x_7 \,. \qquad (20)$$

$$\dot{x}_6 = k_8 x_5 - k_7 x_6 \,. \qquad (21)$$

$$\dot{x}_7 = k_9 x_5 - k_{10} x_7 \,. \qquad (22)$$

where $x_i, i = 1,\ldots,7$ are the mole fractions of the chemical species, and the rate constants (k_i) are cubic functions of the catalyst blend $u(t)$:

$$k_i = c_{i1} + c_{i2} u + c_{i3} u^2 + c_{i4} u^3 \,,\; i = 1,\ldots,10 \qquad (23)$$

The values of the coefficients c_{ij} are given in [7]. The upper and lower bounds on the mass fraction of the hydrogenation catalyst are: $0.6 \le u(t) \le 0.9$, and the initial vector of mole fraction is $\mathbf{x}[0] = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}^T$. This is a continuous process operated in steady state, so that 'time' in equations (16)-(23) is equivalent to travel time and thus length along the reactor. The optimal control problem is to find the catalyst blend along the length of the reactor, which in the control problem formulation is considered at times $0 \le t \le t_f$ where the final effective residence time $t_f = 2000 g \cdot h / mol$ such that the concentration in the reactor is maximised:

$J = x_7(t_f) \times 10^3$. Esposito and Floudas [18] found recently 300 local minima of this problem, so this is a challenging multimodal optimal control problem as the number of intervals N=10 was used. A variable step size four order Runge-Kutta integration method with a relative tolerance of 10^{-8} was applied to solve the dynamic equations (16-22).

5 Singular optimal control of the Park-Ramirez bioreactor

One optimal control problem that has a singular optimal solution was used to test the modified DE algorithm [8]. In this problem the goal is to maximize the production of protein. The system is described by the following differential equations:

$$\dot{x}_1 = g_1(x_2 - x_1) - \frac{x_1}{x_5} u . \tag{24}$$

$$\dot{x}_2 = g_2 x_3 - \frac{x_2}{x_5} u . \tag{25}$$

$$\dot{x}_3 = g_3 x_3 - \frac{x_3}{x_5} u . \tag{26}$$

$$\dot{x}_4 = -g_4 g_3 x_3 - \frac{m - x_4}{x_5} u . \tag{27}$$

$$\dot{x}_5 = u . \tag{28}$$

where $g_1 = \dfrac{4.75 g_3}{0.12 + g_3}$, $g_2 = \dfrac{21.88 x_4}{(x_4 + 0.4)(x_4 + 62.5)}$, $g_3 = \dfrac{x_4 \exp(-5.01 x_4)}{0.10 + x_4}$, $g_4 = 58.75 g_2^2 + 1.71$.

The state variable x_1 represents amount of secreted protein [unit culture volume L^{-1}], x_2 denotes the total protein amount [unit culture volume L^{-1}], x_3 means culture cell density [g L^{-1}], x_4 culture glucose concentration [g L^{-1}], and x_5 the culture volume [L]. The control $u(t)$ represents the rate at which glucose is fed into the reactor [Lh^{-1}]. The secretion rate constant is given by g_1, the protein expression rate is calculated by g_2, the specific growth rate by g_3 and the biomass to glucose yield is estimated by g_4. The optimal control problem consists in the maximization of the amount of the secreted protein in a given time $t_f = 15h$. Therefore the performance index is given by $J = x_1(t_f) x_5(t_f)$. The control input satisfying the constraints $0 \le u(t) \le 2.5$ and the system initial conditions are $x(0) = [0,0,1.0,5.0,1.0]$. The dynamic model (eqns. 24-28) was programmed in the Matlab-Simulink environment. A C-MEX file containing the dynamic equations was implemented in order to speed up the simulations. A variable step size four-order Runge-Kutta integration method with a relative tolerance of 10^{-8} was applied. The DE algorithm was initialized randomly from the control's domain. Since DE algorithms are probabilistic methods the optimizations were repeated 30 times. The problem was solved for N=100 variables.

6 Results and discussion

Multimodal Optimal Control Problem

The parameters for the standard DE algorithm were: NP=25, F=0.9, CR=0.0. Instead of using as a convergence criteria the difference between worst and best solution the value to-reach (VTR) condition as suggested in [12] was used. The value to-reach (VTR) was defined as 10.09415 since it is known that the global optimum for this problem is 10.0942. The maximum number of function evaluations ($E_{\max}$) was defined as 15000. The total number of trials was thirty. Table 1 sumarizes the observed behaviour of the four boundary constraints handling strategies taking into account the number of functions evaluations. It is apparent that the clipping method performed much better since it has the lesser Q_m measure, and also the lesser AES and ENES than the others strategies. Random reinitialization was the worst strategy. Strategies that used bounce-back were a little higher than clipping technique but considerably more efficient than random reinitialization.

Table 1. Statistics calculated to evaluate the four boundary constraints-handling strategies in the DE/rand/1/bin algorihtm on a multimodal optimal control problem.

	Clipping technique	Random reinitialization	Bounce-back	Averaged bounce-back
Mean	3059.2	9561.6	3893.3	3348.3
Std	346.9	407.4	157.0	236.6
ENES	3059.2	9924.2	3893.3	3865.5
AES	3059.2	9561.6	3893.3	3348.3
C	3059.2	9561.6	3893.3	3348.3
P_C	100%	93.3	100%	96.6
Q_m	30.5	102.4	39.8	34.6

These results are according to remarks made by Price et al. [12]. An statistical test of significance among the four strategies was carried out. The null hypothesis: means are equal, was rejected in all cases at the significance level 0.05. Figure 1 shows the optimal control trajectory found by DE algorithm using the four handling constraints strategies. Almost no differences among them can be detected. The smoother operator was not applied to this problem. Numerical values for the best solution found by the algorithm using each constraints-handling strategy are showed in Table 2.

Table 2. Numerical values corresponding to the optimal controls calculated by DE/rand/1/bin algorithm using four boundary contraints-handling strategies.

