This volume contains revised versions of 30 accepted papers. The papers are structured into the following five sections:

- Control Systems (10 papers),
- Signal Processing and Pattern Recognition (4 papers),
- Intelligent and Evolutionary Systems (5 papers),
- Computer Networks and Communications (4 papers),
- Algorithms and Applications (7 papers).

This volume can be interesting for researchers and students in computer science, control systems, and also for persons who are interested in cutting edge themes of information technologies.

ISSN: 1870-4069
www.ipn.mx
www.cic.ipn.mx

SEP

Instituto Politécnico Nacional
"La Técnica al Servicio de la Patria"

Advances in computing science and control

Grigori Sidorov
Mireya García-Vázquez
(Eds.)

Vol. 59

Centro de Computación en Educación
Instituto Politécnico Nacional

RESEARCH IN COMPUTING SCIENCE

ISSN: 1870-4069

Advances in computing science and control

Grigori Sidorov
Mireya García-Vázquez
(Eds.)

Vol. 59

RCS
Research in Computing Science

Table of Contents

I Control Systems

Modeling control systems with fuzzy differential equation 3
Mauricio Odremán Vera, Nohé R. Cázarez-Castro

Localización de conjuntos compactos invariantes para el modelo de crecimiento de un tumor cancerígeno 13
C. Plata-Ante, K.E. Starkov, L.N. Coria

On synchronization for Sprott Systems 23
Ramon Ramirez-Villalobos, Luis N. Coria, Luis T. Aguilar, Konstantin E. Starkov

Bounding the domain containing all compact invariant sets of a AIDS model related cancer 35
C. Plata-Ante, K.E. Starkov, L.N. Coria

Bounding the long-time dynamics of a tumor immune-evasion model 45
P.A. Valle, K.E. Starkov, L.N. Coria

On upper bounds for compact invariant sets of nonlinear bladder cancer system with BCG immunotherapy 56
K.E. Starkov, D. Gamboa

Robust sliding mode control for large scale wind turbine for power optimization 65
Jován O. Mérida, Luis T. Aguilar, Jorge A. Dávila

On the control of input–constrained boost DC–to–DC power converters 77
Jorge Guzmán-Guemez, Javier Moreno-Valenzuela

Position/Force control using sliding mode with H_∞ attenuator to reduce rebounds in a mechanical system with a position constraint 87
Raul Rascón, Joaquín Alvarez, Luis T. Aguilar

A control scheme for the tracking control of the Furuta pendulum 99
Carlos Aguilar-Avelar, Javier Moreno-Valenzuela

II Signal processing and pattern recognition

Fourier transform profilometry in 3-Dimensions with matlab programming 113
A. Nava-Vega, J. A. Araiza, E. Luna

Sistema de tiempo-real para el procesamiento robusto de señales de voz usando filtrado local adaptativo 121
Andrés J. Cuevas-Romano, Yuma Sandoval-Ibarra, Victor H. Díaz-Ramírez, Andrés Calvillo-Téllez

Reconocimiento facial robusto usando filtros de correlación diseñados a través de optimización combinatoria 133
Sergio Pinto-Fernández, Alejandra Serrano-Trujillo, Víctor H. Díaz-Ramírez, Leonardo Trujillo Reyes

Evidencia de mejora en los sistemas de reconocimiento basados en iris, utilizando esquemas adaptados de fusión de imágenes 145
Juan M. Colores-Vargas, Mireya S. García-Vázquez, Alejandro A. Ramírez-Acosta, Héctor M. Pérez-Meana

III Intelligent and evolutionary systems

A comparison of predictive measures of problem difficulty for classification with Genetic Programming 159
Yuliana Martínez, Leonardo Trujillo, Edgar Galván-López, Pierrick Legrand

Algoritmos PSO y DE aplicados al problema de inestabilidad en sistemas multiagentes nómadas 171
Alejandro Sosa, Víctor Zamudio, Rosario Baltazar, Carlos Lino, Miguel Angel Casillas, Marco Sotelo

Three metaheuristics solving a scheduling problem in a RIA environment 183
Adriana Perez-Lopez, Rosario Baltazar, Martín Carpio, Arnulfo Alanis

Emotions characterization over EEG analysis: a survey 194
Adrian R. Aguiñaga, Miguel Ángel López Ramírez , Arnulfo Alanís Garza, Rosario Baltazar

Series de Tiempo Difusas aplicadas al pronóstico de la remuneración por la fabricación del calzado 205
Marisol Gutiérrez, Luis Ernesto Mancilla Espinoza, Alfonso Gutiérrez Lugo, Marco A. Gutiérrez

IV Computer networks and communications

Consideraciones para el control de congestión en redes inalámbricas de sensores utilizando la optimización crosslayer 219
Raymundo Buenrostro-Mariscal, Juan Iván Nieto-Hipólito, María Cosio-León, Mabel Vazquez-Briseno, Juan de Dios Sánchez-López

Diseño de un permutador para un decodificador turbo 3GPP LTE de tasa variable 231
Rodríguez Aguiñaga Adrian, Sánchez Adame Moisés, Calvillo Téllez Andrés

Análisis comparativo de las características de radiación de antenas monopolo de microcinta para dispositivos móviles basados en tecnología 4G - LTE 241
Rainer García Aldama, José Luis Medina Monroy, Ricardo Arturo Chávez Pérez

Privacy Threat Warning in eHealth Events Wireless Sensor Networks Transmissions 253
Maria de los Angeles Cosio León, Juan Iván Nieto Hipólito, Raymundo Buenrostro Mariscal, Mabel Vazquez Briseno

V Algorithms and applications

Identificación de riesgos de sequías y sismos al noroeste de Baja California utilizando Sistemas de Información Geográfica 267
Michelle Hallack-Alegría, Mario González-Durán, Mauricio Peregrina-Llanes

Design and harware implementation of Memory-Polynomial Model based on DSP Board 277
J. R. Cárdenas Valdez, J. A. Galaviz Aguilar, A. Calvillo Téllez, C. Gontrand, J. C. Núñez Pérez

GPU-based parallel solution for a phase field model 284
Juan J. Tapia, Rigoberto Alvarado, Fernando A. Villalbazo

Sistema de monitoreo y control de dispositivos en hogares usando plataformas Arduino 295
D. Dueñas, V. Hernández, J. Iriarte, A. Cervantes, H. Vejar, J. López

Diseño en VHDL de un transceptor de la interfaz de línea digital E1 y su implementación en un FPGA 302
Yudith Florencia Gonzalez Padilla, Topacio Osuna Altamirano, Josué López Leyva.

Corrosion in controlsystems decrease the lifetime of the electronic devices of the industrial plants of Mexicali, BC, Mexico 314
Gustavo López Badilla, Benjamín Valdez Salas, Michael Schorr Wiener

Electronic system to save water due to their decrease in Mexicali by the coating of the all american canal in USA 326
Gustavo López Badilla, Elizabeth Romero Samaniego, Sandra Luz Toledo Perea, Miriam Maleni García Castrellon, Luis Alberto Gameros Rios

Author Index 337
Índice de autores

Editorial Board of the Volume 338
Comité editorial del volumen

Modeling control systems with fuzzy differential equation

Mauricio Odremán Vera[1] and Nohé R. Cázarez-Castro[2]

[1] Doctorado en Ciencias de la Ingeniería
Posgrado en Ciencias de la Ingeniería
Departamento de Ingeniería Eléctrica y Electrónica
Instituto Tecnológico de Tijuana, Tijuana, BC, México
[2] Posgrado en Ciencias de la Ingeniería
Departamento de Ciencias Básicas,
Instituto Tecnológico de Tijuana, BC, México
mauricio.odreman@tectijuana.edu.mx, nohe@tectijuana.edu.mx
Paper received on 15/07/12, Accepted on 18/10/12.

Abstract. Fuzzy automatic control system synthesis is suggested for objects described with fuzzy differential equations with crisp coefficients. An illustrative numerical example solving fuzzy differential equation is presented.

Keywords:
Fuzzy differential equations , H-difference , Control System , Stability

1 Introduction

In engineering, dealing with uncertainties in control system design is a common problem in different branches of industry, this due to uncertainties that must to be added to an object model a priori or a posteriori.

Fuzzy modeling is a common way to consider that uncertainties, and this fuzzy models are based in Zadeh's [1–3] pioneer work. There exist wide literature [4–19] on fuzzy control, but these literature is based in Mamdani and Takagi-Sugeno type models, and do not consider a fuzzy differential equation modeling approach.

In this paper is presented an approach of modeling control systems with fuzzy differential equations.

2 Basic concepts and fuzzy differential equation

In this section, we give definitions which will used throughout the paper. See [20, 21]. Let A be a nonempty set. A fuzzy set u in A is characterized by its membership function $u : A \to [0, 1]$. Then $u(x)$ is interpreted as the degree of membership of a element x in the fuzzy set u for each $x \in A$.

Let $\mathbb{R}_F^n$ be the space of the all compact and convex fuzzy sets on $\mathbb{R}^n$.

Definition 1. *For* $u, v \in \mathbb{R}_F^n$, $(u \oplus v)(x) = \sup_{x_1+x_2=x} min\{u(x_1), v(x_2)\}$

Mireya García-Vázquez, Grigori Sidorov (Eds.).
Advances in Computing Science and Control.
Research in Computing Science 59, 2012, pp. 3- 12

The metric structure is given by the Hausdorff distance. [21]

Definition 2. *For $u \in \mathbb{R}_F^n$. The α-cut is the set $[u]^\alpha = \{s \in \mathbb{R}^n : u(s) \geq \alpha\}$, $0 < \alpha < 1$.*

Definition 3. *Let $u, v \in \mathbb{R}_F^n$. If there exists $w \in \mathbb{R}_F^n$ such that $u = v \oplus w$, then w is called the H-difference of u and v and it is denote by $u \ominus v$.*

Definition 4. *Let $F : T \to \mathbb{R}_F^n$ and $t_0 \in T$. The function F is said to be diferentiable at t_0 if:*
(I) an element $F'(t_0) \in \mathbb{R}_F^n$ exist such that, for all $h > 0$ sufficiently near 0, there are $F(t_0 + h) \ominus F(t_0)$, $F(t_0) \ominus F(t_0 - h)$ and the limits

$$\lim_{h \to 0^+} \frac{F(t_0 + h) \ominus F(t_0)}{h} = \lim_{h \to 0^+} \frac{F(t_0) \ominus F(t_0 - h)}{h}$$

are equal to $F'(t_0)$.
(or)
(II) there is an element $F'(t_0) \in \mathbb{R}_F^n$ exist such that, for all $h < 0$ sufficiently near 0, there are $F(t_0 + h) \ominus F(t_0)$, $F(t_0) \ominus F(t_0 - h)$ and the limits

$$\lim_{h \to 0^-} \frac{F(t_0 + h) \ominus F(t_0)}{h} = \lim_{h \to 0^-} \frac{F(t_0) \ominus F(t_0 - h)}{h}$$

are equal to $F'(t_0)$.

Note that if F is differentiable in the first form (I), then it is not differentiable in the second form (II) and viceverse.

Theorem 1. *Let $F : T \to \mathbb{R}_F^n$ and $[F(t)]^\alpha = [F_L^\alpha(t), F_R^\alpha(t)]$, for each $\alpha \in [0, 1]$.Then (i) if F is differentiable in the fist form (I) then $F_L^\alpha(t), F_R^\alpha(t)$ are differentiable functions and*

$$[F'(t)]^\alpha = [(F_L^\alpha(t))', (F_R^\alpha(t))'], \tag{1}$$

(ii) if F is differentiable in the fist form (II) then $F_L^\alpha(t), F_R^\alpha(t)$ are differentiable functions and

$$[F'(t)]^\alpha = [(F_R^\alpha(t))', (F_L^\alpha(t))']. \tag{2}$$

Consider the fuzzy differential equation with crisp coefficients:

$$\tilde{X}^{(n)} + a_1 \tilde{X}^{(n-1)} + \ldots + a_n \tilde{X}(t) = k_{ob} \tau(t) \tag{3}$$

where a_i $i = 1 \ldots n$, and k_{ob} denote the constant coefficients that are crisp numbers. $\tilde{X}(t)$ denote an unknown fuzzy self-valued output variable. $\tau(t)$ is a fuzzy self-valued control action, t is time. $\tilde{X}^{(i)}$ are the i-th derivatives.

The fuzzy function $\tilde{X}(t)$ has the following properties:

$$(X(t))^\alpha = [X_L^\alpha(t), X_R^\alpha(t)] \tag{4}$$

$$\tilde{X}(t) = \cup_{\alpha\in(0,1]}\alpha X^{\alpha}(t); \forall\alpha \in (0,1] \tag{5}$$

$$(X^{\alpha}(t))^{i} = [(X_L^{\alpha}(t))^{(i)}, (X_R^{\alpha}(t))^{(i)}] \tag{6}$$

$$(\tilde{X}^{(i)}(t)) = \bigcup_{\alpha\in(0,1]} \alpha[(X_L^{\alpha}(t))^{(i)}, (X_R^{\alpha}(t))^{(i)}] \tag{7}$$

Let the control action $\tau(t)$ is formed subject to fuzzy functions $X^{(i)}(t).i = 1\ldots n$

$$\tau(t) = -\sum_{j=0}^{r} k_{p_j} X^{(j)}(t) \tag{8}$$

where $k_{p_j}\ j = 1\ldots r$ denote the parameters of controller tuning and are crisp numbers.

In most cases the stability degree is the main index of performance for fuzzy control systems [22] .Taking this into account, the problem of synthesis of control system dynamic object (3) can be formulated as follows.

It requires to determine such control (8) type, that transfers the dynamic object, described by fuzzy differential equation, from a given fuzzy initial state

$$\tilde{X}(t_0) = \tilde{X}_0, \tilde{X}^{(i)}(t_0) = \tilde{X}_0^{(i)}, (i = 1\ldots n-1) \tag{9}$$

to finite state

$$\tilde{X}(T) = \tilde{0} \tag{10}$$

providing maximal stability degree:

$$J = \max_{k_{pi}\in K_{pi}} \{-\mathrm{Re}\lambda_{\alpha}(a_1, a_2, \ldots a_n, k_{p1}, k_{p2}, \ldots k_{pm}\} \tag{11}$$

where

$$K_{pi} = [K_{pi}^{\min}, K_{p_i}^{\max}], K_{p_i}^{\min} \geq 0, i = 1\ldots r. \tag{12}$$

3 Synthesis of fuzzy control system

The formulated problem of synthesis of fuzzy control system (controller) (3), (8-12) of a high order can be solver on a base of the iterative approach [22]. That is, beginning from some $k_{p_i}^0$ for every parameter k_{p_i}, adding some definite $\triangle k_{p_i}$ and fixing transient processes in system (3), (8), that satisfy initial and finite conditions (9) and (10), the certain fuzzy solution sets are determined. Then we can choose an optimal solution from this set, i.e the optimal value k_{p_i}, that provides the maximal stability degree [22].

In practice many technological objects (including robotic manipulator as object of automatic control) are described by differential equations as a rule of the second or the third order. Taking into account the following, the formulated problem of parametric synthesis of control (3), (8-12) can be solved analytically by a method like suggested in [22].

Let an order of fuzzy differential equation, describing a control system be $n = 2$, the order of fuzzy controller (8) is $r = 0$, taking this into account the characteristic equations of fuzzy control system will be described by the following fuzzy differential equation:

$$\tilde{X}'' + a_1\tilde{X}' + (a_2 + k_{ob}k_{p0})\tilde{X}(t) = \tilde{0} \tag{13}$$

$$\tilde{X}(t_0) = \tilde{X}_0, \tilde{X}'(t_0) = \tilde{X}'_0, \tag{14}$$

here 0 is a fuzzy zero. Then based on [23, 24] we can write the following expressions for fuzzy differential equation (17), and also for initial conditions (14):

$$(X_L^\alpha)''(t) + a_1(X_L^\alpha)'(t) + (a_2 + k_{ob}k_{p0})X_L(t) = 0_L^\alpha \tag{15}$$

$$(X_R^\alpha)''(t) + a_1(X_R^\alpha)'(t) + (a_2 + k_{ob}k_{p0})X_R(t) = 0_R^\alpha \tag{16}$$

Is should be noted that an interval $0^\alpha = [0_L^\alpha, 0_R^\alpha]$, which is and α-cut of the fuzzy zero, is sufficiently small and, in particular case, can be singleton.

From the convergence condition of solution of differential equations, or providing of control system stability and maximal performance criterion, stability degree of the tuning parameter k_{p_0} of fuzzy controller (8) can be determined in accordance with [22].

4 Damping analysis

In examples consider fuzzy differential equation

$$\tilde{X}'' + a_1\tilde{X}' + (a_2 + k_{ob}k_{p0})\tilde{X}(t) = \tilde{0} \tag{17}$$

with $X' = [-0.001, 0.001]$, $\tilde{0} = [-0.0001, 0.0001]$ and $X_0 = [2, 4]$.

4.1 Overdamped case

With $a_1 = 4.5$, $a_2 = 0.95$, $k_{ob} = 20$, $k_{p0} = 1.7$, then we obtain

$$(X_L^\alpha)''(t) + 4.5(X_L^\alpha)'(t) + 4.35X_L(t) = -0.001, X_L = 2, X' = -0.0001 \tag{18}$$

$$(X_R^\alpha)''(t) + 4.5(X_R^\alpha)'(t) + 4.35X_L(R) = 0.001, X_R = 4, X' = 0.0001 \tag{19}$$

and solving (18-19), results:

$$X_L(t) = 3.66551e^{-1.4059028t} - 1.66551e^{-3.0940972t} - 0.0000229885 \tag{20}$$

and

$$X_R(t) = 7.33078e^{-1.4059028t} - 3.33101e^{-3.0940972t} + 0.0000229885, \tag{21}$$

which response is depicted in Fig. 1.

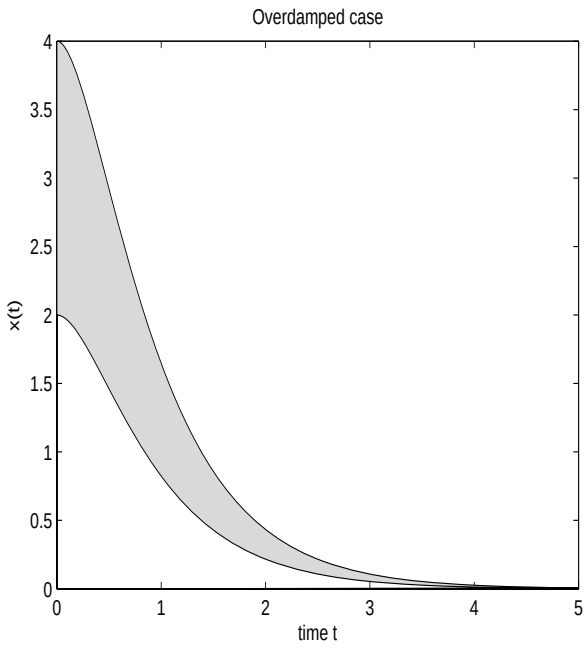


Fig. 1. Over damped case

4.2 Critically damped case

With $a_1 = 4.5$, $a_2 = 1$, $k_{ob} = 20$, $k_{p0} = 1.5$, then we obtain

$$(X_L^{\alpha})''(t) + 4(X_L^{\alpha})'(t) + 4X_L(t) = -0.001, X_L = 2, X' = -0.0001 \tag{22}$$

$$(X_R^{\alpha})''(t) + 4(X_R^{\alpha})'(t) + 4X_R(t) = 0.001, X_R = 4, X' = 0.0001 \tag{23}$$

and solving (22-23), results:

$$X_L(t) = 2.00025e^{-2t} + 4.0004te^{-2t} - 0.00025 \tag{24}$$

and

$$X_R(t) = 3.999755e^{-2t} + 7.9996te^{-2t} + 0.00025, \tag{25}$$

which response is depicted in Fig. 2.

4.3 Underdamped case

With $a_1 = 2.4$, $a_2 = 0.95$, $k_{ob} = 20$, $k_{p0} = 1.7$, then we obtain

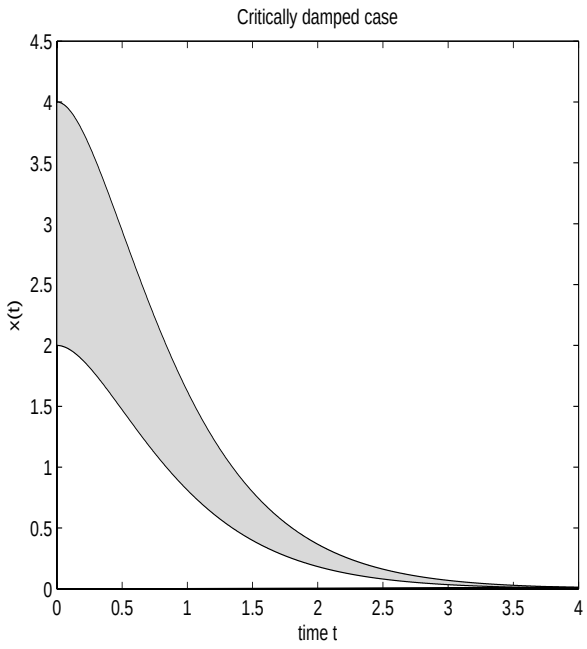


Fig. 2. Critically damped case.

$$(X_L^{\alpha})''(t) + 2.4(X_L^{\alpha})'(t) + 4.35X_L(t) = -0.001, X_L = 2, X' = -0.0001 \quad (26)$$

$$(X_R^{\alpha})''(t) + 2.4(X_R^{\alpha})'(t) + 4.35X_R(t) = 0.001, X_R = 4, X' = 0.0001 \quad (27)$$

and solving (26-27), results:

$$X_L(t) = 2.00023e^{-1.2t}\cos(1.7058722t) + 1.40701e^{-1.2t}\sin(1.7058722t) - 0.000229885 \quad (28)$$

and

$$X_R(t) = 2.81371e^{-1.2t}\cos(1.7058722t) + 3.99977e^{-1.2t}\sin(1.7058722t) + 0.000229885, \quad (29)$$

which response is depicted in Fig. 3.

5 The case study

Consider fuzzy differential equation

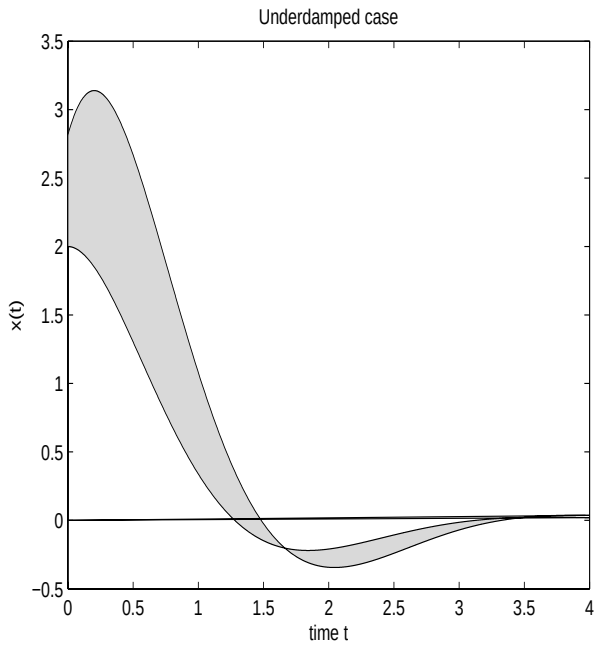


Fig. 3. Underdamped case.

$$\tilde{X}'' + a_1\tilde{X}' + a_2\tilde{X}(t) = \tilde{0} \tag{30}$$

with $X' = [-0.001, 0.001]$, $\tilde{0} = [-0.0001, 0.0001]$, using (15)-(16),

$$(X_L^\alpha)''(t) + a_1(X_L^\alpha)'(t) + a_2X_L(t) = 0_L^\alpha \tag{31}$$

$$(X_R^\alpha)''(t) + a_1(X_R^\alpha)'(t) + a_2X_R(t) = 0_R^\alpha \tag{32}$$

is obtained.

The motion is determined by the roots of the characteristic equation of (30).
If the roots are real and unequal, the motion is overdamped, if the roots are real and equal, the motion is critically damped, finally, if the roots are conjugate complex numbers, the motion is underdamped.

In Fig. 4 , a small block of mass m is attached to one end of a spring, and the other end of the unstretched spring is attached to a fixed wall. Assuming that the block is acted on by a force of friction that opposes the motion.

If the block weigh is $32lb$, the spring constant $k = 36lb/ft$, and the resistance coefficient is $b = 13$, the resulting system equation is:

$$\tilde{X}'' + 13\tilde{X}' + 36\tilde{X}(t) = \tilde{0} \tag{33}$$

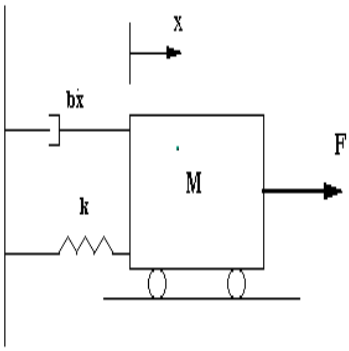


Fig. 4. mass-damper-spring

with $X' = [-0.001, 0.001]$, $\tilde{0} = [-0.0001, 0.0001]$ and $X_0 = [1.9, 2.1]$, and solving (33), results on a overdamped motion because

$$X_L(t) = 3.7802e^{-4t} - 1.6802e^{-9t} \tag{34}$$

and

$$X_R(t) = 3.4198e^{-4t} - 1.5198e^{-9t} \tag{35}$$

and the system's response is depicted in Fig. 5.

6 Conclusions

The approach reported in this paper helps in the modeling of uncertainties directly in fuzzy differential equations, and the reported cases of study gives an approximation that extends control engineering theory to the fuzzy case according with the need of dealing with those uncertainties.

The resulting family of solutions to the fuzzy differential equations gives a footprint of uncertainty, like the defined in [25], which helps to conclude that the results reported in this paper are according with the type-2 fuzzy [25] logic theory of the computing with words paradigm. Finally, the formulas for computing second order fuzzy differential equation was obtained.

References

1. Zadeh, L.A.: The concept of a linguistic variable and its application to approximate reasoning - i. Inf. Sci. **8**(3) (1975) 199–249
2. Zadeh, L.A.: The concept of a linguistic variable and its application to approximate reasoning - ii. Inf. Sci. **8**(4) (1975) 301–357
3. Zadeh, L.A.: The concept of a linguistic variable and its application to approximate reasoning - ii. Inf. Sci. **9**(1) (1975) 43–80

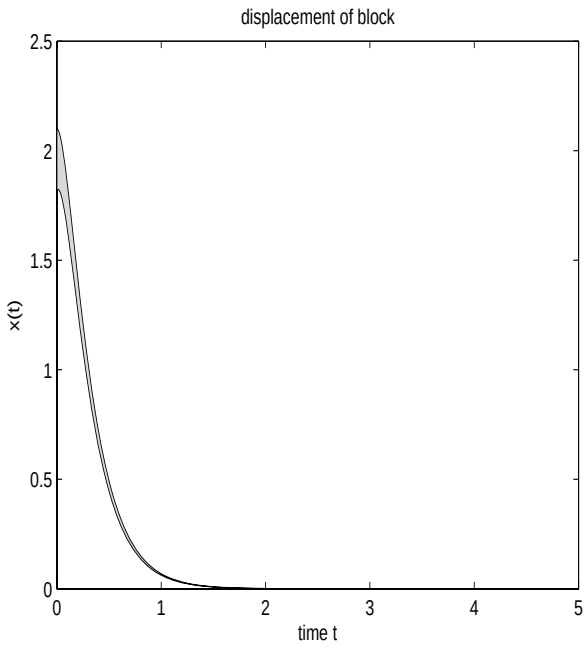


Fig. 5. Block displacement

4. Lim, C., Hiyama, T.: Application of fuzzy logic control to a manipulator. Robotics and Automation, IEEE Transactions on **7**(5) (oct 1991) 688 –691
5. Newton, R., Xu, Y.: Neural network control of a space manipulator. Control Systems, IEEE **13**(6) (dec. 1993) 14 –22
6. Llama, M., Kelly, R., Santibañez, V.: Stable computed-torque control of robot manipulators via fuzzy self-tuning. Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on **30**(1) (feb 2000) 143 –150
7. Yoo, B.K., Ham, W.C.: Adaptive control of robot manipulator using fuzzy compensator. Fuzzy Systems, IEEE Transactions on **8**(2) (apr 2000) 186 –199
8. Tsai, C.H., Wang, C.H., Lin, W.S.: Robust fuzzy model-following control of robot manipulators. Fuzzy Systems, IEEE Transactions on **8**(4) (aug 2000) 462 –469
9. Peng, L., Woo, P.Y.: Neural-fuzzy control system for robotic manipulators. Control Systems, IEEE **22**(1) (feb 2002) 53 –63
10. Sun, Y.L., Er, M.J.: Hybrid fuzzy control of robotics systems. Fuzzy Systems, IEEE Transactions on **12**(6) (dec. 2004) 755 – 765
11. Santibañez, V., Kelly, R., Llama, M.: Global asymptotic stability of a tracking sectorial fuzzy controller for robot manipulators. Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on **34**(1) (feb. 2004) 710 – 718
12. Santibañez, V., Kelly, R., Llama, M.: A novel global asymptotic stable set-point fuzzy controller with bounded torques for robot manipulators. Fuzzy Systems, IEEE Transactions on **13**(3) (june 2005) 362 – 372

13. Merchan-Cruz, E., Morris, A.: Fuzzy-ga-based trajectory planner for robot manipulators sharing a common workspace. Robotics, IEEE Transactions on **22**(4) (aug. 2006) 613 –624
14. Chen, C.S.: Supervisory adaptive tracking control of robot manipulators using interval type-2 tsk fuzzy logic system. Control Theory Applications, IET **5**(15) (13 2011) 1796 –1807
15. Meza, J.L., Santibanez, V., Soto, R., Llama, M.A.: Fuzzy self-tuning pid semiglobal regulator for robot manipulators. Industrial Electronics, IEEE Transactions on **59**(6) (june 2012) 2709 –2717
16. Castillo, O., Aguilar, L., Cazarez, N., Cardenas, S.: Systematic design of a stable type-2 fuzzy logic controller. Applied Soft Computing **8**(3) (2008) 1274 – 1279.
17. Cazarez-Castro, N., Aguilar, L., Castillo, O., Rodriguez, A.: Optimizing type-1 and type-2 fuzzy logic systems with genetic algorithms. Research in Computing Science **39** (October 2008) 131–153
18. Cazarez-Castro, N.R., Aguilar, L.T., Castillo, O.: Fuzzy logic control with genetic membership function parameters optimization for the output regulation of a servomechanism with nonlinear backlash. Expert Systems with Applications **37**(6) (2010) 4368 – 4378
19. Cazarez-Castro, N.R., Aguilar, L.T., Castillo, O.: Designing type-1 and type-2 fuzzy logic controllers via fuzzy lyapunov synthesis for nonsmooth mechanical systems. Engineering Applications of Artificial Intelligence **25**(5) (2012) 971 – 979
20. Puri, M.L., Ralescu, D.A.: Differentials of fuzzy functions. Journal of Mathematical Analysis and Applications **91**(2) (1983) 552 – 558
21. Kaleva, O.: Fuzzy differential equations. Fuzzy Sets and Systems **24**(3) (1987) 301 – 317
22. Zeynalov E. R., Jafarov P. S., J.S.M.: Synthesis of fuzzy controllers for astatic objects, described by differential equations with fuzzy coefficients, Antalya, Turkey (2007) 290 – 295
23. Kaleva, O.: A note on fuzzy differential equations. Nonlinear Analysis: Theory, Methods and Applications **64**(5) (2006) 895 – 900
24. Li, Z.: Fuzzy Chaotic Systems. Springer (2006)
25. Mendel, J.: Uncertain Rule-Based Fuzzy Logic Systems: Introduction and New Directions. Prentice Hall, Upper Saddle River, NJ (2001)

Localización de conjuntos compactos invariantes para el modelo de crecimiento de un tumor cancerígeno

C Plata-Ante, KE Starkov, LN Coria*

Centro de investigación y desarrollo de tecnología digital, CITEDI-IPN
*Instituto Tecnológico de Tijuana
Tijuana, B.C.
cplata@citedi.mx, konst@citedi.mx, *luis.coria@gmail.com
Paper received on 16/07/12, Accepted on 17/10/12.

Abstract. En este documento se describe un método para la localización de conjuntos compactos invariantes que es aplicado a un modelo matemático que representa el comportamiento de un tumor cancerígeno. El resultado final de localización corresponde a la intersección de planos con diferentes orientaciones que están presentes en el primer octante debido a las características matemáticas - biológicas del sistema. Esto se realizó por medio de la aplicación del método de condiciones de extrema y el teorema iterativo para la obtención los resultados. Con ayuda de simulaciones numéricas se muestra como algunos conjuntos compactos invariantes están ubicados en el dominio de localización.

Palabras clave: Cáncer, conjuntos compactos invariantes, sistema biológico.

1 Introducción

Una de las enfermedades mortales que pueden afectar al ser humano desde su estado fetal hasta la edad adulta, es el cáncer. Cuando se padece cáncer, las células normales sufren una mutación que altera su estructura biológica y causa un gran daño al organismo. Por esta razón una gran cantidad de recursos humanos y financieros son utilizados en investigaciones con la finalidad de atacar ésta enfermedad [1], [2]. Actualmente el modelado de sistemas biológicos [3], [4] ayuda a identificar el comportamiento del crecimiento de un tumor cancerígeno y la interacción con las células del sistema inmune [5].

En este documento se estudia el modelo de la regresión y progresión espontánea de un tumor, fenómeno que se produce por la interacción entre las células cancerígenas y las células del sistema inmune llamadas asesinos naturales, siendo un sistema presa-depredador como fue descrito por Sarkar y El-Gohary en [6], [3], respectivamente. Las células cancerígenas que están presentes en el organismo, tomando la forma de un tumor o neoplasia son atacadas por las células del sistema inmunológico con el fin de destruirlas y hacer que el tumor cancerígeno desaparezca del organismo, aunque en ocasiones las células tumorales emigran a otro tejido por un proceso llamado metástasis, característica que se presenta en los tumores malignos, haciendo esta que este proceso resulte más complejo. El modelo matemático de este sistema [7] está dado por:

Mireya García-Vázquez, Grigori Sidorov (Eds.).
Advances in Computing Science and Control.
Research in Computing Science 59, 2012, pp. 13- 22

$$\begin{aligned}\dot{x}_1 &= 1 + a_1 x_1 - a_1 x_1^2 - x_1 x_2 \\ \dot{x}_2 &= a_2 x_2 x_3 - a_3 x_2 \\ \dot{x}_3 &= -a_4 x_3^2 - a_5 x_2 x_3 + (a_4 - a_6) x_3\end{aligned} \tag{1}$$

donde la variable x_1 corresponde a la densidad de población de la presa, células del tumor; y $x_i; i = 2, 3$ corresponde a la densidad del depredador, células de caza y células en reposo (dependiendo de la actividad que estén realizando las células del sistema inmunológico), respectivamente. Cuando el cuerpo detecta células que tienen alguna alteración (células tumorales), el sistema inmunológico actúa activando células llamadas asesinos naturales (linfocitos T y macrófagos). Estas células se encargan de cazar las células tumorales para posteriormente destruirlas, mientras que el resto de las células del sistema inmune se encuentra en reposo. Cuando las células de caza llevan a cabo el proceso de destrucción de las células tumorales liberan una substancia y después mueren. Dicha sustancia hace que las células que están en reposo se activen convirtiéndose en células de caza, las cuales, atacan células cancerígenas, representando con esto la dinámica del sistema. El documento se organiza como sigue: en la siguiente sección se muestran algunas notaciones matemáticas y conceptos para realizar el análisis de la localización de conjuntos compactos invariantes. Después se proponen algunas funciones localizadoras y se realizan algunas simulaciones numéricas para visualizar los resultados. Finalmente se muestran las conclusiones de esta investigación.

2 Preliminares matemáticos y notaciones

Con el objetivo de encontrar un dominio de localización se utilizó el método general descrito en [8], [9] y [10]. Los resultados útiles para el desarrollo de esta investigación se muestran a continuación.

Considerando el siguiente sistema no lineal:

$$\dot{x} = f(x) \tag{2}$$

donde f es un C^∞ -campo vectorial diferenciable-, aquí $x \in \mathbf{R}^n$ es el -vector de estado-. Si se toma $h(x)$ que sea C^∞ -función diferenciable la cual no es un primera integral de 2-. Se toma por $h|_B$ la restricción de h en el conjunto $B \subset \mathbf{R}^n$. Por $S(h)$ se denota el conjunto $\{x \in \mathbf{R}^n \mid L_f h(x) = 0\}$ donde $L_f h$ de 2 está dada por $L_f h(x) = \frac{\partial h}{\partial x} f(x)$. Los valores $h_{\inf} = \inf\{h(x) \mid x \in S(h)\}$ y $h_{\sup} = \sup\{h(x) \mid x \in S(h)\}$ serán usados durante el análisis de localización.

Teorema 1. Cada conjunto compacto invariante Γ de 2 está contenido en el conjunto de localización:

$$K(h) = \{h_{\inf} \leq h(x) \leq h_{\sup}\}.$$

La función h aplicada en este teorema será llamada función localizadora. Es evidente que si todos los conjuntos compactos invariantes están localizados en los conjuntos Q_1 and Q_2, con $Q_1; Q_2 \subset \mathbf{R}^n$,, entonces estarán localizados también $Q_1 \cap Q_2$.

El teorema iterativo se muestra a continuación:

Teorema 2. Sea $h_m(x), m = 0, 1, 2, \ldots$ una secuencia de funciones localizadoras de C^∞ ($\mathbf{R}^n$). Los conjuntos:

$$K_0 = K(h_0), \quad K_m = K_{m-1} \cap K_{m-1,m}, \quad m > 0,$$

con

$$\begin{aligned} K_{m-1,m} &= \{x : h_{m,inf} \leq h_m(x) \leq h_{m,sup}\}, \\ h_{m,sup} &= \sup_{S(h_m)\cap K_{m-1}} h_m(x), \\ h_{m,inf} &= \inf_{S(h_m)\cap K_{m-1}} h_m(x), \end{aligned}$$

contiene todos los conjuntos compactos invariantes del sistema 2 y

$$K_0 \supseteq K_1 \supseteq \cdots \supseteq K_m \supseteq \ldots .$$

Por el sentido biológico y matemático de cada uno de las estados que componen el sistema (1), los conjuntos compactos invariantes se presentarán en el primer octante (octante positivo) R^3_+, y cualquier localización que este fuera de esta región no será útil para este análisis. Además, todos los parámetros del sistema son positivos. También, para simplificación de notaciones $K\left(h\right) = K\left(h\right) \cap \mathbb{R}^3_+$.

3 Análisis de localizaciones de un modelo de un tumor cancerígeno

Para encontrar el dominio de localización del sistema, se proponen diferentes funciones localizadoras lineales y racionales. La intención de proponer estas funciones es implementar el método desarrollado por el Dr. Krishchenko y el Dr. Starkov, de acuerdo a los análisis desarrollados se muestra una ventaja con respecto a otros métodos debido a que no se restringe la propuesta de funciones y la utilización del método iterativo permite mejorar las cotas. A continuación se muestran los resultados obtenidos.

3.1 Resultados de localización por medio de funciones lineales

En esta sección se presentan diferentes funciones que da resultado a una localización compuesta por la intersección de diferentes planos.

Se propone la función

$$h_1 = \frac{a_5}{a_2} x_2 + x_3$$

Al calcular $L_f h$ de la función después de despeja x_2 de la fórmula para obtener la ecuación para $h_1|_{S(h_1)}$:

$$h_1|_{S(h_1)} = \frac{a_5}{a_2} \left(\frac{-a_4 x_3^2 + x_3 \left(a_4 - a_6\right)}{\dfrac{a_5}{a_2} a_3} \right) + x_3$$

Entonces el límite superior de la función esta dado por

$$h_1|_{S(h_1)} < \frac{(a_4 - a_6 + a_3)^2}{4a_3a_4} = h_1 \sup$$

Teorema 3**:** *Todos los conjuntos compactos invariantes del sistema* (1) *están localizados en el conjunto*

$$K(h_1) = \left\{ \frac{a_5}{a_2}x_2 + x_3 \leq \frac{(a_4 - a_6 + a_3)^2}{4a_3a_4} \right\} \tag{3}$$

La segunda función localizadora que se propone es

$$h_2 = x_3$$

Considerando $L_f h = 0$, se tiene $S(h_2) \cap \mathbb{R}^3_+$ respecto a x_3 y se obtiene:

$$h|_{S(h)\cap\mathbb{R}^3_+} = 1 - \frac{a_6}{a_4} - \frac{a_5x_2}{a_4}.$$

Por las características biológicas, considerando $x_2 > 0$, y el límite de x_3 se tiene que:

$$h|_{S(h)\cap\mathbb{R}^3_+} \leq 1 - \frac{a_6}{a_4} = h_2 \sup$$

Teorema 4**:** *Si*

$$a_4 > a_6$$

todos los conjuntos compactos invariantes están localizados en el conjunto:

$$K(h_2) = \left\{ x_3 \leq 1 - \frac{a_6}{a_4} \right\}.$$

Después con la finalidad de encontrar un límite inferior para x_1, se propone la función

$$h_3 = x_1$$

y tenemos:

$$S(h_3) = \left\{ L_f h = 1 + a_1x_1 - a_1x_1^2 - x_1x_2 = 0 \right\}$$

Considerando el resultado en 3, se toma x_2 de la fórmula, entonces

$$x_2 \leq \frac{a_2\rho}{a_5}$$

donde

$$\rho := \frac{(a_4 - a_6 + a_3)^2}{4a_3a_4}$$

utilizando el teorema iterativo, se sustituye el límite de x_2 y se obtiene $K(h_1)$ en $S(h_3)$:

$$S(h_3) \cap K(h_1) \cap \mathbb{R}^3_+ \subset \left\{ |x_1 - \gamma| \geq \pm\sqrt{\frac{1}{a_1} + \gamma^2} \right\}$$

donde

$$\gamma := \frac{a_1 a_5 - a_2 \rho}{2 a_1 a_5}$$

Teorema 5**:** *Si*

$$a_1 a_5 - a_2 \rho > 0$$

el conjunto

$$K(h_3) = \left\{ x_1 \geq \sqrt{\frac{1}{a_1} + \gamma^2} + \gamma \right\}$$

contiene todos los conjuntos compactos invariantes.

Otra función localizadora es

$$h_4 = x_1 + \frac{a_5}{a_2} x_2 + x_3$$

después si $L_f h = 0$ se tiene

$$S(h_4) \cap \mathbb{R}^3_+ \subset \left\{ x_2 \leq \frac{a_2 \left(a_1 x_1 - a_4 x_3^2 - a_1 x_1^2 + a_4 x_3 - a_6 x_3 + 1 \right)}{a_3 a_5} \right\}$$

Teorema 6**:** *El conjunto de localización de conjuntos compactos invariantes para (1) es*

$$K(h_4) = \left\{ x_1 + \frac{a_5}{a_2} x_2 + x_3 \leq \frac{\left(a_3 + a_1 \right)^2}{4 a_1 a_3} + \frac{\left(a_3 + a_4 - a_6 \right)^2}{4 a_3 a_4} + \frac{1}{a_3} \right\}.$$

La última función lineal propuesta es

$$h_5 = x_1 - x_2$$

se calcula la $L_f h_5$ y se tiene $S(h_5)$ respecto a x_2

$$S(h_5) \cap \mathbb{R}^3_+ \subset \left\{ -x_2 \leq -\frac{a_1}{a_3} x_1^2 + \frac{a_1}{a_3} x_1 + \frac{1}{a_3} \right\}.$$

Teorema 7**:** *Todos los conjuntos compactos invariantes están localizados en*

$$K(h_5) = \left\{ x_1 - x_2 \leq \frac{\left(a_3 + a_1 \right)^2}{4 a_1 a_3} + \frac{1}{a_3} \right\}.$$

3.2 Resultados de localización por medio de funciones racionales

Entre las funciones racionales se propone

$$h_8 = \frac{x_2}{x_1}.$$

Se calcula la $L_f h$ y se tiene $S(h_8)$ respecto a x_2

$$S(h_8) \cap \mathbb{R}^3_+ \subset \left\{ x_2 \leq a_3 + \frac{1}{x_1} + a_1 \right\}$$

Utilizando el teorema iterativo, se tiene $K(h_5)$ en $h_8|_{S(h_8)\cap\mathbb{R}^3_+}$

$$S(h_8) \cap K(h_5) \cap \mathbb{R}^3_+ \subset \left\{ h_8|_{S(h_8)\cap\mathbb{R}^3_+} \leq a_1 + a_3 + \frac{4a_1a_3}{(a_3+a_1)^2 + 4a_1} \right\}$$

Teorema 8**:** *Todos los conjuntos compactos invariantes están contenidos en*

$$K(h_8) = \left\{ \frac{x_2}{x_1} \leq a_1 + a_3 + \frac{4a_1a_3}{(a_3+a_1)^2 + 4a_1} \right\}.$$

Se propone la función

$$h_9 = \frac{x_3}{x_2}$$

su $S(h_9)$ está dada por

$$S(h_9) \cap \mathbb{R}^3_+ \subset \left\{ x_2 \leq \frac{a_3 + a_4 - a_6}{a_5} \right\}$$

Aplicando el teorema iterativo con $K(h_3)$ y después sustituyendo en $h_9|_{S(h_9)}$.

$$S(h_9) \cap K(h_3) \cap \mathbb{R}^3_+ \subset \left\{ h_9|_{S(h)\cap\mathbb{R}^3_+} \leq \frac{a_5\,(a_4 - a_6)}{a_4\,(a_3 + a_4 - a_6)} \right\}.$$

Teorema 9**:** *Si*

$$a_4 > a_6,$$

entonces la localización de todos los conjuntos compactos invariantes están contenidos en

$$K(h_9) = \left\{ \frac{x_3}{x_2} \leq \frac{a_5\,(a_4 - a_6)}{a_4\,(a_3 + a_4 - a_6)} \right\}.$$

Otra función que se propone es

$$h_{10} = \frac{x_2}{x_3}$$

la $S(h_{10})$ se describe por

$$S(h_{10}) \cap \mathbb{R}^3_+ \subset \{(a_2 - a_4)\,x_3 = a_3 + a_4 + a_6 + a_5 x_2\}$$

Considerando el límite x_2 de $K(h_1)$ y aplicando el teorema iterativo se tiene que

$$S(h_{10}) \cap K(h_1) \cap \mathbb{R}^3_+ \subset \left\{ h_{10}|_{S(h_{10})\cap\mathbb{R}^3_+} \leq \frac{a_2\rho\,(a_2 - a_4)}{a_5\rho_5} \right\}.$$

Teorema 10**:** *Si*

$$a_2 > a_4$$

entonces todos los conjuntos compactos invariantes están localizados en

$$K(h_{10}) = \left\{ \frac{x_2}{x_3} \leq \frac{a_2\rho\left(a_2 - a_4\right)}{a_5\rho_5} \right\}.$$

Se propone la función

$$h_{11} = \frac{x_1}{x_2}$$

se calcula su $L_f h_{11}$, y la $S(h_{11})$ se describe por

$$S(h_{11}) \cap \mathbb{R}^3_+ \subset \left\{ x_1 \leq \sqrt{\frac{1}{a_1} + \frac{\left(a_1 + a_3\right)^2}{4a_1^2}} + \frac{a_1 + a_3}{2a_1} \right\}$$

Sustituyendo $S(h_{11})$ en h_{11} y usando el teorema iterativo con x_2 de $K(h_1)$, se tiene

$$S(h_{11}) \cap K(h_1) \cap \mathbb{R}^3_+ \subset \left\{ h_{11}|_{S(h)\cap\mathbb{R}^3_+} \leq \frac{a_5\rho_6}{a_2\rho} \right\}$$

donde

$$\rho_6 := \sqrt{\frac{1}{a_1} + \frac{\left(a_1 + a_3\right)^2}{4a_1^2}} + \frac{a_1 + a_3}{2a_1}$$

Teorema 11**:** *Todos los conjuntos compactos invariantes están contenidos en*

$$K(h_{11}) = \left\{ \frac{x_1}{x_2} \leq \frac{a_5\rho_6}{a_2\rho} \right\}.$$

3.3 Simulaciones numéricas

A partir de las funciones obtenidas en la sección anterior, en la Tabla 1 se presentan algunas funciones localizadoras que son válidas de acuerdo a determinados parámetros con la finalidad de representar gráficamente los resultados obtenidos.

Con estas localizaciones se simula el modelo con los parámetros utilizados en [7]: $a_1 = 2.5$, $a_2 = 4.5$, $a_3 = 0.6$, $a_4 = 3.5$, $a_5 = 2$ y $a_6 = 0.1$ y las condiciones de densidad iniciales $x_1(0) = 0.5$, $x_2(0) = 1$ y $x_3(0) = 0.5$.

Con la dinámica obtenida y evaluados los parámetros en cada una de las localizaciones , el valor numérico de cada una está dado por:

$$K(h_1) = \{0.44x_2 + x_3 \leq \rho =: 1.\,904\,8\} \qquad K(h_2) = \{x_3 \leq 0.971\,43\}$$
$$K(h_3) = \{x_1 \geq 0.369\,18\} \qquad K(h_4) = \{x_1 + 0.44x_2 + x_3 \leq 5.\,173\,1\}$$
$$K(h_9) = \left\{ \tfrac{x_1}{x_2} \leq 0.351\,31 \right\} \qquad K(h_{11}) = \left\{ \tfrac{x_3}{x_2} \leq 0.485\,71 \right\}$$

Table 1. Principales resultados de las funciones localizadoras para el sistema (1).

Dominio de la Localización	Condición
$K(h_1) = \left\{ \frac{a_5}{a_2} x_2 + x_3 \leq \rho := \frac{(a_3+a_4-a_6)^2}{4a_3a_4} \right\}$	$*$
$K(h_2) = \left\{ x_3 \leq 1 - \frac{a_6}{a_4} \right\}$	$a_4 > a_6$
$K(h_3) = \left\{ x_1 \geq \sqrt{\frac{1}{a_1} + \frac{(a_1a_5-a_2\rho)^2}{4a_1^2a_5^2}} + \frac{(a_1a_5-a_2\rho)}{4a_1a_5} \right\}$	$*$
$K(h_4) = \left\{ x_1 + \frac{a_5}{a_2} x_2 + x_3 \leq \frac{(a_3+a_1)^2}{4a_1a_3} + \frac{1}{a_3} + \rho \right\}$	$*$
$K(h_9) = \left\{ \frac{x_1}{x_2} \leq \frac{a_5}{2a_1a_2\rho} \left(\sqrt{4a_1 + (a_1+a_3)^2} + a_1 + a_3 \right) \right\}$	$*$
$K(h_{11}) = \left\{ \frac{x_3}{x_2} \leq \frac{a_5(a_4-a_6)}{a_4(a_3+a_4-a_6)} \right\}$	$a_4 > a_6$

El conjunto de localizaciones se muestran en la Figura 1, intersecciones entre funciones que se presentan en la Tabla 1. En las simulaciones se presenta la dinámica del sistema y la región en la que se encuentra contenida; $K(h_1) \cap K(h_2) \cap K(h_3) \cap K(h_4) \cap K(h_9) \cap K(h_{11})$.

Se debe cumplir la condición $a_4 > a_6$ para que la región de localización de conjuntos compactos del modelo (1) esté definida por la Figura 1, lo que significa que la tasa de crecimiento de las células en reposo tiene que ser mayor a su muerte natural. La característica de este sistema es que los parámetros se pueden variar con la finalidad de realizar diferentes análisis con el sistema.

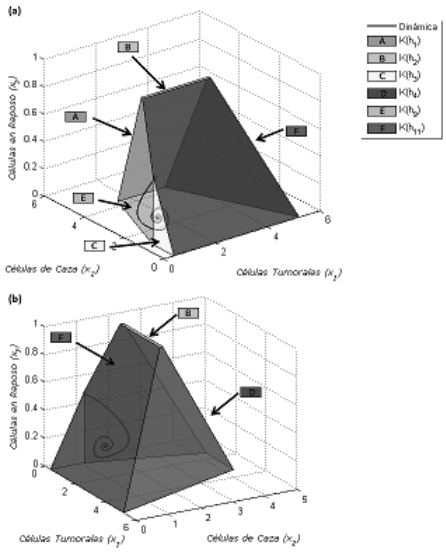


Fig. 1. Localización de conjuntos compactos invariantes; a) punto de equilibrio en x_1, x_2 y x_3; b) punto de equilibrio en x_2, x_3 y x_4.

Una de las cotas que es importante tener el control para poder disminuirla es el cancer, x_1. El resultado de localización $K(h_3)$ que involucra a x_1 determina que siempre habrá un nivel de células tumorales en el paciente, sin embargo uno de los objetivos a largo plazo de la terapia del cáncer es prevenir la progresión, invasión y metástasis de células cancerosas. Aunque se ha demostrado que es difícil de matar o eliminar todas las células del cáncer última del cuerpo con fines terapéuticos, es posible que con la combinación correcta de los fármacos se puede inducir un estado de latencia, que en efecto podría convertir el cáncer en una enfermedad crónica, pero controlada.

4 Conclusiones

En el análisis del *modelo del tumor cancerígeno* se logró obtener una región en $\mathbb{R}^3_+$ que encerrará la dinámica de los sistemas bajo la condición de que las variables y los parámetros son positivos. Se utilizaron funciones lineales y posteriormente, con el propósito de encontrar una región en la cual se encuentran los conjuntos compactos invariantes, se propusieron funciones racionales hasta obtener una figura formada por la intersección de diferentes planos. Por las características de este sistema (1), los parámetros involucrados se pueden variar con el propósito de conocer qué es lo que sucedería en un paciente al sufrir cambios en el organismo, por ejemplo cuando existe liberación sustancias que hacen que las células del sistema inmune ataque a las cancerígenas o manipular la cantidad de células tumorales que se presentan en el organismo y saber cuál es la dinámica que se establece debido a que es un sistema presa-

depredador. Esto lleva a que la localización obtenida será dinámica, ya que depende directamente de los parámetros del sistema.

References

1. De Pillis, L., Radunskaya, A.: A mathematical tumor model with immune resistance and drug therapy: an optimal control approach. Computational and Mathematical Methods in Medicine **3** (2001) 79–100
2. Menchón, S., Ramos, R., Condat, C.: Modeling subspecies and the tumor-immune system interaction: Steps toward understanding therapy. Physica A: Statistical Mechanics and its Applications **386** (2007) 713–719
3. Sarkar, R., Banerjee, S.: Cancer self remission and tumor stability–a stochastic approach. Mathematical biosciences **196** (2005) 65–81
4. Kuang, Y., Nagy, J., Elser, J.: Biological stoichiometry of tumor dynamics: mathematical models and analysis. Discrete and Continuous Dynamical Systems Series B **4** (2004) 221–240
5. El-Gohary, A.: Chaos and optimal control of equilibrium states of tumor system with drug. Chaos, Solitons & Fractals **41** (2009) 425–435
6. El-Gohary, A.: Chaos and optimal control of cancer self-remission and tumor system steady states. Chaos, Solitons & Fractals **37** (2008) 1305–1316
7. El-Gohary, A., Alwasel, I.: The chaos and optimal control of cancer model with complete unknown parameters. Chaos, Solitons & Fractals **42** (2009) 2865–2874
8. ALEXANDER, P., KONSTANTIN, E.: Localization of compact invariant sets of nonlinear systems with applications to the lanford system. International Journal of Bifurcation and Chaos **16** (2006) 3249–3256
9. Krishchenko, A., Starkov, K.: Localization of compact invariant sets of the lorenz system. Physics Letters A **353** (2006) 383–388
10. Krishchenko, A., Starkov, K.: Estimation of the domain containing all compact invariant sets of a system modelling the amplitude of a plasma instability. Physics Letters A **367** (2007) 65–72

On synchronization for Sprott systems

Ramon Ramirez-Villalobos[1], Luis N. Coria[1], Luis T. Aguilar[2], and Konstantin E. Starkov[2]

[1] Instituto Tecnologico de Tijuana
Blvd. Industrial y Av. ITR Tijuana S/N, 22500, Mesa Otay, Tijuana, B.C., Mexico
[2] Centro de Investigacion y Desarrollo de Tecnologia Digital CITEDI-IPN
Av. Instituto Politecnico Nacional 1310, Mesa de Otay, 22510 Tijuana, B.C., Mexico
{ramon.rmzv,luis.coria}@gmail.com,{laguilar,konst}@citedi.mx
Paper received on 22/09/12, Accepted on 22/10/12.

Abstract. In the paper we shown the results of three non-linear observers that solve the synchronization problem for a system proposed by Munmuangsaen et al. (2011). Each of these observers has particular properties. First, a non-linear observer with linear output injection is presented, which guarantees asymptotic stability. With this observer we can affect the damping of transient response of error convergence. Second, we present a Thau observer. The synchronization error of this observer is bounded. Finally, we present a sliding-mode observer. This observer has the property to bring the error dynamics to zero in finite-time. We perform two robustness tests in order to know the behaviour of each proposed observer under variations of parameters and disturbances. We provide results of numerical simulations as an illustration of error dynamics and convergence of trajectories of GS11 system-observer synchronization.
Keywords: Chaos synchronization, linear injection, observer, Thau observer, sliding-modes.

1 Introduction

Chaos synchronization phenomenon in a master-slave formalism has been intensively studied in the literature. Several synchronization methods using chaotic systems have been developed since the works by Pecora and Carol [1] in 1990, and by Cuomo et al. [2] in 1993. They constructed a master-slave scheme where the slave is a modified copy of the master system. Since the work by Nijmeijer and Mareels [3] the synchronization problem can be viewed from a perspective of control theory and solved employing non-linear observers. The idea is to use the output vector of the master system to control the slave system. Then we can synchronize the master system and slave system, this method turns out to be one of the most efficient for chaos synchronization [4],[5].

We can found some papers devoted to study the chaotic systems reported by Sprott [11] in the sense of solution of synchronization problem. The Sprott systems present a particular structure and have a practical application in secure communications. Since they can be implemented by simple circuits whose properties can be predicted and controlled with very high accuracy. Xiao-Dong et al. [12] present a terminal sliding-mode controller to realize finite-time synchronization of Sprott circuits. This controller renders the closed loop system robust with respect to bounded disturbances and parametric

Mireya García-Vázquez, Grigori Sidorov (Eds.).
Advances in Computing Science and Control.
Research in Computing Science 59, 2012, pp. 23- 34

variations. A PID controller was developed via the EP (Evolutionary Programming) algorithm by Hsin-Chieh et al. [13], the performance criterion utilized is the IAE (Integrated Absolute Error). Using the EP algorithm, the optimal control gains of PID controller are derived such that the performance index of IAE is as minimal as possible. Bai et al. [14] described a straightforward technique that consists in adding control variables in the observer in order to obtain a linear error dynamics.

In the literature, we can find different types of observers used to synchronize chaotic systems and they are applied to secure communication systems. As an example, Wang and Ge [6] present an adaptive observer via backstepping design. Teh-Lu et al. [7] present an adaptive observer, it is constructed to synchronize the master system whose dynamics are subjected to the system's disturbances and/or some unknown parameters. Perruquetti et al. [5] present a sliding-mode observer for non-linear systems that can be transformed in a linear canonical form up to an output injection. Thau observer [8] is a model-based observer which reconstructs state variables of a class of nonlinear systems [9] An alternative of Thau observer is presented by Starkov et al. [10] and its extension for the case of systems possessing a positively invariant domain. Each observer has specific properties, it which are useful for secure communication systems based in a master-slave scheme. An adaptive observer estimates state variables and unknown parameters of the master system. Moreover, the sliding-mode observer has the property to bring the error dynamics to zero in finite-time.

Recently, Munmuangsaen et al. [15] present a jerk function, exhibiting a chaotic behaviour for different non-linear functions. This model is given by the following expression:

$$\dddot{x} + \ddot{x} + x = f(\dot{x}); \tag{1}$$

where $f(\dot{x})$ is the non-linear function required for chaos. To find chaotic solutions, Munmuangsaen et al. [15] employed a numerical search procedure and have been found twelve non-linear functions (labeled by 'GS1' to 'GS12'), listed in Table 1.

Table 1. Functions $f(\dot{x})$ that produces chaos.

Case	$f(\dot{x})$
GS1	$\pm 0.1\exp(\mp\dot{x})$
GS2	$\pm\exp(\mp\dot{x}-2)$
GS3	$\pm 5.1\cos(\pm\dot{x}+0.1)$
GS4	$\pm 0.2\tan(\mp\dot{x})$
GS5	$\pm sign(1\mp 4\dot{x})$
GS6	$\pm\dot{x}^2 - 0.2\dot{x}^3$
GS7	$\pm\frac{1}{(\dot{x}\pm 2)^2}$
GS8	$-5\dot{x}\pm\lvert 1\pm 5\dot{x}\rvert$
GS9	$\pm\frac{0.4}{\lvert\pm\dot{x}+1\rvert}$
GS10	$\pm\frac{1}{\lvert\pm\dot{x}+1\rvert^{0.5}}$
GS11	$\pm 4sen(\pm\dot{x}+1)-\dot{x}$
GS12	$\pm\cosh(\dot{x})-0.6\dot{x}$

The purpose of our paper is to present results of three non-linear observers design for GS systems from [15]. According with Moon [16] for a dynamical system to display chaotic behaviour it has to be either nonlinear. Commonly, the nonlinearity in chaotic systems are polynomial terms. We choose the GS systems because they present a different structure than classical chaotic systems. The GS systems have four linear terms (or five terms in two cases) and one non-linear term. The nonlinearity of GS system can be polynomial, exponential, trigonometric or rational function. Also, this systems are easy to construct with electronic components and scaling over a wide range of frequencies [11],[17],[18]. For all the above features, these system may be applied in future researches on the development of secure communication systems.

The paper is organized as follows. A non-linear observer with linear output injection is presented in Section 2. With this observer we can affect the damping of transient response, choosing an appropriate value of gain matrix. This observer guarantees asymptotic stability for all systems proposed by Munmuangsaen et al. (2011). In Section 3 we present a Thau observer design for GS11 system using inequality from [10]. Further, in Section 4, we present a sliding-mode observer for GS11 system which have the property to bring the error estimation to zero in finite-time. Conclusions are presented in Section 5.

2 Non-linear observer with linear output injection for GS systems

Observers with linear output injection can be used in dynamic systems whose nonlinear part depends only upon the output, see [3]. The error of this observer has a linear dynamic and the convergence time of this observer can be tuned by the proper choice of the eigenvalues of the gain matrix.

Firstly, is necessary to present (1) in a state-space representation. Choosing state variables $x_1 = x$; $x_2 = \dot{x}$; $x_3 = \ddot{x}$; Eq. (1) can be expressed by the following state-space form:

$$\begin{aligned} \dot{x}_1 &= x_2; \\ \dot{x}_2 &= x_3; \\ \dot{x}_3 &= f(x_2) - x_1 - x_3. \end{aligned} \tag{2}$$

Since the non-linear term in (2) depends on x_2 then we choose $y = x_2$ as output vector in order to design the non-linear observer with linear output injection.

Also, the nonlinearity is not a smooth function in some cases, but this does not affect the design; this is because a dissipative chaotic flow has the property that trajectories are attached to a bounded region of state-space by strange attractor and its divergence is negative [19],[20]. Since (15) is dissipative with constant state-space contraction of -1. Then $|f(x_2)| \leq \gamma |x_2|$ form some constant $\gamma > 0$. That means the solutions for all initial conditions are well defined on $\Re^3$.

The non-linear observer with linear output injection for Eq. (2) proposed for GS systems is described by the following equations:

$$\begin{aligned}\dot{\hat{x}}_1 &= \hat{x}_2 + k_1(y - \hat{x}_2);\\ \dot{\hat{x}}_2 &= \hat{x}_3 + k_2(y - \hat{x}_2);\\ \dot{\hat{x}}_3 &= f(x_2) - \hat{x}_1 - \hat{x}_3 - k_3(y - \hat{x}_2);\end{aligned} \tag{3}$$

with $k_1; k_2; k_3 \in \Re$.

The observer error is defined as $e = x - \hat{x}$. Dynamics of the observer error is given by following equation:

$$\dot{e} = \begin{bmatrix} 0 & 1-k_1 & 0 \\ 0 & -k_2 & 1 \\ -1 & -k_3 & -1 \end{bmatrix} e. \tag{4}$$

In order to guarantee than trajectories of estimated states converge to real state,we found a gain matrix $K = [k_1, k_2, k_3]$ such that error dynamics is asymptotically stable.

Choosing $\mu_i = -1, i = 1, 2, 3$, as desired poles in the left half of the complex plane, the gain matrix is $K = [-2.375, 3.5, 3.25]$.

Numerical simulation of synchronization between non-linear observer with linear output injection and GS system taking $f(x_2) = 4\sin(x_2 + 1) - 2.2x_2$, are shown in Figs. 1 and 2. The initial conditions of the two systems are different, the initial condition of (1) is $x_i(0) = 0$ $(i = 1, 2, 3)$, the initial condition of (3) is $\hat{x}_i(0)$ $(i = 1, 2, 3)$.

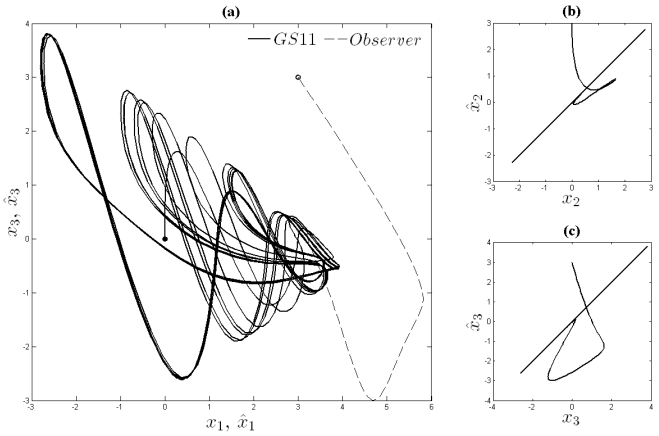


Fig. 1. Synchronization of GS systems:(a) Comparison of dynamics of GS system and non-linear observer with liner output injection, (b-c) Synchronization between real states and estimated states.

We can see that trajectories of non-linear observer with linear output injection converge to chaotic attractor of GS system (Fig. 1(a)) and that estimated states are synchronized with real states (Fig. 1(b-c)). The error dynamics of master-slave synchronization

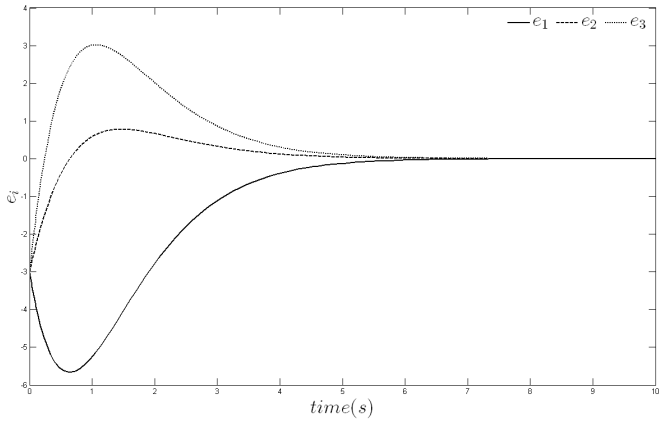


Fig. 2. Error dynamics of master-slave synchronization of non-linear observer with linaer output injection and GS system considering $f(x_2) = 4\sin(x_2 + 1) - 2.2x_2$.

converge to zero in approximately 8 seconds, as it is shown in Fig. 2. If we desire to speed up the convergence time is necessary to choose an appropriate gain matrix.

3 Thau observer design for GS11 system

In this section we present a Thau observer design for GS11 system, where the non-linear functions is defined as $f(x_2) = b\sin(x_2 + 1) - cx_2$. We consider the case GS11 because is a Lipschitz function.

Thau observer [8], has been constructed for a state estimation of non-linear systems. This observer relates the linear part and non-linear part of a dynamic system. The gain of Thau observer can be obtained if the synchronization error converges to a bounded zone [9].

Before to start the Thau observer design, is necessary to make a change of variable in system (2) in order to shift the equilibrium point to the origin of state-space and satisfy $g(0) = 0$. With $\bar{x}_1 = x_1 - x_1^*$ the system (2) can be rewritten as:

$$\dot{x} = Ax + g(x) = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -1 & -c & -1 \end{bmatrix} x + \begin{bmatrix} 0 \\ 0 \\ b\sin(x_2 + 1) \end{bmatrix}; \tag{5}$$

with $x = [\bar{x}_1, x_2, x_3]$; $b = 4$ and $c = 2.2$.

Computing the Jacobi matrix of $g(x)$ we can obtain the Lipschitz constant ℓ, as result, we get:

$$\ell \leq ||g'(x)|| \leq b = 4. \tag{6}$$

We chose the output vector $y = [x_1, x_2]^T$. In this case the matrix $A_0 = A - KC$ of Thau observer is given by:

$$A_0 = \begin{bmatrix} -k_1 & 1-k_4 & 0 \\ -k_2 & -k_5 & 1 \\ -1-k_3 & -c-k_6 & -1 \end{bmatrix}. \tag{7}$$

We assign parameters of the feedback matrix as $k_3 = -1$; $k_4 = k_2 + 1$; $k_6 = -(c+1)$. The Thau observer will be asymptotically stable if A_0 is symmetric and satisfies [10]

$$\frac{||A_0||^{n-1}}{|det(A_0)|} \leq \frac{2}{\ell}. \tag{8}$$

Choosing $k_1 = 15$; $k_2 = 0$; $k_5 = 10$; and using (6) the inequality (8) can be satisfied. Hence the Thau observer is constructed:

$$\begin{aligned} \dot{\hat{\bar{x}}}_1 &= \hat{x}_2 + k_1(\bar{x}_1 - \hat{\bar{x}}_1) + k_4(x_2 - \hat{x}_2); \\ \dot{\hat{x}}_2 &= \hat{x}_3 + k_2(\bar{x}_1 - \hat{\bar{x}}_1) + k_5(x_2 - \hat{x}_2); \\ \dot{\hat{x}}_3 &= \pm b\sin(\pm\hat{x}_2 + 1) - \bar{x}_1 - c\hat{x}_2 - \hat{x}_3 \\ &\quad + k_3(\bar{x}_1 - \hat{\bar{x}}_1) + k_6(x_2 - \hat{x}_2). \end{aligned} \tag{9}$$

Numerical simulations have realized in order to show the synchronization of Thau observer and GS11 system, see Fig. 3 and 4.
The initial conditions of (5) is $[x_1(0), x_2(0), x_3(0)] = [-3, 0, 0]$ and the initial condition of Thau observer is $[\hat{x}_1(0), \hat{x}_2(0), \hat{x}_3(0)] = [0, 3, 3]$.

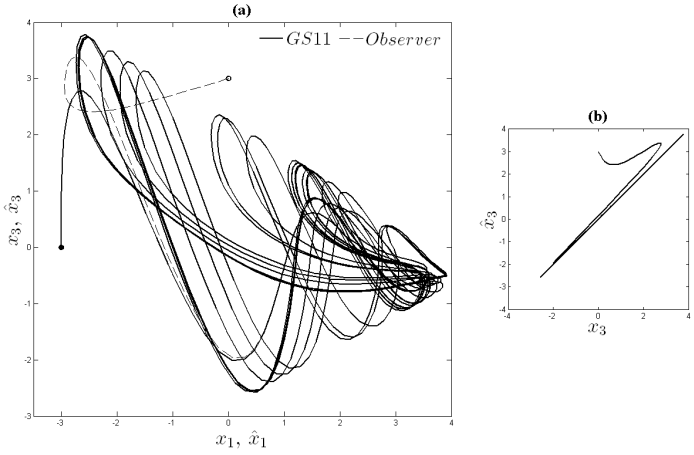


Fig. 3. Synchronization of GS systems:(a) Comparison of dynamics of GS system and Thau observer, (b) Synchronization between real states and estimated states.

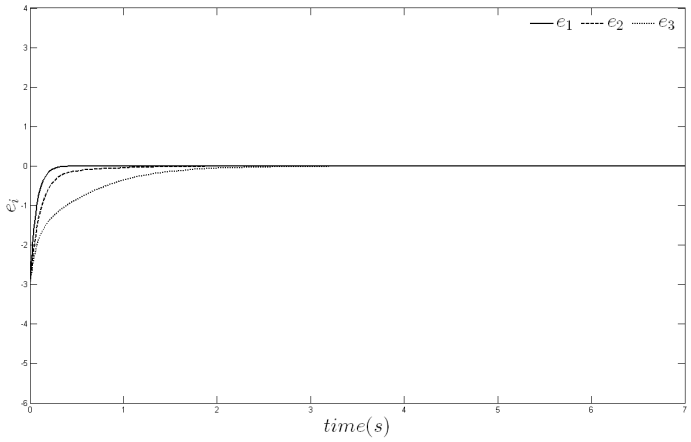


Fig. 4. Error dynamics of master-slave synchronization of Thau observer and GS11 system.

We can see in Fig. 3(a) that dynamic of (9) converge to chaotic attractor of (5) and in Fig. 1(b) that $\hat{x}_3$ is synchronized with x_3. Fig. 4. The error dynamics of master-slave synchronization converge to zero in approximately 4 seconds. This observer ensures asymptotic stability, but if we desire to speed up the convergence time is necessary to use some auxiliary technique.

4 Sliding-mode observer design for GS11 Sprott system

Sliding-mode observer is useful for many reasons: we have reduced observation error dynamics, for a finite-time convergence for all the observables states, to design under some conditions an observer for non-smooth system and robustness under parameter variations [21].

We propose the following observer for GS11 system:

$$\begin{aligned}
\dot{\hat{x}}_1 &= \hat{x}_2 + L_1 sign(y - \hat{x}_1);\\
\dot{\hat{x}}_2 &= \hat{x}_3 + L_2 sign(y - \hat{x}_1);\\
\dot{\hat{x}}_3 &= \pm b \sin(\pm \hat{x}_2 + 1) - \hat{x}_1 - c\hat{x}_2 - \hat{x}_3 - L_3 sign(y - \hat{x}_1);\\
y &= x_1;
\end{aligned} \tag{10}$$

with $L_1; L_2; L_3 > 0$.

The observer error is defined as $e = x - \hat{x}$. Dynamics of observer error is described by the following equation:

$$\begin{aligned}
\dot{e}_1 &= e_2 - L_1 sign(e_1);\\
\dot{e}_2 &= e_3 - L_2 sign(e_1);\\
\dot{e}_3 &= b \sin\theta \mp b \sin\hat{\theta} - e_1 - ce_2 - e_3 - L_3 sign(e_1);
\end{aligned} \tag{11}$$

where $\theta = x_2 + 1$; $\hat{\theta} = \hat{x}_2 + 1$.

In order to verify that trajectories of estimated states converge to real states, the following candidate Lyapunov function is proposed:

$$V(e) = \frac{1}{2} e^T A e + [aL_1 + (ac+1)L_2 + L_3]\,|e_1| > 0 \tag{12}$$

where:

$$A = \begin{bmatrix} 1 & a & 0 \\ a & ac+1 & 1 \\ 0 & 1 & a \end{bmatrix}; \quad 1 < a < c; \quad e = \begin{bmatrix} e_1 \\ e_2 \\ e_3 \end{bmatrix}. \tag{13}$$

The time derivative along the trajectory of the system is:

$$\begin{aligned} \dot{V}(e) = &-(L_1 + aL_2)|e_1| - (c-a)e_2^2 - (a-1)e_3^2 - \beta L_1 \\ &+b(e_2 + ae_3)(\sin\theta - \sin\hat{\theta}) - (L_2 + aL_3)sign(e_1)e_3; \end{aligned} \tag{14}$$

with $\beta = aL_1 + (ac+1)L_2 + L_3$.

Since $sin(x)$ and $sign(x)$ are bounded functions, then Eq. (27) can be expressed as:

$$\begin{aligned} \dot{V}(e) \leq &-(L_1 + aL_2)|e_1| - (c-a)e_2^2 - (a-1)e_3^2 - \beta L_1 \\ &+2b|e_2| + (2ab + L_2 + aL_3)|e_3| \end{aligned} \tag{15}$$

Using the norm properties $-x^T Q x < -\lambda_{\min}\{Q\}||x||_2^2$ and $||x||_1 \leq \sqrt{n}||x||_2$, $\dot{V}(e)$ can be written:

$$\begin{aligned} \dot{V}(e) \leq &-(L_1 + aL_2)|e_1| - \min\{(c-a),(a-1)\}||e_x||_2^2 \\ &+\sqrt{2}\max\{2b, 2ab + L_2 + aL_3\}||e_x||_2 - \beta L_1 \end{aligned} \tag{16}$$

where $e_x = [|e_2|, |e_3|]$. If its possible to satisfy the following condition

$$\beta L_1 > \sqrt{2}\max\{2b, 2ab + L_2 + aL_3\}||e_x||_2; \tag{17}$$

we can guarantee local asymptotic stability. Finally, by applying Theorem 38 from [21] can be concluded that error dynamics converges to zero in finite-time.

We have realized numerical simulations in order to show the convergence of proposed sliding-mode to GS11 system as shwon Figs. 5 and 6. The initial condition of GS11 system is $[x_1(0), x_2(0), x_3(0)] = [0, 0, 0]$, the initial condition of proposed sliding-mode observer is $\hat{x}_1(0) = 3$ $(i = 1, 2, 3)$ and $L_1 = 4$, $L_2 = 10$, $L_3 = 10$.

Fig. 5(a) shows that trajectories of sliding-mode observer converge to chaotic attractor of GS11 system and Fig. 5(b-c) show the synchronization between real states and estimated states. We can see Fig. 6 shows that error dynamics of master-slave synchronization converge to zero in approximately 8 seconds. This observer may only guarantee convergence in infinite time.

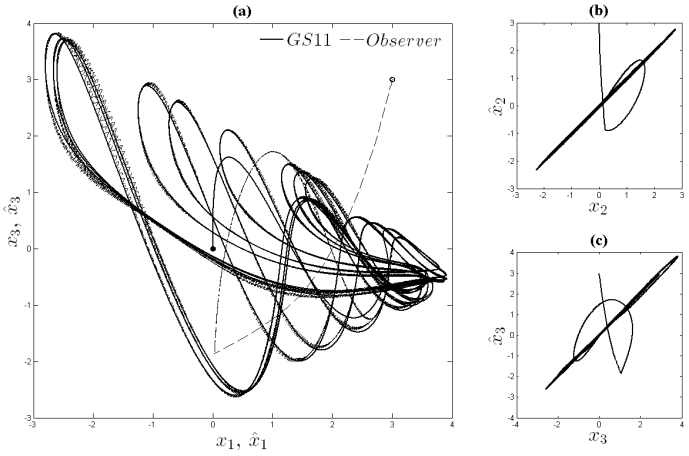


Fig. 5. Synchronization of GS systems: (a) Comparison of dynamics of GS system and sliding-mode observer, (b-c) Synchronization between real states and estimated states.

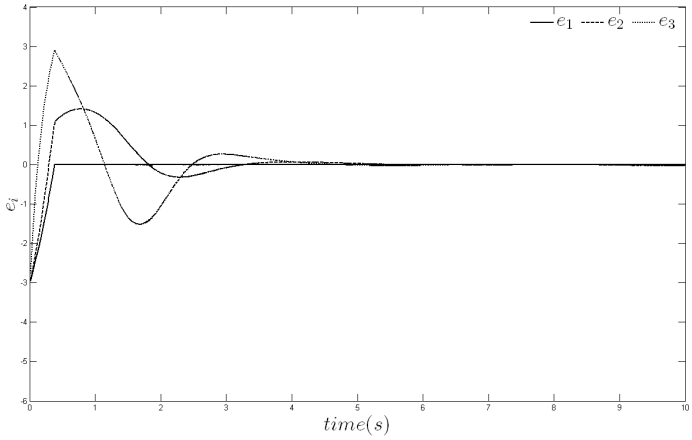


Fig. 6. Error dynamics of master-slave synchronization of Sliding-mode observer and GS11 system.

5 Robustness testing

Robustness test have realized in order to know the behaviour of proposed observer under variations of parameter and disturbances of the coupling signal. We made variations of parameter c of non-linear function of each observer and added a uniformly distributed noise (bound for both sides) in the coupling signal. The numerical simulations were carried out from the same initial conditions used to show the synchronization of observers and GS11 system in the previous sections.

5.1 Variations of parameter

We made variations of parameter c of non-linear function of each observer. The parameter variation performed was $c + \varepsilon$. The Table 2 shows the synchronization error rate of each observer, considering $\varepsilon = 0.1$ and $\varepsilon = 0.5$.

Table 2. Test of variation of parameter c.

	$\varepsilon = 0.1$	$\varepsilon = 0.5$
Linear injection	5.86%	13.11%
Thau observer	13.48%	29.77%
Slindig-mode observer	12.96%	27.97%

In this test we can not make variations of parameter b of non-linear term $f(x_2)$, this introduce a scaling factor of amplitude that implies a modification of the structure of the master-slave scheme.

We provide a discussion of results in the conclusions of this paper.

5.2 Disturbances of the coupling signal

The second test was to introduce a uniformly distributed noise δ (bound for both sides) in the coupling signal, in order to simulate disturbances in the communication channel. The Table 3 shows the synchronization error rate of each observer, considering $|\delta| \leq 0.5$ and $|\delta| \leq 1$.

Table 3. Perturbation test.

	$\|\delta\| \leq \|0.5\|$	$\|\delta\| \leq \|1\|$
Linear injection	27.89%	54.81%
Thau observer	6.12%	8.03%
Slindig-mode observer	5.39%	9.01%

A discussion about the results of this test are provided in the conclusion of this paper.

6 Conclusion

In this paper we have presented a three non-linear observers design in order to solve the chaos synchronization problem of some chaotic systems. Also robustness test have realized in order to know the behaviour under variation of parameters and disturbances in the coupling signal.

We designed a non-linear observer with linear output injection that guarantees asymptotic stability for all GS systems (GS1-GS12). This observer measures only one state variable. If its chosen an appropriate gain matrix we can modify transient response characteristics of the error convergence, e.g., damping, peak time, settling time. Also, we can use different non-linear terms listed in Table 1, this does not affect the observer design. In addition, a Thau observer is designed for GS11 system. This observer measures two state variables. To construct this observer is required to choose an appropriate feedback matrix K such that matrix A_0 is a stable and symmetric matrix and know the Lipschitz constant. Further, we designed a sliding-mode observer for GS11 system. This observer measures only one state variable. We can guarantee local asymptotic stability if a condition is satisfied and conclude that finite-time convergence of error dynamics by applying a theorem.

From the observer designs and robustness testing we notice that each proposed observer has benefits if they are applied in a scheme of secure communication system. The non-linear observer with output linear injection is a good alternative to implement in a scheme of secure communications systems with low noise levels. Presents robustness to variations of parameter. Also, we can modify the transient response characteristics of error convergence, according to requirements of communication channel and electronic circuit specifications of the system implementation. The Thau observer and sliding-mode observer are a good alternative to schemes of secure communications systems with disturbances in communication channel. Both observer have strong robustness to disturbances in communication channel. But, the measurement of two state variables of Thau observer may be a disadvantage. And the implementation of sliding-mode observer presents chattering phenomenon, then is necessary to find a method to reduce or eliminate it.

We consider as future work to implement the non-linear observer with linear output injection in some chaotic masking application with low noise levels, considering all cases of non-linear functions. Also, to implement the sliding-mode observer in a scheme of secure communication system with disturbances in communication channel.

References

1. Pecora, L.M., Caroll, T.L.: Synchronization in chaotic systems. Physical Review Letters. 64, 821–824 (1990)
2. Cuomo, K.M., Oppenheim, A.V., Strogatz, S.H.: Circuit implementation of synchronized chaos with applications to communications. Physical Review Letters. 71, 65–68 (1993)
3. Nijmeijer, H., Mareels, I.: An observer looks at synchronization. IEEE Transactions on circuits and systems-I: Fundamentar theory and applications. 44(10), 882–890 (1997).
4. Jiang, G.P., Tang, W.K.S., Chen, G.: A simple global synchronization criterion for coupled systems. Chaos, Solitons and Fractals. 15, 925–935 (2003).
5. Perruquetti, W., Floquet, T., Moulay, E.: Finite time observers: application to secure communication. IEEE Transactions on Automatic Control. 53(1), 356–360 (2008).
6. Wang, C., Ge, S.S.: Adaptive synchronization of uncertain chaotic systems via backstepping design. Chaos, Solitons and Fractals. 12, 1199–1206 (2001).
7. Teh-Lu, L., Shin-Hwa, T., Aguilar, L.T.: Adaptive synchronization of chaotic systems and its application to secure communications. Chaos, Solitons and Fractals. 11(9), 1387–1396 (2000).