Control	Clipping technique	Random reinitialization	Bounced back	Averaged bounce-back
1	0.6656	0.6661	0.6662	0.6661
2	0.6733	0.6735	0.6735	0.6736
3	0.6764	0.6763	0.6763	0.6762
4	0.9000	0.9000	0.9000	0.9000
5	0.9000	0.9000	0.9000	0.9000
6	0.9000	0.9000	0.9000	0.9000
7	0.9000	0.9000	0.9000	0.9000
8	0.9000	0.9000	0.9000	0.9000
9	0.9000	0.9000	0.9000	0.9000
10	0.9000	0.9000	0.9000	0.9000

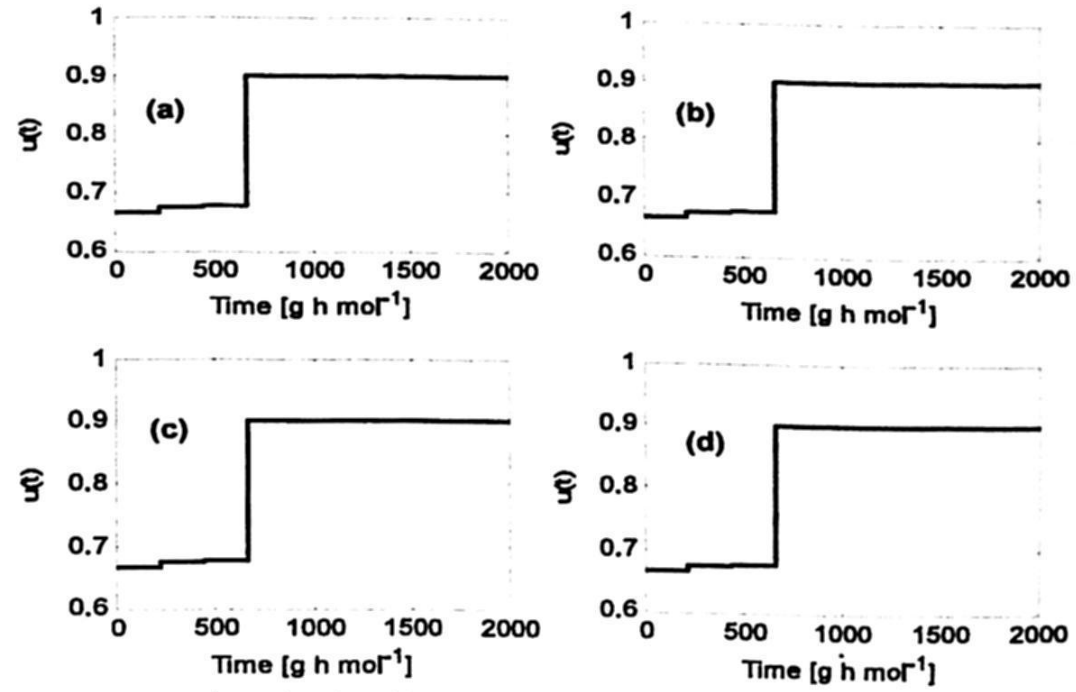


Fig. 1. Optimal trajectories found by the DE/rand/1/bin algorithm for the multimodal problem. a) Clipping technique. b) Bounce-back approach. c) Averaged bounce-back. d) Random reinitialization.

Singular Optimal Control of Park-Ramirez Bioreactor

The parameters for the standard DE algorithm were: population size 100 individuals, scaling factor for mutation F=0.6 and probability of crossover CR=0.9. The total number of trials was thirty. The converge criterion was the VTR=32.47, which has been reported before as the global optimum for this problem [8] using Iterative Dynamic Programming using in case of $N = 100$ parameters. Table 3 shows the performance of the DE/rand/1/bin algorithm using the four strategies regarding the number of function evaluations required to solve this problem.

Table 3. Statistics calculated to evaluate the four boundary constraints-handling strategies in the DE/rand/1/bin algorihtm on a singular optimal control problem.

	Clipping technique	Random reinitialization	Bounce-back	Averaged bounce-back
Mean	64607.0	139233.0	73610.0	69500.0
Std	7584.1	8417.0	6629.9	7399.4
ENES	64607.0	139233.0	73610.0	69500
AES	64607.0	139233.0	73610.0	69500
C	64607.0	139233.0	73610.0	69500
P_C	100%	100%	100%	100%
Q_m	646.0	1392.3	736.1	695.0

As in case of the multimodal problem again the clipping approach was the most efficient since it required lower Q_m-measure ENES and AES than the other strategies. The random reinitialization strategy was the worst method taking into account several statistics. An statistical test of significance among the four strategies was carried out. The null hypothesis: means are equal, was rejected in all cases at the significance level 0.05. Figure 2 shows the optimal control trajectory found by DE algorithm using the four handling constraints strategies. Only small differences in the optimal trajectories due to the effect of the smoother operator are observed.

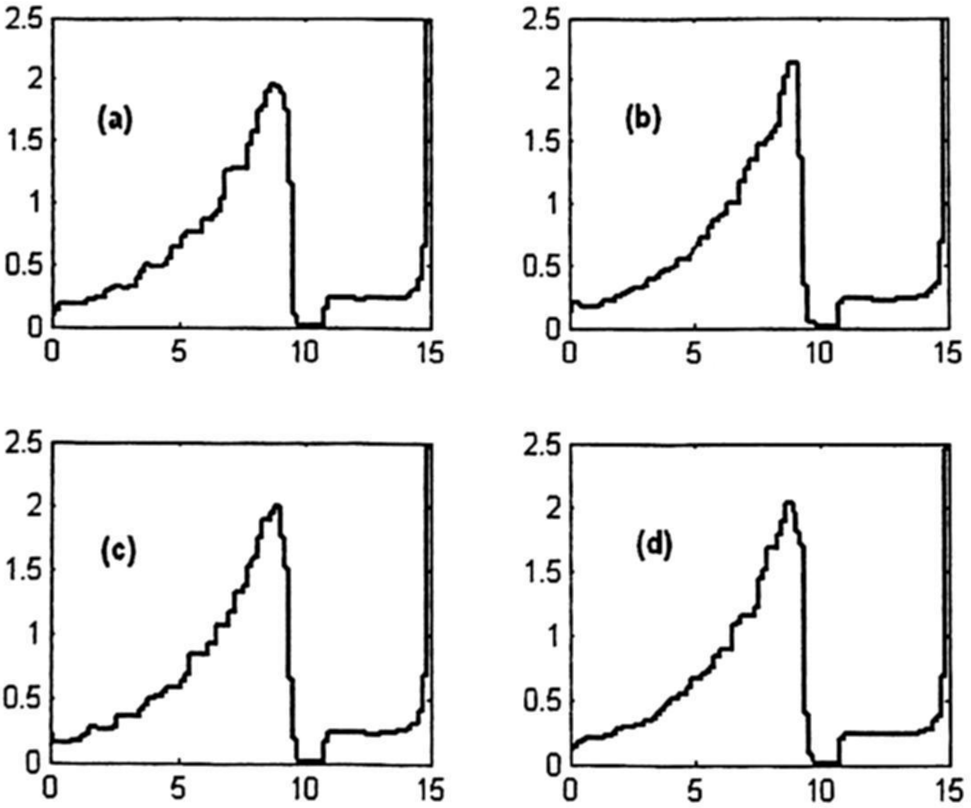


Fig. 2. Optimal trajectories found by the DE/rand/1/bin algorithm for the singular arc control problem. a) Clipping technique. b) Bounce-back approach. c) Averaged bounce-back. d) Random reinitialization.

7 Conclusions

A highly multimodal optimal control problem was used to test four boundary constraints strategies on the performance of the standard differential evolution algorithm (DE/rand/1/bin). Results showed although clipping technique, random reinitialization, bounce-back and averaged bounce back converged to the global optimum, the clipping and bounce-back approaches are more efficient regarding the required number of function evaluations. In addition an singular optimal control problem that appears in the dynamic optimization of bioreactors was solved. Again results showed that the four boundary constraints strategies worked out well, however clipping and bounce-back approaches solved the problem more efficiently.

References

1. Park, S., Ramirez, W. F.: Optimal production of secreted protein in fed-batch reactors. AIChE Journal 34 (9) (1988) 1550-1558.
2. Park, S., Ramirez, W. F.: Dynamics of foreign protein secretion from *Saccharomyces cerevisiae*. Biotechnology and Bioengineering 33 (1989) 272-281.
3. Menawat, A., Mutharasan, R., Coughanowr, D. R.: Singular optimal control strategy for a fed-batch bioreactor: numerical approach. AIChE Journal 33 (5)(1987) 776-783.
4. Roubus, J.A., de Gooijer, C.D., van Straten, G., van Boxtel, A.J.B.: Comparison of optimization methods for fed-batch cultures of hybridoma cells, Bioproc. Eng. 17 (1997) 99-102.
5. Roubos, J.A., van Straten, G., van Boxtel, A.J.B.. An evolutionary strategy for fed-batch bioreactor optimization: concepts and performance, Journal of Biotechnology 67 (1999) 173-187.
6. Luus, R.. On the application of Iterative Dynamic Programming to Singular Optimal Control problems. IEEE, Transactions on Automatic Control 37 (11) (1992)
7. Luus, R. Iterative Dynamic Programming. Chapman & Hall/CRC., Boca Raton, (2000).
8. Tholudur, A., Ramirez, W.F.: Obtaining smoother singular arc policies using a modified iterative dynamic programming algorithm. International Journal of Control, 68(5) (1997)1115-1128.
9. Salomon, R.: Re-evaluating genetic algorithm performance under coordinate rotation of benchmark functions. A survey of some theoretical and practical aspects of genetic algorithms. BioSystems 39 (1996) 263-278.
10. Storn, R., Price, K.: Differential Evolution- a simple and efficient heuristic for global optimization over continuous spaces. Journal of Global Optimization 11 (1997) 341-359.
11. Price K., V.: An Introduction to Differential Evolution. In Corne, D., Dorigo, M., and Glover, F. (eds.). New Ideas in Optimization. Mc Graw Hill (1999).
12. Price K. V., Storn R.M., Lampinen J.A.: Differential Evolution, A practical approach to global optimization. Springer-Verlag. Berlin (2005).

13. Feoktistov V.: Differential Evolution, In search of solutions. Springer-Verlag. New York (2006).
14. Lopez-Cruz I., Rojano-Aguilar A.: Differential Evolution algorithms to solve optimal control problems efficiently. Research on Computing Science 16 (2005): 271-280.
15. Kirk, D.E.: Optimal control theory. An introduction. Dover Publications, Inc. New York. (1998).
16. Bryson, A.E.: Dynamic Optimization. Addison-Wesley. (1999).
17. Lopez-Cruz, I., van Willigenburg, G., van Straten, G. : Differential evolution algorithms for multimodal optimal control problems. Applied Soft Computing 3 (2) (2004) 97-122.
18. Esposito W.R., Floudas, Ch.A.: Deterministic global optimisation in nonlinear optimal control problems. Journal of global optimisation 17 (2000) 97-126.

Combinatorial Fuzzy Logic Controller Design

Arturo Tellez, Luis Villa, Heron Molina,
Oscar Camacho and Romeo Urbieta

Centro de Investigación en Computación, Instituto Politécnico Nacional, Juan de Dios Batiz
Ave. s/n, Nueva Industrial Vallejo, Mexico City, Mexico
http://www.microse.cic.ipn.mx/

Abstract. This paper presents the architecture development of a Fuzzy Logic Controller (FLC), using combinatorial design implemented on a Field Programmable Gate Array (FPGA). This architecture is based on combinatorial basic modules that enable to increase and improve the entire system performance, by means of replication technique, which is widely used in computer architecture, and help to fit the particular application needs. There have been so many FLC implementations since the first hardware one appeared [4] which used complex designs with sequential circuits because of the high hardware resource and delay time costs about combinatorial design. But recent FPGA technology let us use fast combinatorial circuits for complex designs with parallelism for increasing the FLC performance and it is possible to take it up again as a practical way to build FLC for any process, approaching the fast prototyping advantages and easing the scaling to increase the control accuracy.

1 Introduction

There has been a large increment in the use of FLC in many science and industry fields early, so the implementations in several technologies, from the use of fuzzy microcontrollers until reprogrammable ways, as FPGA means.

An FLC can be implemented in software easily and executed in a microprocessor, a microcontroller, or a general purpose computer. Though software- based FLC are cheaper and flexible, there are some difficulties when control systems require high data processing.

The use of FPGA has been profitable when we talk about versatility to make any digital design by means of costs and design time. With fuzzy logic application in high velocity systems of industry, especially in real time systems, as image processing, robotics, automotive, and other industries, it is necessary to develop new faster implementations. FPGA came from the Applied Specific Integrated Circuit (ASIC) industry as an economic and flexible option, for processing data at competitive velocity and opening a wide door for researchers to make their own designs in short time.

In principle, the implementation of FLC is not based on the mathematic model of the plant, but this kind of system is very effective to control a process where the transfer function is not known, instead the control action is based on the extern influence and simple decisions based on a knowledge base acquired with experience, the same

© G. Sidorov, M. Aldape, M. Martinez, S. Torres. (Eds.)
Advances in Computer Science and Engineering.
Research in Computing Science 42, 2009, pp. 371-382

Received 21/03/09
Accepted 27/04/09
Final version 08/05/09

way a human would do it, exploiting the heuristic ability. A FLC let the engineers to integrate human reasoning in computing systems.