8. Thau, F.E.: Observing the states of nonlinear dynamic system. International Journal of Control. 17, 471–479 (1973).
9. Martinez-Guerra, R., Yu., W., Cisneros-Saldaña, E.: A nre model-free sliding observer to synchronization problem. Chaos, Solitons and Fractals. 36, 1141–1156 (2008).
10. Author: On synchronization of chaotic systems based on the Thau observer design. Communications in Nonlinear Science and Numerical Simulation. 17, 17–28 (2012).
11. Sprott, J.C.: A new class of chaotic circuit. Physics Letters A. 266(1), 19–23 (2000).
12. Xiao-Dong, X., Wan-li, Y., Su-Wen, Z.: Finite time synchronization of Sprott circuits with uncertain parameters. International Conference on Advanced Computer Control. 693–696 (2009).
13. Hsin-Chieh, C., Jen-Fuh, C., Jun-Juh, Y., Teh-Lu, L.: EP-based PID control design for chaotic synchronization with application in secure communication. Expert Systems with Applications. 34(2), 1169–1177 (2008).
14. Bai, E., Lonngren, E., Sprott, J.C.: On the synchronization of a class of electronic circuits that exhibit chaos. Chaos, Solitons and Fractals. 13(7), 1515–1521 (2002).
15. Munmuangsaen, B., Srisuchinwong, B., Sprott, J.C.: Generalization of simplest autonomous chaotic system. Physics Letters A. 375, 1445–1453 (2011).
16. Moon, F.: Chaotic and Fractal Dynamics. Springer-Verlag New York, LLC. (1990).
17. Sprott, J.C.: Simple chaotic systems and circuits. American Journal of Physics. 68(8), 758–763 (2009).
18. Sprott, J.C.: A new chaotic jerk. IEEE Transactions on Circuits and Systems-II: Express Briefs. 58(4), 240–243 (2011).
19. Barboza, R.: Dynamics of a hyperchaotic Lorenz system. International Journal of Bifurcations and Chaos. 17(12), 4285–4294 (2007).
20. Munmuangsaen, B., Srisuchinwong, B.: Elementary chaotic snap flows. Chaos, Solitons and Fractals. 44, 995–1003 (2011).
21. Perruquetti, W., Barbot, J.: Sliding mode control in engineering. Marcel Dekkerm, Inc. (2002)

Bounding the domain containing all compact invariant sets of a AIDS model related cancer

C. Plata-Ante, K.E. Starkov, L.N. Coria*

Centro de Investigación y Desarrollo de Tecnología Digital, CITEDI-IPN
*Instituto Tecnológico de Tijuana
Tijuana, B.C.
cplata@citedi.mx, konst@citedi.mx, *luis.coria@gmail.com
Paper received on 21/09/12, Accepted on 23/10/12.

Abstract. The suffering of two diseases like AIDS and cancer is a deadly danger to the life of any person. In this paper we analyzed a model which describes the dynamics between the immune system of four nonlinear differential equations where each state represents cell populations: cancer cells, healthy cells T-CD4+, infected cells T-CD4+ and HIV-1 free virus. The final localization domain is obtained by the method of Localization of Compact Invariant Sets, which consist in first order extreme conditions, and the iterative theorem. We proposed some linear and rationals localizing functions which intersections enclosed all compact invariant sets of the system under study. Finally, numerical simulations are presented in order to show the localization results.

Keywords: Cancer, HIV/AIDS, compact invariant sets, biological system.

1 Introduction

The mathematical models have been used years ago to analyze some physical models like biological models. Have been modeling different biological systems to observe its behavior, since models that represents a predator-prey system [1], viral infection [2], different types of cancer as brain tumours [3], bladder cancer [4], diabetes [5], hematologic disorders (leukemia) [6] and models that help in disease with some parameters of control through some treatment, for example in tumor immunotherapy [7]; or models showing the immune system response against some disease. Through this type of model has been observed the dynamic between the Human Immunodeficiency Virus (HIV) and the immune response [8], [9]. According to the number of deaths, some of these diseases represent a serious threat, being the main heart disease, cancer, respiratory diseases, among others [10]. The HIV/AIDS is one of the main causes of death in the world due it generate a serious deterioration of the immune system. When the count of cells T-CD4+ reaches a low amount, the body goes to the next level of infection is AIDS, which is the final stage of HIV where the body can contract various infections and diseases, including cancer. Cancer is characterized by show uncontrolled growth of abnormal cells leading to a tumor, which is one cause of death in AIDS patients.

Given the impact of these diseases on society, in this paper we analyze a mathematical model wich describes an interation of HIV/AIDS and immune system by the

Mireya García-Vázquez, Grigori Sidorov (Eds.).
Advances in Computing Science and Control.
Research in Computing Science 59, 2012, pp. 35- 44

method called Localization of Compact Invariant Sets. This model has been of interest to some researchers, e.g. in [11], where authors analyze the system varying parameters. The results of these analyze is show a deep similarity the clinical observations from different types of cancer in patients infected with HIV-1. However, to our knowledge there are no reported for the system 1. Our analysis is important because it allow to know the region where are located different dynamics in the system.

The following model was developed by Lou *et al*. [12] based on the model of Perelson *et al.* [13], that describes the dynamics between immune system cells and the HIV virus. This system considered two routes of spread for HIV *in vivo*, besides entering a cell as a free infected particle, can be also be passed during cell-cell contact. Additionally to these two variables, immune system and HIV, in the next system are considered cancer cells in their analysis. About cancer, these cells proliferate on different way as normal cells. Finally the immune system can recognize between cancer cell and normal cells to play its role, protect. The model is represented by the following differential equations

$$\begin{aligned}
\frac{dC(t)}{dt} &= r_1 C(t)\left(1 - \frac{C(t)+T(t)+I(t)}{m}\right) - k_1 C(t)T(t); \\
\frac{dT(t)}{dt} &= s + r_2 T(t)\left(1 - \frac{C(t)+T(t)+I(t)}{m}\right) - \mu_T T(t) \\
&\quad -k_2 T(t)V(t) - k_3 T(t)I(t); \\
\frac{dI(t)}{dt} &= k_2 T(t)V(t) + k_3 T(t)I(t) - \mu_I I(t); \\
\frac{dV(t)}{dt} &= q\mu_I I(t) - \delta V(t);
\end{aligned} \tag{1}$$

where $C(t)$ represents the concentration of cancer cells, $T(t)$ is the concentration healthy cells, $I(t)$ is the concentration infected cells and $V(t)$ is the concentration of virus HIV-1 free.

According to the literature, the number of healthy cells that become to cancerous cells is very small compared with the uncontrolled proliferation of cancer cells [14]. With respect to T cells, if they encounter some specific antigens can stimulate its growth. However, even when these cells are stimulated or highly proliferated, the total number of T cells in the body remains bounded. The term $k_2 T(t)V(t)$ in (1) represents the rate at which the free virus infects T-CD4+ cells. Given to the biological characteristics of the system, parameters and states are positive. For the development of this analysis, nomenclature of each state was replaced by x_1, x_2, x_3, y x_4, respectively. The document is organized as follows: in the next section is shown some mathematical notations and concepts for the analysis of localization of compact invariant sets. After that, we propose some localizing functions and show in numerical simulations. Finally, we present the conclusion of this research.

2 Some mathematical preliminaries and notations

The localization method is used to define the region of the state space where are located all compact invariant sets, this analysis is useful for understanding the dynamics of the system. In order to find a localization we used a general method described in [15], [16] y [17]. The useful results from these papers are described below

Considering the following nonlinear system as:

$$\dot{x} = f(x) \tag{2}$$

where f is a C^∞ -differentiable vector field-, here $x \in \mathbf{R}^n$ is the -state space vector-. Let $h(x)$ be a C^∞ -differentiable function such that h is not the first integral of (2)-. By $h|_B$ we denote the restriction of h on a set $B \subset \mathbf{R}^n$. By $S(h)$ we denote the set $\{x \in \mathbf{R}^n \mid L_f h(x) = 0\}$ where $L_f h$ of (2) is given by $L_f h(x) = \frac{\partial h}{\partial x} f(x)$. Now we define $h_{\inf} = \inf\{h(x)|x \in S(h)\}$ and $h_{\sup} = \sup\{h(x)|x \in S(h)\}$ that will be used in this analysis.

Theorem 1. Each compact invariant set Γ of (2) is contained in the localization set:

$$K(h) = \{h_{\inf} \leq h(x) \leq h_{\sup}\}.$$

The function h applied here is called localizing. It is evident that if all compact invariant sets are located in sets Q_1 and Q_2, with $Q_1; Q_2 \subset \mathbf{R}^n$, then they are located in the set $Q_1 \cap Q_2$ as well.

The iterative theorem is described as shown below:

Theorem 2. Let $h_m(x), m = 0, 1, 2, \ldots$ be a sequence of functions C^∞ ($\mathbf{R}^n$). The sets:

$$K_0 = K(h_0), \quad K_m = K_{m-1} \cap K_{m-1,m}, \quad m > 0,$$

with

$$\begin{aligned} K_{m-1,m} &= \{x : h_{m,inf} \leq h_m(x) \leq h_{m,sup}\}, \\ h_{m,sup} &= \sup_{S(h_m) \cap K_{m-1}} h_m(x), \\ h_{m,inf} &= \inf_{S(h_m) \cap K_{m-1}} h_m(x), \end{aligned}$$

contain all compact invariant sets of the system (2) and

$$K_0 \supseteq K_1 \supseteq \cdots \supseteq K_m \supseteq \ldots .$$

Due the biological and mathematical sense of each state that compose the system (1), the compact invariant sets will be present in R^4_+, and any localization outside of this region will not be useful for this analysis. In addition, system parameters are positive. Also, for the sake of simplicity of notations $K(h) = K(h) \cap \mathbb{R}^4_+$ and $S(h) = S(h) \cap \mathbb{R}^4_+$.

3 Localizations analysis of the AIDS model related cancer

In order to find the localization domain, we proposed different localizing functions among linear and rational. From our expertise of the localization analysis some special functions can be applied but no better results are found; especially with quadratic and high order functions, we obtain trivial results respecting of our previous calculations. The goal of proposing these functions is to implement the method developed by Dr. Krishchenko and Dr. Starkov, according to the analysis developed shows an advantage over other methods because the proposal does not restrict the use of functions and iterative method improves the bounds.

3.1 Localization results by linear functions

In this section we present different linear functions delimiting the domain of each of state variables.

1. We proposed the following function

$$h_1 = x_1; \tag{3}$$

and its $L_f h_1$ and from this is determinate $S(h_1)$,

$$S(h_1) \subset \left\{ \frac{r_1}{m} h_1|_{S(h_1)} = r_1 - \frac{x_2 + x_3}{m} - k_1 x_2 \right\}.$$

Then we obtained $h_1|_{S(h_1)}$ and we get $S(h_1) \subset \left\{ h_1|_{S(h_1)} \leq m \right\}$.
Therefore, all compact invariant sets are contained in

$$K(h_1) = \{ x_1 \leq m \}.$$

2. The second function that we proposed is

$$h_2 = x_2;$$

by determining $L_f h_2$ is obtained $S(h_2)$:

$$S(h_2) = \left\{ (\mu_T - r_2)\, x_2 = -r_2 x_2 \left(\frac{x_1 + x_2 + x_3}{m} \right) - k_2 x_2 x_4 - k_3 x_2 x_3 + s \right\};$$

solving with respect to x_2 to get $h_2|_{S(h_2)}$ and simplifying the expression, we have

$$S(h_2) \subset \left\{ h_2|_{S(h_2)} \leq \frac{s}{\mu_T - r_2} \right\}.$$

Then we can state that if $\mu_T - r_2 > 0$, all compact invariant sets are located in

$$K_1(h_2) = \left\{ x_2 \leq \frac{s}{\mu_T - r_2} \right\}.$$

3. Developing the product situated on the right-expression $S(h_2)$ we have

$$S(h_2) = \left\{ \begin{array}{c} (\mu_T - r_2)\, x_2 = -m^{-1}(r_2 x_2^2 + r_2 x_1 x_2 + r_2 x_2 x_3) \\ -x_2(k_2 x_4 - k_3 x_3) + s \end{array} \right\};$$

and we rewrite respect to x_2

$$S(h_2) \subset \left\{ h_2|_{S(h_2)} \leq \frac{\sqrt{4r_2 sm + (\mu_T - r_2)^2 m^2} - (\mu_T - r_2)m}{2r_2} \right\}.$$

From the later formulae, all compact invariant sets are in the set

$$K_2(h_2) = \left\{ x_2 \leq \frac{\sqrt{4r_2 sm + (\mu_T - r_2)^2 m^2} - (\mu_T - r_2)m}{2r_2} \right\}.$$

4. In order to find a bound of x_3 we proposed the following function

$$h_3 = x_2 + x_3; \tag{4}$$

Calculating the Lie derivative and is obtained $S(h_3)$

$$S(h_3) = \left\{ -\frac{1}{m} r_2 x_2^2 + r_2 x_2 - \mu_T x_2 - \mu_I x_3 - \frac{1}{m} r_2 x_1 x_2 - \frac{1}{m} r_2 x_2 x_3 + s = 0 \right\};$$

solving respect x_3 of the expression to get $h_3|_{S(h_3)}$. Simplifying is rewritten as

$$S(h_3) \subset \left\{ \begin{array}{c} h_3|_{S(h_3)} \leq -\frac{r_2}{\mu_I m}\left(x_2 - \frac{m(\mu_I + r_2 - \mu_T)}{2r_2}\right)^2 \\ + \frac{m(\mu_I + r_2 - \mu_T)^2}{4\mu_I r_2} + \frac{s}{\mu_I} \end{array} \right\}.$$

Is necessary to calculate the $h_{3\,\mathrm{sup}}$ from $S(h_3)$ Then we can ensure that all compact invariant sets are located in the set

$$K(h_3) = \left\{ x_2 + x_3 \leq \frac{m(\mu_I + r_2 - \mu_T)^2}{4\mu_I r_2} + \frac{s}{\mu_I} \right\}. \tag{5}$$

5. Finally within a linear function is proposed $h_4 = x_4$.
Calculating its Lie derivative and obtaining

$$S(h_4) = \left\{ x_4 = \frac{q\mu_I x_3}{\delta} \right\}.$$

Applying the iterative theorem with (5) and using its limit for x_3, we have

$$S(h_4) \cap K(h_3) \subset \left\{ x_4 \leq \frac{q}{\delta}\left[\frac{m(\mu_I + r_2 - \mu_T)^2}{4r_2} + s \right] \right\};$$

then all compact invariant sets are located in

$$K(h_4) = \left\{ x_4 \leq \frac{q}{\delta}\left(\frac{(\mu_I + r_2 - \mu_T)^2 m}{4r_2} + s \right) \right\}. \tag{6}$$

3.2 Localization results by rational functions

After having proposed linear functions, we proposed some rational functions. With this kind of functions we expect a refinement of the localization region.

1. The first proposed rational function is

$$h_5 = \frac{x_4}{x_3}.$$

Calculating the $L_f h_5$ which is subsequently developed to obtain $h_5|_{S(h_5)}$

$$S(h_5) = h_5|_{S(h_5)} = \frac{1}{\delta - \mu_I}\left(q\mu_I - k_2\frac{x_2x_4^2}{x_3^2} - k_3\frac{x_2x_4}{x_3}\right), \quad \delta - \mu_I > 0;$$

we found that the supreme value is given by h_5

$$S(h_5) \subset \left\{h_5|_{S(h_5)} \leq \frac{q\mu_I}{\delta - \mu_I}\right\}.$$

Then as a result we get that If $\delta - \mu_I > 0$, then all invariant compact sets are contained in

$$K(h_5) = \left\{\frac{x_4}{x_3} \leq \frac{q\mu_I}{\delta - \mu_I}\right\}.$$

2. The next function is given by

$$h_6 = \frac{x_2}{x_1}. \tag{7}$$

Calculating the $L_f h_6$ and $S(h_6)$ is describe by

$$S(h_6) \subset \left\{\begin{array}{c} \frac{1}{x_1}\left(s + r_2x_2\left(1 - \dfrac{x_1+x_2+x_3}{m}\right) - \mu_T x_2 - k_2x_2x_4 - k_3x_2x_3\right) \\ -\dfrac{x_2}{x_{1^2}}\left(r_1x_1\left(1 - \dfrac{x_1+x_2+x_3}{m}\right) - k_1x_1x_2\right) = 0 \end{array}\right\};$$

applying the iterative theorem with (6), considering the bound of x_4 is assigned to

$$\gamma := \frac{q}{\delta}\left(\frac{(\mu_I + r_2 - \mu_T)^2 m}{4r_2} + s\right);$$

we get

$$S(h_6) \cap K(h_4) \subset \left\{\begin{array}{c} \left(\dfrac{r_1}{m} + k_1 - \dfrac{r_2}{m}\right)x^2 + (r_2 - \mu_T - r_1 - k_2\gamma)\,x_2 \\ + \left(\dfrac{r_1}{m} - \dfrac{r_2}{m}\right)x_1x_2 + \left(\dfrac{r_1}{m} - \dfrac{r_2}{m} - k_3\right)x_2x_3 + s = 0 \end{array}\right\}. \tag{8}$$

Whether it satisfies that

$$\frac{r_1}{m} + k_1 - \frac{r_2}{m} > 0; \tag{9}$$

the term of x_2 is

$$r_2 - \mu_T - r_1 - k_2\gamma < 0; \tag{10}$$

respecting x_1x_2 is

$$\frac{r_1}{m} - \frac{r_2}{m} > 0; \tag{11}$$

and x_2x_3

$$\frac{r_1}{m} - \frac{r_2}{m} - k_3 > 0; \tag{12}$$

then solving (8) in order to find a bound to x_2, we have

$$S(h_6) \cap K(h_4) \subset \left\{ x_2 \geq \frac{s}{r_1 + \mu_T + k_2\gamma - r_2},\ r_1 + \mu_T + k_2\gamma - r_2 > 0 \right\}.$$

After replacing the bound obtained for x_2 in function and using the iterative theorem (3), we have

$$S(h_6) \cap K(h_4) \cap K(h_1) \subset \left\{ h_6|_{S(h_6)} \geq \frac{s}{(r_1 + \mu_T + k_2\gamma - r_2)\, m} \right\}.$$

Therefore, if the inequalities (9), (10), (11) and (12) then all invariant compact sets are contained in the set

$$K(h_6) = \left\{ \frac{x_2}{x_1} \geq \frac{s}{(r_1 + \mu_T + k_2\gamma - r_2)\, m} \right\}.$$

3.3 Numerical simulations

Based on the functions obtained in the previous section, the Table 3.3 present some localizing functions valid according to certain parameters, in order to show with figures the results obtained. In order to perform the simulations we consider the following values on the parameters (13):

$$\begin{array}{lll} \delta = 3 & k_1 = 3.43 \times 10^{-4} & k_2 = 2.4 \times 10^{-5} \\ k_3 = 2.4 \times 10^{-5} & m = 1500 & \mu_T = 0.02 \\ \mu_I = 0.24 & q = 200 & r_1 = 0.5 \\ r_2 = 0.019 & s = 10 & \end{array} \tag{13}$$

Table 1. Localization result for the system (1).

Localization domain	Condition	Numerical values
$K(h_1) = \{x_1 \leq m\}$	*	$x_1 \leq 1500$
$K_2(h_2) = \left\{ x_2 \leq \frac{\sqrt{4r_2sm+(\mu_T-r_2)^2m^2}-(\mu_T-r_2)m}{2r_2} \right\}$	$\mu_T > r_2$	$x_2 \leq 756.23$
$K(h_3) = \left\{ x_2 + x_3 \leq \frac{m(\mu_I+r_2+\mu_T)^2}{4\mu_I r_2} + \frac{s}{\mu_I} \right\}$	*	$x_2 + x_3 \leq 4730.8$
$K(h_4) = \left\{ x_4 \leq \gamma := \frac{q}{\delta} \left(\frac{(\mu_I+r_2-\mu_T)^2m}{4r_2} + s \right) \right\}$	*	$x_4 \leq 75693$
$K(h_5) = \left\{ \frac{x_4}{x_3} \leq \frac{q\mu_I}{\delta-\mu_I} \right\}$	$\delta > \mu_I$	$\frac{x_4}{x_3} \leq 17.391$
$K(h_6) = \left\{ \frac{x_2}{x_1} \geq \frac{s}{(r_1+\mu_T+k_2\gamma-r_2)m} \right\}$	$r_1 > r_2 > k_3$	$\frac{x_2}{x_1} \geq 2.3 \times 10^{-3}$

The analysis result of the localizing functions is given by the region bounded, for each state variable there is a supreme where is located all compact invariant sets of the system (1).

In the following Figure 1 shows the dynamics of the system along of the results of localization given by $K(h_1) \cap K_2(h_2) \cap K(h_3) \cap K(h_4) \cap K(h_5) \cap K(h_6)$.

The analysis result of the localizing functions is given by the region bounded, for each state variable there is a supreme where is located all compact invariant sets of the system (1).

In the following Figure 1 shows the dynamics of the system along of the results of localization given by $K(h_1) \cap K_2(h_2) \cap K(h_3) \cap K(h_4) \cap K(h_5) \cap K(h_6)$.

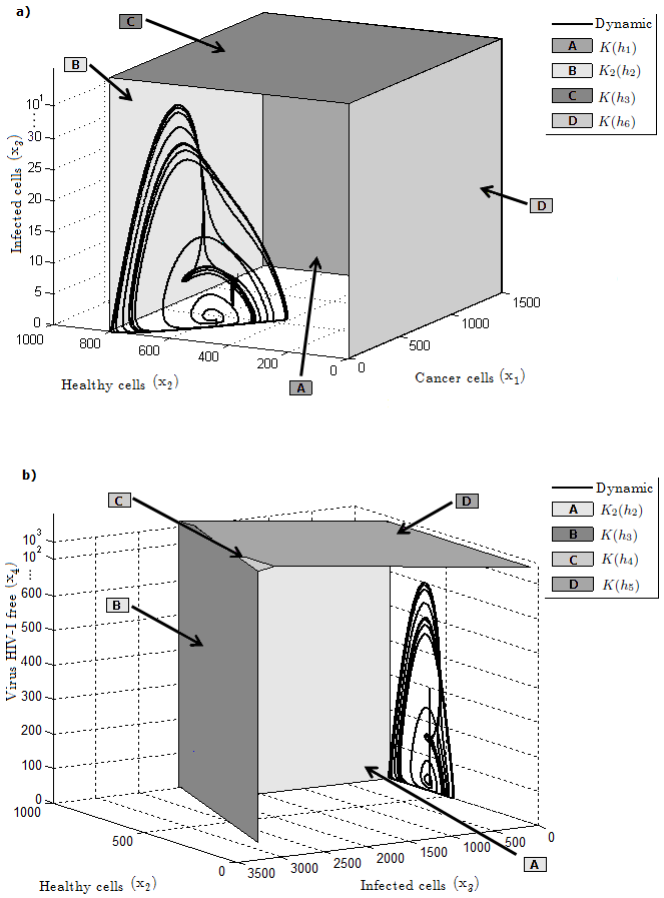


Fig. 1. Localization of compact invariant sets; a) x_1, x_2 and x_3; b) x_2, x_3 and x_4.

4 Conclusions

In the analysis of localization of compact invariant sets for the AIDS model related cancer we achieve a region in the domain $\mathbb{R}^4_+$ that encloses the dynamics system under the condition that the variables and parameters are positive. We used linear functions and subsequently we proposed rational functions to sharp the localization region. Numerical simulations allow to obtain a graphical interpretation of the including compact invariant sets and the localization behavior. The localization results give us information about how to manipulate the parameters, according to their biological characteristics can be manipulated in order to decrease the boundary of the system in-large. One of the parameter that can be modified, corresponds to the generation of T-CD4 + cells, repre-

sented by s. Note that according the medical implications described in the literature, the uncontrolled increase of this parameter can take the patient to a stable state to a critical state, causing autoimmune diseases.

References

1. El-Gohary, A., Al-Ruzaiza, A.: Chaos and adaptive control in two prey, one predator system with nonlinear feedback. Chaos, Solitons and Fractals **34** (2007) 443–453
2. Wanga, K., Wangb, W., Liu, X.: Viral infection model with periodic lytic immune response. Chaos, Solitons and Fractals **28** (2006) 90–99
3. Kansal, A., Torquato, S., Harsh, G., et al: Simulated brain tumor growth dynamics using a three-dimensional cellular automaton. Journal of Theoretical Biology **203** (2000) 367–382
4. Reznikoff, C., Belair, C., Yeager, T., Savelieva, E., et al: A molecular genetic model of human bladder cancer pathogenesis. Seminars in Oncology **23** (1996) 571–584
5. Boutayeb, A., Chetouani, A.: A critical review of mathematical models and data used in diabetology. BioMedical Engineering OnLine **5** (2006) 43
6. Haurie, C., Dale, D., Mackey, M.: Cyclical neutropenia and other periodic hematological disorders: A review of mechanisms and mathematical models. American Society of Hematology **92** (1998) 2629–2640
7. Kirschner, D., Panetta, J.: Modeling immunotherapy of the tumor-immune interaction. Journal of mathematical biology **37** (1998) 235–252
8. Cooper, L.: Theory of an immune system retrovirus. Proccedings of the National Academy of Sciences of the United States of America **83** (1986) 9159–9163
9. De Boera, R., Perelson, A.: Target cell limited and immune control models of HIV infection: a comparison. Journal of theoretical biology **190** (1998) 201–214
10. Murphy, S., Xu, J., Kochanek, K.: Deaths: Preliminary data for 2010. National Vital Statistics Reports **60** (2012)
11. Wen, Q., Lou, J.: The global dynamics of a model about hiv-1 infection in vivo. Ricerche di matematica **58** (2009) 77–90
12. Lou, J., Ruggeri, T., Ma, Z.: Cycles and chaotic behavior in an AIDS-related cancer dynamic model in vivo. Journal of Biological System **15** (2007) 149–168
13. Perelson, A., Krischner, D., De Boera, R.: Dynamics of HIV infection of CD4+ T-cells. Mathematical Biosciences **114** (1993) 81–125
14. Levy, J.: HIV and the pathogenesis of AIDS. Wiley-Blackwell (2007)
15. Krishchenko, A., Starkov, K.: Localization of compact invariant sets of nonlinear systems with applications to the lanford system. International Journal of Bifurcation and Chaos **16** (2006) 3249–3256
16. Krishchenko, A., Starkov, K.: Localization of compact invariant sets of the lorenz system. Physics Letters A **353** (2006) 383–388
17. Krishchenko, A., Starkov, K.: Estimation of the domain containing all compact invariant sets of a system modelling the amplitude of a plasma instability. Physics Letters A **367** (2007) 65–72

Bounding the long-time dynamics of a tumor immune-evasion model

P.A. Valle [1], K.E. Starkov [1], Luis N. Coria [2]

[1] Centro de Investigación y Desarrollo de Tecnología Digital, CITEDI-IPN
[2] Instituto Tecnológico de Tijuana, ITT
Tijuana, B.C.
pvalle@citedi.mx, konst@citedi.mx, luis.coria@gmail.com
Paper received on 04/10/12, Accepted on 25/10/12.

Abstract. In this document, we present results concerning the boundary for a localizing domain that contains all compact invariant sets for a tumor immune-evasion model. This model consists of a system of four nonlinear ordinary differential equations which describes the dynamics between tumor cells, immune effectors cells, the immuno-stimulatory cytokine Interleukin 2 and the suppressive cytokine TGF-β. The boundary for the final localizing set is expressed with some algebraic inequalities depending on the model parameters. This domain is important in the study of mathematical models which describes the dynamics of certain diseases because it provides important information about its long-time behavior, i.e. the location of equilibrium points, chaotic attractors, limit cycles, periodic orbits, homoclinic orbits and heteroclinic orbits. Our results are obtained by using two methods, one called *Localization of compact invariant sets*, which is based on first order extremum conditions, and the *Iterative theorem*. Finally, numerical simulations of dynamics of tumor growth are fulfilled in order to illustrate the localizing bounds.

Keywords: Boundary, Localization, Compact invariant sets, Iterative theorem, Cancer, Immune system.

1 Introduction

The design of mathematical models for biological systems began with the interaction of mathematics and biology; these models started to play a major role in the field of medicine by helping to better understand the evolution and spread of certain diseases, specially those whose study has generated an extremely complex problem to physicians through the years, e.g. : HIV-AIDS [1], [2], [3], hepatitis C [4], [5], H1N1 influenza [6], [7], tuberculosis [8], and some others. Furthermore, it should be noted that in practice it is necessary to develop *in vivo* experiments in order to determine the effect of different treatments. Therefore, numerical simulations of the models can help to diminish the amount of these experiments [9]. For this reason, a group of diseases of particular interest is cancer.

Cancer is a group of over 100 diseases characterized by uncontrolled proliferation of abnormal cells. These cells spread through the body and interfere with vital functions by invading tissues and organs, this process is called metastases and can lead to death

Mireya García-Vázquez, Grigori Sidorov (Eds.).
Advances in Computing Science and Control.
Research in Computing Science 59, 2012, pp. 45- 55

of the individual [10]. Despite the fact that overall mortality rates have declined in recent years, it remains as a major cause of illness and death worldwide in both men and women [11]. Although, over time there have been developed different types of treatments for this disease, only surgery, radiotherapy and chemotherapy have been accepted by medical society as conventional treatments. Nevertheless, the human body's immune system has proven to have the potential to fight cancer [12]. Therefore, its interaction with the immune system has been of particular interest in the scientific community and it is well documented: [13], [14], [15], [16], [17], among others. Some mathematical models also describe the effect of certain treatments, such as chemotherapy and biotherapy, with the aim to provide physicians a tool that allows them to plan more scientifically the schedules for therapies [15], [18], [19].

Simultaneously with mathematical modeling, some methods for analyzing dynamic systems have been adapted in order to study mathematical models of biological systems. Such models require a different approach from those used to analyze other types of systems such as: physical, chemical, electronic, artificial, and others. This is because the state variables of biological systems describe the interaction of large populations such as: cells, proteins, antibodies, viruses, bacteria, swarms, and some others. Moreover, mathematical models of biological systems are much more complex and their evolution over time is slow, in the case of tumor growth this evolution may take several years [20].

Two methods of particular interest which have been used to study the global dynamics of biological systems, see [21], [22], [23], are the so called *Localization of compact invariant sets*, which is based on first order extremum conditions, and *Iterative theorem*. These methods allow us to define a domain in the state space in which all compact invariant sets of a dynamical system are located. This domain is important in the study of biological systems because it provides important information about the location of equilibrium points, chaotic attractors, limit cycles and periodic, homoclinic and heteroclinic orbits. Moreover, since the localization domain depends on system parameters, in some cases, it is possible to propose conditions to reduce its bounds to such degree that the only possible dynamic will be an equilibrium point. This equilibrium point can be interpreted as healthy state for an individual affected by a disease such as cancer.

Therefore, since the analysis of biological systems play an important role in oncology, in this paper we study the dynamics of a tumor immune-evasion mathematical model, which describes the dynamics among four populations: effector cells ($\dot{x}$), cancer cells ($\dot{y}$), cytokine interleukin 2 ($\dot{z}$) and cytokine TGF-β_1 ($\dot{w}$):

$$\begin{aligned}
\dot{x} &= \frac{cy}{1+\gamma w} - \mu_1 x + \left(\frac{xz}{g_1+z}\right)\left(p_1 - \frac{q_1 w}{q_2+w}\right); \\
\dot{y} &= ry\left(1-\frac{y}{b}\right) - \frac{axy}{g_2+y} + \frac{p_2 wy}{g_3+w}; \\
\dot{z} &= \frac{p_3 xy}{(g_4+y)(1+\alpha w)} - \mu_2 z; \\
\dot{w} &= \frac{p_4 y^2}{\tau_c^2+y^2} - \mu_3 w.
\end{aligned} \tag{1}$$

This tumor immune-evasion model, according to [24], can help doctors to better understand the evolution of a malignant tumor, its mechanisms of immune evasion and its interaction with effector cells. The main feature of this model is that it takes into consideration the secretion from the tumor of the cytokine Transforming growth factor - β_1 (TGF-β_1) which at: counter immuno-stimulating properties of IL-2, preventing tumor detection by the immune system, reducing the expression of antigens on cancer cells and inhibit activation and expansion of cytotoxic T cells and B cells, *prevents the destruction of the malignant tumor*. In addition, cytokine TGF-β_1 possesses angiogenic properties, which benefits the development and metastasis of malignant tumors [25].

The main objective of this paper is to establish the existence and form a compact invariant domain in the space R^4_+ for the tumor immune-evasion model (1). The importance of this domain lies in the fact that at being a compact invariant set any trajectory that enters into this domain will remain in it for all future time i.e. trajectories will not diverge exponentially. Biologically, this implies that concentrations of cytokines and cells described by system (1) will not increase uncontrollably, which would affect negatively the patient health.

The paper is organized as follows: section 2 presents mathematical preliminaries concerning the method for the localization of compact invariant sets, section 3 shows our localization results obtained by means of linear and nonlinear localizing functions, in section 4 we present numerical simulations in order to illustrate our final localizing domain, in section 5 we give a description about the biological implication of our results, section 6 presents the conclusions and finally the reader can see the references used in the development of this document.

2 Localization of compact invariant sets

By *Localization of compact invariant sets* we mean the calculation of the domain on the state space where all compact invariant sets are located. These compact invariant sets are presented under certain conditions in any specific mathematical model. The relevance of this analysis is because it is useful to study the long-term dynamics of the system. The general localization method of compact invariant sets of a nonlinear system was described in [26], [27]. In this section we present useful results. Let us consider a

nonlinear system with the form:

$$\dot{x} = f(x); \tag{2}$$

where f is a continuous vectorial function for a $C^\infty-$differentiable vector field; $x \in \mathbf{R}^n$ is the state vector. Let $h(x)$ be a C^∞ such that h is not the first integral of (2). By $h|_B$ we denote the restriction of h on a set $B \subset \mathbf{R}^n$. The function h used in this statement is called localizing. By $S(h)$ we denote the set $\{x \in \mathbf{R}^n \mid L_f h(x) = 0\}$, where $L_f h(x)$ represents the Lie derivative of (2) and is given by: $L_f h(x) = \dfrac{\partial h}{\partial x} f(x)$. Let us define $h_{\inf} := \inf\{h(x) \mid x \in S(h)\}$; $h_{\sup} := \sup\{h(x) \mid x \in S(h)\}$.

2.1 General theorem

The general theorem concerning the localization of all compact invariant sets of a dynamical system establishes the following:

Theorem 2.1. *Each compact invariant set Γ of (2) is contained in the localization set $K(h) = \{h_{\inf} \leq h(x) \leq h_{\sup}\}$.*

If we consider the location of all compact invariant sets inside the domain $U \subset \mathbf{R}^n$ we have the localization set $K(h) \cap U$, with $K(h)$ defined in **Theorem 2.1**. It is evident that if all compact invariant sets are located in the sets Q_1 and Q_2, with $Q_1; Q_2 \subset \mathbf{R}^n$, then they are located in the set $Q_1 \cap Q_2$ as well.

2.2 Non existence condition

Suppose that we are interested in the localization of all compact invariant sets located in some subset Q of the state space $\mathbf{R}^n$. We formulate

Proposition 2.1. *If $Q \cap S(h) = \emptyset$ then the system (2) has no compact invariant sets located in Q.*

2.3 Iterative theorem

A refinement of the localization set $K(h)$ is realized with help of the iteration theorem stated as follows.

Theorem 2.2 *Let $h_m(x), m = 0, 1, 2, \ldots$ be a sequence of infinitely differentiable functions. Sets*

$$K_0 = K(h_0), \quad K_m = K_{m-1} \cap K_{m-1,m}, \quad m > 0,$$

with

$$\begin{aligned} K_{m-1,m} &= \{x : h_{m,inf} \leq h_m(x) \leq h_{m,sup}\}, \\ h_{m,sup} &= \sup_{S(h_m) \cap K_{m-1}} h_m(x), \\ h_{m,inf} &= \inf_{S(h_m) \cap K_{m-1}} h_m(x), \end{aligned}$$

contain any compact invariant set of the system (2) and

$$K_0 \supseteq K_1 \supseteq \cdots \supseteq K_m \supseteq \ldots .$$

3 Main localization results

In this section we present results for linear and nonlinear localizing functions. The intersections of these regions make the localization of all compact invariant sets for the tumor immune-evasion model shown above. Since variables x, y, z and w in model (1) represent concentrations with biological sense, we examine compact invariant sets only inside the positive domain:

$$R_+^4 = \{x > 0, y > 0, z > 0, w > 0\};$$

which is also consider a compact invariant set. In addition all parameters of this model are supposed to be positive. Also, for the simplicity of notations we consider the following: $S(h) = S(h) \cap R_+^4$ and therefore $K(h) = K(h) \cap \mathbf{R}_+^4$.

3.1 Localization by means of linear functions

In order to obtain a localizing set that provides *the supreme value for the secretion of the cytokine TGF-β_1* by the tumor, we propose the following localizing function $h_1 = w$; from which by calculating its Lie derivative and performing the corresponding operations we can obtain the next set

$$S(h_1) = \left\{ \mu_3 w = p_4 \left(1 - \frac{\tau_c^2}{\tau_c^2 + y^2} \right) \right\};$$

now, we can define the following

Theorem 3.1: *The supreme value for the concentration of the cytokine TGF-β_1 is given by w_{max} in the localizing set:*

$$K(h_{11}) = \left\{ w_{min} = 0 \leq w \leq \frac{p_4}{\mu_3} = w_{max} \right\}. \tag{3}$$

Now, we propose the localizing function $h_2 = y$ in order to establish a *supreme value for the concentration of cancer cells,* from which by calculating its the Lie derivative and applying the iterative theorem with the localizing set (3) we can obtain the next

$$S(h_1) \cap K(h_{11}) \subset \left\{ h_{2|_{S(h_1)}} \leq b \left(1 + \frac{p_2 p_4}{(\mu_3 g_3 + p_4) r} \right) \right\};$$

then, according to calculations performed we can have the following

Theorem 3.2: *The maximum concentration of cancer cells is given by the value y_{max} on the localizing set:*

$$K(h_{21}) = \left\{ y_{min} = 0 \leq y \leq b \left(1 + \frac{p_2 p_4}{(\mu_3 g_3 + p_4) r} \right) = y_{max} \right\}. \tag{4}$$

Now, in order to obtain a *supreme value for the dynamics of the state variable corresponding to the effector cells*, we take the localizing function $h_3 = x$; from which by calculating its Lie derivative and applying the iterative theorem with the sets (3) and (4) we can obtain the set

$$S(h_3) \cap K(h_{11}) \cap K(h_{21}) \subset \left\{ x \leq \frac{cy_{max}}{\mu_1 - p_1} \right\};$$

and if the next condition is fulfilled

$$\mu_1 > p_1; \tag{5}$$

we can define the next

Theorem 3.3: *If condition (5) is fulfilled, then the maximum concentration of effector cells at the tumor site is given by the value x_{max} in the next localizing set:*

$$K(h_{31}) = \left\{ x_{min} = 0 \leq x \leq \frac{cy_{max}}{\mu_1 - p_1} = x_{max} \right\}. \tag{6}$$

Now, trying to obtain the *supreme value for the cytokine IL-2 concentration at the tumor site* we take the localizing function $h_4 = z$; from which by calculating its Lie derivative, applying the iterative theorem by using the sets (3), (4) and (6), and if condition (5) is fulfilled we can obtain the following

$$S(h_4) \cap K(h_{11}) \cap K(h_{21}) \cap K(h_{31}) \subset \left\{ z \leq \frac{1}{\mu_2} \frac{p_3(x_{max})}{1 + \alpha(w_{min})} \left(1 - \frac{g_4}{g_4 + y_{max}} \right) \right\};$$

then, we can get the next

Theorem 3.4: *If condition (5) is fulfilled, then the maximum concentration of IL-2 at the site of the tumor is given by the value z_{max} in the next localizing set:*

$$K(h_{41}) = \left\{ z_{min} = 0 \leq z \leq \frac{p_3 x_{max} y_{max}}{\mu_2 (g_4 + y_{max})} = z_{max} \right\}. \tag{7}$$

3.2 Localization by means of nonlinear functions

Nonlinear localizing functions are used to reduce the domain conformed by the intersection of the sets obtained by linear functions. The relevance in reducing this domain remains in the fact that it would be easier to find the different dynamics of a system by numerical simulations. In addition, the decrease of the bounds allows us to better understand the system dynamics in the long term. Below we show results obtained with the nonlinear localizing function $h_5 = xy$; from which by calculating its Lie derivative and by using the localizing sets (3), (4) in order to apply the iterative theorem we can get the following

$$S(h_5) \cap K(h_{11}) \cap K(h_{21}) \subset \left\{ xy \leq \frac{b}{r} \left(cy_{max} - \beta_1 x^2 + \beta_2 x \right) \right\};$$

where

$$\beta_1 := \frac{a}{g_2 + y_{max}} \quad \text{and} \quad \beta_2 := \frac{p_2 p_4}{\mu_3 g_3 + p_4} + p_1 + r - \mu_1;$$

then, according to the sign of β_2 we can define the following:

Theorem 3.5: *If* $\beta_2 > 0$; *then the localizing set is given by*

$$K(h_{51}) = \left\{ xy \leq \frac{b}{r} \left(\frac{\beta_2^2}{4\beta_1} + c y_{max} \right) \right\}. \tag{8}$$

Theorem 3.6: *If* $\beta_2 \leq 0$; *then the localizing set is given by*

$$K(h_{52}) = \left\{ xy \leq \frac{bcy_{max}}{r} \right\}. \tag{9}$$

4 Numerical simulations

The final compact localizing set in which all compact invariant sets of (1) are located is given by the intersections of the following sets:

$$K(h_{11}) \cap K(h_{21}) \cap K(h_{31}) \cap K(h_{41}) \cap K(h_{51}).$$

Now, since all localizing sets depend on system parameters we use specific values in order to get our results. The periodic orbit and the boundaries illustrated in Figure 1 were obtained by using the following values in system parameters: $\mu_1 = 0.03$; $p_1 = 0.01$; $g_1 = 2 \times 10^7$; $c = 0.005$; $q_1 = 0.1121$; $q_2 = 2 \times 10^6$; $\gamma = 10$; $r = 0.18$; $b = 1 \times 10^9$; $a = 1$; $g_2 = 1 \times 10^5$; $p_2 = 0.27$; $g_3 = 2 \times 10^7$; $p_3 = 5$; $g_4 = 1 \times 10^3$; $\mu_2 = 10$; $\alpha = 1 \times 10^{-3}$; $\mu_3 = 10$; $\tau_c = 1 \times 10^6$; $p_4 = 0.1204$.