Besides, FPGA-based design let us create, probe and modify easily and quickly any quantity of digital systems and implement dedicated systems that helps to speed up other non-real time systems.

From this point of view, it has been developed a large quantity of FLC architectures, derived from Computing Architecture. These architectures are classified by its processing way. There are sequential, combinatorial [9], parallel, pipelined and mixed models. Everyone has its own advantages and disadvantages and designers must choose which processing fits their design needs. Some designers used these architectures with other techniques that help to improve the general performance in complex algorithms. Arithmetic operations like multiplication and division are very expensive from the point of view of timing and used resources in a reconfigurable device; that is why these arithmetic operations represent a dare for designers. Some designers prefer to implement these operations to calculate a parameter of the FLC every time it is necessary [8]; this technique is called Runtime Computation (RTC). But some designs use extern elements like memories, sometimes called Look Up Tables (LUT), to calculate FLC parameters by anticipation; this another technique is called Look Up Computation (LUC) and represents a good way to improve the timing; some of this systems use supervising computing to set up the FLC configuration on line and recalculate new parameters [5:7].

However, these two techniques have some advantages and disadvantages too: LUC approach has the advantage of improving the timing critically, but has the disadvantage of using lots of external memory and a complicate way of recalculating and actualizing all the parameters of the FLC. By contrast, RTC has the disadvantage of being expensive when implementing the design on a reconfigurable device, because all the arithmetic calculation is made on line, and they need to use lots of resources on the reprogrammable device; but this represents an advantage, because it is not needed to actualize an entire LUT.

It is a dare to play with these architectures and techniques to make a balanced FLC design, by which it is necessary to change the way of designing algorithms to describe a FLC. Then, the feasibility of implementing a FLC on FPGA depends on the choice of an optimum algorithm that spend low time processing and the least quantity of used resources of the device as possible, without increasing the complexity in order to make a suitable upgrading.

This paper shows a practical approach of FLC combinatorial architecture in order to make simple construction modules and easy upgrading using a reprogrammable device, FPGA.

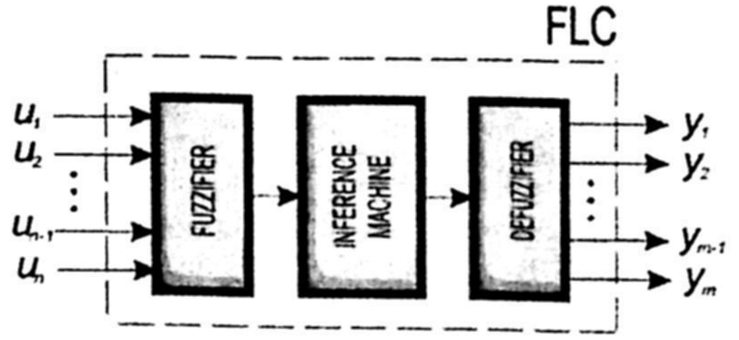


Fig. 1. Fuzzy Logic Controller (FLC). Assume as the system inputs and as the system outputs. There are inputs per outputs

2 System Description

The purpose of this section is to show a practical way of designing efficient FLC, which consists of eight steps and let us implement it easily on FPGA. Assume as the inputs to the FLC and as the outputs. Fig. 1 shows a FLC which consists of three basic stages: Fuzzification, Inference Machine and Defuzzification. The *Fuzzification* stage consists of a set of fuzzifiers, one per input that converts every crisp input into several fuzzy values or membership values. The *Inference Machine* contains the behaviour of the FLC and it is built with MIN- MAX modules; these modules are interconnected according to the fuzzy rule set inspired by one expert and decide which action will be taken based on the fuzzy values obtained from the fuzzification stage. These rules have simple inferences of the type IF- THEN. Also, the *Defuzzification* converts these inferred values onto crisp values, by means of statistical calculations, which represents the control action over the actuator.

The advantages of hardware design in FPGA, using HDL let us build any system in short time and design or redesign when needed. Next steps are required for build a FLC:

1. Establish whatever the designer want to control and which variables will be related to get it.
2. Define the number of inputs and outputs of the FLC based on the last step.
3. Define the number of membership functions or fuzzy sets for each input and output based on the last step and define their shape based on the process characteristics and operation range of the FLC (discourse universe).
4. Set the FLC configuration by means of the fuzzy inference rules according to the wished operation and based on the expert knowledge about the process.
5. Build the fuzzifier with simple membership functions simply by replication (trapezoidal, triangular, S, Z).
6. Build the inference machine based on step 4, by means of MIN- MAX modules using the building steps shown in Sect. 2.2.
7. Once inference machine is ready, build the defuzzification stage by means of multiplication and division modules using parallelism.

8. Finally, FLC can be implemented on FPGA.

These steps are related to a combinatorial architecture divided in several modules according to three RTC basic stages for a FLC, which will be described in the next subsections.

For the FLC implementation it was used VHDL, Xilinx ISE 6.3i, Mentor Graphics Modelsim Xilinx Edition III 6.0a. It is used Xilinx Spartan 3 XC3S200–5FT256 FPGA Starter Kit. In order to verify the FLC performance, it is necessary to make a simulation using the Fuzzy Toolbox of MATLAB and build a control system with SIMULINK.

2.1 Fuzzification

The fuzzification stage comprises a set of fuzzifiers attached to every input variable; each one parallel from the others and their performance does not depend on the others either.

We assume that all membership functions shape will be triangular, trapezoidal, S and Z, because they are the easiest to implement in hardware. Regardless the shape, the fuzzifier processing is based on the isosceles triangular shape, which comprises of several circuits like comparators (<, > and =), MUX, adders, subtracts, a multiplier and a divider, each one with a size of bits, as shown in the Fig. 2. This VHDL module converts a crisp digital value into a membership digital value, which it is supposed to be previously converted by an ADC way, according to the input parameters: the CENTER and the APERTURE of the triangle. These two parameters of the membership functions accomplish the RTC technique in order to make the online adaptation and the FLC tuning. Once it is known the aperture it is possible to calculate the triangle slope using a divider module. Parallel with this process, the module calculates the exact position in the triangle based on the input parameter and the aperture; because of being an isosceles triangle it is necessary to know which region of the triangle the input will address (left to the center or right to the center) and the MUX drives this result to the multiplication module. Whenever the slope is ready, it is multiplied by the MUX selection and then it is obtained the membership value of this fuzzy set. The last MUX drives the results by means of the input value parameter and the center of the triangle.