Figure 1.(a) shows the localizing domain concerning the variables x (Effector cells), y (Cancer cells) and z (Cytokine IL-2) and Figure 1.(b) shows the localizing domain concerning the variables x (Effector cells), y (Cancer cells) and w (Cytokine TGF-β_1).

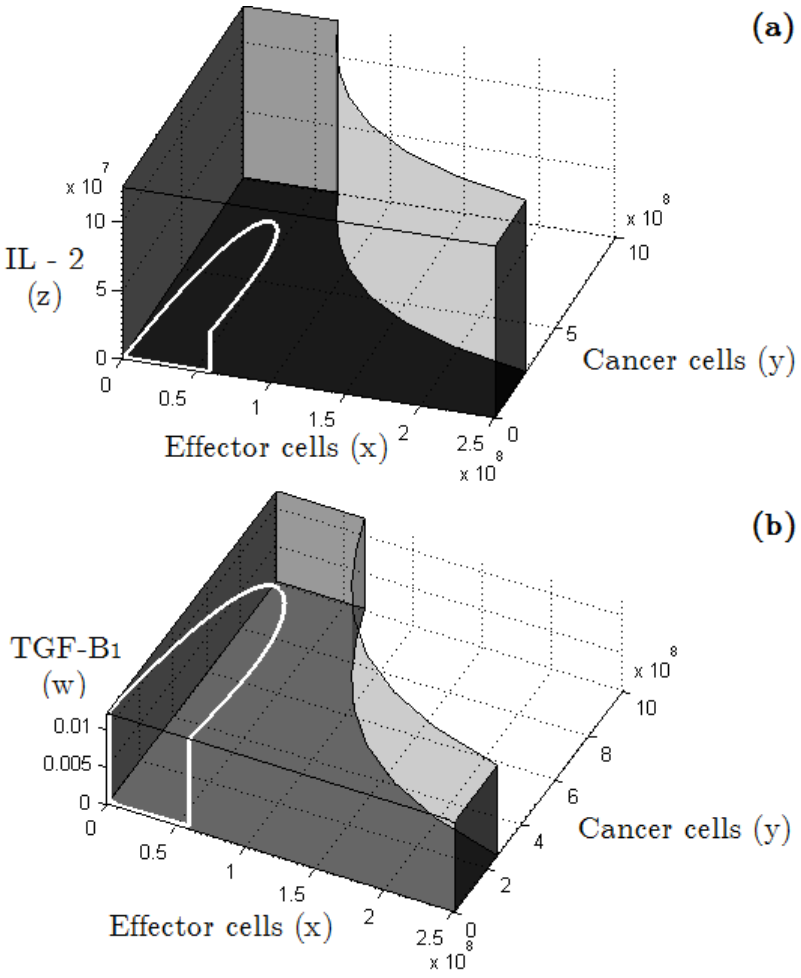


Fig. 1. Localizing domain for the tumor immune-evasion model (1):(a) Dynamics of the state variables x, y y z. (b) Dynamics of the state variables x, y y w.

5 Biological implications

The dynamical properties in long-time behavior of a specific system become observable if we can find their bounds. The existence of the localizing sets $K\left(h_{31}\right)$ and $K\left(h_{41}\right)$ (which represent effector cells and IL-2 concentration respectively) depends on the condition (5) $(\mu_1 > p_1)$. This condition means that the mortality rate of effector cells is greater than its proliferation rate. Although this condition implies deterioration in the health of the patient, it is important to analyze it. We have found in the literature that the condition $\mu_1 > p_1$ may occur in a patient due to the following scenarios:

– **Tumor defense mechanisms.** These mechanisms affect the lifetime of immune cells and are generated by the genetic instability of cancer cells [28], [25], some of which may contribute to fulfill the condition (5) are:

1. Some tumors have levels of antigens too low to be detected by the immune system which can induce apoptosis in T cells due to the lack of warning signals to alert the immune system. This phenomenon can lead to immune tolerance to cancer cells.
2. Production by the tumor of immune inhibitory substances such as TGF-β_1. This protein performs several functions within the cell including the process of apoptosis.
3. Induction in proliferation of suppressor T cells by the malignant tumor.

– **Treatments such as chemotherapy and radiotherapy.** These play an important role in immune system deficiency because they affect the patient ability to generate T cells, which decreases the number of white blood cells and weakens the immune system. Furthermore, this contributes to make the patient more susceptible to acquire various types of infections [29], [30], [31], [32] y [33].

6 Conclusions

Our approach can be compared with the results obtained by Kirschner and Tsygvintsev (2009) in [34], where the authors use quasi-Lyapunov functions in order to establish some bounds for a cancer immunotherapy mathematical model. Nevertheless, in [35] Starkov and Coria (2012) use the Localization of compact invariant sets method and the Iterative theorem in order to establish a compact domain where all compact invariant sets of the cancer immunotherapy model are located; also they give some conditions for a tumor free equilibrium point, it is important to say that this conditions depend only on the system parameters. The advantages of using localizing functions instead of quasi-Lyapunov functions remain in the fact that these functions do not have the same limitations, e.g. localizing functions does not have to be positive definite and its derivative does not have to be negative definite, in order to get useful information the Lie derivative of the localizing function needs to have a definite sign, with which we are able to define a supreme or an infinite value for the bounds according with **Theorem 2.1**. Additionally, we can obtain an improvement of the bounds if it is possible to use the **Theorem 2.2**.

The tumor immune evasion system (1) is an extension of the cancer immunotherapy model presented by Kirschner and Tsygvintsev and we were able to establish a compact domain for all compact invariant sets of (1) by applying linear and nonlinear localizing functions which intersection makes. Nevertheless, since system (1) does not have any treatment parameters, i.e. the cellular immunotherapy and the external administration of IL-2 from the cancer immunotherapy system, it is not possible to give conditions for a tumor free equilibrium point; also the boundaries given in section 3 are not as manipulable as the ones presented in [35].

Now, we present some general conclusions about our results:

– We were able to define supreme values for each of the state variables of the biological system under study, these values are given by the localizing sets: $K\left(h_{i1}\right)$

and $K\left(h_{5j}\right); i=1,2,3,4; j=1,2$; which define the compact domain in the state space where all compact invariant sets of the system (1) are located.
- The existence of the localizing sets concerning the supreme values of effector cells ($K\left(h_{31}\right)$) and IL-2 ($K\left(h_{41}\right)$) concentrations depends on the condition (5): $\mu_1 > p_1$, see discussion in the previous section.
- Localizing sets $K\left(h_{51}\right)$ and $K\left(h_{52}\right)$ (obtained through the nonlinear localizing function h_5), allows to reduce the localizing region in the xy plane. Additionally, the function h_5 provides a general overview of the interaction between effector cells and cancer cells i.e. as the concentration of effector cells increases the upper bound for cancer cells concentration decreases and viceversa. Therefore, it became obvious that in order to completely eradicate the malignant tumor it is necessary to maintain a sufficient amount of effector cells in the tumor site.
- The existence of the sets: $K\left(h_{11}\right), K\left(h_{21}\right), K\left(h_{51}\right)$ and $K\left(h_{52}\right)$ does not depend on any condition that may have conflict with the biological sense of the system parameters.

References

1. Craig, I., Xia, X.: Can HIV/AIDS be controlled? applying control engineering concepts outside traditional fields. Control Systems Magazine, IEEE **25** (2005) 80–83
2. Law, M., Prestage, G., Grulich, A., Van de Ven, P., Kippax, S.: Modelling the effect of combination antiretroviral treatments on HIV incidence. AIDS **15** (2001) 1287–1294
3. Lima, V., Johnston, K., Hogg, R., Levy, A., Harrigan, P., Anema, A., Montaner, J.: Expanded access to highly active antiretroviral therapy: a potentially powerful strategy to curb the growth of the HIV epidemic. Journal of Infectious Diseases **198** (2008) 59–67
4. Dahari, H., Lo, A., Ribeiro, R., Perelson, A.: Modeling hepatitis C virus dynamics: Liver regeneration and critical drug efficacy. Journal of theoretical biology **247** (2007) 371–381
5. Neumann, A., Lam, N., Dahari, H., Davidian, M., Wiley, T., Mika, B., Perelson, A., Layden, T.: Differences in viral dynamics between genotypes 1 and 2 of hepatitis C virus. Journal of Infectious Diseases **182** (2000) 28–35
6. Keeling, M., Danon, L.: Mathematical modelling of infectious diseases. British medical bulletin **92** (2009) 33–42
7. Tracht, S., Del Valle, S., Hyman, J.: Mathematical modeling of the effectiveness of facemasks in reducing the spread of novel influenza a (H1N1). PloS one **5** (2010) e9018
8. Feng, Z., Castillo-Chavez, C., Capurro, A.: A model for tuberculosis with exogenous reinfection. Theoretical Population Biology **57** (2000) 235–247
9. Bellomo, N.: Modeling Complex Living Systems: A Kinetic Theory and Stochastic Game Approach. Birkhäuser (2008)
10. Britannica, E.: Cancer. Encyclopædia Britannica 2009 Student and Home Edition (2009.)
11. Siegel, R., Naishadham, D., Jemal, A.: Cancer statistics, 2012. CA: A Cancer Journal for Clinicians **62** (2012) 10–29
12. Oldham, R., Dillman, R.: Principles of Cancer Biotherapy. Fifth edn. Springer (2009)
13. Bunimovich-Mendrazitsky, S., Shochat, E., Stone, L.: Mathematical model of BCG immunotherapy in superficial bladder cancer. Bulletin of mathematical biology **69** (2007) 1847–1870
14. Bunimovich-Mendrazitsky, S., Gluckman, J., Chaskalovic, J.: A mathematical model of combined bacillus calmette-guerin (BCG) and interleukin (IL)-2 immunotherapy of superficial bladder cancer. Journal of Theoretical Biology **277** (2011) 27–40

15. de Pillis, L., Gu, W., Fister, K., Head, T., Maples, K., Murugan, A., Neal, T., Yoshida, K.: Chemotherapy for tumors: An analysis of the dynamics and a study of quadratic and linear optimal controls. Mathematical Biosciences **209** (2007) 292–315
16. El-Gohary, a., Alwasel, I.: The chaos and optimal control of cancer model with complete unknown parameters. Chaos, Solitons and Fractals **42** (2009) 2865–2874
17. Lou, J., Ruggeri, T., Ma, Z.: Cycles and chaotic behavior in an AIDS-related cancer dynamic model in vivo. Journal of Biological Systems **15** (2007) 149–168
18. dOnofrio, A.: A general framework for modeling tumor-immune system competition and immunotherapy: Mathematical analysis and biomedical inferences. Physica D: Nonlinear Phenomena **208** (2005) 220–235
19. Castiglione, F., Piccoli, B.: Cancer immunotherapy, mathematical modeling and optimal control. Journal of Theoretical Biology **247** (2007) 723–732
20. Kirschner, D., Panetta, J.: Modeling immunotherapy of the tumor – immune interaction. Journal of Mathematical Biology **37** (1998) 235–252
21. Starkov, K., Coria, L.: Bounding the domain of some three species food systems. In: Analysis and Control of Chaotic Systems. Volume 2. (2009) 193–198
22. Starkov, K.E., Coria, L., Valle, P.A.: Bounding the long-time dynamics of a cancer immunotherapy model. Dynamics Days Europe XXXI; Oldemburg, Germany (2011)
23. Coria, L., Starkov, K.E., Plata, C.: Localización de conjuntos compactos invariantes para un modelo de un tumor cancerígeno. CIINDET 2011; Cuernavaca Morelos, México (2011)
24. Arciero, J., Jackson, T., Kirschner, D.: A mathematical model of tumor-immune evasion and sirna treatment. DISCRETE AND CONTINUOUS DYNAMICAL SYSTEMS SERIES B **4** (2004) 39–58
25. Ruddon, R.W.: Cancer Biology. Fourth edn. OXFORD (2004)
26. Krishchenko, A.: Estimations of domains with cycles. Computers & Mathematics with Applications **34** (1997) 325–332
27. Krishchenko, A., Starkov, K.: Localization of compact invariant sets of the lorenz system. Physics Letters A **353** (2006) 383–388
28. Clark, W.R.: In Defense of Self. How the Immune System Really Works. Oxford University Press (2008)
29. Cukier, D.: Coping With Chemotherapy and Radiation. McGraw-Hill Professional (2004)
30. Institute, N.C.: Chemotherapy and You. U.S. Department of Health and Human Services and National Institutes of Health (2011)
31. Lu, J.J., Brady, L.W.: Radiation Oncology: An Evidence-Based Approach. Springer (2008)
32. Lyss, A.P., Fagundes, H., Corrigan, P.: Chemotherapy and Radiation For Dummies. John Wiley and Sons (2011)
33. Perry, M.C.: The Chemotherapy Source Book. Fourth edn. Lippincott Williams and Wilkins (2008)
34. Kirschner, D., Tsygvintsev, A.: On the global dynamics of a model for tumor immunotherapy. Mathematical Biosciences and Engineering **6** (2009) 573–583
35. Starkov, K., Coria, L.: Global dynamics of the kirschner-panetta model for the tumor immunotherapy. Nonlinear Analysis: Real World Applications **In Press** (2012)

On upper bounds for compact invariant sets of nonlinear bladder cancer system with BCG immunotherapy

K.E. Starkov and D. Gamboa

Centro de investigacin y desarrollo de tecnologa digital, CITEDI-IPN
Av. del Parque 1310, Mesa de Otay, Tijuana, BC, Mexico.
fax: 52 66231388; phone 52 66231344
konst@citedi.mx, gamboa@citedi.mx
Paper received on 21/09/12, Accepted on 22/10/12.

Abstract. In this work upper bounds for all variables of the nonlinear bladder cancer system with BCG immunotherapy are derived. These bounds characterize ultimate health conditions in the ideal situation of infinite time interval and may be used in studies of global dynamics. Some nonexistence conditions of compact invariant sets are presented as well.
Key words: Upper Bounds, Bladder Cancer, Immunotherapy, Compact Invariant Sets

1 Introduction

The dynamic of cancerous tumor growth and immunological response can be described by a nonlinear model. The analysis of these models are necessary to understand in short and long time behavior how the tumor invades surrounding organs (this processes is called metastases) or simply growth *in situ*. Analysis of nonlinear mathematical biological ordinary differential equation (ODE) system is relatively a new area to investigate, the advantage over the PDE mathematical biological systems which is most common to find these analysis in literature is the simplicity that carry at the moment of visualize a whole framework with the main variables. Analyzing a high degree mathematical biological ODE system can present a better interaction framework of the whole system, and with a proper anti-tumor treatment may reduce as many side effects possible if a key variable is identified. In order to reduce cancerous tumor significantly several treatments must by combine. Latency may occur if a non proper serious care is taken, see e.g. paper [1].

Bladder Cancer is a tumor that growths in the inner surface of the bladder, it originated at first in the epithelium as *carcinoma in-situ (*CIS*)* and after some time the tumor begins to invade the superficial muscle layer. According to the TNM classification of Cancer if tumor is not eradicated by surgery on T2a stage the tumor will penetrated into the deep muscle leading to a more aggressive treatment. Cystectomy consist in removal of the bladder which is an aggressive treatment,this is usually made when the tumor has growth up to stage T3a.

If no surgery or treatment are applied up to stage T3a, the tumor next stage is invasion of the surrounding organs , for example lungs, liver, bone, etc. If metastasis is detected no treatment or surgery can be able to eradicate the tumor, see e.g. paper [2].

Mireya García-Vázquez, Grigori Sidorov (Eds.).
Advances in Computing Science and Control.
Research in Computing Science 59, 2012, pp. 56- 64

Hence we can say that in a short time of period the tumor is *in-situ* and there is absence of metastasis therefore it can be removed with local surgery, better known as transurethral resection (TUR). This procedure consist on identifying cancerous regions in the bladder cavity and remove the damage tissue by making small cuts heating the circular point of the probe. The probe is introduced directly by the urethra. In this case to avoid latency, immunotherapy has proved to be more effective with less side effects in comparison with quimiotherapy after a tumor surgery has been made. The immunotherapy BCG (Bacillus-Calmette-Guerin) -*an attenuated strain of Mycobacterium bovis (M. bovis) used for anti tuberculosis immunization* - clinically has proven to be a reliable procedure to avoid latency after TUR surgery in 50-70 [3]. Immunotherapy is a treatment which enhances the immune response against a pathogen, in this case cancer cells are known to be a pathogen because a set of cancerous cells make a cancerous tumor.

The mathematical model under study was presented by Svetlana in 2007, [4], where it describe the interaction between tumor cells within the bladder, immune system response and BCG immunotherapy with a system of nonlinear ODEs. To describe the model slightly, we have that the model is divided into two subpopulation of tumor cells, we have those that have been infected (T_i) by BCG (B) and those who have not been infected by BCG (Tu), but the total tumor cells is given by the sum of T_u and T_i. The effector cells (E) are the set of the immune system response cells (APC, natural killer cells, lymphocyte-activated killer) defined by one single term. Also the model describe two types of tumor growth behavior, exponential and logistic. Is called exponential growth because the rate proportional to the product $T_u r$ is constant , hence there is no death of tumor cells as a result of self-limiting competition for resources such as oxygen and glucose , thus the product $-p_2 T_u B$ where p_2 is the coefficient of infection rate of tumor cells by BCG will continue growing at the rate $T_u r$ indicating that the BCG treatment has a low effect on the tumorous cells. Is called logistic growth when the tumor carrying capacity (β^{-1}) may be viewed as the maximum carry capacity of the tumor where is prevail the competition for resources (β).

In this paper localizations analysis of compact invariant sets for the bladder cancer system and bounds for the bounded positively invariant domain were made under the logistic tumorous growth cells due to the complexity that encompasses the dynamics of all the variables. The model is presented below:

$$\begin{aligned}
\frac{dB}{dt} &= -\mu_1 B - p_1 EB - p_2 BT_u + b \\
\frac{dE}{dt} &= -\mu_2 E + \alpha T_i + p_4 EB - p_5 ET_i \\
\frac{dT_i}{dt} &= -p_3 ET_i + p_2 BT_u \\
\frac{dT_u}{dt} &= T_u(-p_2 B + r[1 - \beta T_u])
\end{aligned} \tag{1}$$

On the model we can see that the BCG immunotherapy have a parameter b which is a constant dose of the treatment BCG, then some free BCG binds to tumorous cells at a coefficient p_2 infecting them but other BCG is lost at a rate p_1 due to the effector cells recognized them as pathogens cells. The coefficient μ_1 indicates the rate of BCG decay. From here we can say that b is our free parameter that can be manipulated in

order to be considered as a proper doses of BCG with lower side effects if bounds are apply.

The effector cells on the other hand have two complex dynamic, the first one is to deal with BCG at a rate p_4 and the second one is to eliminate the tumor infected cells by BCG at a rate p_5. The μ_2 coefficient is the effector cells mortality rate and α be the rate of effector stimulations due to infected tumor cells. The advantage of the effector cells on eliminate first the BCG is that strengthens their cells against these tumorous cells. The Ti is attacked at a rate p_3 by the effectors strengthens cells and continuously infected by BCG at a rate p_2 on new regions of the tumor area uninfected, see e.g. paper [Svetlana, 2007].

The method of *Localization analysis of Compact Invariant Sets* is a powerful mathematical tool for finding trapping regions given a localizing function, this is important because it gives a solution to the localization problem of all compact invariant sets that the system have. Localizing functions are tricky to find using a general methodology because each localizing function is customized for each system under study. In literature some of these function were apply on physic systems [5], mechanical systems [6], chaotic systems [7],[8] for mention some few of them. The model (1) differs from other models in literature, see e.g. paper [9] by the fact of considering the interaction terms of αT_i and $p_4 EB$ instead of analyzing the explicit effects of one citokine.

In this work we present result concern to the solution to the localization problem of all compact invariant contained in the positive orthant for the bladder cancer with BCG immunotherapy under a logistic growth and a nonexistence conditions where the boundaries positive domains are presented with some inequalities depending of the parameters of the system.

2 Biological sense of the system,its parameters

Svetlana in 2007 presents for the first time the system (1) where the parameters value are compiled from peer-reviewed mathematical models of cancer growth and immunotherapy BCG. The list of all the parameters of the system (1) are presented in Table 1,[4]. There main idea of using values from literature was to obtain a generic qualitative results that are intrinsic to the models structure. This leads to results that are not tied to any specific growth rates if mathematical analysis is made. In our case the biological interpretation once is located all compact invariant sets will depend of the parameters value of the system (1). Notwithstanding the parameter ranges are realistic and agree with values from the literature this leads to our results to have biologic sense after our mathematical analysis. The parameter range for tumor growth rate presented in Table 1 is presented in vitro, from mice or from humans. In this case is consider the in vitro parameter, proposed by Aranha in 2000, [10]. If a parameter is obtained in a laboratory under an artificial control environment, isolated from living organisms or systems but artificially maintained in a test tube for continuous surveillance is called in vitro parameter. According to Andrea in 2004 studies in vitro leads to a greater understanding with all the variables of the bladder cancer system. The system 1 which has a complex dynamics between the immunotherapy BCG and the tumor growth, the in vitro parameter value will be for the tumor growth. [11]

3 Some preliminaries and necessary notations

We consider a $C^\infty-$ differentiable system

$$\dot{x} = F(x), \tag{2}$$

with $x \in \mathbf{R}^n$, $F(x) = (F_1(x), \ldots, F_n(x))^T$ and $F_i(x) \in C^\infty(\mathbf{R}^n)$, $i = 1, \ldots, n$.

Let $h(x) \in C^\infty(\mathbf{R}^n)$ be a function such that h is not the first integral of the system (2).

The function h is used in the solution of the localization problem of compact invariant sets and is called a localizing function. Suppose that we are interested in the localization of all compact invariant sets located in some set $N \subset \mathbf{R}^n$ where N is an invariant set for the system (2) or a domain. By $S(h)$ we denote the set $\{x \in \mathbf{R}^n : L_F h(x) = 0\}$, where $L_F h(x)$ is a Lie derivative with respect to F. Further, we define $h_{\inf}(N) := \inf\{h(x) \mid x \in N \cap S(h)\}$, $h_{\sup}(N) := \sup\{h(x) \mid x \in N \cap S(h)\}$.

Proposition 1. *If $N \cap S(h) = \emptyset$ then the system (2) has no compact invariant sets located in N.*

Theorem 1. *For any $h(x) \in C^\infty(\mathbf{R}^n)$ all compact invariant sets of the system (2) located in N are contained in the set defined by the formula*

$$K(N) = \{x \in N : h_{\inf}(N) \leq h(x) \leq h_{\sup}(N)\}$$

as well.

Theorem 2. *Let $h_m(x), m = 1, 2, \ldots$ be a sequence of functions from $C^\infty(\mathbf{R}^n)$. Sets*

$$K_1 = K_{h_1}, K_m = K_{m-1} \cap K_{m-1,m}, m > 1,$$

with

$$\begin{aligned} K_{m-1,m} &= \{x : h_{m,\inf} \leq h_m(x) \leq h_{m,\sup}\}, \\ h_{m,\sup} &= \sup_{S_{h_m} \cap K_{m-1}} h_m(x), \\ h_{m,\inf} &= \inf_{S_{h_m} \cap K_{m-1}} h_m(x), \end{aligned}$$

contain all compact invariant sets of the system (2) and $K_1 \supseteq K_2 \supseteq \cdots \supseteq K_m \supseteq \ldots$.

4 Main result: polytopic localization for all compact invariant sets

In order to study the system (1) we took the a dimensionless blader cancer system from Svetlana in 2007. The system is below:

$$\begin{aligned}\dot{x} &= x(-1 - p_1 y - p_2 w) + b \\ \dot{y} &= y(-\mu + p_4 x - p_5 z) + \alpha z \\ \dot{z} &= -p_3 yz + p_2 xw \\ \dot{w} &= w(-p_2 x + r - r\beta w)\end{aligned} \tag{3}$$

where x represent the treatment BCG and the parameter b is the continuous doses of BCG, y represent the set of effector cells (APC, natural killer cells, lymphocyte-activated killer), z represents the tumor infected cells and w is the tumor uninfected cells via endocytosis after administration of the BCG in the bladder. All variables are considered in the positive orthant $\mathbf{R}^4_+ = \{\, x > 0; y > 0; z > 0; w > 0\}$ because of its biological nature. Also, let $\mathbf{R}^4_{+,0}$ be the closure of $\mathbf{R}^4_+ : \mathbf{R}^4_{+,0} = \{x \geq 0; y \geq 0; z \geq 0; w \geq 0\}$.

By using results in Section 2 and the localizing function $h_1 = x$ we can derive the localization set

$$K_1(h_1) = \{x \leq b\}$$

From this we concluded that the upper bound for the treatment will depend of the BCG doses administrated, according to Svetlana in 2007 a optimal doses of BCG is with less side effects.

Next, by using results in Section 2 and the localizing function $h_2 = w$ we can obtain the localization set

$$K_2(h_2) = \left\{ w \leq \frac{1}{\beta} \right\}$$

Now let

$$\frac{\mu}{\alpha} \leq \frac{p_3}{p_5} \tag{4}$$

Then by using results in Section 2 and the localizing function $h_3 = z - qy$ we can obtain the localization set

$$K_3(h_3) = \left\{ z - qy \leq \frac{p_2 p_5 b}{p_3 \alpha \beta} \right\},$$

with $q \in \left[\frac{\mu}{\alpha}, \frac{P_3}{P_5} \right]$.

Further, by using results in Section 2 and the localizing function $h_4 = z + w$ we can obtain the localization set

$$K_4(h_4; q) := \left\{ z \leq z_{\max}(q) := \frac{1}{\beta} + \frac{bp_2}{2q\alpha\beta} + \sqrt{\frac{b^2 p_2^2}{4q^2\alpha^2\beta^2} + \frac{rq}{4\beta p_3}} \right\}$$

under condition (4).

The best bound $z_{\max}(q_{\min})$ may be found from the solution of the minimization problem

$$\min\left\{z_{\max}(q); q \in \left[\frac{\mu}{\alpha}, \frac{P_3}{P_5}\right]\right\}$$

and we introduce

$$K(h_4) := K(h_4; q_{\min})$$

Now let

$$\mu - p_4 b > 0 \tag{5}$$

Finally, by using results in Section 2 and the localizing function $h_5 = y$ we can obtain the localization set

$$K(h_5) = \left\{y \leq y_{max} := \frac{\alpha z_{\max}}{\mu - P_4 b}\right\}$$

As a result, we come to

Theorem 3. *Suppose that conditions 4 and 5 hold. Then all compact invariant sets are located in the set*

$$K := \cap_{i=1;2;4;5} K(h_i)$$

5 Conditions of non existence of compact invariant sets in $\mathbf{R}^4_{+,0} \cap \{w > 0\}$

Conditions of non existence in the domain $\mathbf{R}^4_{+,0} \cap \{w > 0\}$ can be derived by exploiting the next rational localization function.

$$h_6 = \frac{x}{w}$$

We have

Theorem 4. *Let*

$$(1 + r)^2 > 4 p_2 b \tag{6}$$

Then there are no compact invariant sets contained in $\mathbf{R}^4_{+,0} \cap \{w > 0\}$.

We notice that the condition (6) holds according with parameters from Table 1 in [4].

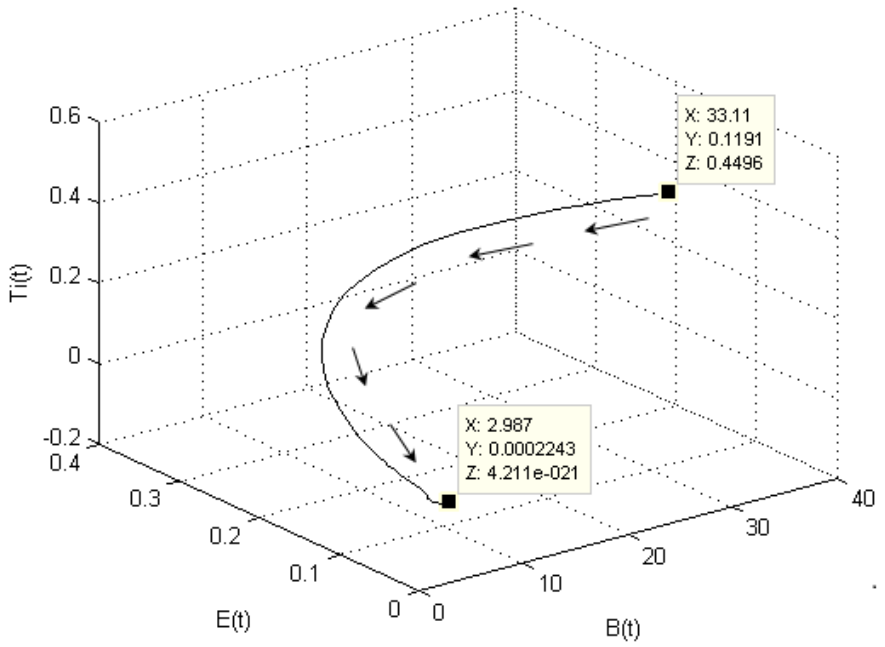


Fig. 1. Trajectory converging to an equilibrium point (b, 0,0,0).

6 Simulations

In order to simulate we took the parameters values of Table 1 in [4]. First we simulate the trajectory around one of the equilibrium point (b,0,0,0) presented also in [4], where the equilibrium point will depend of the value of the treatment. In Figure 1 is presented the flow of the trajectory to the equilibrium point. The initial conditions are (33.11 1.778 0.1191 0.4496).

In Figure 2 in presented the the same trajectory inside the polytope, this is, restricted by the upper bound of the parameters x,z,w.

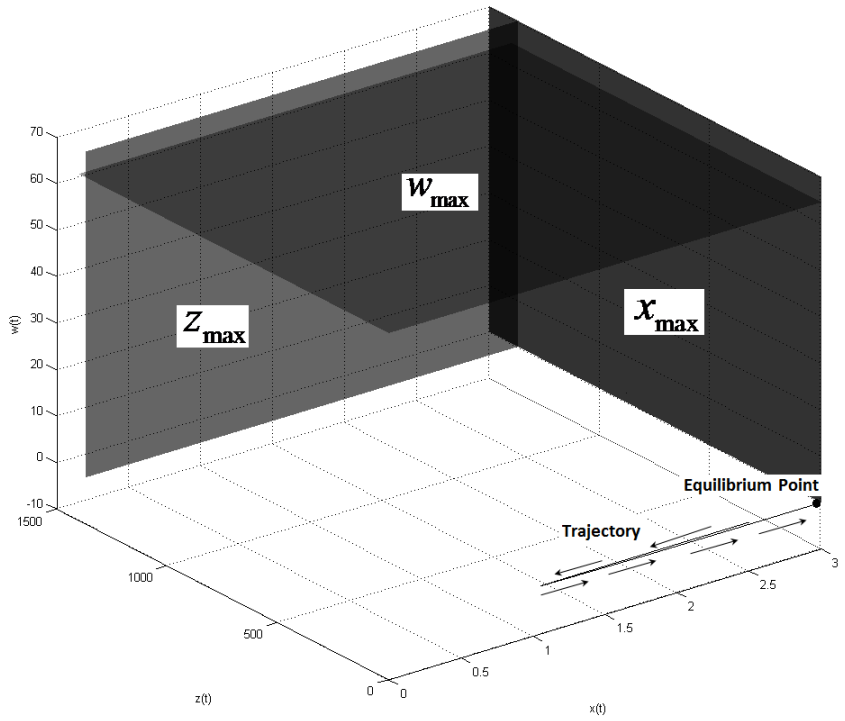


Fig. 2. Trajectory inside the positive polytope, $x_{max}, z_{max}, w_{max}$

7 Biological sense

The biological interpretation concerned to the mathematical analysis are the following:

1. The upper bound of the treatment BCG indicates the maximum carrying capacity of BCG infection to the tumor uninfected cells in the inner bladder.
2. The upper bound of the tumor logistic growth rate under the parameter value in vitro present the maximum carrying capacity of tumor growth of continuous feeding of glucose and oxygen for survival.
3. The upper bound of the effector cells will depend of the maximum bound of tumor infected cells by BCG, in order to eradicate the tumor in the bladder. In this case negligible the side effects by BCG.

8 Concluding remarks

In this paper we present results concerning upper bounds for all variables of the system (3) which characterize ultimate health conditions in the ideal situation of infinite time interval. These bounds are useful in studies of global dynamics of (3) and should be complemented by a proof that the system (3) has no escaping to infinity trajectories in $\mathbf{R}^4_+$. The corresponding work now is in a process.

References

1. Eftimie, R., Bramson, J, L., D., E.D.J.: Interactions between the immune system and cancer: a brief review of non-spartial mathematical models. Bulleting of Mathematical Biology (2010)
2. Babjuk, M., Oosterlinck, W., Sylvester, R., Kaasinen, E., Böhle, A., Palou-Redort, J.: Eau guidelines on non-muscle-invasive urothelial carcinoma of the bladder. Elsevier European Urology **54** (2008) 303–314
3. Braasch, M.R., Bohle, A., O'Donnell, M.A.: Intravesical instillation treatment of non–muscle-invasive bladder cancer. Elsevier European Urology **8** (2009) 549–555
4. Svetlana, B.M., Eliezer, S., Lewi, S.: Mathematical model of bcg immunotherapy in superficial bladder cancer. Bulletin of Mathematical Biology (2007)
5. Starkov, K.E.: Bounds for a domain containing all compactinvariantsets of the system describing the laser–plasma interaction. Chaos, Solitons and Fractals **39** (2009) 1671–1676
6. Coria, L.N., Starkov, K.E.: Bounding a domain containing all compactinvariantsets of the permanent-magnet motor system. Communications in Nonlinear Science and Numerical Simulation **14** (2009) 3879–3888
7. Krishchenko, A.P., Starkov, K.E.: Localization of compact invariant sets of the lorenz system. Physics Letters A **353** (2006) 383–388
8. Krishchenko, A.P., Starkov, K.E.: Estimation of the domain containing all compact invariant sets of a system modelling the amplitude of a plasma instability. Physics Letters A **367** (2007) 65–72
9. Kirschner, D., Panetta, J.: Modelling immunotherapy of the tumor-immune interaction. J. Math.Biol. **37** (1998) 235–252
10. Aranha, O., Wood, D., Sarkar, F.: Ciprofloxacin mediated cell growth inhibition, s/g2-m cell cycle arrest, and apoptosis in a human transitional cell carcinoma of the bladder cell line. Clin. Cancer Res. **6** (2000) 891–900
11. BÖHLE, A., BRANDAU, S.: Immune mechanisms in bacillus calmette-guerin immunotherapy for superficial bladdercancer. The Journal of Urology **170** (2003) 964–969

Robust sliding mode control for large scale wind turbine for power optimization

Jován O. Mérida[1], Luis T. Aguilar[1], and Jorge A. Dávila[2]

[1]Instituto Politécnico Nacional. Avenida del Parque 1310 Mesa de Otay, Tijuana, 22510 Baja California, México
[2]Instituto Politécnico Nacional. Av. Ticomán 600, Col. San José Ticomán, Delegación Gustavo A. Madero, 07340 Méx. D.F., México.
jmeridao0900@ipn.mx, laguilarb@ipn.mx
jadavila@ipn.mx
Paper received on 22/09/12, Accepted on 15/10/12.

Abstract. Wind turbine control must covered different objectives such as: power, speed, and load control to achieve a low cost of energy. In this paper, the problem of designing an output feedback control scheme of variable speed wind energy conversion system without wind speed measurement is addressed. The control objective is to track a wind speed profile to operate the wind turbine in maximum power extraction while reducing mechanical loads. In order to bring some improvement a combination of sliding-mode state feedback torque controller with wind speed estimator are derived. In order to validate the mathematical model and evaluate the performance of proposed controller in presence of disturbances and measurement noise, we used Matlab-Simulink. Simulation results show that the proposed control strategy is effective in terms of tracking speed, power extraction and load reductions in comparison with existing controllers.

Keywords: Renewable energy, nonlinear control, robust control, sliding mode, wind turbines, observer.

1 Introducción

As a result of population expansion and increased global integration, has been a great growth in energy consumption. This supposes a risk for the depletion of natural resources, this has caused the increase in demand of renewable energy generation systems [13]. Wind energy is currently one the fastest growing renewable energy technologies in the world implemented in over 80 countries [17]. The worldwide installed capacity of wind power for 2011 grew by 20.3%. The WWEA published this year the last updated version for wind turbines installed worldwide, with a total installed capacity of 237016 MW, enough to cover a 3% of the world's electricity defendant [17].

Wind turbines present great challenges at scientific level because they are complex systems, nonlinear and are subjected to parameters uncertainties, unmodeled dynamics and unknown disturbances. Ongoing research is focused on increasing energy efficiency and reducing mechanical stress. One solution is the use of advanced control strategies that enhance the performance of the turbine, this allows better use of resources of the

Mireya García-Vázquez, Grigori Sidorov (Eds.).
Advances in Computing Science and Control.
Research in Computing Science 59, 2012, pp. 65- 76

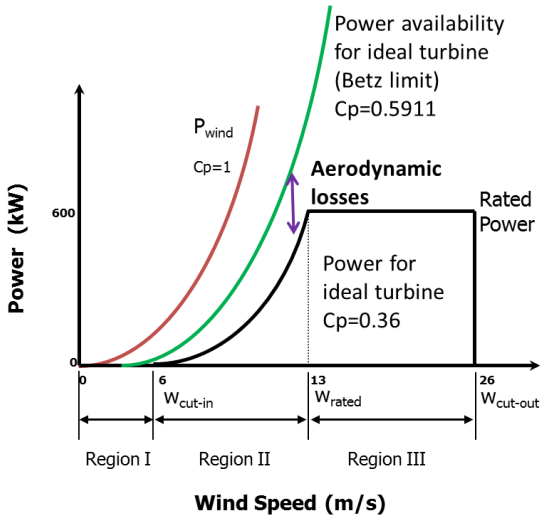


Fig. 1. Power curve for the CART.

turbine, increasing the lifetime of mechanical and electrical components, earning higher returns.

There are two primary types of horizontal-axis wind turbines: fixed speed and variable speed [15]. In this work we choose the variable speed because although the fixed speed system is easy to build and operate, does not have the ability that the variable speed system has in energy extraction, up to a 20-30% increase over fixed speed [15]. Wind turbine controller objectives depend on the operation area [16]. Variable speed wind turbine operation can be divided into three operating regions (Fig. 1):

- Region I: Below cut-in wind speed.
- Region II: Between cut-in wind speed and rated wind speed.
- Region III: Between rated wind speed and cut-out wind speed.

In region I wind turbines do not run, because power available in wind is low compared to losses in turbine system. Region II is an operational mode where it is desirable that the turbine capture as much power as possible from the wind, this because wind energy extraction rates are low and the structural loads are relatively small. Generator torque provides the control input to vary the rotor speed, while the blade pitch is held constant. Region III is encountered when the wind speeds are high enough that the turbine must limit the fraction of the wind power captured such that safe electrical and mechanical loads are not exceeded. If wind speeds exceed contains the region III, the system will make a forced stop the machine, protecting it from aerodynamic loads excessively high. Generally the rated rotor speed and power output are maintained by the blade pitch control with the generator torque constant at its rated value.

Region II is considered in the present work. Several control strategies have been proposed in the literature, mostly based on linear time-invariant models. This, for several

reasons. First, linear control theory is a well-developed topic while nonlinear control theory is less developed and difficult to implement. Second, most wind turbine control systems, to date, is based on linear control theory, thus the implemented wind turbine controllers are the based on linearized models [6]. Classical controllers have been extensively used, particularly PI control [11]. Another method commonly used is PID controllers. These PID controllers are used in conjunction with gain-scheduled to accommodate to variations in the wind [12]. Linear methods such as LQ, LQG and H_∞ are studied in [5]. Although some of these classical methods have been successfully applied, they are limited and problematic when extended to consider multiple controlled variables, such as controlling tower vibration, rotor speed, and blade vibration simultaneously (see, e.g., [16, 9, 18, 4]). Recently, nonlinear control of wind turbines has been of interest to the scientific community. A first-order sliding-mode controller for power regulation in Region III is developed in [2], demonstrating the viability and effectiveness of the control strategy. Beltran et al. [3] extended the control to Region II and III in conjunction with a Maximum Power Point Tracking algorithm, results show that the proposed control strategy is effective in terms of power capture and regulation reduction of the drive train mechanical stresses and output power fluctuations.

In this paper a strategy of sliding mode control for the problem of tracking the optimal rotor speed is developed. This control strategy presents attractive features, such as, robustness to parametric uncertainties of the turbine, robustness with respect to unknown disturbances and model uncertainties.

The paper is organized as follows. In Section 2 the wind turbine model and problem formulation is presented. The wind speed estimator is given in Section 3. The robust control design is provided in Section 4. Performance of the proposed controller is given in Section 5 through simulations. Section 6 presents some conclusions.

2 Wind Turbine Model and Problem Statement

2.1 Mathematical Model

The aerodynamic power captured by the rotor is given by the nonlinear expression [8]

$$P_a = \frac{1}{2}\rho\pi R^2 C_p(\lambda,\beta)v^3 \tag{1}$$

where v is the wind speed, ρ is the air density, and R is the rotor radius. The efficiency of the rotor blades is denoted as C_p, which depends on the blade pitch angle β, or the angle of attack of the rotor blades, and the tip speed ratio λ, the ratio of the blade tip linear speed to the wind speed. The parameters β and λ affect the efficiency of the system. The coefficient C_p is specific for each wind turbine. The relationship of tip speed ratio is given by

$$\lambda = R\frac{\omega_r}{v}. \tag{2}$$

The turbine estimated $C_p - \lambda - \beta$ surface, derived from simulation is illustrated in Fig. 2(a). This surface was created with the modeling software WTPerf [7], which uses blade-element-momentum theory to predict the performance of wind turbines [8]. The

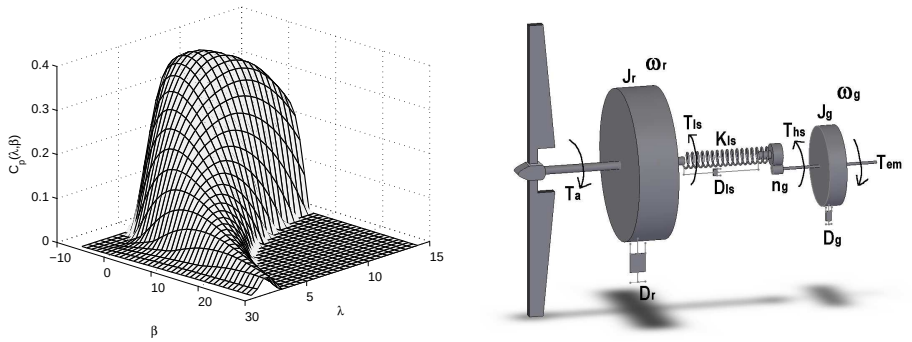


Fig. 2. (a) Power coefficient curve; (b) Two-mass model.