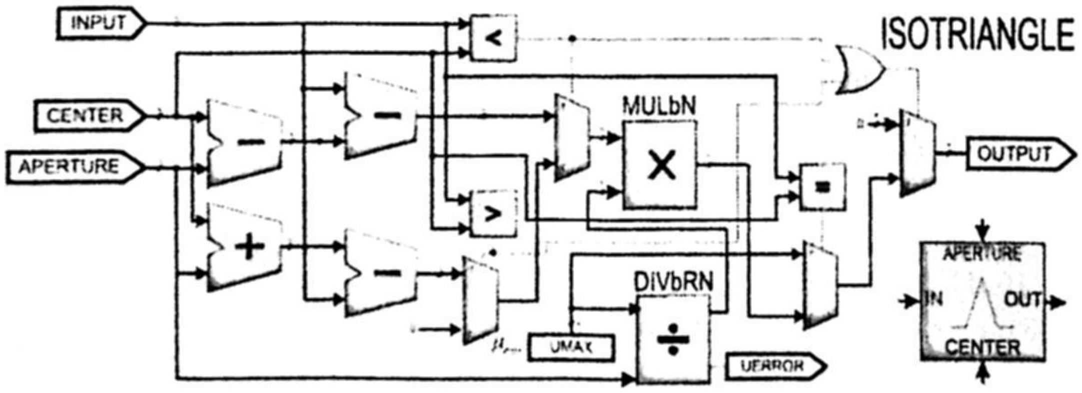


Fig. 2. Isosceles triangular membership function shape. This module uses combinatorial arithmetic operations like multiplication and division.

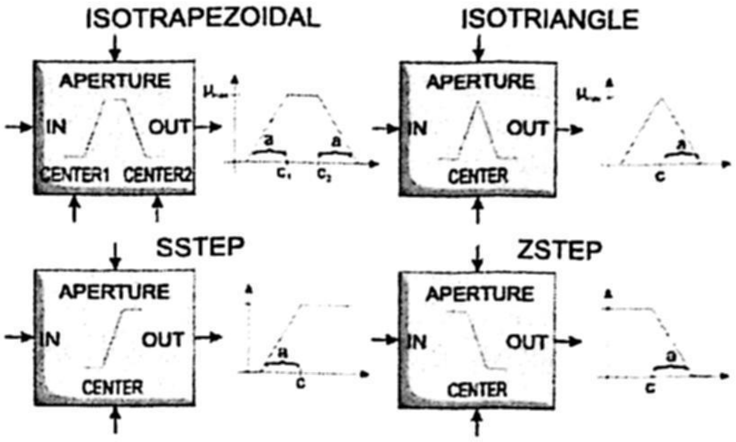


Fig. 3. Several hardware suitable membership functions: symmetric triangle and trapezoid and S and Z functions.

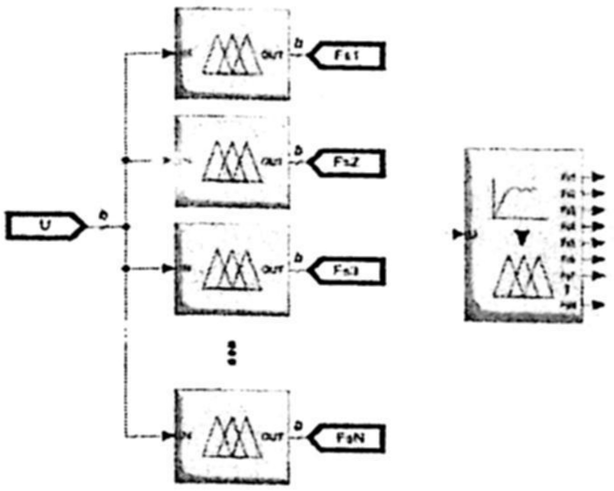


Fig. 4. Fuzzifier for a single input. Note that every fuzzy set is tied to the same input.

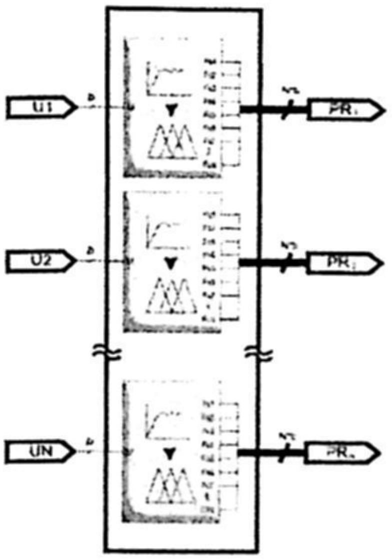


Fig. 5. Fuzzification Stage comprises with several fuzzifiers per input variable

The critical calculation in this module is the slope because uses arithmetic division operation. This combinatorial circuit calculate the slope whenever the aperture is changed; otherwise it is not calculated and the module uses the last calculated slope. Once the slope is known, the next critical calculation is the multiplication, and this operation represents now the bottleneck of this module (only when aperture has not any change).

Trapezoidal, S and Z membership function shapes have similar components. As mentioned before, a fuzzifier is compounded of several fuzzy sets described for membership functions, interconnected in parallel having the same input values. These functions may have several shapes as shown in Fig. 3 and the interconnection seems like it follows in Fig. 4.

Every fuzzifier has one input and several outputs, depending of the choice of the designer. Also, every input should have a fuzzifier and the interconnection seem like in Fig. 5.

The results of every fuzzifier represent the premises of the inference rule set of the FLC. Next section describes the inference machine construction according to a set of steps using Mamdani operation.

2.2 Inference Machine

Let us define a *premise* as the input data involved with the control, it means that an involved input will be considered to decide which control action will be taken. A *consequence* is a result of the inference, the output data of inference machine, it means the decision that FLC will take based on the premises.