WTPerf simulation was performed to obtain the operating parameters for the CART (Controls Advanced Research Turbine). The wind turbine considered in this study is variable speed one, in which the rotor speed increases and decreases with changing wind speed, producing electricity with a variable frequency. (Fig. 2(b)). Fig. 2(a) indicates that there is one specific λ at which the turbine is most efficient. From (1) and (2), one can note that the rotor efficiency is highly nonlinear and makes the entire system a nonlinear system. The efficiency of power capture is a function of the tip speed ratio and the blade pitch. The power captured from the wind follows the relationship

$$P_a = T_a \omega_r \tag{3}$$

where

$$T_a = \frac{1}{2}\rho\pi R^3 \frac{C_p(\lambda,\beta)}{\lambda} v^2 \tag{4}$$

is the aerodynamic torque which depends nonlinearly upon the tip speed ratio. A variable speed wind turbine generally consists of an aeroturbine, a gearbox, and a generator, as shown in Fig. 2(b). The wind turns the blades generating an aerodynamic torque T_a, which spin a shaft at the speed ω_r. The low speed torque T_{ls} acts as a braking torque on the rotor. The gearbox, which increases the rotor speed by the ratio n_g to obtain the generator speed ω_g and decreases the high speed torque T_{hs}. The generator is driven by the high speed torque T_{hs} and braked by the generator electromagnetic torque T_{em} [1]. The mathematical model of the two-mass wind turbine, can be described as follows

$$\begin{aligned} J_r\dot{\omega}_r &= T_a(\omega_r,\beta,v) - K_{ls}(\theta_r - \theta_{ls}) - D_{ls}(\omega_r - \omega_{ls}) - D_r\omega_r \\ J_g n_g\dot{\omega}_g &= -T_{em}n_g + K_{ls}(\theta_r - \theta_{ls}) + D_{ls}(\omega_r - \omega_{ls}) - D_g n_g\omega_g \end{aligned} \tag{5}$$

where ω_{ls} is the low shaft speed, θ_r is the rotor side angular deviation, θ_{ls} is the gearbox side angular deviation, $J_r > 0$ is the rotor inertia, $J_g > 0$ is the generator inertia, $D_r > 0$ is the rotor external damping, $D_g > 0$ is the generator external damping, D_{ls}

Table 1. One-mass model parameters

Notation	Numerical value	Units
R	21.650	m
ρ	1.308	kg/m^3
J_t	3.920×10^5	kg m^2
D_t	400	Nm/rad/s
H	36.600	m
n_g	43.165	

is the low speed shaft damping, and K_{ls} is the low speed shaft stiffness. Assuming an ideal gearbox with transmission n_g

$$n_g = \frac{\omega_g}{\omega_{ls}} = \frac{T_{ls}}{T_{hs}}. \tag{6}$$

If a perfectly rigid low speed shaft is assumed, $\omega_r = \omega_{ls}$, a single mass model of the turbine can then be considered, upon using (6) and (5), one gets:

$$J_t\dot{\omega}_r = T_a(\omega_r, \beta, v) - D_t\omega_r - T_g \tag{7}$$

where $J_t = J_r + n_g^2 J_g$, $D_t = D_r + n_g^2 D_g$, and $T_g = n_g T_{em}$ are the turbine total inertia, turbine total external damping, and generator torque in the rotor side, respectively. The parameters of the model are given in Table 1. Those parameters are based on the CART which is a two-bladed, teetered, active-yaw, upwind, variable speed, variable pitch, horizontal axis wind turbine which is located at the National Wind Technology Center in Colorado. The nominal power is 600 kW, the rated wind speed of 13 m/s, a cut out wind speed of 26 m/s, and it has a maximum power coefficient $C_{pmax} = 0.3659$. The rated rotor speed is 41.7 rpm. The pitch system can pitch the blades up to 18 deg/s with pitch accelerations up to 150 deg/s^2 [19]. The required constraints for torque and rotor speed are 162 kNm and 58 rpm respectively. The gearbox is connected to an induction generator via the high speed shaft, and the generator is connected to the grid via power electronics. In this work we will ignore the power electronics control and an ideal performance will be assumed [18].

2.2 Problem Statement

The main objective in the region II is to maximize the power extracted from the wind. While energy is captured from the wind, the aerodynamic power should be maximized below rated wind speed. In (2) the tip speed ratio can be altered to include the optimized points shown in (8), this leads to a unique maximum point (see (9)) that corresponds to a maximum power production, that is

$$\lambda_{opt} = R\frac{\omega_{ropt}}{v}, \tag{8}$$

$$C_p(\lambda_{opt}, \beta_{opt}) = C_{p\max}. \tag{9}$$

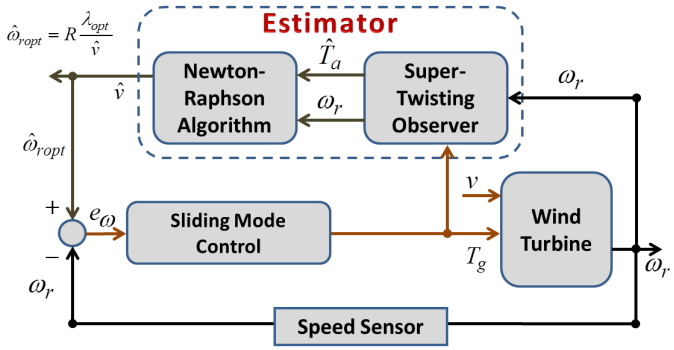


Fig. 3. Proposed control scheme.

To maximize the extracted energy the maximum rotor efficiency must be maintained during operation. For this β is fixed to β_{opt} and ω_{ropt} must change depending on the wind speed variations

$$\omega_{ropt} = R\frac{\lambda_{opt}}{v}. \tag{10}$$

Then, the objective control is to find a control law T_g to track the optimal rotor speed ω_{ropt} while the loads on the turbine are reduced. The controller should take into consideration the nonlinear nature of the wind turbine behavior, the flexibility of drive train, as well as the turbulent nature of the wind. There are numerous generator torque controllers. In the aim of making a comparison between the proposed and a existing control law, a brief description of this last one is given below. In [1] the next nonlinear static state feedback control (NSSFC) is done.

$$T_g = T_a - D_t\omega_r - J_t\dot{\omega}_{ropt} - J_t a_0 e_\omega. \tag{11}$$

The control technique presented above shows a main drawback: it is not so robust with respect to perturbation. The proposed control strategy, therefore, shall overcome this problem in order to have a better performance.

3 Wind Speed Estimation

An estimator of the wind speed is developed using the wind turbine itself as a measuring device as it is crucial for deriving the optimal rotor speed ω_{ropt}. As is shown in Fig. 3, the estimator is composed of two blocks [1].

1) A first block, that allows to estimate, from the rotor speed measurement ω_r and the generator control torque T_g, the value of the aerodynamic torque T_a.

2) A second block with, as input, the estimate of the aerodynamic torque $\hat{T}_a$. The block output is the effective wind speed estimate $\hat{v}$.

3.1 Aerodynamic Torque Estimation

In order to construct a robust output feedback controller that recovers the performance of the state feedback, Super-Twisting Observer (STO) is designed [14], which is able to reconstruct the state and uncertain functions. These estimated values are used in the controller instead of the true ones. The controller requires aerodynamic torque which cannot be easily measured. According to wind turbine dynamics the rotor speed ω_r and aerodynamic torque T_a are related by (7), the unknown term is T_a and ω_r is the measured variable. It is clear that the relative degree between them is one so the structural requirement to implement sliding-mode observer, in this case to the relative degree is one, which allows to reconstruct T_a. To estimate T_a by means of the measurement of the rotor speed ω_r, the following observer is proposed

$$\begin{aligned} \dot{\hat{\omega}}_r &= \hat{x} - \frac{1}{J_t}\left(D_t\omega_r + T_g\right) - k_1\left|e_{\omega_r}\right|^{\frac{1}{2}} sign\left(e_{\omega_r}\right) - k_2 e_{\omega_r} \\ \dot{\hat{x}} &= -k_3 sign\left(e_{\omega_r}\right) - k_4 e_{\omega_r} \end{aligned} \tag{12}$$

where $e_{\omega_r} = \hat{\omega}_r - \omega_r$ is the measurement error, $k_1 - k_4$ are adjustable gains, and $\hat{T}_a = J_t\hat{x}$ is the estimation of the aerodynamic torque. Choosing the parameter of the observer (12) according to [14] then its state and unknown term converge in finite time to ω_r and T_a respectively.

3.2 Wind Speed Computation

The estimate of the wind speed $\hat{v}$ is related to the one of $\hat{T}_a$ by the following equation:

$$\hat{T}_a - \frac{1}{2}\rho\pi R^3 C_q\left(\frac{\hat{\omega}_r R}{\hat{v}}\right)\hat{v}^2 = 0 \tag{13}$$

where $C_q(\hat{\lambda}) = C_q(\hat{\lambda}, \beta_{opt})$ is a tabulated function of $\hat{\lambda}$. In order to use a numerical method for (13) solved with respect to $\hat{v}$, this function is interpolated with a polynomial in λ

$$C_q(\lambda) = \sum_{i=0}^{n} \alpha_i \lambda^i. \tag{14}$$

The Newton-Raphson algorithm is then used to calculate $\hat{v}$. This value is exploited to deduce the optimal rotor speed $\hat{\omega}_{ropt} = \lambda_{opt}\hat{v}/R$.

4 Robust Control with Estimator

In this section a combination of sliding mode controller (SMC) with a estimator is presented (Fig. 3). The proposed controller will track the wind speed in order to achieve $\hat{\omega}_{ropt}$. A sliding manifold is chosen as follows

$$e_\omega = \hat{\omega}_{ropt} - \omega_r \tag{15}$$

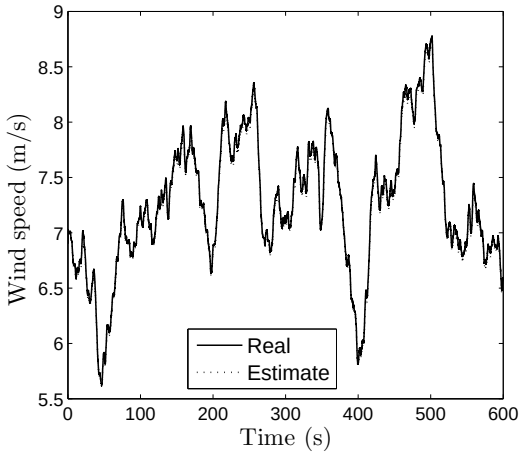


Fig. 4. Wind speed profile of $v_m = 7$ m/s mean value.

where e_ω is the rotor speed error. Here the controller is developed to achieve robust speed tracking. For that purpose we impose a first-order dynamics to e_ω

$$\dot{e}_\omega + c_0\omega_r = 0 \tag{16}$$

where $c_0 > 0$ then developing (16) one gets

$$J_t\dot{\hat{\omega}}_{ropt} + J_t c_0 e_\omega + D_t\omega_r + T_g - \hat{T}_a = 0. \tag{17}$$

The following controller (18) is designed for (17)

$$T_g = \hat{T}_a - D_t\omega_r - J_t\dot{\hat{\omega}}_{ropt} - J_t c_0 e_\omega - J_t k_s sign(e_\omega) \tag{18}$$

where $k_s > 0$.

5 Simulation Results

The wind speed is described as a slowly varying average wind speed superimposed by a rapidly varying turbulent wind speed. The model of the wind speed v at the measured point is

$$v = v_m + v_t \tag{19}$$

where v_m is the mean value and v_t is the turbulent component. The wind field was generated following [10]. The turbulence v_t is being modeled as a 2nd order, linear process

$$\begin{aligned} \dot{w}_1 &= w_2 \\ \dot{w}_2 &= -\frac{p_1 + p_2}{p_1 p_2} w_2 - \frac{1}{p_1 p_2} w_1 + \frac{k}{p_1 p_2} e \end{aligned} \tag{20}$$

where $e \in \mathcal{N}(0,1)$ is a noise process with intensity $k/(p_1 p_2)$, p_1, p_2, k are parameters depending on the mean wind speed.

This turbine was modeled in Matlab-Simulink. Simulations were performed under the next operating conditions: in presence of a constant additive control input disturbance of 500 Nm, an additive measurement noise on ω_r with a SNR around 7 dB, and wind speed profile of v_m =7 m/s with turbulence intensity of 15%. As seen in Fig. 4, the strategy gives a good wind speed estimation, this allows to get a better rotor speed reference.

The results show that with the proposed approach, power increases slightly and decreases the loads when is compared with the control (11). The rotor speed (Fig. 5) with the proposed controller tracks a little more closely the optimal rotor speed ω_{ropt} leading to more power capture, also the dynamic characteristics improve with slightly lower mechanical stresses as illustrated in Fig. 6. Figs. 7 and 8 show that the power extraction obtained with our strategy is better.

6 Conclusions

This paper addresses the problem of power generation control in variable speed wind turbines. The objectives are: synthesizing a robust controller to maximize the energy extracted from the wind, while ensuring reduction of mechanical loads. To this end, a strategy of sliding mode control with an estimate of the wind speed was proposed. The developed estimator allows the estimation of the aerodynamic torque as well as the effective wind speed from noisy measurements. The proposed controller provides a suitable compromise between conversion efficiency and mechanical stresses, also has a better perturbation rejections in comparison with existing controllers. The control strategy has been validated with an aeroelastic wind turbine simulator and the results shown the feasibility of the proposed strategy.

References

1. Nonlinear Control of Variable Speed Wind Turbines without wind speed measurement (Dec 2005)
2. Beltran, B., Ahmed-Ali, T., Benbouzid, M.: Sliding mode power control of variable speed wind energy conversion systems. In: Electric Machines Drives Conference, 2007. IEMDC '07. IEEE International. vol. 2, pp. 943 –948 (May 2007)
3. Beltran, B., Ahmed-Ali, T., El Hachemi Benbouzid, M.: Sliding mode power control of variable-speed wind energy conversion systems. Energy Conversion, IEEE Transactions on 23(2), 551 –558 (June 2008)
4. Bianchi, F.D., de Battista, H., Mantz, R.J.: Wind turbine control systems: principles, modelling and gain scheduling design. Advances in Industrial Control, Springer-Verlag, London Limited, 1st edn. (2007)
5. Bossanyi, E.A.: Individual blade pitch control for load reduction. Wind Energy 6(2), 119–128 (2003), http://dx.doi.org/10.1002/we.76
6. Boukhezzar, B., Lupu, L., Siguerdidjane, H., Hand, M.: Multivariable control strategy for variable speed, variable pitch wind turbines. Renewable Energy 32(8), 1273–1287 (2007)

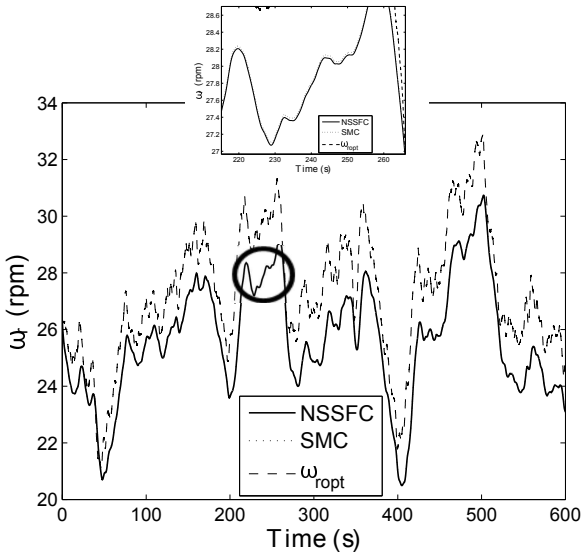


Fig. 5. Closed-loop system responses: rotor speed.

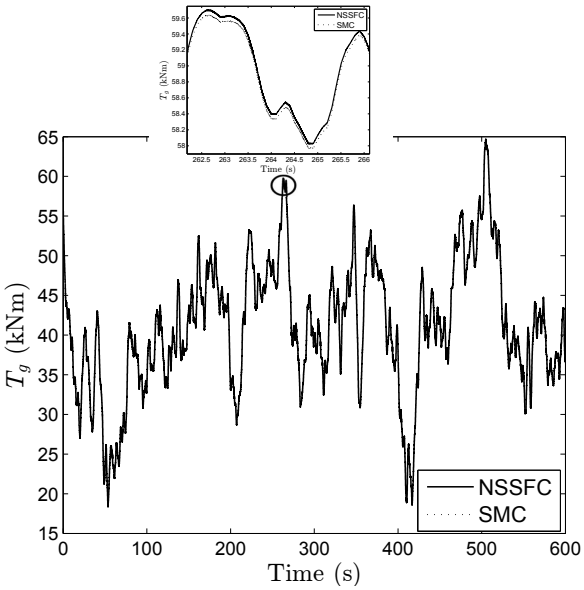


Fig. 6. Closed-loop system responses: generator torque.

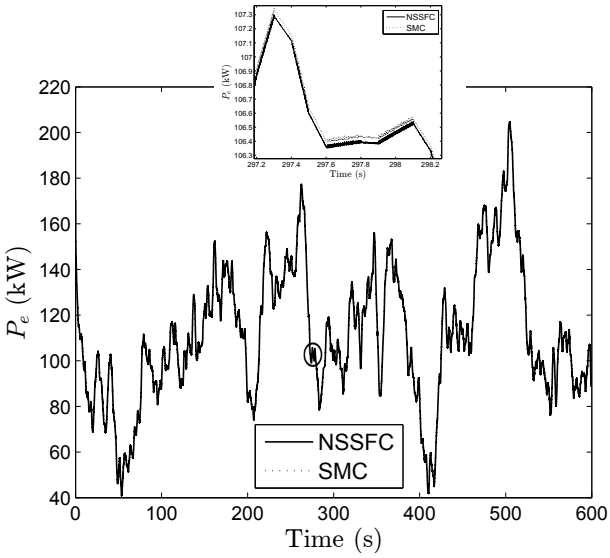


Fig. 7. Closed-loop system responses: electrical power.

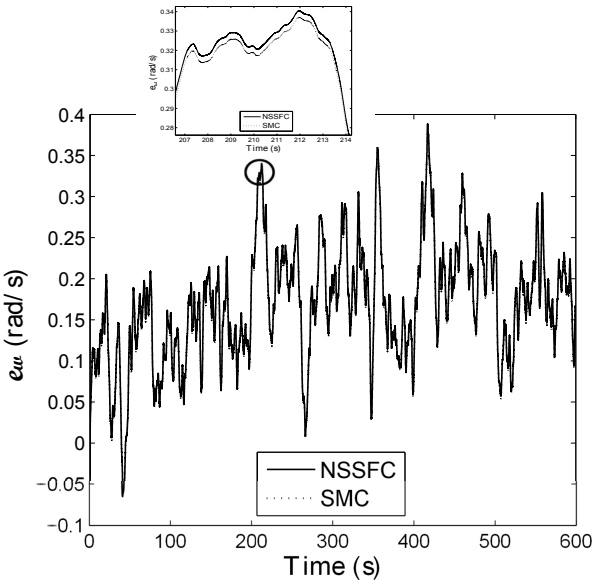


Fig. 8. Closed-loop system responses: rotor speed error.

7. Buhl, M.: NWTC design codes WTPerf. [Online]. Available: http://wind.nrel.gov/designcodes/simulators/wtperf/ (2009)
8. Burton, T., Jenkins, N., Sharpe, D., Bossanyi, E.: Wind Energy Handbook. John Wiley & Sons, 2da edn. (2011)
9. Grimble, M.: Horizontal axis wind turbine control: comparison of classical, LQG and H_∞ designs. Dynamics and Control 6(2), 143–161 (April 1996)
10. Hammerum, K.: A fatigue approach to wind turbine control. Me thesis, Informatics and Mathematical Modelling, Technical University of Denmark, DTU, Richard Petersens Plads, Building 321, DK-2800 Kgs. Lyngby (2006), http://www2.imm.dtu.dk/pubdb/p.php?4980
11. Hand, M.M., Balas, M.J.: Non-linear and linear model based controller design for variable-speed wind turbines. In: 3rd ASME/JSME Joint Fluids Engineering Conference. San Francisco, California (1999)
12. Hansen, M., Hansen, A., Larsen, T., Øye, S., Sørensen, P., Fuglsang, P.: Control design for a pitch regulated, variable speed wind turbine. Tech. Rep. Risø-R-1500(EN), Risø National Laboratory, Denmark (2005)
13. Masters, G.M.: Renewable and efficient electric power systems. John Wiley & Sons (2004), http://books.google.com.mx/books?id=NFb_mM580nAC
14. Moreno, J.A.: A linear framework for the robust stability analysis of a generalized super-twisting algorithm. In: Proc. 6th. Int. Conf. Elect. Eng., Comp. Sci. and Aut. Cont. (CCE 2009). pp. 12–17. Mexico (Nov 2009)
15. Ofualagba, G., Ubeku, E.U.: Wind energy conversion system- wind turbine modeling. In: IEEE Power and Energy Society General Meeting - Conversion and Delivery of Electrical Energy in the 21st Century. pp. 1–8 (July 2008)
16. Pao, L.Y., Johnson, K.E.: Control of wind turbines. IEEE Control Systems Magazine 31(2), 44–62 (April 2011)
17. The World Wind Energy Association: World wind energy report 2011. Tech. rep. (2012)
18. Thomsen, S.: Nonlinear Control of a Wind Turbine. Me thesis, Lyngby: Informatik og Matematisk Modellering, Danmarks Tekniske Universitet (2006)
19. Wright, A.D., Fingersh, L.J.: Advanced control design for wind turbines. part I: Control design, implementation, and initial tests. Tech. Rep. NREL/TP-500-42437, NREL (March 2008)

On the control of input–constrained boost DC–to–DC power converters ⋆

Jorge Guzmán–Guemez and Javier Moreno–Valenzuela⋆⋆

Instituto Politécnico Nacional–CITEDI,
Av. del Parque 1310, Mesa de Otay, Tijuana, B.C., Mexico
{jguzman, moreno}@citedi.mx
http://www.citedi.mx
Paper received on 22/09/12, Accepted on 21/10/12.

Abstract. Boost DC–to–DC (direct current–to–direct current) power converters are useful in many applications as part of devices used at home and industry. The theoretical and practical study of boost DC–to–DC power converters has prompted the attention of many researchers. In this paper, a new controller for Boost DC–to–DC power converters is proposed. The new scheme takes into account that the duty cycle is constrained to physically admissible values. The analysis of the closed–loop trajectories provides the conclusion that output voltage regulation is achieved in asymptotic form. Our theoretical results are supported by using numerical simulations and real–time experiments.

1 Introduction

The problem of regulating the output voltage of a boost DC–to–DC (direct current–to–direct current) power converter has attracted the attention of many control researchers for several years now. Besides its practical relevance, the system is an interesting theoretical case study because it is a switched device whose averaged dynamics are described by a bilinear second order non–minimum phase system with saturated input [1].

In order to provide a degree of robustness to compensate uncertainties in the load, supply voltage and unmodeled disturbances, many control algorithms have devised to achieve voltage output regulation.

A brief literature review is provided in the next. In the textbooks [2] and [3] a number of algorithms for the boost DC–to–DC power converter are analyzed, but none of them deals with the practical situation that the duty cycle input should be into admissible values. The work of Spinetti, Fossas and Biel in [4] presented an interesting controller which requires feedback of the inductor current and output voltage. The scheme proposed there provides a very fast convergence of the output voltage to the desired one. However, two main disadvantages are detected in this controller: the assumption that all the boost parameters should be known and the assumption that the control input may take any real value, which is not possible in practice. In the paper by Karagiannis,

⋆ Work supported by CONACyT, project number 176587, and SIP–IPN, Mexico.
⋆⋆ Author to whom correspondence should be addressed.

Mireya García-Vázquez, Grigori Sidorov (Eds.).
Advances in Computing Science and Control.
Research in Computing Science 59, 2012, pp. 77- 86

Astolfi and Ortega in [5] proposed an algorithm that leaves aside the inductor current feedback. A certain degree of robustness is provided thanks to adaptation of the supply voltage. Other advantage of this scheme is that stability is guaranteed even if duty cycle saturation occurs.

In spite of the fact that the algorithms proposed in the literature achieve the control objective, only a few of them takes into account the practical limitation that the percentage of duty cycle $u(t)$ is limited to

$$u(t) \in [0, 1). \tag{1}$$

In other words, no much attention has been devoted to design control schemes that guarantee output voltage regulation of a input–saturated boost converter. See for instance the work in [5].

In this paper, a new controller is proposed. The new scheme takes into account the physical constraint that the duty cycle input $u(t)$ should satisfy (1). The analysis of the closed–loop trajectories provides the conclusion that output voltage regulation is achieved in asymptotic form. Our theoretical results are supported by using numerical simulations and real–time experiments.

In ideal conditions (assuming that all the boost DC–to–DC converter parameters are known and that there are not disturbances) the new scheme guarantees global convergence of the output voltage to the desired one, while the generated duty cycle control input stays into the practical admissible limits.

Although in presence of model disturbances, the algorithm is not able to achieve output voltage regulation, which is corroborated by simulation and experiment, the proposed methodology is promising since several extensions can be developed to improve its performance and robustness.

The present document is organized as follows. Section 2 deal with the boost DC–to–DC dynamic model. The new scheme and the analysis of the closed–loop trajectories are presented in Section 3. Section 4 is devoted to numerical tests, while Section 5 is concerned to real–time experimental results. Finally, some concluding comments are drawn is Section 6.

2 Boost DC–to–DC converter model and control goal

2.1 Boost DC–to–DC model converter

Consider the switch–regulated boost converter circuit of Figure 1. The positive quantity E represents the external voltage supply, $i(t)$ is the current through inductor $L > 0$, $v(t)$ is the voltage through capacitor C, and R the load resistor. The signal u_s takes values in the discrete set $\{0, 1\}$.

In order to represent this switch–regulated circuit, the following average non–linear equation system can be obtained using circuit analysis via Kirchhoff laws

$$L\frac{di}{dt} = -[1-u]v + E, \tag{2}$$

$$C\frac{dv}{dt} = [1-u]i - \frac{v}{R}, \tag{3}$$

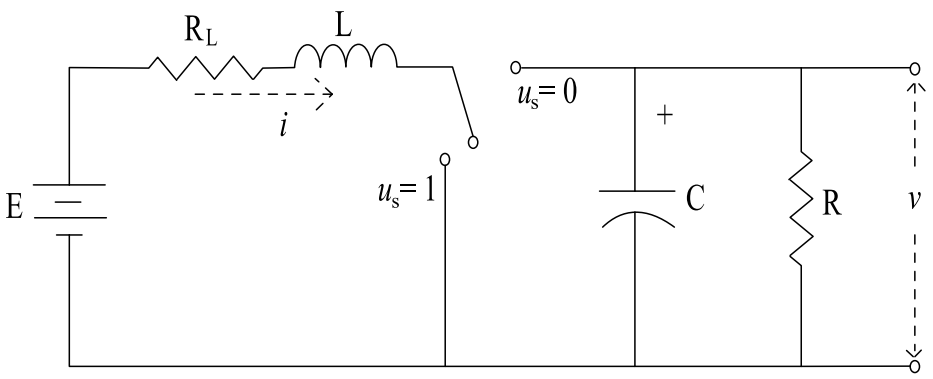


Fig. 1. The boost converter circuit.

where $u(t)$ is a continuous control signal representing the duty cycle percentage of the PWM circuit controlling the switch. See the textbooks [2] and [3] for further details in the modeling of the boost DC–to–DC converter.

2.2 Control goal

It is easy to show that an equilibrium point of the system (2)–(3), assuming a constant duty cycle percentage $u(t) = u_e$, is given by

$$i = i_d = \frac{v_d}{[1 - u_e]R},$$
$$v = v_d = \frac{E}{1 - u_e}.$$

Solving for u_e, we have

$$u_e = 1 - \frac{E}{v_d},$$

where v_d is the desired voltage. In practice, the actual duty cycle percentage $u(t)$ should satisfy

$$0 < u_e < 1,$$

therefore $v_d > E$. Notice that $u(t) = u_e = 1$ results in an undefined equilibrium point i_d and v_d.

The control goal consists in the specification of the desired output voltage $v_d > E$ and the design of a control law $u(t)$ satisfying the constraint (1) such that the output voltage achieves

$$\lim_{t \to \infty} v(t) = v_d, \tag{4}$$

while the inductor current $i(t)$ remains bounded.

3 Proposed scheme, analysis and extensions

3.1 Proposed scheme

In consideration to the control goal established above, the controller proposed in this paper is given by

$$u = 1 - sat\left(\frac{E}{V_d} + \gamma[V_d e_i - i_d e_v]\right), \tag{5}$$

where $\gamma > 0$ is a constant,

$$e_i = i - i_d, \tag{6}$$
$$e_v = v - V_d, \tag{7}$$

are the current error and voltage error, respectively, and

$$sat(z) = \begin{cases} 1 - \xi_m, \text{ if } z < \xi_m, \\ 1 - z, \text{ if } \xi_m \leq z \leq \xi_M, \\ 1 - \xi_M, \text{ if } z > \xi_M, \end{cases} \tag{8}$$

with

$$0 < \xi_m < \xi_M < 1.$$

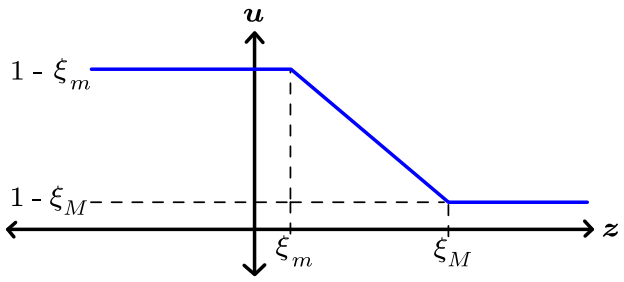


Fig. 2. Profile of the function $sat(z)$.

The profile of the saturation function $sat(z)$ in (8) is depicted in Figure 2. It is worthwhile to notice that by using the definition of the saturation function (8), it is possible to show that the duty cycle percentage $u(t)$ satisfies the following inequality

$$1 - \xi_M \leq u(t) \leq 1 - \xi_m.$$

Therefore, the constraint (1) is satisfied for all $t \geq 0$.

It is noteworthy that the number ξ_m and ξ_M are chosen so that

$$u = 1 - sat\left(\frac{E}{V_d}\right) = 1 - \frac{E}{V_d} = u_e,$$

which is always satisfied for

$$0 < \xi_m \leq \frac{E}{v_d} \leq \xi_M < 1. \tag{9}$$

3.2 Analysis

Let us define

$$x = v_d e_i - i_d e_v \tag{10}$$

in order to simplify the notation.

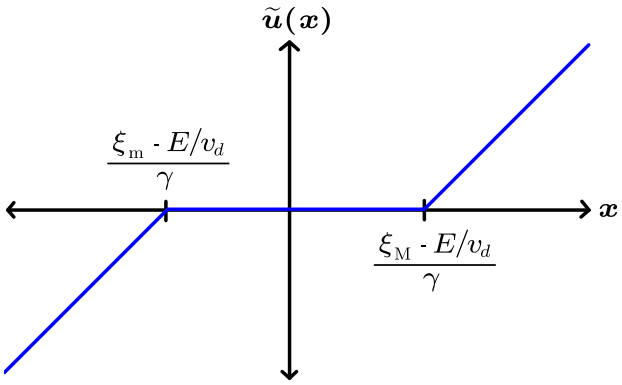


Fig. 3. Profile of the function $\tilde{u}(x)$.

Substituting (6)–(7) and the duty cycle percentage input $u(t)$ in (5) into the boost converter dynamics (2)–(3), the closed–loop system can be written as

$$L\frac{de_i}{dt} = -\gamma x[e_v + v_d] + \tilde{u}(x)[e_v + v_d] - \frac{E}{v_d}e_v, \tag{11}$$

$$C\frac{de_v}{dt} = \gamma x[e_i + i_d] - \tilde{u}(x)[e_i + i_d] + \frac{E}{v_d}e_i - \frac{1}{R}e_v, \tag{12}$$

where

$$\tilde{u}(x) = \frac{E}{v_d} + \gamma x - sat\left(\frac{E}{v_d} + \gamma x\right) \tag{13}$$

is a dead zone–type nonlinear function. The profile of the function $\tilde{u}(x)$ is illustrated in Figure 3. Even more, $\tilde{u}(x)$ can be explicitly given as

$$\tilde{u}(x) = \begin{cases} \frac{E}{v_d} + \gamma x - \xi_M, & \text{for } \frac{E}{v_d} + \gamma x > \xi_M, \\ 0, & \text{for } \xi_m \leq \frac{E}{v_d} + \gamma x \leq \xi_M, \\ \frac{E}{v_d} + \gamma x - \xi_m, & \text{for } \frac{E}{v_d} + \gamma x < \xi_m. \end{cases} \tag{14}$$

See Figure 3 for a plot of the function $\tilde{u}(x)$ in (14).

It is possible to show that the state space origin $[e_i\ e_v]^T = [0\ 0]^T$ is an equilibrium point of the closed–loop (11)–(12).

Now, let us introduce the following Lyapunov function candidate

$$W = \frac{1}{2}Le_i^2 + \frac{1}{2}Ce_v^2, \tag{15}$$

which is positive definite and radially unbounded. The time–derivative of W along of the closed–loop system trajectories (11)–(12) is given by

$$\dot{W} = -\gamma x^2 - \frac{e_v^2}{R} + \tilde{u}(x)x, \tag{16}$$

which is obtained after some direct algebra.

Now, in order to find a upper bound on $\dot{W}$ let us write the product $\tilde{u}(x)x$ explicitly as follows

$$\tilde{u}(x)x = \begin{cases} [\frac{E}{v_d} - \xi_M]x + \gamma x^2, & \text{for } x > \frac{\xi_M - E/v_d}{\gamma}, \\ 0, & \text{for } \frac{\xi_m - E/v_d}{\gamma} \leq x \leq \frac{\xi_M - E/v_d}{\gamma}, \\ [\frac{E}{v_d} - \xi_m]x + \gamma x^2, & \text{for } x < \frac{\xi_m - e/v_d}{\gamma}. \end{cases} \tag{17}$$

Similarly, the profile of $\tilde{u}(x)x$ is shown in Figure 4.

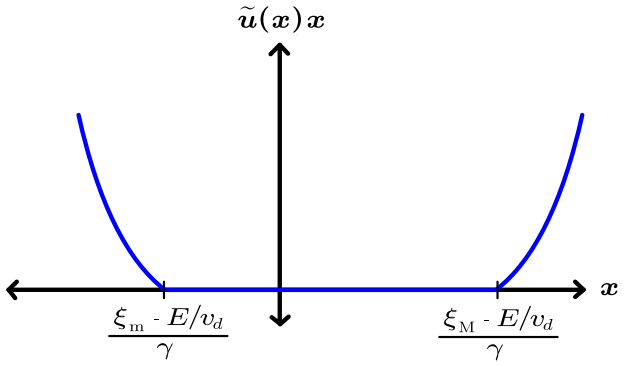


Fig. 4. Profile of the function $\tilde{u}(x)x$.

Taking into account the explicit definition of the function $\tilde{u}(x)x$ in (17), the time derivative of Lyapunov function candidate $\dot{W}$ in (16) can be expressed as

$$\dot{W} = \begin{cases} [\frac{E}{v_d} - \xi_M]x - \frac{e_v^2}{R}, & \text{for } x > \frac{\xi_M - E/v_d}{\gamma}, \\ -\gamma x^2 - \frac{e_v^2}{R}, & \text{for } \frac{\xi_m - E/v_d}{\gamma} \leq x \leq \frac{\xi_M - E/v_d}{\gamma}, \\ [\frac{E}{v_d} - \xi_m]x - \frac{e_v^2}{R}, & \text{for } x < \frac{\xi_m - E/v_d}{\gamma}, \end{cases} \tag{18}$$

Notice that in the sector $x > \frac{\xi_M - E/v_d}{\gamma}$, where $x > 0$, the inequality

$$[\frac{E}{v_d} - \xi_M]x < 0$$

is accomplished because condition (9). Therefore, $\dot{W} = [\frac{E}{v_d} - \xi_M]x - \frac{e_v^2}{R} \leq -\frac{e_v^2}{R}$. Proceeding in a similar form in the other two sectors, it is possible to show that a upper bound of $\dot{W}$ in (18) is given by

$$\dot{W} \leq -\frac{e_v^2}{R}, \ \forall \ \begin{bmatrix} e_i \\ e_v \end{bmatrix} \in \mathbb{R}^2, \tag{19}$$

which is globally negative semidefinite, and implies that the state space origin $[e_i \ e_v]^T = [0 \ 0]^T$ is stable in the Lyapunov sense [6]. Besides, W in (15) is a Lyapunov function of the system (11)–(12)

Finally, since the system (11)–(12) is autonomous, by invoking LaSalle's theorem, the proof that the state space origin $[e_i \ e_v]^T = [0 \ 0]^T$ is globally asymptotically stable can be achieved. Then,

$$\lim_{t \to \infty} \begin{bmatrix} e_i(t) \\ e_v(t) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \tag{20}$$

and the control goal (4) is satisfied.

4 Numerical simulations

In order to illustrate the above analysis, simulations have been carried out assuming that the boost converter is a continuous system and the controller $u(t)$ in (5) is implemented in discrete time with a sample time $T_s = 0.0001$ [s], which produces a sampled–data system.

The following values were chosen: $L = 5$ [mH], $C = 12$ [μF] and $R = 182$ [Ω]. Besides, we selected $\xi_m = 0.1$, $\xi_M = 0.9$ and $\gamma = 0.1$.

The initial conditions of the system were

$$i(0) = 0.0598 \text{ [A] and } v(0) = 9.7440 \text{ [V]}. \tag{21}$$

We carried out a simulation with a duration of 0.01 [s], specifying a desired voltage v_d = 15 $[V]$. The results are observed in Figure 5, where the time evolution of the output voltage $v(t)$, the inductor current $i(t)$ and the duty cycle percentage $u(t)$ are appreciated. The capacitor voltage $v(t)$ reach v_d in approximately 0.003 [s], enough time for $u(t)$ to reach a practically constant value.

Now, results shown in Figure 6, consider that $v_d(t)$ is a square periodic signal defined as

$$v_d(t) = \begin{cases} 15 \text{ [V]}, & \text{for } 0 \leq t \leq 0.5T_r, \\ 20 \text{ [V]}, & \text{for } 0.5T_r \leq t \leq T_r, \end{cases} \tag{22}$$

with $T_r = 1.0$ [s] defining the period of v_d. In this case the simulation had a duration of 5 [s]. Similarly, Figure 6 depicts the time evolution of the output voltage $v(t)$, the inductor current $i(t)$ and the duty cycle percentage $u(t)$. It is observed that the output voltage $v(t)$ converges to v_d.

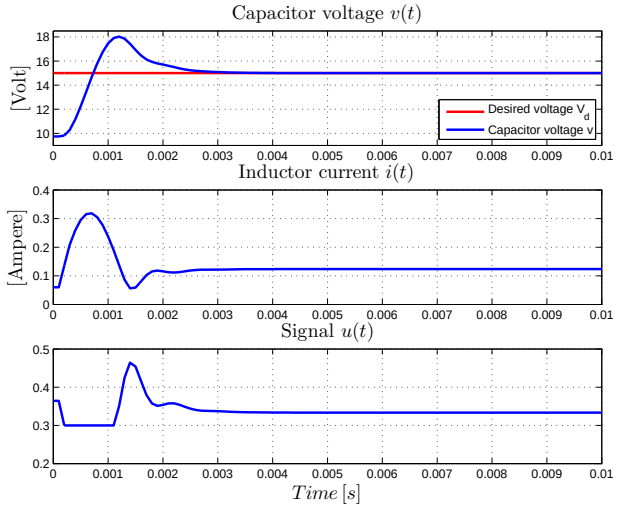


Fig. 5. Simulation: Capacitor voltage $v(t)$, inductor current $i(t)$ and control signal $u(t)$ for $0 \leq t \leq 0.01$ [s] and $v_d = 15$ [V].

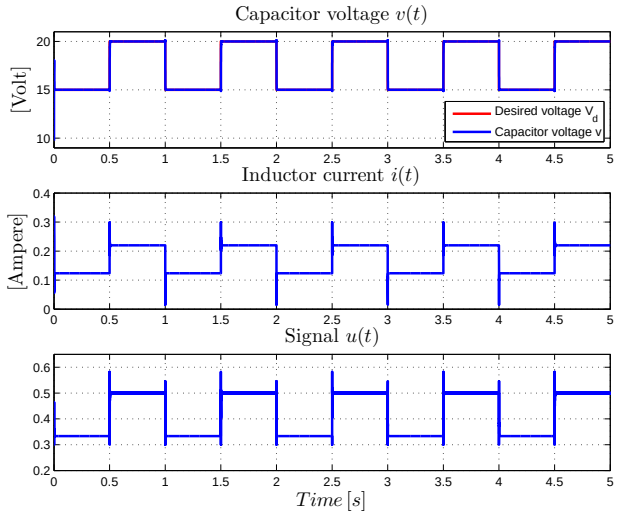


Fig. 6. Simulation: Capacitor voltage $v(t)$, inductor current $i(t)$ and control signal $u(t)$ for $0 \leq t \leq 5$ [s] and $v_d(t)$ in (22).

5 Experimental results

Laboratory experiments have taken place under similar conditions than in the numerical simulations. The switching frequency of the PWM was 50 $[kHz]$. We choose $\xi_m = 0.1$, $\xi_M = 0.9$ and γ = 0.5.

Figure 7 shows that the rise time is almost 0.1 [s], much more time than the one shown in the simulation results; compare Figure 5 with respect to Figure 7.

The control signal $u(t)$ is saturated in very short periods of time. However, there is no any operation problem since the output voltage $v(t)$ and the inductor current $i(t)$ remain bounded for all time.

We also present in Figure 8 the output voltage $v(t)$ and the inductor current $i(t)$ when $v_d(t)$ in (22) is used. Besides, Figure 8 shows also the time evolution of the control input $u(t)$.

In Figure 8 it is appreciated a steady state error of 0.35 [V] when $v_d = 15$ [V], and 0.70 [V] when $v_d = 20$ [V]. In addition to the discrete implementation of the controller, there are other model uncertainties and disturbances, such as deviations in the values of the inductance L, capacitance C and load resistance R. Besides, the inductor has a significative resistance which also causes the steady state error.

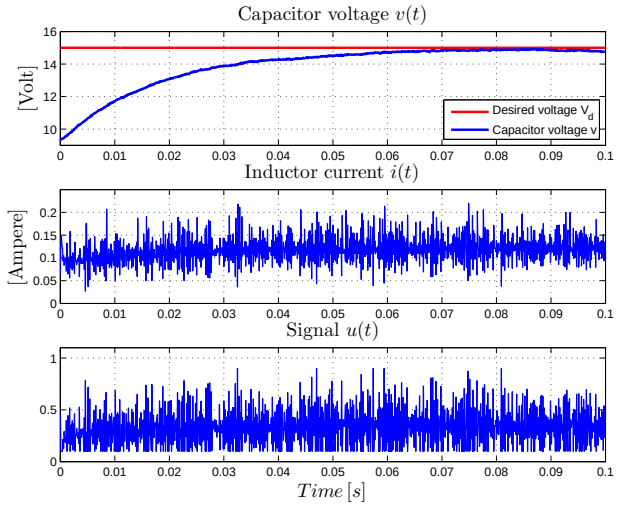


Fig. 7. Experiment: Capacitor voltage $v(t)$, inductor current $i(t)$ and control signal $u(t)$ for $0 \leq t \leq 0.01$ [s] and $v_d = 15$ [V].

6 Conclusions

A new control scheme was proposed in this paper. The new controller is based in a hard saturation function in order to keep the duty cycle percentage into admissible physical

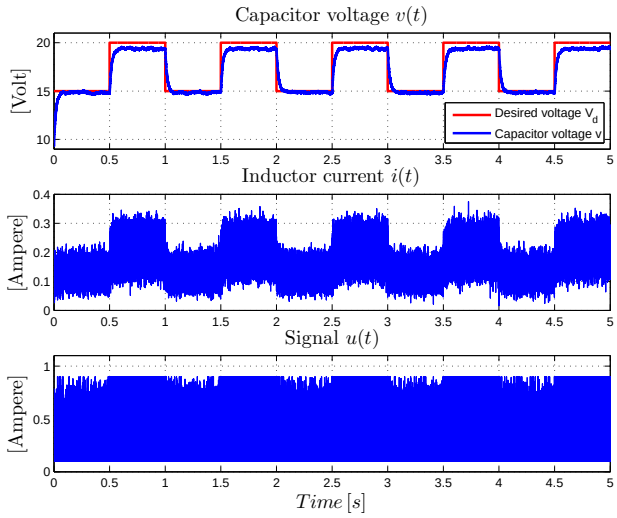


Fig. 8. Experiment: Capacitor voltage $v(t)$, inductor current $i(t)$ and control signal $u(t)$ for $0 \leq t \leq 5$ [s] and $v_d(t)$ in (22).

values. In simulation results, $v(t)$ reaches v_d, while in experimental results $v(t)$ is close to v_d. As explained before, the reason for this situation is that in experiment the system is affected by disturbances. In order to improve the performance and robustness, the introduction of a dynamic extension, such as an integral action, is being explored currently.

References

1. Rodriguez, H., Ortega, R., Escobar, G.: A Robustly Stable Output Feedback Saturated Controller for the Boost DC–to–DC Converter. In: Proc. of the 38th Conference on Decision and Control, Phoenix, USA, pp. 2100–2105, December (1999).
2. Ortega, R., Loría, A., Nicklasson, P. J., Sira–Ramirez, H.: *Pasivity–Based Control of Euler–Lagrange Systems*. Springer–Verlag, London, 1998.
3. Sira–Ramírez, H. Silva–Ortigoza, R.: *Control Design Techniques in Power Electronics Devices*. Springer–Verlag, London, 2006.
4. Spinetti, M., Fossas, E., Biel, D.: Stability analysis of a Lyapunov-based controlled boost converter. Proceedings of the 48th IEEE Conference on Decision and Control. pp. 6544–6548 (2009).
5. Karagiannis, D., Astolfi, A., Ortega, R.: Two results for the adaptive output feedback stabilization of nonlinear systems. Automatica 39, pp. 857–866 (2003).
6. Khalil, H. K..: *Nonlinear Systems*. Springer–Verlag, London, 2002.

Position/Force control using sliding mode with H_∞ attenuator to reduce rebounds in a mechanical system with a position constraint

Raul Rascón[a], Joaquín Alvarez[a] and Luis T. Aguilar[b]

[a] *CICESE Research Centre, Electronics and Telecommunication Department, P.O. BOX 434944, San Diego, CA 92143-4944, (emails: rrascon@cicese.edu.mx, and jqalvar@cicese.mx)*;
[b] *Instituto Politécnico Nacional, Avenida del parque 1310 Mesa de Otay, Tijuana México 22510 (email: laquilar@citedi.mx).*
Paper received on 24/09/12, Accepted on 21/10/12.

Abstract. This paper focuses on the problem of the control of a three degrees-of-freedom mechanical system, subject to constraints on the position, dry friction Dahl type and external disturbances. It is proposed a controller using sliding mode with an H_∞ attenuator to solve a position/force regulation problem. It is proved using Lyapunov tools that the nonlinear system has a local equilibrium asymptotically stable and achieves zero steady-state position error even in the presence of certain disturbances and dynamical friction. As well is given a parameter tuning that could reduce the number of rebounds between the end-efector and the position constraint. Furthermore, the controller attenuates other external perturbations and model discrepancies. The results obtained are illustrated with experiments.

1 Introduction

This paper focuses on the regulation problem for a mechanical system with a position constraint. The methodology of sliding mode with H_∞ attenuator is applied to solve the problem in question. Some recently references using this methodology can be found in [1–3]. Basically is a sliding mode controller that involves an H_∞ control design on the sliding surface. The purpose of this design is the elimination of disturbances and parametric uncertainties through the sliding mode control, otherwise, the H_∞ design will attenuate the parametric uncertainties and disturbances, by this way the trajectories will remain bounded around the reference.

The methodology to solve the regulation problem in a mechanical system under unilateral constraints was previously addressed by [4]. In [5] was constructed a PID controller type for robots under constraints, where a $\mathcal{H}_2/\mathcal{H}_\infty$ control is proposed to attenuate the influence of disturbances and uncertainties.

In [6] was designed an integral control using nonlinear H_∞ control and sliding modes for a permanent magnet synchronous motor. Another example of this type of controllers can be found in [2] where a sliding mode control with an H_∞ approach is utilized for output control. At the same time in [7], it is presented a mixed H_∞-sliding mode controller used to control a magnetic levitation system.

Furthermore [8] used an strategy of H_∞ and sliding modes applied to a current control problem in a switched converter. Finally [9] analyzed and designed an integral

Mireya García-Vázquez, Grigori Sidorov (Eds.).
Advances in Computing Science and Control.
Research in Computing Science 59, 2012, pp. 87- 98

sliding mode controller combined with H_∞ to control systems with unmatched perturbations.

The present paper considers that the dynamical system is nonlinear. Additionally, the motion of the system is affected by unknown disturbances, and the available state measurements are incomplete and imperfect.

To prove the stability of the controlled mechanical system we use quadratic functions; some references can be found in [10–13]. These quadratic functions allow us to ensure that the trajectories converge asymptotically to the desired position, and prove the convergence to the sliding surface in finite time.

In this study we combine the robustness properties of sliding mode with H_∞ control to design a controller which is capable to handle the above mentioned factors and thereby yielding a good performance on real systems. Experiments confirm the validity of the theoretical analysis.

The paper is organized as follows: In Section 2 it is defined the problem statement. In Section 3, the design procedure is considered. The develop of the sliding mode with H_∞ attenuator control is addressed in Section 4. Stability analysis is approached in Section 5. In Section 6 addresses the issue of H_∞ synthesis. Experiments are offered in Section 7. Finally in Section 8 some conclusions are discussed.

2 Problem Statement

The main concern of this work is the regulation control design, and its stability analysis, of a mechanical system subject to a position constraint (see Figure 1). This is a system, described by a lagrangian model. It can display an important dynamical behaviour like rebounds, due to collisions with the constraint, which may risk the integrity of a mechanical device. Hence, we design the controller with the aim, besides of having a good regulation, to reduce the presence of this phenomenon.

The equations of motion of the open-loop constrained mechanical system can be expressed as

$$M(q)\ddot{q} + C(q,\dot{q})\dot{q} + G(q) + F(\dot{q}) = \tau + \tau_c(q) + w(t) \tag{1}$$

where $q(t), \dot{q}(t), \ddot{q}(t) \in \mathbf{R}^3$ represent the displacement, velocity and acceleration of the rotational links of the mechanical system; $M(q) \in \mathbf{R}^{3\times 3}$ is the inertia matrix, which is symmetric and positive definite for all $q \in \mathbf{R}^3$; $C(q,\dot{q})\dot{q}$ is the vector of centripetal and Coriolis forces; $G(q) \in \mathbf{R}^3$ is the vector of gravitational forces; $\tau_c(q) \in \mathbf{R}^3$ is the torque generated by the spring by making contact with constraint; $\tau \in \mathbf{R}^3$ are the control inputs, and $w(t) = [w_1(t), w_2(t), w_3(t)]^T \in \mathbf{R}^3$ are unknown external disturbances. $F(\dot{q}) \in \mathbf{R}^3$ is the vector of frictional forces, which are represented as a combination

$$F_i = \sigma_{0i}\dot{q}_i + F_{di}, \quad i = 1,2,3 \tag{2}$$

of viscous friction $\sigma_{0i}\dot{q}_i$ and Dahl friction F_{di} which is governed by the following dynamic model:

$$\dot{F}_{di} = \sigma_{1i}\dot{q}_i - \sigma_{1i}|\dot{q}_i|\frac{F_{di}}{F_{ci}} + w_{2i}, \tag{3}$$

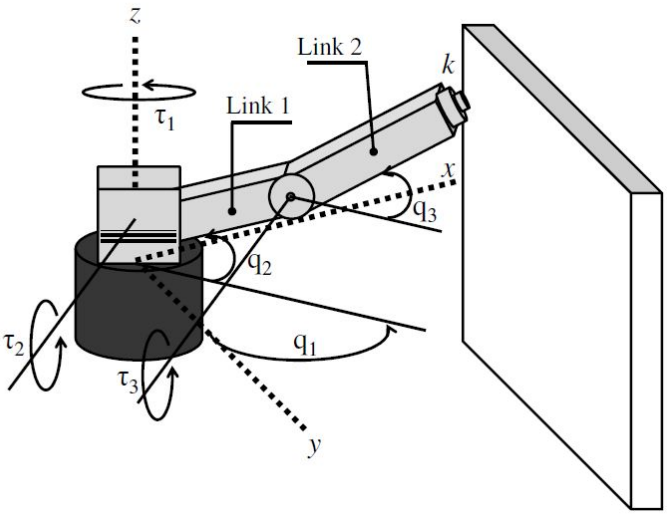


Fig. 1. Pegasus robot with three degrees-of-freedom and position constraint

where $\sigma_{0i} > 0$, $\sigma_{1i} > 0$, and $F_{ci} > 0$ are the viscous friction coefficient, stiffness coefficient and Coulomb friction level, respectively, corresponding to the ith manipulator joint; w_{2i} is an external disturbance which is involved to account for inadequacies of the frictional model.

The Dahl model (3) describes the spring-like behaviour during stiction and is essentially Coulomb friction with a lag in the change of the friction force when the motion direction is changed. Since the Coulomb friction is only a function of the displacement and the sign of the velocity, this dynamic model appears nonsmooth (see for details).

Clearly, the above component-wise relations can be rewritten in the vector form

$$F = \sigma_0 \dot{q} + F_d, \tag{4}$$

$$\dot{F}_d = \sigma_1 \dot{q} - \sigma_1 \text{diag}\{|\dot{q}_i|\} F_c^{-1} F_d + w_2, \tag{5}$$

where $F = \text{col}\{F_i\}$, $F_d = \text{col}\{F_{di}\}$, $x = \text{col}\{q_i\}$, $\sigma_0 = \text{diag}\{\sigma_{0i}\}$, $\sigma_1 = \text{diag}\{\sigma_{1i}\}$, $F_c = \text{diag}\{F_{ci}\}$, $w_2 = \text{col}\{w_{2i}\}$, the notations diag and col are used to denote a diagonal matrix and a column vector, respectively. The Euclidean position between the origin of the inertial frame of the robot and the constraint is given by x_0, since the constraint is located along the x axis, the position of the end effector of the robot with respect to x axis is given by x_r, at the same time the position of the end effector with respect to y axis is given by y_r, and it is z_r with respect to z, which are denoted as $R(t) = [x_r(t), y_r(t), z_r]^T \in \mathbf{R}^3$. A spring is added at the end effector tip, which is punctual with a stiffness coefficient k, which acts as a force sensor. One way to represent the force generated by the spring is using the Hooke law $F = kx_r$.

An impact is generated between the end effector of the robot and the constraint when $x_r \geq x_0$ where $x_r = [l_1 \cos(q_2) + l_2 \cos(q_3)] \cos(q_1)$. The impact generates forces of equal magnitude and opposite directions between the robot and the constraint. Specifically, the impact force acting in the spring $F_c \in \mathbf{R}$ is defined as follows

$$F_c = \frac{k}{2} \left(x_r - x_0 + |x_r - x_0| \right). \tag{6}$$

The impact force acting on the links of the robot produce a torque denoted by $\tau_c(x_r, q) \in \mathbf{R}^3$, given by

$$\tau_c = -F_c \begin{bmatrix} [l_1 \cos(q_2) + l_2 \cos(q_3)] \cos(q_1) \\ [l_1 \cos(q_2) + l_2 \cos(q_3)] \sin(q_1) \\ l_1 \sin(q_2) + l_2 \sin(q_3) \end{bmatrix}. \tag{7}$$

The above term completes the model (1) for the three degrees-of-freedom Pegasus robot constrained on the position.

3 Design Procedure

The control objective is to find a control law $\tau \in \mathbf{R}^3$, which depends on the desired force at the spring F_d (through the desired position x_{d1} along the x axis), the joint positions (q_1, q_2, q_3), the reference position x_0, the joint velocities $(\dot{q}_1, \dot{q}_2, \dot{q}_3)$, and the joint desired positions (q_{d1}, q_{d2}, q_{d3}) such that the system in closed loop satisfies

$$\lim_{t\to\infty} |q_1(t) - q_{d1}| = 0, \quad \lim_{t\to\infty} |q_2(t) - q_{d2}| = 0, \quad \lim_{t\to\infty} |q_3(t) - q_{d3}| = 0 \tag{8}$$

in spite of the upper bounded disturbance $\sup_t \|w_1(t)\| \leq N \in \mathbf{R}^3$, where the H_∞ control attenuates the influence of another external disturbances w_0, $w_2 \in \mathbf{R}^3$. Given the fact that $F_d = k(x_{d1} - x_0)$, with $F_d \geq 0$ and $x_{d1} = [l_1 \cos(q_{d2}) + l_2 \cos(q_{d3})] \cos(q_{d3})$, by substituting x_{d1} into F_d it is obtained q_{d2}

$$q_{d2} = \arccos\left(\frac{F_d}{l_1 k \cos(q_{d1})} + \frac{x_0}{l_1 \cos(q_{d1})} - \frac{l_2 \cos(q_{d3})}{l_1} \right) \tag{9}$$

Can be shift the equilibrium point of (1) to the origin by introducing the transformation based on the following

$$x_1 = \int_0^t x_2(t)dt, \; x_2 = [q_1 - q_{d1}, q_2 - q_{d2}]^T, \; x_3 = [\dot{q}_1, \dot{q}_2]^T, \; x_4 = [F_{d1}, F_{d2}]^T. \tag{10}$$

hence the state space equations are as follows

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= x_3 \\ \dot{x}_3 &= M^{-1}(x_2 + q_d)[-C(x_2 + q_d, x_3)x_3 - G(x_2 + q_d) \\ &\qquad - \sigma_0 x_3 - x_4 + \tau_c(x_2 + q_d) + u + w_1] \\ \dot{x}_4 &= \sigma_1 x_3 - \sigma_1 \text{diag}\{|x_{3i}|\} F_c^{-1} x_4 + w_2. \end{aligned} \tag{11}$$

where diag$\{q_d\}$, diag$\{\sigma_0\}$, diag$\{\sigma_1\}$, diag$\{w_1\}$, diag$\{w_2\}$, and diag$\{F_c\} \in \mathbf{R}^{3\times 3}$.

4 Sliding Mode Control using H_∞ Attenuator

Let us consider the following sliding surface

$$s = \nu x_1 + \mu x_2 + x_3 - \int_0^t u_\infty dt \tag{12}$$

where u_∞ is an H_∞ control which operates locally around the equilibrium point of system (11), also, the sliding surface (12) is a dynamical variable.

The control law which ensures that trajectories reach the sliding manifold is given by

$$\begin{aligned} u = {} & C(x_2 + q_d, x_3)x_3 + \sigma_0 x_3 + x_4 - \tau_c(x_2 + q_d) + G(x_2 + q_d) \\ & -M(x_2 + q_d)\left[x_2 + x_3 - u_\infty + \lambda s + \beta \text{sign}(s)\right]. \end{aligned} \tag{13}$$

The proposed control law will be acting at all time $t \geq 0$, that is, when the system is in free or in constrained motion (in contact with the constraint). The parameters diag$\{\lambda\}$ and diag$\{\beta\} \in \mathbf{R}^{3\times 3}$ have positive values which will be tuned to ensure the motion of the trajectories be driven toward the sliding surface.

Due the sliding surface (12)is a dynamical variable, it will be added as another state, this leads to the extended systems

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= x_3 \\ \dot{x}_3 &= -x_2 - x_3 - \lambda s - \beta \text{sign}(s) + M^{-1}(x_2 + q_d)w_1 \\ \dot{x}_4 &= \sigma_1 x_3 - \sigma_1 \text{diag}\{|x_{3i}|\}F_c^{-1}x_4 + w_2 \\ \dot{s} &= \nu x_2 + \mu x_3 + M^{-1}(x_2 + q_d)[-C(x_2 + q_d, x_3)x_3 - \sigma_0 x_3 - x_4 \\ & \qquad + \tau_c(x_2 + q_d) - G(x_2 + q_d) + u + w_1] - u_\infty. \end{aligned} \tag{14}$$

5 Stability Analysis

Theorem 1 *Let the system (14) through the controller governed by (12), (13), considering the condition* $\sup_t |w_{1i}(t)| \leq N_i$ *for all time and a constant value* $N_i > 0$ *and* $w_{2i}(t) = 0$*, with* $i = 1, 2, 3$*. Then the trajectory* q *in system (14) is asymptotically stable.*

Proof. from substituting the control law (13), the closed-loop system takes the form

$$\begin{aligned} s^T \dot{s} &= s^T \left(-\lambda s - \beta \tfrac{s}{\|s\|} + M^{-1}(x_2 + q_d) \sum_{i=1,2,3} N_i \right) \\ &\leq -\lambda \|s\|^2 - \left(\lambda_{min}\{\beta\} - \lambda_{max}\{M^{-1}(x_2 + q_d)\} \sum_{i=1,2,3} N_i \right) \|s\|. \end{aligned}$$

Can be conclude the existence of sliding modes on the surface $s = x_1 + x_2 + x_3 - \int_0^t u_\infty dt$ while the condition $\lambda_{min}\{\beta\} - \lambda_{max}\{M^{-1}x_2 + q_d)\} \sum_{i=1,2,3} N_i > 0$ remains

valid. Also, we can demonstrate finite time convergence of the trajectories to the surface $s = 0$ using the quadratic functionrática

$$V(s) = s^T s. \tag{15}$$

and compute its time derivative along the solutions of (14),

$$\begin{aligned}
\dot{V}(s(t)) &\leq -2s^T \lambda s - 2\left(\beta - M^{-1}(x_2 + q_d) \sum_{i=1,2,3} N_i\right) \|s\| \\
&\leq -2\left(\lambda_{min}\{\beta\} - \lambda_{max}\{M^{-1}(x_2 + q_d)\} \sum_{i=1,2,3} N_i\right) \|s\| \\
&= -2\left(\lambda_{min}\{\beta\} - \lambda_{max}\{M^{-1}(x_2 + q_d)\} \sum_{i=1,2,3} N_i\right) \sqrt{V(s(t))}.
\end{aligned} \tag{16}$$

From (16) it follows that

$$V(t) = 0 \quad \text{para} \quad t \geq t_0 + \frac{\sqrt{V(t_0)}}{\left(\lambda_{min}\{\beta\} - \lambda_{max}\{M^{-1}(x_2+q_d)\} \sum\limits_{i=1,2,3} N_i\right)} = t_f. \tag{17}$$

Hence, $V(t)$ converges to zero in finite time and, in consequence, a motion along the manifold $s = [0, 0, 0]^T$ occurs in the discontinuous system (14). Thus, in the following developments, it will be assumed that system (14) is in sliding mode, so $s = \dot{s} = 0$ for $t \geq t_f$. From (12) it is shown that the dynamics of system (14) once on sliding mode, are described by

$$\begin{aligned}
\dot{x}_1 &= x_2 \\
\dot{x}_2 &= x_3 \\
\dot{x}_3 &= -\nu x_2 - \mu x_3 + u_\infty \\
\dot{x}_4 &= \sigma_1 x_3 - \sigma_1 \text{diag}\{|x_{3i}|\} F_c^{-1} x_4 + w_2.
\end{aligned} \tag{18}$$

Lemma 1 *Let the system (18), considering the input $u_\infty = 0$. Then trajectories (x_2, x_3) are asymptotically stable.*

$$V(x_2, x_3) = (x_2 + x_3)^T (x_2 + x_3) + 2x_2^T x_2 + x_3^T x_3 \tag{19}$$

where $V(x_2, x_3)$ is a positive definite and also radially unbounded function. The time derivative of $V(x_2, x_3)$ along the trajectories of (18) with the input $u_\infty = 0$ is given by

$$\begin{aligned}
\dot{V}(x_2, x_3) &= 6x_2^T \dot{x}_2 + 2x_2^T \dot{x}_3 + 2x_3^T \dot{x}_2 + 4x_3^T \dot{x}_3 \\
&= -2x_2^T \nu x_2 - 4x_3^T \mu x_3 + 6x_2^T x_3 - 2x_2^T \mu x_3 - 4x_2^T \nu x_3 + 2x_3^T x_3 \\
&= -2x_2^T \nu x_2 - 4x_3^T \mu x_3 + x_2^T (6 - 2\mu - 4\nu) x_3 + 2x_3^T x_3 < 0.
\end{aligned} \tag{20}$$

By choosing the constants of the main diagonal of matrix
$(\text{diag}\{6\} - \text{diag}\{2\mu\} - \text{diag}\{4\nu\})$ making it a zero matrix in compliance with $\text{diag}\{\mu\} > 1/2 \in \mathbf{R}^{2\times 2}$, and $\text{diag}\{\nu\} > 0 \in \mathbf{R}^{2\times 2}$, can be assured while the system remains in $s = 0$, that the trajectories (x_2, x_3) of the system (18) using $u_\infty = 0$ converge to zero at $t \to \infty$.

Thus, the regulation problem for x_2 in the deviation system (18) can formally be stated as a nonlinear H_∞-control problem.

In the sequel, the investigation will be confined to the H_∞ position regulation problem, where

1. The output to be controlled is given by
$$z = \rho \begin{bmatrix} 0 \\ x_2 \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} u_\infty \tag{21}$$
with a positive weight coefficient ρ.
2. The position x_2 is the available measurement and are corrupted by the error vector $w_0(t) \in \mathbf{R}^3$.
$$y = x_2 + w_0, \tag{22}$$

The H_∞ control problem in question is thus stated as follows. Given the system representation (18)-(22) and a real number $\gamma > 0$, it is required to find (if any) a causal dynamic feedback controller
$$u_\infty = K(\xi) \tag{23}$$
with internal state $\xi \in \mathbf{R}^1 2$ such that the undisturbed closed-loop state x_2 is uniformly asymptotically stable around the origin and its $\mathfrak{L}_2$ gain is locally less than γ, i.e., inequality
$$\int_0^T \|z(t)\|^2 dt < \gamma^2 \int_0^T \|w(t)\|^2 dt \tag{24}$$
is satisfied for all $T > 0$ and all piecewise continuous functions
$w(t) = [w_0(t), w_1(t), w_2(t)]^T$ for which the corresponding state trajectory of the closed-loop system (18), initialized at the origin, remains in some neighbourhood of this point.

6 H_∞ Synthesis

The above H_∞ control problem in question is the nonlinear H_∞ control problem for nonsmooth systems, modelled by equations of the form
$$\begin{aligned} \dot{x} &= f_1(x) + f_2(x) + g_1(x)w + g_2(x)u \\ z &= h_1(x) + k_{12}(x)u \\ y &= h_2(x) + k_{21}(x)w \end{aligned} \tag{25}$$
where $x \in \mathbf{R}^n$ is the state space vector, $u \in \mathbf{R}^m$ is the control input, $w \in \mathbf{R}^r$ are unknown disturbances, $z \in \mathbf{R}^l$ is the output to be controlled, $y \in \mathbf{R}^p$ are the measurements available in the system. Adapting (18) to the form of (25) leads to
$$f_1(x) = \begin{bmatrix} x_2 \\ x_3 \\ -x_2 - x_3 \\ \sigma_1 x_3 \end{bmatrix}, \qquad f_2(x) = \begin{bmatrix} 0 \\ 0 \\ 0 \\ -\sigma_1 \text{diag}\{|x_{3i}|\} F_c^{-1} x_4 \end{bmatrix}, \tag{26}$$

$$g_1(x) = \begin{bmatrix} 0_{3\times3} & 0_{3\times3} & 0_{3\times3} \\ 0_{3\times3} & 0_{3\times3} & 0_{3\times3} \\ 0_{3\times3} & M^{-1}(x_2+q_d) & 0_{3\times3} \\ 0_{3\times3} & 0_{3\times3} & I_{3\times3} \end{bmatrix}, \quad g_2(x) = \begin{bmatrix} 0_{3\times3} \\ 0_{3\times3} \\ I_{3\times3} \\ 0_{3\times3} \end{bmatrix}, \tag{27}$$

$$h_1(x) = \rho \begin{bmatrix} 0_{3\times1} \\ x_2 \end{bmatrix}, \quad h_2(x) = x_2 + q_d,$$
$$k_{12}(x) = \begin{bmatrix} I_{3\times3} \\ 0_{3\times3} \end{bmatrix}, \; k_{21}(x) = \begin{bmatrix} I_{3\times3} \; 0_{3\times6} \end{bmatrix} \tag{28}$$

6.1 Local solution to the H_∞ problem

The following local analysis involve the linear H_∞ control problem for the following system

$$\begin{aligned} \dot{x} &= A_1 x + B_1 w + B_2 u \\ z &= C_1 x + D_{12} u \\ y &= C_2 x + D_{21} w \end{aligned} \tag{29}$$

where

$$\begin{aligned} A_1 = \tfrac{\partial f_1(0)}{\partial x} + \tfrac{\partial f_2(0)}{\partial x}, \quad B_1 = g_1(0) \quad B_2 = g_2(0) \\ C_1 = \tfrac{\partial h_1(0)}{\partial x}, \qquad D_{12} = K_{12}(0) \\ C_2 = \tfrac{\partial h_2(0)}{\partial x}, \qquad D_{21} = K_{21}(0). \end{aligned} \tag{30}$$

The system (18) must fulfill the stabilizability and detectability conditions around u, and y, respectively. Under these assumptions, the following conditions are necessary and sufficient for a solution of the linear problem to exist (see [14]).

A1 There exists a bounded positive semidefinite symmetric solution of

$$PA_1 + A_1^T P + C_1^T C_1 + P\left[\frac{1}{\gamma^2} B_1 B_1^T - B_2 B_2^T\right] P = 0 \tag{31}$$

such that the matrix $\left[A_1 - \left(B_2 B_2^T - \gamma^{-2} B_1 B_1^T\right) P\right]$ has all eigenvalues with negative real part.

A2 There exists a bounded positive semidefinite symmetric solution of

$$AZ + ZA^T + B_1^T B_1 + Z\left[\frac{1}{\gamma^2} P B_2 B_2^T P - C_2 C_2^T\right] Z = 0 \tag{32}$$

where $A = A_1 + (1/\gamma^2) B_1 B_1^T P$, such the matrix $\left[A - Z(C_2^T C_2 - \gamma^{-2} P B_2 B_2^T P)\right]$ has all eigenvalues with negative real part. The above equations (31) and (32) are known as Riccati equations.

By the bounded real lemma [15], conditions A1 and A2 ensure that there exists a positive constant ϵ_0 such that the system of the perturbed Riccati equations

$$P_\varepsilon A_1 + A_1^T P_\varepsilon + C_1^T C_1 + P_\varepsilon \left[\frac{1}{\gamma^2} B_1 B_1^T - B_2 B_2^T\right] P_\varepsilon + \varepsilon I = 0 \tag{33}$$

$$A_\varepsilon Z_\varepsilon + Z_\varepsilon A_\varepsilon^T + B_1 B_1^T + Z_\varepsilon \left[\frac{1}{\gamma^2} P_\varepsilon B_2 B_2^T P_\varepsilon - C_2 C_2^T\right] Z_\varepsilon + \varepsilon I = 0 \tag{34}$$

has a unique positive definite symmetric solution $(P_\varepsilon, Z_\varepsilon)$ for each $\varepsilon \in (0, \varepsilon_0)$ where $A_\varepsilon = A_1 + \left(1/\gamma^2\right) B_1 B_1^T P_\varepsilon$.

Equations (33) and (34) are subsequently utilized to derive a local solution of the H_∞ -control problem as in (25). Let conditions A1 and A2 hold be satisfied, and let $(P_\varepsilon, Z_\varepsilon)$ be the corresponding positive definite solution of (33) and (34) under some $\varepsilon > 0$. Then the output feedback

$$\dot{\xi} = f_1(\xi) + f_2(\xi) + \left[\frac{1}{\gamma^2} g_1(\xi) g_1^T(\xi) - g_2(\xi) g_2^T(\xi)\right] P_\varepsilon \xi + Z_\varepsilon C_2^T \left[y - h_2(\xi)\right] \tag{35}$$

$$u_\infty = -B_2^T(\xi) P_\varepsilon \xi \tag{36}$$

is a local solution of the H_∞-control problem.

7 Experiments

Performance issues and robustness properties of the proposed sliding mode controller with H_∞ attenuator (13) have been tested in the three degrees-of-freedom platform called Pegasus robot as in Figure 2. Since only the states $[q_1, q_2, q_3]^T$ measurements are available, the H_∞ filter (35) was applied to have access to the remaining states.

The experiments were carried out using the Pegasus robot, simulink from MatLab® and the data acquisition board SENSORAY 626 to be used as interphase between the computer and the robot, as well the force sensor utilized was the FC2231 from Measurement Specialties^{TM} with a measurement range from 0-50 Lbf. The parameters of the Pegasus robot made by Amatrol are shown in Table 1. Initial conditions, controller gains and external disturbances are displayed in Table 2.

8 Conclusions

It was developed a fully practical framework for sliding mode control involving H_∞ control methodology. The afore mentioned design procedure has been shown to be eminently suited to solving a position/force regulation problem for a mechanical system with friction and a position constraint. To facilitate exposition, the friction model chosen for treatment has been confined to the Dahl model augmented with viscous friction. The sliding mode-H_∞ output regulation synthesis proposed is suited to globally solve the regulation problem when the inequality $\lambda_{min}\{\beta\} - \lambda_{max}\{M^{-1}x_2 + q_d)\} \sum_{i=1,2,3} N_i > 0$ is satisfied, even in the presence of disturbances $\sup_t |w_{1i}(t)| \leq N_i$ for all time and a

Table 1: Pegasus robot parameters.

Notation	Description	Value	Units
l_1	Length of the link 1	0.297	m
l_2	Length of the link 2	0.297	m
m_1	Mass of the link 1	0.38	kg
m_2	Mass of the link 2	0.34	kg
I_1	Inertia 1	0.000243	kg m^2
I_2	Inertia 2	0.000068	kg m^2
I_3	Inertia 2	0.000015	kg m^2
g	Gravity	9.80665	m/s^2
l_{c1}	Length to the centre of mass: Link 1	0.1485	m
l_{c2}	Length to the centre of mass: Link 2	0.1485	m

Table 2: Initial conditions, controller gains and external disturbances

Notation	Description	Value	Units
$q_1(0)$	Joint position 1	22.5	grad
$\dot{q}_1(0)$	Joint velocity 1	0	grad/s
$q_3(0)$	Joint position 3	45	grad
$\dot{q}_3(0)$	Joint velocity 3	0	rad/s
$s(0)$	Sliding motion	[0,0,0]	
λ	Controller gains	diag{40,30,30}	$1/s^2$
μ	Controller gains	diag{7,9,9}	$1/(m.s)$
β	Controller gains	diag{1,1,1}	N.m
γ	Controller gains	diag{9,16,16}	$1/(m.s^2)$
F_d	Desired force at the end effector	25	N
q_{d1}	Desired joint position 1	0	rad
q_{d3}	Desired joint position 3	0	rad
w_1	Disturbance set on link 1	$0.1\sin(t)$	N.m
w_2	Disturbance set on link 2	$0.1\cos(t)$	N.m
w_3	Disturbance set on link 3	$0.1\cos(t)$	N.m

Fig. 2. Pegasus robot configured to have a position constraint.

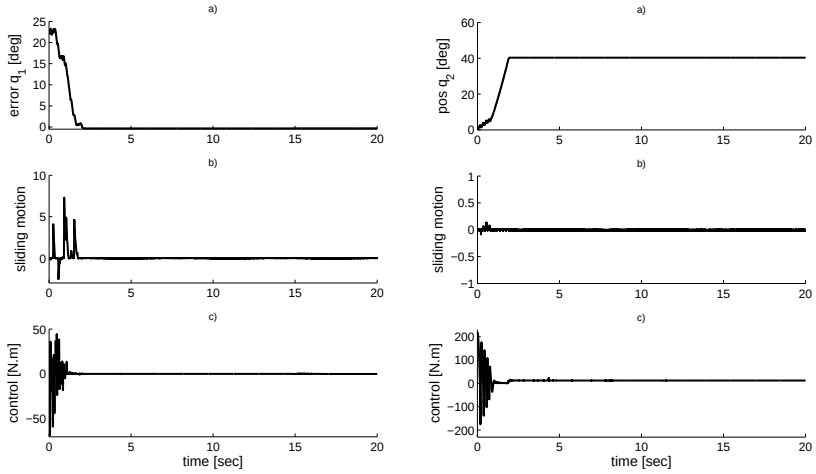


Fig. 3. Values for the joints q_1 and q_2.

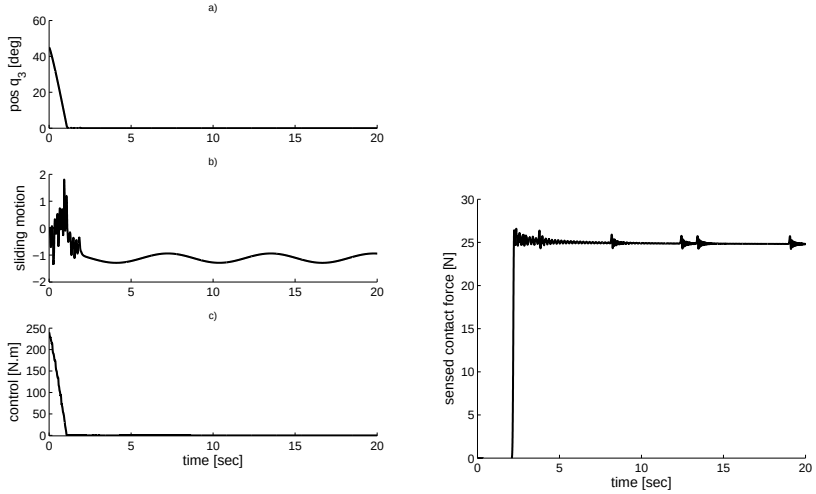


Fig. 4. Values for the joint q_3 and the sensed contact force on the end-effector.

constant value $N_i > 0$, with $i = 1, 2, 3$, whenever this inequality is not satisfied the controller will attenuate the disturbances and dead zone model discrepancies, but it is enough to increase the elements of the matrix diag$\{\beta\}$ in order to fulfill the inequality and turn it into an asymptotically stable system.

The experimental platform consists in a robot manipulator of three degrees-of-freedom named Pegasus which operates under constrained conditions, and it has a force sensor mounted at the end-effector. This platform has a transmission made of chains and gears, which presents a considerable backlash phenomenon in each joint.

The controller gains were chosen heuristically, therefore, it is possible that another gain values yield better results. Effectiveness of the design procedure has been supported by experiments in the platform of the Pegasus robot configured to have a position constraint.

References

1. Lian, J., Zhao, J.: Robust h-infinity integral sliding mode control for a class of uncertain switched nonlinear systems. Journal of Control Theory and Applications **8** (2010) 521–526
2. Castaños, F., Fridman, L.: Dynamic switching surfaces for output sliding mode control: An approach. Automatica **47** (2011) 1957 – 1961
3. Ghafari-Kashani, A., Faiz, J., Yazdanpanah, M.: Integration of non-linear $\mathcal{H}_\infty$; and sliding mode control techniques for motion control of a permanent magnet synchronous motor. Electric Power Applications, IET **4** (2010) 267 –280
4. Brogliato, B., S.I.N., Orhant, P.: On the control of finite-dimensional mechanical systems with unilateral constraints. IEEE Transactions on Automatic Control **42** (1997) 200–215
5. Tseng, C.S.: Mixed $\mathcal{H}_2/\mathcal{H}_\infty$ adaptive tracking control design for uncertain constrained robots. Asian Journal of Control **7** (2005) 296–309
6. Ghafari-Kashani, A.R., F.J., Yazdanpanah, M.: Integration of non-linear $h\infty$ and sliding mode control techniques for motion control of a permanent magnet synchronous motor. IET Electr. Power Appl. **4** (2010) 267–280
7. Shen, J.C.: $\mathcal{H}_\infty$ control and sliding mode control of magnetic levitation system. Asian Journal of Control **4** (2002) 333 –340
8. Vidal-Idiarte, E., Martinez-Salamero, L., Calvente, J., Romero, A.: An $\mathcal{H}_\infty$ control strategy for switching converters in sliding-mode current control. IEEE Transactions on Power Electronics **21** (2006) 553 – 556
9. Castaños, F., Fridman, L.: Analysis and design of integral sliding manifolds for systems with unmatched perturbations. IEEE Transactions on Automatic Control **51** (2006) 853 – 858
10. Paden, B., Sastry, S.: A calculus for computing filippov's differential inclusion with application to the variable structure control of robot manipulators. IEEE Transactions on Circuits and Systems **34** (1987) 73–81
11. Shevitz, D., Paden, B.: Lyapunov stability theory of nonsmooth systems. IEEE Transactions on Automatic Control **39** (1994) 1910–1914
12. Kazerooni, H.: Contact instability of the direct drive robot when constrained by a rigid environment. IEEE Transactions on Automatic Control **35** (1990) 710–714
13. Branicky, M.: Multiple lyapunov functions and other analysis tools for switched and hybrid systems. IEEE Transactions on Automatic Control **43** (1998) 475–482
14. Aguilar, L., Orlov, Y., Acho, L.: Nonlinear $\mathcal{H}_\infty$ control of nonsmooth time varying systems with application to friction mechanical manipulators. Automatica **39** (2003) 1531–1542
15. Anderson, B., Vreugdenhil, R.: Network analysis and synthesis. Englewood Cliffs, Prentice Hall, NJ (1973)

A control scheme for the tracking control of the Furuta pendulum*

Carlos Aguilar–Avelar and Javier Moreno–Valenzuela**

Instituto Politécnico Nacional–CITEDI,
Av. del Parque 1310, Mesa de Otay, Tijuana, B.C., Mexico
{caguilar, moreno}@citedi.mx
http://www.citedi.mx
Paper received on 24/09/12, Accepted on 25/10/12.

Abstract. The purpose of this document is to introduce a new control scheme which is based in the feedback linearization technique. The error signal is defined as a vector of dimensation two, with first element defined as the difference between a differentiable time–varying signal and the arm position, while the second element is defined as the negative of the pendulum position. The control problem consists in designing a control input to keep the error trajectories ultimately bounded. The proposed controller is tested by means of numerical simulations and real–time experiments, which support its practical viability.

1 Introduction

Underactuated mechanical systems are systems which have more degrees-of-freedom than actuators to control. Their uses are common on a lot of applications as robot mobile systems, vehicles used on space or under the sea (with special features that increases difficulties to control) or because of the mathematical model used for control design as where joint flexibility is included in the model. The Furuta pendulum is a well–known underactuated mechanical system used by many control researchers to test linear and non linear techniques[1]. This mechanism consists in an arm rotating in the horizontal plane and pendulum rotating in the vertical plane. The system has only one actuator that provides torque $\tau \in \mathbb{R}$ at the arm.

Feedback linearization is a commonly control technique used in non linear systems; see for example [2]. The basic idea of the feedback linearization control is to define a proper measurable output. Then, by computing an appropriated number of time differentiations of the output, the controller is derived so that the resulting closed–loop system is linear and time–invariant. A possible disadvantage of this approach is that the so–called zero dynamics may be unstable.

The purpose of this document in to introduce a new control scheme which is based in the feedback linearization technique. The control objective is to keep the error trajectories uniformly ultimately bounded. Numerical simulations and real–time experiments support the introduced theory and show its practical viability.

* Work supported by CONACyT, project number 176587, and SIP–IPN, Mexico.
** Author to whom correspondence should be addressed.