A fuzzy rule set is the FLC configuration of the simple form

IF *premise* 1 **AND** *premise* 2 **AND, ..., AND** *premise n*
  THEN *conseq* 1 **AND** *conseq* 2 **AND, ..., AND** *conseq m*

For instance: *"IF water is cold AND dirtiness is worst AND charge is heavy THEN washing is hard AND time is long."*

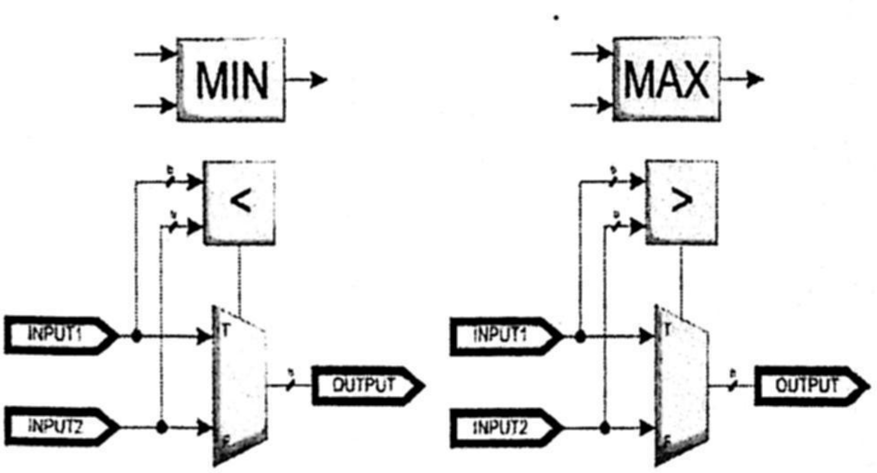


Fig. 6. MIN- MAX modules consist of simple MUXes with comparator circuits.

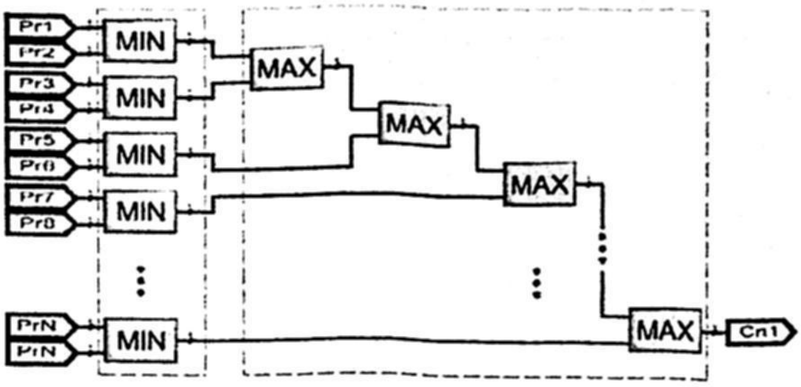


Fig. 7. Inference machine stage construction. All premises collaborate with a consequence.

A Mamdani inference machine consists of MAX- MIN (Fig. 6) modules interconnected according to the fuzzy rule set. This is the main part of a FLC because represents the FLC configuration [2]. Every module is connected the way the rule says and it is necessary to consider some specifications:

a) Identify all related rules that have common consequences within the fuzzy inference rule set. This common consequence rule set is a new subset.
b) Within the common subset obtained from last step, connect all related premises, according to the rule, pair by pair to MIN modules.
c) From all results after the MIN modules, it is necessary to find the maxim value of them using cascade MAX operations. This is the corresponding value of the consequence.
d) Repeat all steps for every consequence.

A MAX- MIN structure of an inference machine has MIN modules in parallel. Unlike the MAX modules are in cascade, as shown in Fig. 7.

2.3 Defuzzification

This defuzzification obtains a crisp output by means of output fuzzy sets, sometimes called *Centroid* method. The calculation of the centroid is made using the membership values $\mu_l(x_1, x_2, \ldots,$, obtained from the inference engine, and the output fuzzy set centers [1]. It is often considered as *singleton* membership function, because of its computational simplicity and because this statistical calculation is independent of the output fuzzy set shapes.

$$x_q^{crisp} = \frac{\sum_{l=1}^{\bar{n}} b_l^q \mu_l(x_1, x_2, \ldots, x_n)}{\sum_{l=1}^{\bar{n}} \mu_l(x_1, x_2, \ldots, x_n)} \tag{1}$$

This defuzzifier needs a division calculation, as seen in the Eq. 1, which results computationally very expensive. It is needed to emphasize that a b multiplication results in a number of bits, which is not practical neither cheap computationally. In order to avoid the b multiplication before the division, so part of Eq. 1 was implemented this way:

$$\sigma_l = \frac{b_l^q}{\sum_{l=1}^{R} \mu_l(x_1, x_2, \ldots, x_n)} \tag{2}$$

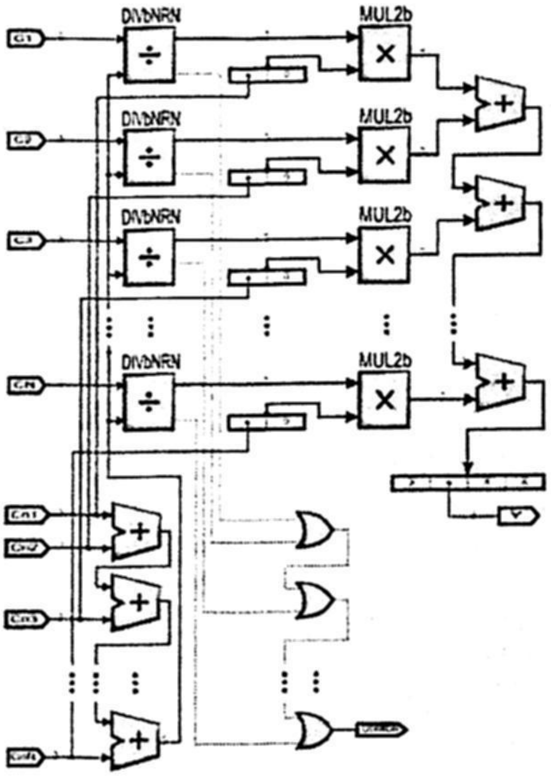


Fig. 8. Defuzzification stage: Centroid. This module performs several division and multiplication operations simultaneously in order to obtain the crisp output.

Where every center is divided by the summation of the membership values obtained from the inference machine. Then, the result (Eq. 2) is multiplied by every membership value obtained from the inference machine. To get this, it was needed to implement a combinatorial non- restoring division [3] modified to obtain a fixed point bits quotient, because $\sigma_i <$, as shown in Fig. 8.

The defuzzifier has two disadvantages: a) uses one division operation per consequence, which results expensive whenever the number of output fuzzy sets increases; b) though it still manages a multiplication after the division, it does not use a division which is more expensive than a multiplication.

Finally, the result is allocated in the third and second octets of the summation result (right end is LSB), integer and fractional corresponding parts.

3 Implementation and Verification

As example of application, a FLC for a DC servo is implemented, as mentioned above, in order to verify the correct performance of the FLC. Now, requirements of the system are explained.

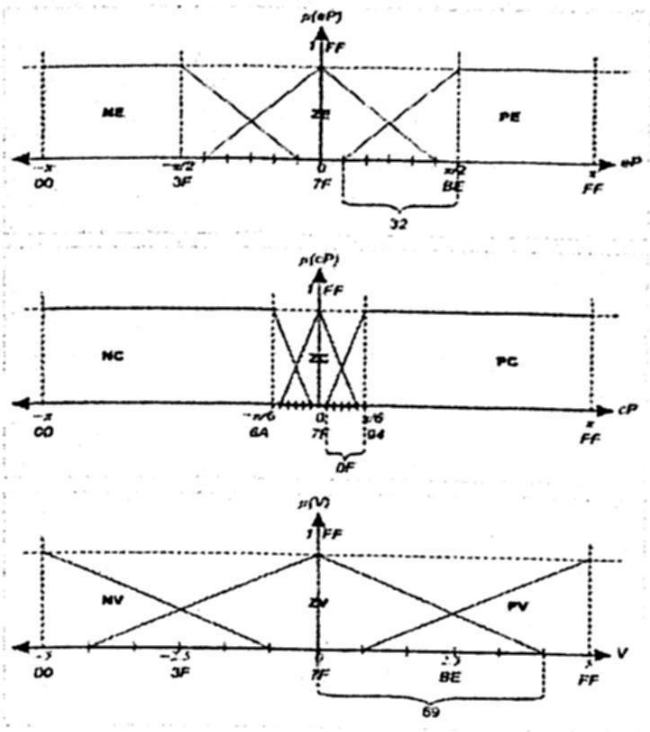


Fig. 9. Initial membership function distribution for DC servo. Initially, distribution of fuzzy sets is arbitrary. It can be modified later.

The control system was built in MATLAB Fuzzy Toolbox first, creating a fuzzy inference system by software (FIS). Then it is chosen a set of representative values or test bench in Table 1 and it is applied to the FIS in order to prove the FLC performance once it is implemented in the Xilinx FPGA kit. There are two versions of the FLC: the first one, named FLC1, has a specific initial set of input membership function parameters (center and aperture) and the second one, named FLC2, has been modified their input parameters in order to observe the influence of online adaptation. Finally every value in the table was compared against the FPGA results. Every obtained result is reflected in Table 1.

Suppose a 2X1 fuzzy system for a DC servo, which uses nine rules because it has three fuzzy sets per input (position error <rads>- eP: NE, ZE, PE. position error change velocity <rads>- cP: NC, ZC, PC) and output (voltage <volts>- V: NV, ZV, PV), shown in the Fig. 9, which are the following:

```
IF eP  is NE AND cP  is NC THEN V  is NV
IF eP  is NE AND cP  is ZC THEN V  is NV
IF eP  is NE AND cP  is PC THEN V  is NV
IF eP  is ZE AND cP  is NC THEN V  is NV
IF eP  is ZE AND cP  is ZC THEN V  is ZV
IF eP  is ZE AND cP  is PC THEN V  is PV
IF eP  is PE AND cP  is NC THEN V  is PV
IF eP  is PE AND cP  is ZC THEN V  is PV
IF eP  is PE AND cP  is PC THEN V  is PV
```

Then, it is provided a test bench which consists of 23 values and describes several situations.

Table 1. Test bench for a DC servo control.

Case	eP	cP		V	
				1	
1	$-\frac{\pi}{2}$	$-\frac{\pi}{6}$	−5.89	−5	−5
2	$-\frac{\pi}{2}$	0	−5.89	−5	−5
3	$-\frac{\pi}{2}$	$\frac{\pi}{6}$	−5.89	−5	−5
4	0	$-\frac{\pi}{6}$	−5.89	−5	−5
5	0	0	0	0	0
6	0	$\frac{\pi}{6}$	5.89	5	5
7	$\frac{\pi}{2}$	$-\frac{\pi}{6}$	5.89	5	5
8	$\frac{\pi}{2}$	0	5.89	5	5
9	$\frac{\pi}{2}$	$\frac{\pi}{6}$	5.89	5	5
10	$-\frac{\pi}{4}$	$-\frac{\pi}{12}$	−2.5	−2.5	−2.4
11	$-\frac{\pi}{4}$	$\frac{\pi}{12}$	0	−0.4	0
12	$\frac{\pi}{4}$	$-\frac{\pi}{12}$	0	0.3	0
13	$\frac{\pi}{4}$	$\frac{\pi}{12}$	2.5	2.5	2.4
14	$-\frac{\pi}{4}$	0	−2.5	−2.4	−2.4
15	0	$-\frac{\pi}{12}$	−2.5	−2.3	−2.4
16	$\frac{\pi}{4}$	0	2.5	2.4	2.4
17	0	$\frac{\pi}{12}$	2.5	2.2	2.4
18	$-\frac{\pi}{4}$	$-\frac{\pi}{6}$	−5.53	−5	−5
19	$-\frac{\pi}{4}$	$\frac{\pi}{6}$	0.59	0.2	0.2
20	$\frac{\pi}{4}$	$-\frac{\pi}{6}$	−0.59	−0.2	−0.2
21	$\frac{\pi}{4}$	$\frac{\pi}{6}$	5.53	5	5
22	$-\frac{\pi}{2}$	$-\frac{\pi}{12}$	−5.53	−5	−5
23	$-\frac{\pi}{2}$	$\frac{\pi}{12}$	−5.53	−5	−5

Cases 1 to 9 represent those circumstances where input data is not member of a pair of membership functions at same time (no overlapping region) for every input, it means, there are 2 active rules at least. Cases 10 to 13 comprises all those circumstances where exist overlapping in every input, it means, there are 4 active rules. Cases 14 to 17 comprise all those circumstances, where there is overlapping in one variable and there is not in the other; and vice versa. Cases 18 to 23 comprises all those circumstances where the input data pertains only to one left or right end membership function in one variable and there is overlapping in the other.

As shown in Table 1, it is possible to observe the influence of overlapping, which may vary the FLC accuracy due to division truncation in defuzzification stage, which may be corrected simply by using both third octet and the second eight bits as the fractional part, as shown in Fig. 8.