Mireya García-Vázquez, Grigori Sidorov (Eds.).
Advances in Computing Science and Control.
Research in Computing Science 59, 2012, pp. 99- 110

The Furuta pendulum dynamics and the control problem formulation are given in Section 2. A new control is proposed in Section 3. There, the closed–loop system and zero dynamics are discussed. In Section 4, by using an original experimental platform which has been accurately identified, the numerical and real–time experimental tests are presented. Finally, some concluding remarks are provided in Section 5.

2 Furuta pendulum dynamic model and control problem formulation

2.1 Model

As previously described, the Furuta pendulum is mechanism consisting in an arm rotating in the horizontal plane and pendulum rotating in the vertical plane. See Figure 1 for a description of the relative joint angle measurements and torque application.

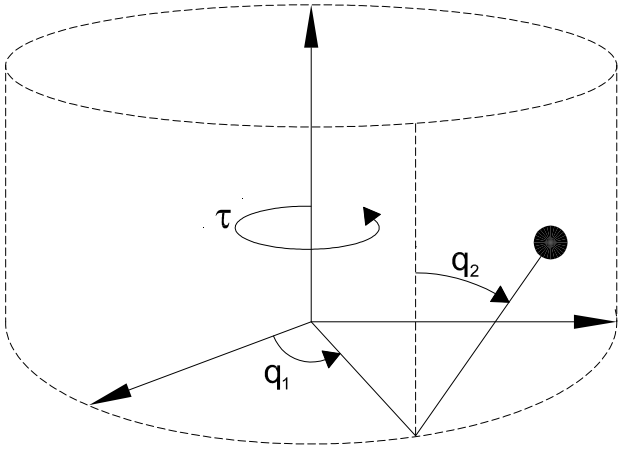


Fig. 1. Furuta pendulum.

The dynamic model of the Furuta pendulum in Euler-Langrange form is written as [1], [4], [5],

$$M(\boldsymbol{q})\ddot{\boldsymbol{q}} + C(\boldsymbol{q}, \dot{\boldsymbol{q}})\dot{\boldsymbol{q}} + \boldsymbol{g}(\boldsymbol{q}) + F_v\dot{\boldsymbol{q}} + \boldsymbol{f}_{cl}(\dot{\boldsymbol{q}}) = \boldsymbol{u}, \tag{1}$$

where

$$\boldsymbol{q} = \begin{bmatrix} q_1 \\ q_2 \end{bmatrix}$$

is the vector of joint position,

$$\boldsymbol{u} = \begin{bmatrix} \tau \\ 0 \end{bmatrix}$$

is the torque input vector, being $\tau \in \mathbb{R}$ the torque input of the arm,

$$M(\boldsymbol{q}) = \begin{bmatrix} \theta_1 + \theta_2 \sin^2(q_2) & \theta_3 \cos(q_2) \\ \theta_3 \cos(q_2) & \theta_4 \end{bmatrix},$$

$$C(\boldsymbol{q}, \dot{\boldsymbol{q}}) = \begin{bmatrix} \frac{1}{2}\theta_2 \dot{q}_2 \sin(2q_2) & -\theta_3 \dot{q}_2 \sin(q_2) + \frac{1}{2}\theta_2 \dot{q}_1 \sin(2q_2) \\ -\frac{1}{2}\theta_2 \dot{q}_1 \sin(2q_2) & 0 \end{bmatrix},$$

$$\boldsymbol{g}(\boldsymbol{q}) = \begin{bmatrix} 0 \\ -\theta_5 \sin(q_2) \end{bmatrix},$$

$$F_v = \begin{bmatrix} f_{v1} & 0 \\ 0 & f_{v2} \end{bmatrix} = \begin{bmatrix} \theta_6 & 0 \\ 0 & \theta_7 \end{bmatrix},$$

$$\boldsymbol{f}_{cl}(\dot{\boldsymbol{q}}) = \begin{bmatrix} f_{c1} \tanh(\beta \dot{q}_1) \\ f_{c2} \tanh(\beta \dot{q}_2) \end{bmatrix} = \begin{bmatrix} \theta_8 \tanh(\beta \dot{q}_1) \\ \theta_9 \tanh(\beta \dot{q}_2) \end{bmatrix},$$

where $M(\boldsymbol{q}) \in \mathbb{R}^{2\times 2}$ is the positive definite inertia matrix and $C(\boldsymbol{q}, \dot{\boldsymbol{q}})\dot{\boldsymbol{q}} \in \mathbb{R}^{2\times 1}$ is the centrifugal and Coriolis torque vector, $\boldsymbol{g}(\boldsymbol{q}) \in \mathbb{R}^{2\times 1}$ is known as the gravitational torque vector, $F_v \in \mathbb{R}^{2\times 2}$ is a diagonal matrix of containing the viscous friction coefficients of each joint, and $\boldsymbol{f}_{cl}(\dot{\boldsymbol{q}}) \in \mathbb{R}^{2\times 1}$ is continuous and differentiable version of the Coulomb friction vector with $\beta > 0$ large enough.

The physical meaning of the input vector $\boldsymbol{u} \in \mathbb{R}^2$ is that the system is equipped with one actuator only, which delivers the torque input $\tau \in \mathbb{R}$.

2.2 Control problem

First, let us define the following signals

$$\boldsymbol{e} = \begin{bmatrix} e_1 \\ e_2 \end{bmatrix} = \begin{bmatrix} q_{d1} - q_1 \\ -q_2 \end{bmatrix} \in \mathbb{R}^2, \tag{2}$$

where $q_{d1}(t)$ is twice differential signal that denotes the desired angle position of the arm.

The control problem consists in designing a controller $\tau \in \mathbb{R}$ such that the error trajectories $\boldsymbol{e}(t) \in \mathbb{R}^2$ satisfies the definition of a uniformly ultimately bounded signal. In other words, the controller should guarantee

$$\|\boldsymbol{e}(0)\| < \alpha \;\; \Rightarrow \;\; \|\boldsymbol{e}(t)\| \leq b \;\; \forall\, t \geq t_0 + T. \tag{3}$$

3 Proposed scheme

3.1 A controller derived from feedback linearization

Let us define the vectors

$$\boldsymbol{h}_v = \begin{bmatrix} h_{v1} \\ h_{v2} \end{bmatrix} = \frac{1}{\det M(q)} \begin{bmatrix} M_{22} z_1 - M_{12} z_2 \\ -M_{21} z_1 + M_{11} z_2 \end{bmatrix}, \tag{4}$$

with

$$z_1 = C_{11}\dot{q}_1 + C_{12}\dot{q}_2 + f_{v1}\dot{q}_1 + f_{c1}\tanh(\beta\dot{q}_1), \tag{5}$$
$$z_2 = C_{21}\dot{q}_1 + C_{22}\dot{q}_2 + g_2 + f_{v2}\dot{q}_2 + f_{c2}\tanh(\beta\dot{q}_2), \tag{6}$$

and

$$\boldsymbol{g}_v = \begin{bmatrix} g_{v1} \\ g_{v2} \end{bmatrix} = \frac{1}{\det M(q)} \begin{bmatrix} M_{22} \\ -M_{21} \end{bmatrix}. \tag{7}$$

Thus, by using the definition of $\boldsymbol{e}$ in (2), the open–loop dynamics can be written as

$$\frac{d}{dt}\begin{bmatrix} e_1 \\ e_2 \end{bmatrix} = \begin{bmatrix} \dot{e}_1 \\ \dot{e}_2 \end{bmatrix}, \tag{8}$$
$$\frac{d}{dt}\begin{bmatrix} \dot{e}_1 \\ \dot{e}_2 \end{bmatrix} = \begin{bmatrix} \ddot{q}_{d1} - h_{v1} - g_{v1}\tau \\ -h_{v2} - g_{v2}\tau \end{bmatrix} \tag{9}$$

Feedback linearization is a control technique commonly used in non linear systems. This approach consists in the transformation of the non linear system into an equivalent system through a proper output signal.

Now, consider

$$\boldsymbol{r} = \dot{\boldsymbol{e}} + \Delta\boldsymbol{e}, \tag{10}$$

where $\Delta = diag\{\Delta_1, \Delta_2\} \in \mathbb{R}^{2\times 2}$ which is positive definite. For the open loop system (8)–(9) we propose the output function

$$y = r_1 + r_2 \tag{11}$$

where r_1 and r_2 are defined in (10).

Equation (11) satisfies

$$\dot{y} = \ddot{e}_1 + \Delta_1\dot{e}_1 + \ddot{e}_2 + \Delta_2\dot{e}_2 \tag{12}$$
$$= \ddot{q}_{d1} - [h_{v1} + h_{v2}] - [g_{v1} + g_{v2}]\tau + \Delta_1\dot{e}_1 + \Delta_2\dot{e}_2 \tag{13}$$

The output equation (11) can be turned into an exponentially convergent signal by using a torque input given by

$$\tau = \tau_{fbl} = \frac{\ddot{q}_{d1} + K_p y - [h_{v1} + h_{v2}] + \Delta_1\dot{e}_1 + \Delta_2\dot{e}_2}{g_{v1} + g_{v2}}, \tag{14}$$

where K_p is a positive constant.

Notice that the controller (14) is valid in a region of the state space where

$$g_{v1} + g_{v2} \neq 0.$$

However, without loss of generality, in this paper we assume that

$$g_{v1} + g_{v2} < 0, \ \forall\, \boldsymbol{q} \in \mathbb{R}^2. \tag{15}$$

Let us notice that the experimental case study presented in this paper satisfies assumption (15); see Section 5.

The closed–loop system can be written as

$$\frac{d}{dt}y = -K_p y, \tag{16}$$

for which we have that

$$\lim_{t\to\infty} y(t) = 0,$$

with exponential convergence rate.

3.2 Zero dynamics

The zero dynamics is obtained assuming that the output signal y and its time–derivative are equal to zero, that is,

$$y = 0, \tag{17}$$

and

$$\dot{y} = 0. \tag{18}$$

In preparation to obtain the zero dynamics, let us first take into account that by virtue of definition (2), the following

$$\begin{aligned} q_1 &= q_{d1} - e_1, \\ \dot{q}_1 &= \dot{q}_{d1} - \dot{e}_1, \\ q_2 &= -e_2, \\ \dot{q}_2 &= -\dot{e}_2, \end{aligned}$$

is satisfied. At the same time

$$\begin{aligned} h_{v1}(q_1, q_2, \dot{q}_1, \dot{q}_2) &= h_{v1}(t, e_1, e_2, \dot{e}_1, \dot{e}_2), \\ h_{v2}(q_1, q_2, \dot{q}_1, \dot{q}_2) &= h_{v2}(t, e_1, e_2, \dot{e}_1, \dot{e}_2), \\ g_{v1}(q_1, q_2) &= g_{v1}(t, e_1, e_2), \\ g_{v2}(q_1, q_2) &= g_{v2}(t, e_1, e_2), \end{aligned}$$

which means that the Furuta pendulum dynamics can be expressed as a function of the time t, the position tracking error $\boldsymbol{e}$, and the velocity tracking error $\dot{\boldsymbol{e}}$.

From the equation (17) and the definition of $\boldsymbol{r}$ in (10), we have that

$$\frac{d}{dt}e_1 + \Delta_1 e_1 = -\frac{d}{dt}e_2 - \Delta_1 e_2 \tag{19}$$

is satisfied.

Besides, by using (18), the fact that

$$\ddot{e}_1 = \ddot{q}_{d1} - \ddot{q}_1 = \ddot{q}_{d1} - h_{v1} - g_{v1}\,\tau\big|_{y=0}\,,$$

expanding the definition of h_{v1} in (4) as

$$h_{v1} = h_{C1} - \frac{\theta_3 \cos(q_2)}{\det(M)}\left[-\theta_5 \sin(q_2)\right]$$

with

$$h_{C1} = \frac{M_{22}}{\det(M)} z_1 - \frac{M_{12}}{\det(M)} \left[C_{21}\dot{q}_1 \right.$$

$$\left. + C_{22}\dot{q}_2 + f_{v2}\dot{q}_2 + f_{c2}\tanh(\beta\dot{q}_2) \right],$$

z_1 in (5), and using the definition $q_2 = -e_2$, it is possible to show that

$$\frac{d^2}{dt^2}e_2 + \kappa(\boldsymbol{q})\Delta_2\frac{d}{dt}e_2 + \kappa(\boldsymbol{q})\frac{\theta_3\cos(q_2)}{\det(M)}\theta_5\sin(e_2) = -\kappa(\boldsymbol{q})[\ddot{q}_{d1} + \Delta_1\dot{e}_1]$$

$$+\kappa(\boldsymbol{q})h_{C1} - \frac{g_{v1}}{g_{v1}+g_{v2}}h_{v2}, \tag{20}$$

where

$$\kappa(\boldsymbol{q}) = \kappa(t,\boldsymbol{e}) = 1 - \frac{g_{v1}}{g_{v1}+g_{v2}}$$

Notice that by virtue of the fact that g_{v1} (see its definition in equation (7)), and the assumption (15), the inequality

$$\kappa(\boldsymbol{q}) > 0,\ \forall\, \boldsymbol{q} \in \mathbb{R}^2,$$

is satisfied.

The system (19) and (20) governs the dynamics of the signals $[e_1\ e_2\ \dot{e}_2]^T \in \mathbb{R}^3$ for $y(t) = \dot{y}(t) = 0$ for all $t \geq 0$, that is, equations (19) and (20) represent a form to describe the zero dynamics.

In agreement with the discussion about the dependency of the Furuta pendulum dynamics, it is clear that

$$\begin{aligned} h_{v1}(t,e_1,e_2,\dot{e}_1,\dot{e}_2) &= h_{C1}(t,e_1,e_2,\dot{e}_1,\dot{e}_2) - \frac{\theta_3\cos(-e_2)}{\det(M)}\left[-\theta_5\sin(-e_2)\right] \\ &= h_{C1}(t,e_1,e_2,\dot{e}_1,\dot{e}_2) - \frac{\theta_3\cos(e_2)}{\det(M)}\left[\theta_5\sin(e_2)\right]. \end{aligned}$$

3.3 Discussion on the control goal

It is possible to show that the trajectories of the zero dynamics (19) and (20) are uniformly ultimately bounded. However, the proof of this fact will be left out for shortening of this paper.

The exponential convergence of the output signal $y(t)$, and the assumption that $q_{d1}(t)$, $\dot{q}_{d1}(t)$, and $\ddot{q}_{d1}(t)$ are bounded, guarantee that the signals $e_1(t)$ and $e_2(t)$ are uniformly ultimately bounded.

Furthermore, there is an ultimate bound of the trajectories $e_1(t)$ and $e_2(t)$ that depends on the Furuta pendulum dynamics and on the control gains Δ_1 and Δ_2.

Fig. 2. Furuta pendulum prototype built at *Instituto Politécnico Nacional–CITEDI* Research Center.

4 Numerical and experimental results

In this section, the real–time implementation of the proposed controller (14) is presented. The experimental tests have been conducted in a Furuta pendulum built at the *Instituto Politécnico Nacional–CITEDI* Research Center. See Figure 2 for picture of the experimental system.

The constant parameters θ_i of the Furuta pendulum model in (1) have been identified by using a procedure based on least squares. These parameters are shown in Table 1, which was obtained by using the filtered dynamic model and the classical least squares identification; see for instance [6] and [7], where identification procedures for mechanical systems are proposed. In the identification process, we considered

$$\beta = 100,$$

which is related to the vector of Coulomb friction $\boldsymbol{f}_{cl}(\dot{\boldsymbol{q}}) \in \mathbb{R}^2$ in the Furuta pendulum model (1).

The identified model was useful to perform numerical simulations of the proposed theory and as previous step to achieve implementations in real–time. Let us notice that

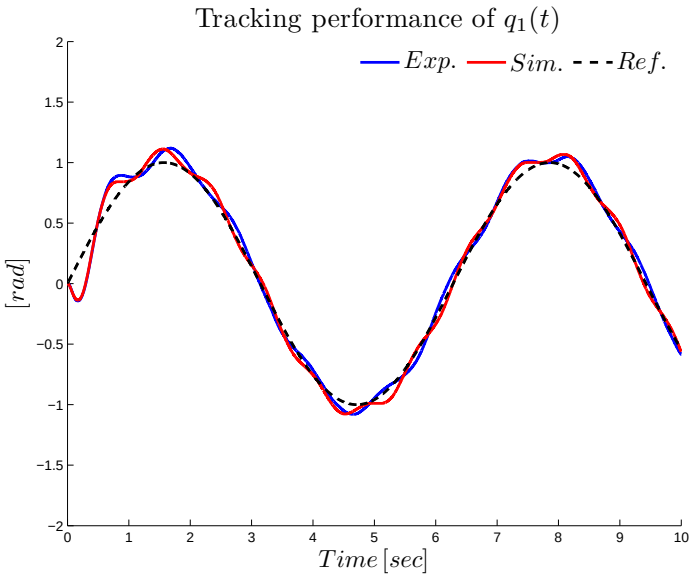


Fig. 3. Simulation and experiment: Time evolution of $q_d(t)$ and $q_1(t)$ obtained by simulation and experiment.

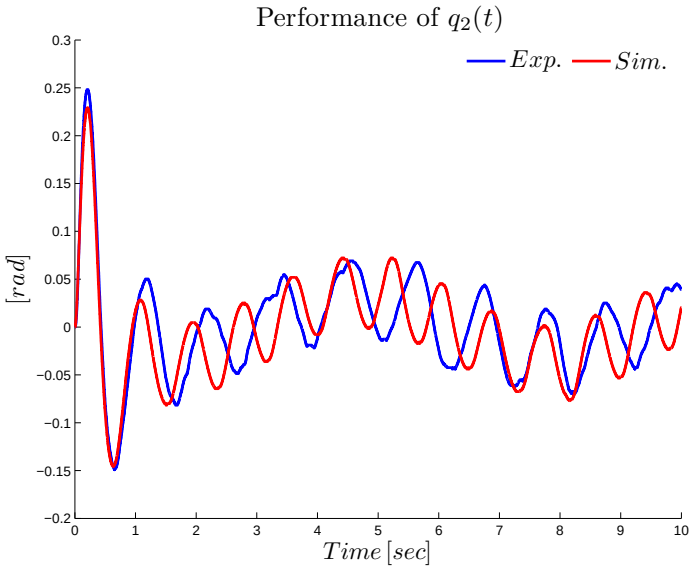


Fig. 4. Simulation and experiment: Time evolution of $q_2(t)$ obtained by simulation and experiment.

Table 1. Numerical values of the Furuta pendulum parameters.

Symbol	Value	Unit
θ_1	0.20959	$\mathrm{Kg}\cdot\mathrm{m}^2\cdot\mathrm{rad}$
θ_2	0.04926	$\mathrm{Kg}\cdot\mathrm{m}^2\cdot\mathrm{rad}$
θ_3	0.06258	$\mathrm{Kg}\cdot\mathrm{m}^2\cdot\mathrm{rad}$
θ_4	0.04539	$\mathrm{Kg}\cdot\mathrm{m}^2\cdot\mathrm{rad}$
θ_5	1.71142	$\mathrm{Kg}\cdot\mathrm{m}^2\cdot\mathrm{rad}$
θ_6	0.08514	$\mathrm{N}\cdot\mathrm{m}\cdot\mathrm{rad/sec}$
θ_7	0.00238	$\mathrm{N}\cdot\mathrm{m}\cdot\mathrm{rad/sec}$
θ_8	0.13738	$\mathrm{N}\cdot\mathrm{m}\cdot\mathrm{rad/sec}$
θ_9	0.02789	$\mathrm{N}\cdot\mathrm{m}\cdot\mathrm{rad/sec}$

to obtain a successful experimental result the applied torque should respect the power limits. Besides, simulations are important to check that stability is still obtained in spite of the fact that the controller is implemented in discrete form and the joint velocity is estimated through discrete differentiation from joint position measurements.

Since the experimental system has been identified, experiments have been compared with respect to numerical simulations.

The desired joint position trajectory $q_{d1}(t)$ for the arm position was defined as

$$q_{d1}(t) = 1.0\sin(t)$$

Besides, concerning the output function y in (11), we used

$$\Delta_1 = 5.0, \tag{21}$$

$$\Delta_2 = 8.0. \tag{22}$$

and $K_p = 10.0$ in the new controller (14).

The initial condition of the experimental system was

$$\begin{bmatrix} q_1(0) \\ q_2(0) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \ [\mathrm{rad}],$$

and

$$\begin{bmatrix} \dot{q}_1(0) \\ \dot{q}_2(0) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \ [\mathrm{rad/s}].$$

4.1 Results

The numerical and experimental results are illustrated in Figures 3–6. In particular, the Figure 3 shows the time evolution of $q_{d1}(t)$ and $q_1(t)$, and the Figure 4 depicts the time evolution of $q_2(t)$. The applied torque $\tau(t)$ and the output signal $y(t)$ are observed in Figures 5 and 6, respectively.

As observed in Figures 3–6, the experimental results and the simulation are very similar, which shows that the identified parameters θ_i are relatively accurate.

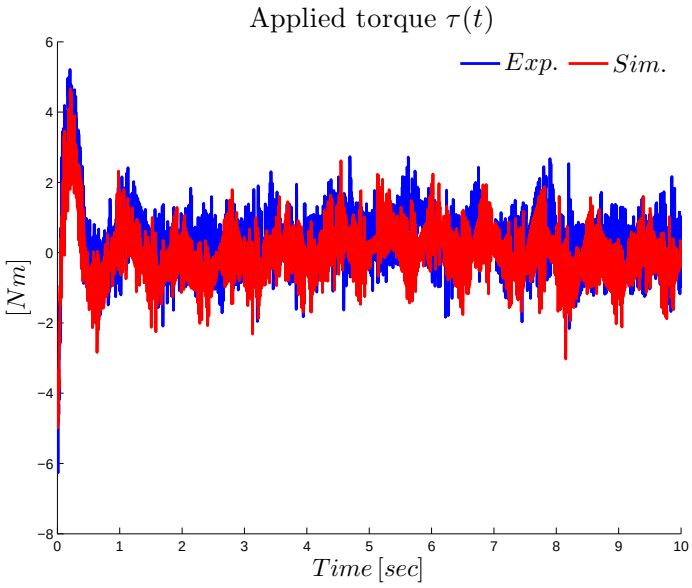


Fig. 5. Simulation and experiment: Time evolution of control input $\tau(t)$.

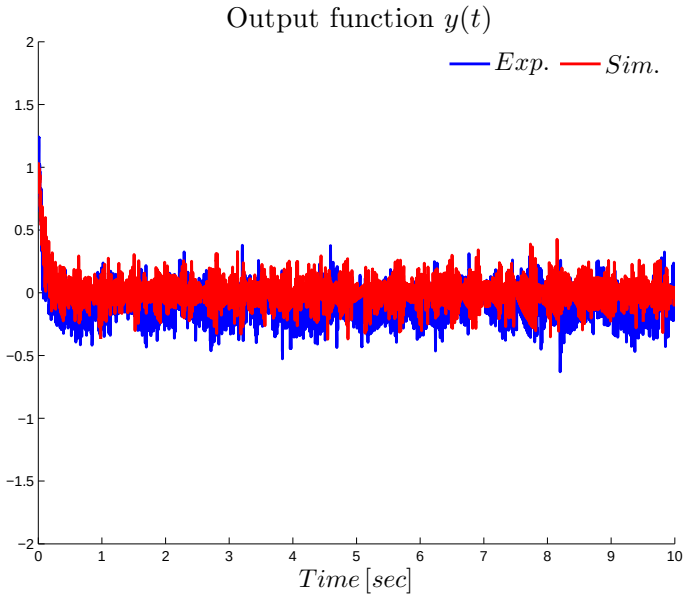


Fig. 6. Simulation and experiment: Time evolution of the output signal $y(t)$.

The control action $\tau(t)$ obtained by simulation and experiment has high frequency components. This is attributed to the velocity estimation obtained by using the dirty derivative algorithm [8]

$$\dot{\boldsymbol{q}}(Tk) \approx \frac{\boldsymbol{q}(Tk) - \boldsymbol{q}(T[k-1])}{T}, \tag{23}$$

with $T = 0.001$ [s] the sampling period and k the integer time index, and the PWM switching of the servo amplifier.

Also, in Figure 6 the response of the output function $y(t)$ in (11) is appreciated for both simulation and experiment. Although theory predicts exponential convergence of $y(t)$, an oscillatory behavior with high frequency components is presented. It is noteworthy that the output function $y(t)$ in (11) depends of the joint velocity $\dot{\boldsymbol{q}}(t) \in \mathbb{R}^2$. The main reason for the oscillatory behavior of $y(t)$ is that the joint velocity $\dot{\boldsymbol{q}}(t)$ is estimated via the noisy dirty derivative algorithm (23) and the relative high value of the gains Δ_1 and Δ_2 in (21) and (22), respectively.

By means of numerical simulation assuming a continuous time implementation of the controller and non quantized position and velocity measurements, the exponential convergence of $y(t)$ was corroborated

5 Concluding remarks

A new controller based on the feedback linearization technique has been introduced in this paper.

The output function $y(t)$ was selected as a linear combination of the position and velocity tracking errors. In fact, the output function $y(t)$ is inspired from the filtered tracking error used in passivity–based controllers for fully actuated mechanical systems; see [3] for example.

The main result consisted in proving that the output function converges to zero in an exponential form, while the trajectories of the zero dynamics are uniformly ultimately bounded. Simulations and real–time experiments confirmed the validity of the main result.

Current work is focused in the application of the proposed ideas to other under actuated mechanical systems. Other possible extension is the use of integral action to define the output function $y(t)$. The addition of integral action can induce the behavior of second order system in the time evolution of the output function $y(t)$. However, this idea is also under research.

References

1. I. Fantoni and R. Lozano, *Non–Linear Control for Underactuated Mechanical Systems*, Springer–Verlag, London, 2002.
2. H. K. Khalil, *Nonlinear Systems*, Prentice Hall, Upper Saddle River, 1996.
3. A. Behal, W. E. Dixon, B. Xian, and D. M. Dawson, *Lyapunov-Based Control of Robotic Systems*, Taylor and Francis, 2009,

4. L. Sciavicco and B. Siciliano, *Modelling and Control of Robot Manipulators*, Springer–Verlag, London, 2000.
5. B. S. Cazzolato and Z. Prime, "On the dynamics of the furuta pendulum", *Journal of Control Science and Engineering*, Vol. 2011, Article ID 528341, 8 pages.
6. P. Logothetis and J. Kieffer, "On the identification of the robot dynamics without acceleration measurements", Internal Report, Faculty of Engineering and Information Technology, Australian National University, 1996. Available at `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.55.8716`
7. M. Gautier and Ph. Poignet, "Extended Kalman filtering and weighted least squares dynamic identification of robot", *Control Engineering Practice*, Vol 9, No. 12, pp. 1361–1372, 2001.
8. R. Kelly, V. Santibáñez and A. Loria, *Control of Robot Manipulators in Joint Space*, Springer–Verlag, London, 2005.

Fourier Transform Profilometry in 3-Dimensions with Matlab programming

[1]A. Nava-Vega, [2]J. A. Araiza, [3]E. Luna.

adriana.nava@uabc.edu.mx, aaraiza@gmail.com, eala@astrosen.unam.mx.

[1] Universidad Autónoma de Baja California
Facultad de Ciencias Químicas e Ingeniería
Calzada Tecnológico 14418, Otay, B.C. C.P. 22390

Instituto Nacional de Astrofísica, Óptica y Electrónica
Departamento de Óptica
Luis Enrique Erro, No.1
Tonanzintla Puebla.

[3] Universidad Nacional Autónoma de México
Instituto de Astronomía
Carretera Tijuana-Ensenada Km. 101
Ensenada, B.C.
Paper received on 04/10/12, Accepted on 25/10/12.

Abstract. A Fourier transform method is used to process images recording with fringe projection technique. In this work, the fringe projection technique is tested in objects to obtain an image series, where each image is processing applying a Fast Fourier Transform algorithm. The image it is processing, enhancement and treating using Matlab programming. It is described the algorithm employed and implemented with matlab programming, the method to record the images, the stages of the experimental recording and how these images are processing to obtain a phase, and finally a recover the surface. The algorithm is tested in simple objects (spherical ball), showing that we have a good results.

Key words: Interferometry, Image processing, Fast Fourier Transform, Profilometry, Fringe projection.

I. Introduction

Profilometry [1] is a non contact method that allows a fast and non destructive, non invasive inspection. There are many methods that use fringe projection and have been studied the last decades, among these methods there is the Fourier Transform Profilometry [2], using in this work, this technique present advantages

Mireya García-Vázquez, Grigori Sidorov (Eds.).
Advances in Computing Science and Control.
Research in Computing Science 59, 2012, pp. 113-120.

because only one image is used into the process of analysis. With a series of strips an object under test is illuminated and the characteristics of the surface are showing with the deformation of the strips, an image with fringes is recording. The surface shapes are recovered using Fourier transform, a filter in frequency domain and finally with the inverse Fourier transform process. Over the years, there has been an increase importance in development and applications of techniques for signal and image processing and some techniques are characterized with a parameter known as phase, in this area of the knowledge; signal magnitude is referred to as amplitude, whereas square magnitude is referred as intensity. In the present work, the phase parameter helps to image processing, related with amplitude and intensity.

This work described the Fourier transform procedure, the algorithm used and its implementation, to apply this Fourier transform method we used Matlab language programming. Experimental setups using fringe projection technique are described, and the obtained results with geometrical object are presented.

I.1 Fourier Transform Profilometry

The method proposed by M. Takeda it is a computer- based technique for automatic 3-D shape measurement, a grating pattern projected onto the object surface is Fourier-transformed and processed in its spatial frequency domain as well as its space signal domain. This technique has a much higher sensitivity than the conventional moiré technique [3] and is capable of fully automatic distinction between a depression and elevation on the object surface.

II. Fast Fourier Transform Procedure. Phase Determination

The phase is obtaining using the Fourier Transform Method, [2]. In optical measurements a fringe pattern can express as,

$$g(x,y) = a(x,y) + b(x,y)\cos[2\pi f_o x + \phi(x,y)], \tag{1}$$

where the phase $\phi(x,y)$ contains the shape of an object under test, and $a(x,y), b(x,y)$ represent variations in the background irradiance and contrast in the image, respectively. A fringe pattern of the form (1) is put into a computer by a CCD camera, image-sensing device that has enough resolution to satisfy the sampling-theory requirement, particularly in x-direction. The input fringe pattern is rewritten as follows,

$$g(x,y) = a(x,y) + c(x,y)\exp(2\pi i f_o x) + c^*(x,y)\exp(-2\pi i f_o x) \tag{2}$$

Where c* represent the complex conjugate, and

$$c(x,y) = (\frac{1}{2})b(x,y)\exp[i\phi(x,y)] \tag{3}$$

Equation (2) is Fourier transformed with respect to x, using a fast-Fourier-transform (FFT) algorithm, and the result is

$$G(f,y) = A(f,y) + C(f - f_o, y) + C^*(f + f_o, y) \tag{4}$$

The capital letters denote the Fourier spectra; f is the spatial frequency in the x-direction. We use either the two spectra on the carrier, say $C(f - f_o)$ and translated it by f_o on the frequency axis toward the origin to obtain $C(f,y)$. Note that the unwanted background variation $a(x,y)$ has been filtered out this stage. Again, using the FFT algorithm, we compute the inverse Fourier transform of $C(f,y)$ with respect to f and it is obtained $c(f,y)$. Then we calculate a complex logarithm of (3),

$$\log[c(x,y)] = \log[\frac{1}{2}b(x,y)] + i\phi(x,y) \tag{5}$$

Now we have the phase $\phi(x,y)$ in the imaginary part completely separate from the unwanted amplitude variation $b(x,y)$. The process it is show in a diagram of the figure 1, where the algorithm wrap.m is sketched and the process order is indicated.

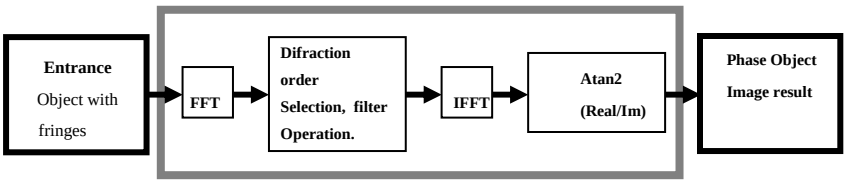


Fig. 1. Diagram of the wrap, unwrapping and phase recover algorithm, showing the stages and steps that are follow to image processing.

II.1 Algorithm implementation

The algorithm was implemented using Matlab language programming [4], logic process follows these steps:

1. An image in fits extension it is read.
2. It is call a wrapping module, that follow the next steps:

- It is obtained the FFT2 of the input image, as result it delivers a complex value.
- The former image is display in logarithm mode, it is obtained a matrix that allows to display the spectrum Fourier modes.
- It is show the image with the modes, zero order and first one (positive and negative) order.

3. It is selected one mode, usually is taking the positive first order.
4. It is call a filter (circle window).
5. At this selected region, it is apply a FFTSHIF function.
6. The result of the formal process, give us a matrix data that take the components of high frequencies to the edges.
7. It is apply the FFT2 at these results.
8. After this step it is come back to the spatial image
9. It is separate and selected the real and imaginary part.
10. It is apply each the ATAN2 function.
11. Finally we get the phase of the image
12. The Takeda method to unwrap the phase is applied.
13. The final image it is display.

III. Experiments

Some experiments were development to test the Fourier transform algorithms to image processing; an experimental setup presented in figure 2 is prepared to generate harmonic fringes programmed into the computer. This computer is used to control the fringe pattern program, fringe patterns are displayed using a Liquid Crystal Display (LCD) projector. It was displayed a fringe pattern created in Matlab programming with resolution of 25 pixels per period, it means, l line/12.5 pixel. The projection goes over an object that changes the straight fringes shape by object's topology. The physical fundamental parameters in the experimental setup are: object-camera distance, LCD- object distance.

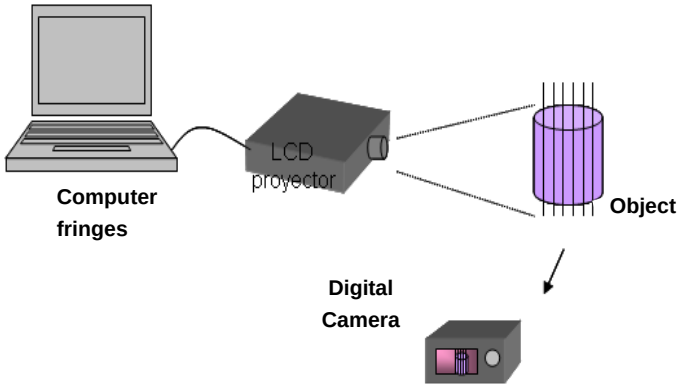
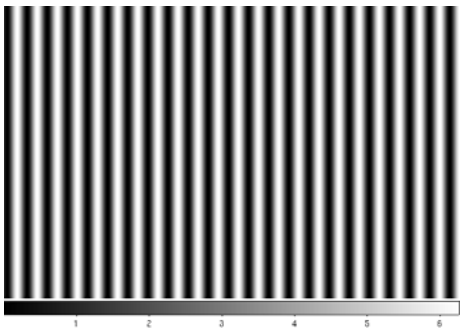


Fig. 2. Experimental Fringe projection set up. The fringes are display with a Liquid Crystal Display (LCD), the object under test is illuminated with a series of strips, controlled with a computer program, finally the image it is acquired with a digital camera.

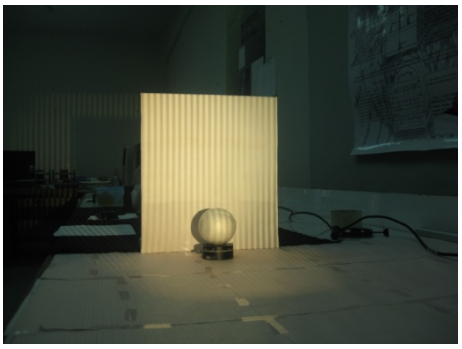
III.1 Experiments and results

It was taken images using the experimental setup described at figure 2; the images were acquiring using fringe projection techniques. The image to process has .fits extension; however, also we work with .jpg extension.

Images were acquired with a digital camera SD 1100IS, with 8.0 mega pixels, at jpg extension, as a part of the process, the image has .fits, extension, as a part of the process implemented at Matlab; besides, to manage the image, we used SAO IMAGE DS9 4.0b9 software. Figure 3 present a sequence of the process, starting from the fringe projection, through the final phase image resultant. At (a) it is presented a fringe pattern without any obstacle, as a reference image, next, at (b) we can see one angle of the experimental setup, showing a sphere with fringes over a background plane and over itself, a close up can see at (c), where it can appreciate the deformation generated by the sphere. At figure3 (d) is showing an image, as a result of processing, presenting a phase map image, we can notice some errors, as dark areas; after some more detailed processing, we can generated one (e) image with better results, along this image we indicating some contour lines and a grid in pixels.



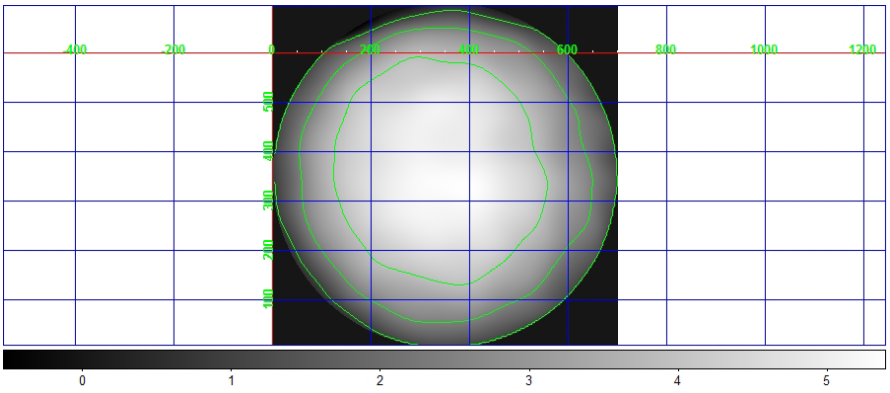
(a)



(b)

(c)

(d)



(e)

Fig. 3. Results generated by the fringe projection technique and image processing with Fourier transform. A sequence of the process is presented, at (a) we have the fringe pattern generated at the computer, (b) a sphere place into the experimental setup with fringe projection, look at the deformation at (c) , follow at background straight line, that adopt the shape of the sphere. At (d) we present the result of the image processing, here we can observe the phase of the object. We can see at (e) Unwrapped phase generated.

Conclusions

In the present work, we describe briefly the fringe profilometry technique, describing the experimental setup and the logic steps that follow a process, to get the image phase, such process used fast Fourier transform procedure. These logic processes were programming in Matlab language. We generated fringe patterns through computer programs, projected with a LCD projector; we obtained images that were processing, generating nice and quite good results to obtain the 3-dimensions topology of a sphere, as an object under test. With these results, we can continue improving the processing to expand the algorithm to automate all the experiments components [5] and expand the applications [6] of the fringe projection and Fourier transforms profilometry, for example, to vision system analysis in three dimensions.

Bibliography

1. Optical Shop Testing, Edited by Daniel Malacara, Jhon Wiley & Sons, Inc. Second Edition, (1991), ISBN: 0-471-52232-5.
2. Mitsuo Takeda, Hideki Ina, and Seiji Kobayashi, “Fourier-transform method of fringe-pattern analysis for computer-base topography and interferometry”, J.Opt.Soc.Am/Vol.72, No.1, January 1982.

3. Oster, G. and Y. Nishijima, “Moiré Patterns,” Sci. Amer, 208(5), 54-63 (May 1963).
4. Rafael C.Gonzalez, Richard E.Woods, “Digital Image Processing”, Prentice Hall, 2002. ISBN: 0-201-18075-8.
5. Mitsuo Takeda and Kazuhiro Mutoh, “Fourier transform profilometry for the automatic measurement of 3-D objects shapes, Vol. 22, No.24, Applied Optics, 2983.
6. Case, S.K., J.A. Jalkio and R.C.Kim, “3-D Vision System Analysis and Design”, in Three-Dimensional Machine Vision, Takeo Kanade, Ed. Kluwer Academic Publishers, Norwell, M.A., 1987, pp. 63-95.

Sistema de tiempo-real para el procesamiento robusto de señales de voz usando filtrado local adaptativo

Andrés J. Cuevas-Romano, Yuma Sandoval-Ibarra, Victor H. Diaz-Ramirez y Andrés Calvillo-Téllez

Instituto Politécnico Nacional - CITEDI, Avenida del parque 1310, Mesa de Otay, Tijuana B.C. 22510, México
{acuevas,sandoval,vhdiaz,calvillo}@citedi.mx
http://www.citedi.mx/
Paper received on 04/10/12, Accepted on 22/10/12.

Resumen Se presenta un sistema de tiempo-real para el procesamiento robusto de voz, implementado en un arreglo de compuertas lógicas programables (FPGA). El sistema es capaz de estimar una señal de voz limpia a partir de una fuente distorsionada usando un algoritmo basado en el calculo de estadísticas de orden prioritario dentro de una ventana deslizante. El algoritmo incrementa la calidad de voz en términos de métricas objetivas, e introduce únicamente ruido musical imperceptible. El sistema de procesamiento es adaptativo ya que puede variar los parámetros del estimador local usado de acuerdo a los cambios temporales de la señal y de la relación señal a ruido local en cada posición. Se presentan los resultados obtenidos con el sistema propuesto en términos de calidad, inteligibilidad e introducción de ruido artificial, los cuales son discutidos y comparados con los resultados obtenidos con el filtrado de Wiener y el algoritmo de sustracción espectral. Se presenta también, la evaluación del desempeño de tiempo-real del sistema implementado en el FPGA.