Also, FLC tuning was made changing the membership function parameters of inputs and outputs. The results in Table 1 show how accuracy increased, expanding the aperture of the membership functions and the position of fuzzy singletons, as you can see in field FLC2 compared with FLC1.

Finally, Table 2 shows all timing and resources in FPGA used for every implemented module built for DC servo FLC example. DC servo FLC needs 84 ns to make an inference. Then, its processing data rate is 11.9 MFLIPS. Upgrading this architecture does not affect considerably the FLC performance, because of the parallelism of this architecture.

4 Conclusions

It was designed an FLC architecture using RTC combinatorial arithmetic modules. In order to get this, it was supplied to designer a practical approach for FLC design, using a study case (DC servo). Those developed VHDL modules were implemented in FPGA and it was possible to verify the FLC performance compared with the FIS simulated with MATLAB. Then, we got a FLC architecture that competes with recent developments and give us a practical fast FLC prototyping.

Table 2. FPGA timing and resource results obtained for DC servo control.

Algorithm	Delay (ns)	LUT
16 bits non-restoring division	48.50	644
Modified 8 bits non-restoring division	28.83	208
8 bits restoring division	28.84	124
8 bits multiplication	13.17	36
Isosceles triangle MF	36.70 14.51	251
S-step MF	36.70	249
Z-step MF	36.70	251
Fuzzifier	37.42	755
Defuzzifier	41.49	677
Mamdani inference machine	19.32	242
MIN-MAX operations	9.36	16
FLC	84.01	2689

Parallelism that exists between the FLC modules improved its general performance. So, this architecture has the capability of grow modularly, adapting to the designer needs without presenting high complexity when upgrading the system. This modularity may be approached using a FIS to VHD language interpreter that simply generates the proper HDL program, using the basic modules presented in this paper, regardless the used technology, based on the MATLAB `*.fis` configuration file.

It was demonstrated that this FLC has adaptation capabilities by having online modifiable registers, being their membership function parameters which provides accuracy. Also, it was demonstrated that combinatorial design is a feasible and practical way for FLC design, obtaining results which are comparable with those reported, approaching the goods of recent FPGA technology.

References

1. Téllez, A. Fuzzy Logic Controller Architecture using Combinatorial Logic, Instituto Politécnico Nacional. Centro de Investigación en Computación. Mexico City, 2008.
2. Patyra, M. J.; Mlynek, D.M.; "Fuzzy logic: implementation and applications;" Wiley; 1996.
3. Oberman, S. F.; Flynn, M. J.; "Division Algorithms and Implementations;" IEEE Transactions on Computers; Aug 1997; Vol 46, No. 8; pp. 833–854.
4. Togai M.; Watanabe H.; "Expert system on a chip: An engine for real–time approximate reasoning;" IEEE Expert Syst. Mag., 1986, pp. 55–62, Volume 1.
5. Vasantha Rani, S.P.J.; Kanagasabapathy, P.; Sathish Kumar, A.; "Digital Fuzzy Logic Controller using VHDL;" INDICON, 2005 Annual IEEE, 11–13 December 2005, pp. 463–466.
6. Singh, S.; Rattan, K.S.; "Implementation of a fuzzy logic controller on an FPGA using VHDL;" Fuzzy Information Processing Society, 2003. NAFIPS 2003. 22nd International Conference of the North American 24–26 July 2003, pp. 110–115.
7. Deliparaschos, K.M.; Nenedakis, F.I.; Tzafestas, S.G.; "A fast digital fuzzy logic controller: FPGA design and implementation;" Emerging Technologies and Factory Automation, 2005. ETFA 2005. 10th IEEE Conference, 19–22 September 2005, Volume 1.
8. Gaona, A.; Olea, D.; Melgarejo, M.; "Sequential Fuzzy Inference System Based on Distributed Arithmetic;" Computational Intelligence for Measurement Systems and Applications, 2003. CIMSA '03. 2003 IEEE International Symposium, 29–31 July 2003, pp. 125–129.
9. Manzoul, M.A.; Jayabharathi, D.; *"Fuzzy Controller on FPGA Chip;"* Fuzzy Systems, 1992., IEEE International Conference, 8–12 March 1992, pp. 1309–1316.

Author Index
Índice de autores

Editorial Board of the Volume

Comité editorial de volumen

Additional Reviewers

Árbitros adicionales

Mohamed Abdel Fattah
Antonio Alarcón Paredes
Majed Alhaisoni
Jesús Almendros
Rosa Ayala
Tristan Behrens
Alejandro Catala
Edgar Armando Catalán Salgado
Cristóbal Costa Soria
Hugo Jair Escalante
Adrián Fernández Martínez
René García Hernández
Alexander Hasenfuss
Michael Koester
Olga Kolesnikova
Manuel Lazo-Cortés
Saturnino Leguizamón
Manuel Llavador
Itzamá López
Miguel R. Luaces
Bella Martínez
Roberto Mercado
Emanuel Montero Reyno
Nahun Enrique Montoya
Alexis Morales Fernández
Adetola Oredope
Nicolás Padilla
Ivo H. Pineda-Torres
Walter Rentaría
Gustavo Rodríguez Gómez
Israel Román Godínez
Samuel Sánchez Islas
Lina Sarmiento Castañeda
Xuanhua Shi
José Francisco Solís
María Josefa Somodevilla
Arturo Téllez Velázquez
Thu Trinh

Impreso en los Talleres Gráficos
de la Dirección de Publicaciones
del Instituto Politécnico Nacional
Tresguerras 27, Centro Histórico, México, D.F.
Mayo de 2009.
Printing 500 / Edición 500 ejemplares.

This volume contains 31 carefully selected papers by 97 authors from seven countries: Belgium, France, Republic of Korea, Mexico, Spain, United Kingdom and Venezuela. These papers present the most recent developments in a range of areas related to computer science and engineering. The papers are arranged into 9 thematic fields:

- Algorithm Theory,
- Natural Language Processing and Knowledge Representation,
- Pattern Recognition and Data Mining,
- Computer Vision,
- Multi-agent Systems and Simulation,
- Computer Networks,
- Digital Signal Processing,
- Computer Architecture and Digital Systems Design,
- Fuzzy Logic and Control

The volume will be useful for researchers, students, and general public interested in the corresponding areas of computer science and engineering.

ISSN: 1870-4069
www.ipn.mx
www.cic.ipn.mx

SEP

INSTITUTO POLITÉCNICO NACIONAL
"La Técnica al Servicio de la Patria"