Keywords: Mejora de voz, filtrado local adaptativo, estadísticas de orden prioritario, FPGA, sistema en tiempo real

1. Introducción

Debido al increíble avance que han registrado los equipos móviles de comunicación, la demanda por contar con técnicas robustas para la mejora de voz en tiempo-real es hoy en día una necesidad importante. La mejora de voz, consiste en incrementar la calidad de la señal en términos de su inteligibilidad y de la reducción del ruido mediante el uso de métricas de desempeño [2,3]. La mejora de voz es un problema difícil ya que las señales son variantes en el tiempo y además, las funciones de ruido que corrompen a las señales pueden tener un comportamiento estadístico no homogéneo, lo que dificulta la separación efectiva de la voz y el ruido [1]. En la actualidad, existe un gran número de técnicas para la mejora de voz, planteadas desde distintos enfoques teóricos. Algunas de estas técnicas, utilizan un solo canal para separar la señal de voz del ruido. Otras estrategias utilizan un arreglo de sensores (micrófonos) en diferentes

Mireya García-Vázquez, Grigori Sidorov (Eds.).
Advances in Computing Science and Control.
Research in Computing Science 59, 2012, pp. 121- 132

posiciones para resolver el problema. Existe un gran número de trabajos exitosos enfocados a la supresión del ruido en sistemas mono-canal. Cuando podemos asegurar que la función de ruido tiene parámetros estadísticos estacionarios, y su densidad espectral es una constante, el algoritmo de sustracción espectral es la mejor opción [4]. En este enfoque, es necesario estimar el espectro del ruido a partir de la señal capturada de tal manera que en promedio, la relación señal a ruido (SNR) de la señal aumenta [4,5]. Por otra parte, cuando el ruido puede ser modelado como un proceso aleatorio estacionario, la estrategia que es recomendable usar es el filtrado de Wiener [1]. Este enfoque consiste en un sistema lineal que estima la señal de voz a partir de la señal ruidosa minimizando el error cuadrático promedio. Debido a que el filtrado de Wiener y el algoritmo de sustracción espectral se implementan en el dominio de la frecuencia, es muy común que ambos métodos introduzcan artefactos artificiales no deseados como el ruido musical; afectando así la inteligibilidad de la voz. Es importante observar que estas estrategias asumen que el ruido tiene parámetros estadísticos constantes a lo largo del tiempo. Sin embargo, esta suposición no es correcta en la gran mayoría de casos, por ejemplo, ante la presencia de condiciones reales, como ruido de calle o murmullos originados por conversaciones de terceras personas [7]. En este caso es necesario la incorporación de estrategias de estimación robusta para el procesamiento de la señal [8]. Un filtro robusto es usualmente diseñado para solucionar un problema estadístico de estimación con optimización de criterios de desempeño. Al tomar en cuenta las características estadísticas de las señales y del ruido se puede construir un algoritmo de filtrado local adaptativo como función de las estadísticas de orden prioritario de la señal capturada, dentro de una ventana deslizante [7]. Al usar este enfoque, el filtro es capaz de suprimir el ruido conservando los detalles finos de la señal. En procesamiento de voz estas características pueden ser de gran utilidad para aumentar la calidad de la voz sin degradar significativamente la inteligibilidad. En años recientes, los diseños de sistemas basados en FPGA han adquirido una gran popularidad debido a la flexibilidad brindada. Estos dispositivos tienen la posibilidad de reconfigurar el hardware interno conforme a las necesidades del programador y permiten tener un sistema completo dentro del chip del FPGA. Las ventajas de tener un sistema basado en FPGA, son que reduce el tamaño físico del sistema, consume poca energía y debido a que los bloques de procesamiento y los controladores están dentro del FPGA se pueden hacer cambios y mejoras sin necesidad de hacer modificaciones físicas al sistema. En este trabajo se propone un sistema de procesamiento en tiempo-real usando un algoritmo localmente adaptativo basado en estadísticas de orden prioritario para la mejora de voz en un FPGA. La estimación de la señal libre de ruido se realiza aplicando un estimador estadístico sobre las muestras de la señal dentro una ventana deslizante. El algoritmo varia el tamaño y contenido de la ventana así como la función de estimación en relación con las estadísticas locales de la señal ruidosa. Esto significa que el sistema propuesto es capaz de estimar la señal libre de ruido empleando un estimador variante con el tiempo sobre una ventana localmente adaptativa. La ventana adaptativa, es un subconjunto de los elementos de la señal dentro de la ventana deslizante los cuales son cercanos de acuerdo a un criterio específico conveniente respecto a un elemento dado [6]. Consecuentemente, el algoritmo propuesto es capaz de adaptarse a fragmentos no estacionarios de la señal y del ruido y estimar una señal mejorada haciendo uso de un estimador variante en el tiempo. Como resultado,

el método mejora la calidad de voz sin tener que sacrificar la inteligibilidad e introduciendo únicamente ruido musical imperceptible. La organización del documento se describe enseguida. En la sección 2, se presenta una breve introducción a la estimación robusta utilizando estadísticas de orden prioritario. También, se presenta la base teórica del estimador adaptativo utilizado para la mejora de voz, y se presenta el diseño del sistema para el procesamiento de voz en tiempo-real. En la sección 3, se presentan los resultados obtenidos con el sistema propuesto en términos de métricas de desempeño. Finalmente, la sección 4 presenta nuestras conclusiones.

2. Estimación robusta empleando estadísticas de orden prioritario

Consideremos una señal **f** que consiste en la superposición de una señal de voz **s** y la función de ruido **b**, como a continuación:

$$\mathbf{f} = \mathbf{s} + \mathbf{b}. \tag{1}$$

En la Ec. (1), $\mathbf{f}, \mathbf{s}$ y $\mathbf{b}$ son vectores de tamaño $N \times 1$ que representan a las secuencias discretas $f(n)$, $s(n)$ y $b(n)$. Para cada posición-i del segmento **f** podemos crear una ventana deslizante $\mathbf{w}_i$ de tamaño S, como a continuación: $\mathbf{w}_i = [f(n) : |n-i| \leq \frac{(S-1)}{2}]^T$. Estamos asumiendo, que el tamaño S es un numero impar, i es el índice de la posición central de la ventana y T representa la transpuesta del vector. El renglón variacional de $\mathbf{w}_i$ se denota como una secuencia unidimensional $\{\mathbf{V}(r); r = 1,2,\ldots,S\}$, cuyos elementos son ordenados de forma ascendente con respecto a sus valores, es decir, $\mathbf{V}(1) \leq \mathbf{V}(2), \leq \ldots \leq \mathbf{V}(S)$. Los valores $\mathbf{V}(r)$ y $r(V)$ son conocidos como la r-ésima estadística de orden prioritario y el rango del valor V, respectivamente. Ambas cantidades pueden obtenerse del histograma de la señal dentro de $\mathbf{w}_i$ [6,10].

Para crear una ventana adaptativa, es necesario obtener un subconjunto de datos alrededor del elemento central de la ventana deslizante. Existen diversos criterios para construir vecindarios adaptativos, uno de los más utilizados para la reducción de ruido es el vecindario-EV [11]. El vecindario-EV se puede construir a partir de la ventana deslizante como a continuación:

$$\mathbf{v}_i = [\mathbf{v}_i(n) = \mathbf{w}_i(n) : \mathbf{w}_i(i) - \varepsilon_v \leq \mathbf{w}_i(n) \leq \mathbf{w}_i(i) + \varepsilon_v]^T, \tag{2}$$

donde ε_v es un valor constante. En la teoría de estimación robusta existen varios tipos de estimadores de localización de parámetros que se pueden utilizar para estimar el valor del elemento central del vecindario. Nuestra meta es utilizar un estimador robusto para separar la señal de voz de la función de ruido a partir de la ventana adaptativa. El estimador L es uno de los estimadores robustos más populares y se basa en la combinación lineal de estadísticas de orden prioritario. Este estimador es muy popular debido a su simplicidad y gran robustez. El estimador L que se calcula sobre el vecindario adaptativo esta dado por [12]

$$y(i) = \mathbf{a}^T \mathbf{v}_i, \tag{3}$$

donde $\mathbf{v}_i$ es la ventana adaptativa y **a** es un vector de coeficientes de peso, ambos de tamaño $S_A \times 1$. Sea $\mathbf{V}_i$ una matriz diagonal de los elementos del vector $\mathbf{v}_i$. Si calculamos

$\mathbf{R} = \mathbf{V}_i \times \mathbf{V}_i^*$, se obtiene $(y(i))^2 = \mathbf{a}^T\mathbf{R}\mathbf{a}$. Sea x_i un valor de referencia que se asume es muy cercano al valor central de la ventana deslizante de la señal libre de ruido. El error cuadrático entre la estimación y_i y el valor x_i esta dado por $(e(i))^2 = \mathbf{a}^T\mathbf{R}\mathbf{a} - 2\mathbf{a}^T\mathbf{r}$, donde $\mathbf{r} = x(i)\mathbf{v}_i$ es un vector de tamaño $S_A \times 1$. Para obtener un estimador libre de sesgo, debe cumplirse la condición $\mathbf{a}^T\mathbf{u} = 1$, donde $\mathbf{u}$ es un vector unitario de tamaño $S_A \times 1$. Aplicando el método de multiplicadores de Lagrange [12], podemos encontrar el vector de coeficientes como a continuación:

$$\mathbf{a} = \mathbf{R}^{-1}\left[\mathbf{r} + \frac{\mathbf{u}(1 - \mathbf{u}^T\mathbf{R}^{-1}\mathbf{r})}{\mathbf{u}^T\mathbf{R}^{-1}\mathbf{u}}\right]. \tag{4}$$

Finalmente, el vector $\mathbf{r}$ puede calcularse como

$$\mathbf{r} = \mathbf{v}_i\left[\mu_{\mathbf{v}} + S\hat{N}R_i\left(f(i) - \mu_{\mathbf{v}}\right)\right], \tag{5}$$

donde $\mu_{\mathbf{v}}$ es el valor promedio de la ventana adaptativa y $S\hat{N}R_i$ es la relación señal a ruido normalizada "[0,1]" de la señal ruidosa en la i-esima posición de $f(n)$.

2.1. Algoritmo de procesamiento de voz usando estimación robusta

En la Tabla 2.1, se muestra el pseudocódigo del algoritmo propuesto usado para la mejora de voz.

Tabla 1. Psedocódigo y complejidad computacional del algoritmo propuesto.

No.	Instrucción	Tiempo	Complejidad
1.	N = tamaño(**f**)	C_1	$O(1)$
2.	S=tamaño($\mathbf{w}_i$)	C_2	$O(1)$
3.	k = 1	C_3	$O(1)$
4.	for i = 1 to N	C_4	$O(N+1)$
5.	$\mathbf{w}_k = [f(n) : \lvert n-k\rvert \leq (S-1)/2]^T$	C_5	$O(NS)$
6.	**rv** = renglonVaracional($\mathbf{w}_i$)	C_6	$O(NS\ log\ S)$
7.	**v** = conjuntoEV(**rv**)	C_7	$O(NS)$
8.	S_A = tamaño(**v**)	C_8	$O(NS)$
9.	$\mathbf{a}_i$ = calcCoef (**v**)	C_{10}	$O(NS)$
10.	for j = 1 to S_A	C_{11}	$O(N(S_A+1))$
11.	$\mathbf{a}_j^T\mathbf{v}_j$	C_{12}	$O(NS_A)$
12.	if(k < N)	C_{13}	$O(N)$
13.	k = k + 1	C_{14}	$O(N-1)$

De acuerdo al pseudocódigo presentado en la Tabla 2.1, la complejidad computacional del algoritmo se puede expresar por $T(n) = O(NS\ log\ S)$. Observemos que la complejidad computacional del algoritmo adaptativo es polinomial y crece con respecto al tamaño S de la ventana deslizante. Si el tamaño de la ventana es pequeño, el tiempo de procesamiento es corto pero se corre el riesgo de realizar una estimación con información insuficiente en cada posición de la ventana. Si la ventana deslizante es grande,

el tiempo computacional crece considerablemente, sin embargo, la estimación realizada en cada posición de la ventana puede llevarse a cabo usando información suficiente, y así, obtener una estimación precisa. Para tener una buena estimación, se recomienda utilizar el tamaño de la ventana deslizante de al menos dos veces el periodo fundamental de la señal de voz.

2.2. Sistema de tiempo-real usando un FPGA

En esta sección se presenta la implementación realizada en el FPGA del algoritmo adaptativo descrito en la sección 2.1. Primero se realiza una descripción de la configuración del hardware interno del FPGA. Posteriormente, se describe la implementación basada en el procesador embebido Nios II [9]. La tarjeta de desarrollo donde se implemento el sistema de procesamiento de voz, es la tarjeta Altera DE2-115 que cuenta con el chip Cyclone IV EP4CE115. Nuestro objetivó, es configurar un sistema capaz de procesar la señal entrante a velocidad alta; es decir, se requiere que el bloque de procesamiento sea lo suficientemente rápido para que el sistema mantenga un flujo de datos de salida constante, que evite pérdidas por traslape de datos. Para tener el mejor desempeño se usó la configuración Nios II/f (fast) que ópera a mayor velocidad a expensas de utilizar un mayor número de elementos lógicos en el FPGA. Para interconectar los núcleos de propiedad intelectual (IP Cores) se usa la herramienta SOPC Builder. Un núcleo de propiedad intelectual es un modulo de hardware prefabricado que se puede usar en diseños específicos para reducir el tiempo de desarrollo. En la Fig. 1 se observa el diagrama del sistema dentro del FPGA generado con SOPC Builder. En este diseño, se usaron los núcleos de Nios II, JTAG, PLL, controlador (entrada/salida) PIO, Audio, pantalla LCD, y SRAM. Como se observa en la Fig. 1 los núcleos fueron interconectados usando el SIF (system interconnect fabric) que es un sistema de recursos lógicos para interconectar los componentes del software SOPC Builder.

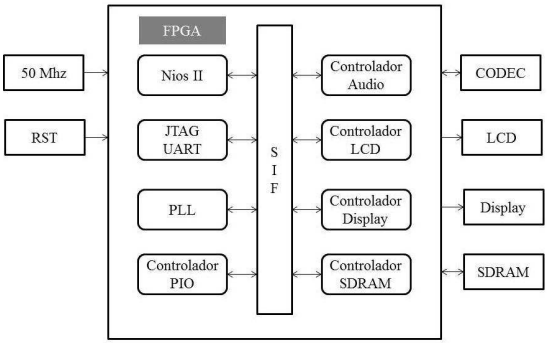


Figura 1. Diagrama a bloques del sistema FPGA usando la herramienta SOPC.

Implementación en Nios II Existen dos tipos de formatos de datos comúnmente usados para el procesamiento digital de señales, punto fijo y punto flotante. Estos formatos

se pueden usar en aplicaciones basadas en FPGAs, sin embargo, en este trabajo se utiliza el formato punto fijo ya que el CODEC de la tarjeta usa datos enteros signados con complemento a dos. El formato punto fijo signado es usado para representar números negativos y positivos, mientras que el formato sin signo solo representa números no negativos. La Tabla 2 muestra el formato numérico usando enteros de 16 y 32 bits. La razón de usar un formato de 16 bits para la implementación es que con un entero de 16 bits se tiene un rango dinámico máximo de 96 dB; que es aproximadamente el rango dinámico del oído humano. En la Tabla 3 se observan algunas conversiones de bits a decibeles.

Tabla 2. Rango dinámico de datos para enteros de 16 y 32 bits.

Formato	Rango
16 bits sin signo	0 a 65,535
16 bits signado	−32,768 a 32,767
32 bits sin signo	0 a 4,294,967,295
32 bits signado	−2,147,483,648 a 2,147,483,647

Tabla 3. Relación entre amplitud y dB.

Amplitud de 16 bits	dB
32,767	0
24,576	−2.4987
16,384	−6.0206
8,192	−12.0412
3,277	−20
1,638	−26.0206
328	−40

La frecuencia de muestreo que se utilizo para procesar las señales de voz es de 8 Khz y los datos son codificados como enteros signados de 16 bits. En la Fig. 2 el CODEC captura la señal de voz entrante a través de la entrada de micrófono (mic) y codifica cada muestra para ser procesada dentro del FPGA. Posteriormente, los datos se dirigen de regreso hacia el CODEC para su reproducción. El algoritmo adaptativo hace su procesamiento en segmentos de datos y se requiere que el sistema haga su procesamiento en alta velocidad, por lo tanto, el bloque de datos no puede ser muy grande ya que tomará una gran cantidad de tiempo para procesar. Adicionalmente, la longitud del segmento no puede ser demasiado pequeña ya que podrían haber interrupciones en el flujo de datos debido a que entrarían mas datos de los que es posible procesar. Es importante notar que el tamaño del bloque de datos es un parámetro ajustable para tener mejor desempeño de latencia o una mejor calidad de procesamiento y depende de la capacidad de procesamiento del sistema. Para la implementación realizada, el tamaño del segmento de datos a procesar es de 200 muestras. Cabe mencionar que 200 datos muestreados a 8KHz/s equivale a un bloque de datos de 25 ms.

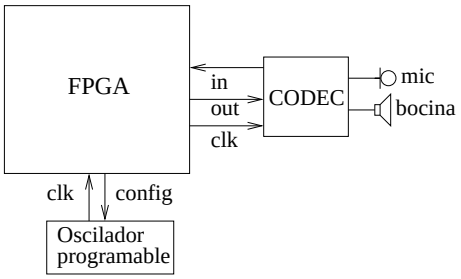


Figura 2. Arquitectura del sistema tiempo-real en el FPGA

3. Evaluación del desempeño del sistema propuesto de tiempo-real

En esta sección se presentan los resultados obtenidos con el sistema propuesto en diferentes implementaciones digitales realizadas. Los resultados obtenidos, son comparados respecto a las siguientes técnicas: algoritmo de sustracción espectral y filtrado de Wiener. La calidad de los resultados esta dada en términos de la métrica de evaluación perceptual de la calidad (PESQ) [15] y la medida de inteligibilidad de tiempo corto objetiva (STOI) [17]. También, se evalúa el índice de artefactos artificiales introducidos por el algoritmo de procesamiento con la métrica SAR (Source Artifacts Ratio), el nivel de distorsión que introducen los algoritmos de procesamiento a la señal con la media SDR (Source to Distortion Ratio), y la capacidad que tienen los algoritmos para suprimir ruido con la métrica SIR (Sourceto Interference Ratio).

3.1. Evaluación del desempeño del algoritmo adaptativo

Para evaluar el desempeño de los algoritmos de procesamiento se usaron los archivos de voz de la base de datos NOIZEUS [14]. Estos archivos, consisten en treinta oraciones producidas por diferentes locutores, capturadas con una frecuencia de muestreo de 8Khz, cuentan con 2^{16} niveles de cuantización y están codificados en formato ".wav". Los archivos de voz se contaminaron con diferentes fuentes de ruido ambiental: ruido de tránsito vehicular, ruido de calle y murmullos con un valor de SNR de 5, 10 y 15 dB. Las señales de voz contaminadas por las fuentes de ruido fueron procesadas con los métodos de substracción espectral, filtrado de Wiener y el algoritmo adaptativo. Cuando la señal de voz esta contaminada con ruido vehicular con un valor de SNR de 15 dB, los parámetros utilizados por el algoritmo propuesto son $S = 65$, $k_1 = 1$ y $k_2 = 0.8$. En la Fig 3 se puede apreciar un ejemplo de una señal procesada con el algoritmo propuesto, en comparación con la señal ruidosa y la señal libre de ruido. También se muestran los espectrogramas de la señal ruidosa y de la señal procesada. La linea negra describe el comportamiento del valor ε_v estimado.

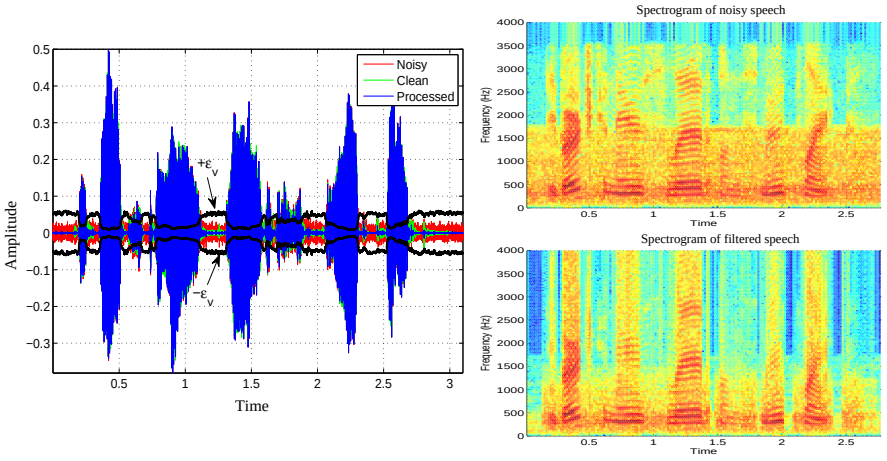


Figura 3. Señal de voz distorsionada con ruido de transito vehicular, su espectrograma y el espectrograma de la señal procesada con el algoritmo propuesto.

Se evaluaron 30 señales de voz con las diferentes funciones de ruido. En las Tablas 4 -12 se muestran los intervalos de confianza con un nivel del 95 % para las diferentes métricas usadas en la evaluación del desempeño de los algoritmos.

Tabla 4. Intervalos de confianza del 95 % cuando la señal de voz esta contaminada con ruido vehicular con 15 dB de SNR.

	Propuesto	**Wiener**	**SpecSub**
SIR	7.72 ± 0.05	19.84 ± 0.31	9.16 ± 0.11
PESQ	2.05 ± 0.01	1.81 ± 0.03	2.08 ± 0.01
STOI	0.75 ± 0.01	0.66 ± 0.01	0.73 ± 0.01
SDR	6.89 ± 0.04	8.51 ± 0.08	7.32 ± 0.07
SAR	15.16 ± 0.06	8.89 ± 0.09	12.47 ± 0.20

Tabla 5. Intervalos de confianza del 95 % cuando la señal de voz esta contaminada con ruido vehicular con 10 dB de SNR.

	Propuesto	**Wiener**	**SpecSub**
SIR	13.70 ± 0.17	18.07 ± 0.78	12.91 ± 0.45
PESQ	2.18 ± 0.04	1.84 ± 0.05	2.21 ± 0.03
STOI	0.80 ± 0.01	0.72 ± 0.01	0.79 ± 0.01
SDR	11.48 ± 0.10	10.28 ± 0.09	10.72 ± 0.11
SAR	15.67 ± 0.10	11.22 ± 0.12	15.37 ± 0.56

Tabla 6. Intervalos de confianza del 95 % cuando la señal de voz esta contaminada con ruido vehicular con 5 dB de SNR.

	Propuesto	**Wiener**	**SpecSub**
SIR	9.33 ± 0.10	19.75 ± 0.50	9.17 ± 0.16
PESQ	2.05 ± 0.01	1.78 ± 0.03	2.05 ± 0.01
STOI	0.75 ± 0.00	0.65 ± 0.01	0.72 ± 0.01
SDR	7.49 ± 0.07	8.50 ± 0.08	7.21 ± 0.07
SAR	12.61 ± 0.07	8.90 ± 0.08	12.17 ± 0.26

Cuando las señales de voz están contaminadas con ruido vehicular el algoritmo que presenta los mejores resultados en términos de calidad (PESQ) es el método de sustracción de espectral, sin embargo, introduce ruido musical muy notorio e introduce un alto nivel de distorsión (ver nivel de SDR). El filtrado de Wiener es el que presenta mejor capacidad para eliminar el ruido (buen nivel de SIR) pero sacrifica inteligibilidad (STOI) y calidad (PESQ), ya que degrada los detalles de la señal. El algoritmo propuesto presenta buenos resultados en términos de calidad (PESQ) preservando la inteligibilidad (STOI), además el porcentaje de distorsión (SDR) es tolerable y no introduce ruido musical perceptible. En las Tablas 7-9 se muestran los intervalos de confianza con un nivel del 95 % de las diferentes métricas de evaluación del desempeño de los algoritmos cuando la señal de voz esta contaminada con ruido de murmullos para un SNR de 15, 10 y 5 dB.

Tabla 7. Intervalos de confianza del 95 % cuando la señal de voz esta contaminada con ruido de murmullo con 15 dB de SNR.

	Propuesto	**Wiener**	**SpecSub**
SIR	18.27 ± 0.14	21.48 ± 0.33	18.20 ± 0.31
PESQ	2.70 ± 0.01	2.23 ± 0.02	2.71 ± 0.02
STOI	0.89 ± 0.00	0.84 ± 0.00	0.90 ± 0.00
SDR	16.51 ± 0.09	12.35 ± 0.08	15.36 ± 0.15
SAR	21.39 ± 0.09	12.98 ± 0.12	19.18 ± 0.70

Tabla 8. Intervalos de confianza del 95 % cuando la señal de voz esta contaminada con ruido de murmullo con 10 dB de SNR.

	Propuesto	**Wiener**	**SpecSub**
SIR	12.90 ± 0.16	18.47 ± 0.61	13.07 ± 0.51
PESQ	2.38 ± 0.02	1.93 ± 0.04	2.39 ± 0.03
STOI	0.82 ± 0.00	0.76 ± 0.01	0.83 ± 0.01
SDR	11.49 ± 0.10	10.32 ± 0.11	10.95 ± 0.13
SAR	17.31 ± 0.14	11.16 ± 0.19	16.09 ± 0.92

Tabla 9. Intervalos de confianza del 95 % cuando la señal de voz esta contaminada con ruido de murmullo con 5 dB de SNR.

	Propuesto	**Wiener**	**SpecSub**
SIR	6.18 ± 0.08	9.68 ± 1.05	6.62 ± 0.41
PESQ	2.17 ± 0.05	1.61 ± 0.07	2.17 ± 0.05
STOI	0.76 ± 0.01	0.64 ± 0.01	0.73 ± 0.01
SDR	5.64 ± 0.08	5.80 ± 0.32	5.50 ± 0.23
SAR	15.90 ± 0.09	8.87 ± 0.20	13.21 ± 0.57

Cuando las señales de voz están contaminadas con ruido de murmullos, el algoritmo propuesto presenta buenos resultados en términos de inteligibilidad (STOI) y calidad (PESQ). Ademas, el porcentaje de distorsión (SDR) es tolerable y el ruido musical introducido por el algoritmo es imperceptible (el mejor valor de SAR). En las Tablas 10-12 se muestran los intervalos de confianza con un nivel del 95 % de las diferentes métricas para evaluar el desempeño de los algoritmos cuando la señal de voz esta contaminada con ruido de calle con un nivel de SNR de 15, 10 y 5 dB.

Tabla 10. Intervalos de confianza del 95 % cuando la señal de voz esta contaminada con ruido de calle con 15 dB de SNR.

	Propuesto	**Wiener**	**SpecSub**
SIR	18.55 ± 0.33	22.62 ± 0.82	18.17 ± 0.57
PESQ	2.51 ± 0.02	2.18 ± 0.04	2.48 ± 0.02
STOI	0.86 ± 0.00	0.79 ± 0.01	0.87 ± 0.00
SDR	16.42 ± 0.17	12.48 ± 0.14	14.93 ± 0.26
SAR	20.66 ± 0.15	13.01 ± 0.23	18.40 ± 0.85

Tabla 11. Intervalos de confianza del 95 % cuando la señal de voz esta contaminada con ruido de calle con 10 dB de SNR.

	Propuesto	**Wiener**	**SpecSub**
SIR	11.96 ± 0.18	17.12 ± 1.27	12.05 ± 0.59
PESQ	2.42 ± 0.04	2.10 ± 0.08	2.43 ± 0.04
STOI	0.81 ± 0.01	0.76 ± 0.02	0.82 ± 0.01
SDR	10.80 ± 0.09	10.45 ± 0.20	10.34 ± 0.11
SAR	17.47 ± 0.28	12.00 ± 0.50	17.32 ± 1.42

Tabla 12. Intervalos de confianza del 95 % cuando la señal de voz esta contaminada con ruido de calle con 5 dB de SNR.

	Propuesto	**Wiener**	**SpecSub**
SIR	6.65 ± 0.11	11.39 ± 1.17	6.68 ± 0.46
PESQ	1.99 ± 0.05	1.48 ± 0.07	1.88 ± 0.05
STOI	0.76 ± 0.01	0.61 ± 0.03	0.70 ± 0.02
SDR	5.98 ± 0.10	6.33 ± 0.20	5.19 ± 0.23
SAR	15.26 ± 0.20	8.93 ± 0.56	13.06 ± 1.61

Podemos observar cuando la señal de voz esta contaminada con ruido de calle, el algoritmo de sustracción espectral presenta buenos resultados en la inteligibilidad (STOI) y calidad (PESQ), pero introduce ruido musical muy notorio y un alto nivel de distorsión (SDR). El filtrado de Wiener tiene mayor capacidad para eliminar el ruido (ver nivel de SIR), sin embargo, distorsiona considerablemente la señal (SDR) e introduce un nivel importante de artefactos artificiales (SAR). El algoritmo adaptativo da buenos resultados en términos de calidad e inteligibilidad, y no introduce ruido musical perceptible. Podemos observar en los resultados obtenidos que el algoritmo propuesto es robusto porque es capaz de adaptarse bien a las características no homogéneas de la señal de entrada así como al comportamiento no estacionario de la funciones de ruido. Además, el sistema propuesto es capaz de eliminar el ruido sin introducir ruido musical perceptible.

3.2. Evaluación del desempeño del sistema en tiempo-real

En esta sección se presenta la evaluación del desempeño del sistema en tiempo-real basado en FPGA. La señal de entrada fue procesada con el algoritmo adaptativo usando los parámetros $S = 35$, $k_1 = 1$ y $k_2 = 0.8$. Para evaluar el desempeño de tiempo-real del sistema se midió latencia del FPGA usando un tono de 600 Hz como señal de entrada. El resultado obtenido es una latencia de 64 ms cuando el filtro esta desactivado, es decir, cuando la señal de entrada al CODEC pasa directamente al bloque de salida a través del procesador NIOS II sin sufrir modificación alguna. Una vez activado el bloque de procesamiento, se introducen 16 ms adicionales para obtener un total de 80 ms de latencia. Estos resultados pueden verse en la Fig.4.

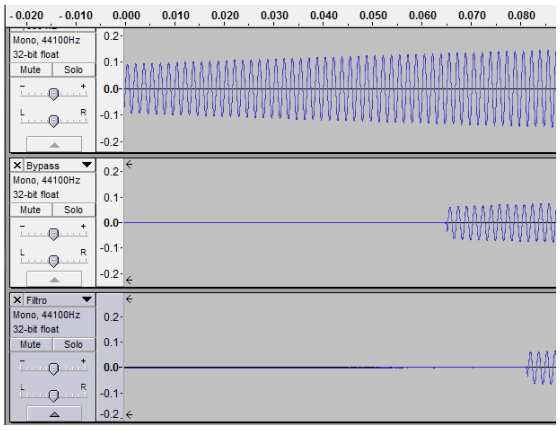


Figura 4. Latencia del sistema: (arriba) tono de entrada de 600 Hz, (centro) salida del sistema sin procesamiento, (abajo) salida del sistema con procesamiento activado.

4. Conclusiones

Se implemento un sistema de tiempo-real basado en un FPGA para la mejora de voz utilizando estimadores robustos de orden prioritario. El sistema es capaz procesar señales de voz a velocidad alta. El algoritmo implementado es capaz de reducir los efectos del ruido sin introducir ruido musical perceptible. En base a las pruebas realizadas, el sistema propuesto es capaz de incrementar la calidad de una señal de voz ruidosa, sin degradar la inteligibilidad e introduciendo bajos niveles de ruido artificial. El sistema implementado en el FPGA, puede ser utilizado para aplicaciones de tiempo-real.

Referencias

1. J Benesty, and S. Makino and J. Chen, *Seepch Enhancement*, Springer Series on Signals and Communitacion Technology, 2005.
2. Philips C. Loizou, *Speech Enhancement: theory and practice*, Taylor & Francis, 2007.
3. Yi Hu and Philips C. Loizou, *Subjective comparision and evaluation of speech enhancement algorithms*, Speech Communication, vol. 49, pp. 588-601, 2007.
4. S. F. Boll, *Suppression of Acoustic Noise in Speech Using Spectral Subtration*, IEEE Transactions Acoustics Speech Singal Process, vol. 27, no. 2, pp. 113-120, 1979.
5. R. J. McAulay, and M. L. Malpass, *Speech Enhacement Using a Soft-Desicion Noise Suppression Filter*, IEEE Transactions Acoustics Speech Singal Process, vol. 28, pp. 137-145, 1980.
6. L. Yaroslavsky and M. Eden, *Fundamentals of digital optics*. Boston:Birkhäuser, 1996.
7. E. Hansler and G. Schmidt, Eds., *Speech and audio processing inadverse environments*, ser. Signals and Communication Technology. Springer, 2008.
8. Jaakko Astola and Pauli Kuosmanen, *Fundamentals of Nolinear Digital Filtering*, Series Edit by Fidker and Phil Mars, 1997.
9. V. L. Richard Boulanger, Ed., *The Audio Programming Book*. The MIT Press, 2011
10. V. Kober and M. Mozerov and J. Alvarez-Borrego and I. A. Ovseyevich, *Rank Image Processing Using Spatially Adaptative Neigborhoods*, Pattern Recognition and Image Analysis. 2001;(3):542-552.
11. Yuma Sandoval Ibarra, Vìctor H. Díaz-Ramírez, and Juan J. Tapia Armenta, Algoritmo de orden localmente-adaptativo para la mejora de señales de voz, Congreso Internacional de Ciencias de Computación CICOMP 10, 2010.
12. Vìctor H. Díaz-Ramírez, and Andres J. Cuevas-Romano, *Speech processing using local adaptive rank-order estimators*, III Encuentro Internacional Académico y de Investigación y VII Encuentro Regional Académico, 2011.
13. H. Hirsch, and D. Pearce, *The Aurora Experimental Framework for the Performance Evaluation of Speech Recognition Systems under Noisy Conditions.* ISCA ITRW ASR2000, Paris, France, September 18-20, 2000.
14. *IEEE Recommended Practice for Speech Quality Measurements* IIEEE Trans. Audio and Electroacoustics, AU-17(3), 225-246, 1969.
15. *Perceptual evaluation of speech quality (PESQ), and objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs.* ITU-T Recommendation P. 862, 2000.
16. *Objective measurement of active speech level* 2000.
17. Cees H.Tall, Richard C. Hendriks, Richard Heusdens, and Jesper Jensen *An Algorithm for Intelligibility Prediction of Time-Frequency Weighted Noise Speech* IEEE Transactions on Audio, Speech and Languaje Processing, vol. 19, 2125-2136, 2011.

Reconocimiento facial robusto usando filtros de correlación diseñados a través de optimización combinatoria

Sergio Pinto-Fernández[*], Alejandra Serrano-Trujillo[*], Víctor H. Díaz-Ramírez[*] y Leonardo Trujillo Reyes[**]

*Instituto Politécnico Nacional – CITEDI, Ave. del Parque 1310
Mesa de Otay, Tijuana B.C., México
**Instituto Tecnológico de Tijuana, Ave. Tecnológico, Fracc. Tomás Aquino,
Tijuana B.C., México
{spinto, aserrano, vhdiaz}@citedi.mx,
leonardo.trujillo.ttl@gmail.com
Paper received on 04/10/12, Accepted on 25/10/12.

Resumen. El desempeño de los filtros compuestos de correlación para el reconocimiento de patrones depende de la adecuada selección de las imágenes de entrenamiento usadas para la síntesis de los filtros. Comúnmente, estas imágenes son elegidas de forma subjetiva por el diseñador en base su experiencia, por lo que no es posible garantizar una selección óptima. En este trabajo, se plantea el uso de un algoritmo evolutivo para la optimización combinatoria de imágenes de entrenamiento para la síntesis de filtros compuestos de correlación, usados en el reconocimiento facial. Dado un conjunto de imágenes disponibles de rostros el algoritmo encuentra la combinación óptima de imágenes de entrenamiento para la síntesis de un filtro de correlación con el mejor desempeño en términos de métricas de calidad. Como resultado, el reconocimiento facial con filtros de correlación mejora sustancialmente. Los resultados obtenidos con el algoritmo propuesto en pruebas de reconocimiento facial son presentados y discutidos en términos del desempeño con métricas objetivas y eficiencia de clasificación.

Palabras clave: Reconocimiento facial, filtros compuestos de correlación, optimización combinatoria, algoritmos evolutivos.

1. Introducción

El desarrollo de algoritmos robustos para el reconocimiento facial se ha mantenido como una línea de investigación de creciente interés, dado a la gran demanda que existe por contar con sistemas automáticos confiables para el reconocimiento de rostros. El reconocimiento facial es ampliamente utilizado en distintas aplicaciones, sin embargo, su principal área de uso se ubica en sistemas de seguridad. El problema del reconocimiento facial consiste en la validación de la identidad de una

Mireya García-Vázquez, Grigori Sidorov (Eds.).
Advances in Computing Science and Control.
Research in Computing Science 59, 2012, pp. 133-144.

persona, usando únicamente una imagen de su rostro. Este problema se puede abordar por distintas técnicas, por ejemplo, a través de sistemas basados en características o por sistemas basados en filtros de correlación. Los sistemas basados en características son quizás los sistemas más utilizados debido a que pueden aplicarse en una gran cantidad de problemas y comúnmente entregan buenos resultados. Sin embargo, su principal debilidad es que su desempeño general depende de la toma de decisiones subjetivas por parte del programador que podrían optimizarse mediante procedimientos formales. En los sistemas basados en filtros de correlación las coordenadas del valor de máxima intensidad en la función de salida del sistema es el estimador de máxima verosimilitud de las coordenadas del patrón objetivo en la escena observada. Comúnmente, un filtro de correlación se diseña mediante la optimización de métricas objetivas de calidad planteadas en términos de los modelos de la señal y el ruido. Un filtro de correlación compuesto, es aquel cuya respuesta al impulso está dada por una combinación de imágenes de entrenamiento que deben ser representativas del patrón objetivo y de sus diferentes versiones distorsionadas esperadas. Los filtros compuestos tienen la ventaja de que pueden construirse con el fin de obtener un desempeño optimizado para un conjunto de patrones fijos a ser identificados y para discriminar un conjunto de objetos falsos conocidos, como el fondo o cualquier otro objeto no deseado. Uno puede notar que el desempeño de un filtro de este tipo dependerá en gran medida de la correcta selección de las imágenes de entrenamiento. Aun más, esta selección puede variar considerablemente dependiendo de la aplicación. Como resultado, no es posible garantizar un óptimo desempeño al no existir un procedimiento formal para tal selección.

Existen propuestas basadas en procesos iterativos, sugeridas con el fin de resolver el problema de la selección de imágenes de entrenamiento [1,2]. Se considera que el conjunto resultante de imágenes permite la síntesis de un filtro compuesto con el mejor desempeño en términos de criterios específicos. Una limitación de las propuestas existentes es que el filtro resultante es incapaz de controlar por completo el plano de correlación a la salida, debido a que sólo se consideran imágenes del patrón objetivo para el entrenamiento. Por lo tanto, será muy probable la aparición de lóbulos laterales en el plano de correlación de salida, que son ocasionados por el fondo o cualquier otro objeto. Consecuentemente, podrían generarse errores por falsas alarmas y una baja capacidad de discriminación. Para evitar esta problemática, el diseño del filtro puede realizarse a través de un algoritmo de adaptación, utilizando en un enfoque de construcción incremental [3]. Un algoritmo de este tipo, busca patrones falsos en el área del fondo para ser rechazados en cada iteración. De esta forma, la capacidad de discriminación del filtro se incrementa monotónicamente hasta llegar a un valor aceptable. Cabe notar que el problema de la selección de las mejores imágenes de entrenamiento para un filtro compuesto es un problema de optimización combinatoria. Esto es, dadas múltiples vistas del patrón objetivo, la meta es realizar la elección de una combinación de tales vistas que, al ser utilizadas en la síntesis de un filtro compuesto, se genere el mejor desempeño con respecto a un criterio determinado, comparada con todas las otras posibles combinaciones. Por otro lado, debe observarse que la cantidad de imágenes de entrenamiento utilizadas para la síntesis de un filtro compuesto no debe incrementarse deliberadamente. Kumar y Prochavsky [4] demostraron que la relación señal

a ruido (SNR) de un filtro compuesto se reduce gradualmente al incrementar la cantidad de imágenes de entrenamiento, lo que en este planteamiento se traduce como la necesidad de obtener un conjunto pequeño de imágenes de entrenamiento que genere el mejor desempeño en términos de métricas determinadas.

En este trabajo se propone el diseño y la implementación de un algoritmo iterativo para la síntesis de filtros de correlación compuestos, optimizados para el reconocimiento facial. Dado un conjunto de imágenes de rostros, el algoritmo busca la combinación óptima de imágenes que sintetice el mejor filtro en términos de métricas de desempeño. Para ejecutar tal búsqueda, el algoritmo utiliza un enfoque evolutivo, evitando así el diseño incremental comúnmente usado.

Este documento se organiza de la siguiente manera. En la sección 2 se presenta una breve descripción de los filtros compuestos de correlación para el reconocimiento de objetos. En la sección 3, se describe la base de datos de rostros utilizada en los experimentos realizados. La sección 4 describe el algoritmo evolutivo propuesto usado para la síntesis de filtros de correlación optimizados para el reconocimiento facial. Los resultados obtenidos con el método propuesto a través de simulaciones por computadora, se presentan en la sección 5. Estos resultados, son analizados y discutidos en términos de confiabilidad en el reconocimiento facial. Finalmente, la sección 6 presenta nuestras conclusiones.

2. Filtros de correlación compuestos

En esta sección, se presenta el diseño de dos de los filtros compuestos de correlación para reconocimiento de objetos más populares: el filtro de altura máxima de correlación promedio (MACH) [5] y el filtro SDF de compromiso óptimo (OTSDF) [6,7]. Estos filtros fueron diseñados para producir picos de correlación delgados y altos en la localización del patrón objetivo en la escena observada, y para producir valores de intensidad cercanos a cero en el área ocupada por cualquier objeto desconocido. La Fig. 1 muestra el diagrama a bloques de un sistema básico de reconocimiento de objetos con filtros de correlación. El procedimiento para la detección consiste en la transformación al dominio de la frecuencia de la imagen de entrada y su espectro es multiplicado por la respuesta en frecuencia de un filtro compuesto. El resultado de esta operación se convierte al dominio espacial obteniéndose así un plano de correlación. Las coordenadas del valor de máxima intensidad, conocido como pico de correlación, indicarán la posición del objetivo.

Sea $\{T\} = \{T_i(\mu,\nu); i = 1,\ldots,N\}$ un conjunto de N distintas imágenes expresadas en dominio de la frecuencia, donde cada una representa una versión distorsionada del patrón a detectar $t(x,y)$. Aquí, $T(\mu,\nu)$ es la transformada de Fourier de $t(x,y)$. Los filtros compuestos deben ser capaces de reconocer al patrón objetivo $t(x,y)$ y a todas sus versiones distorsionadas en $\{T\}$ con sólo una operación de correlación.

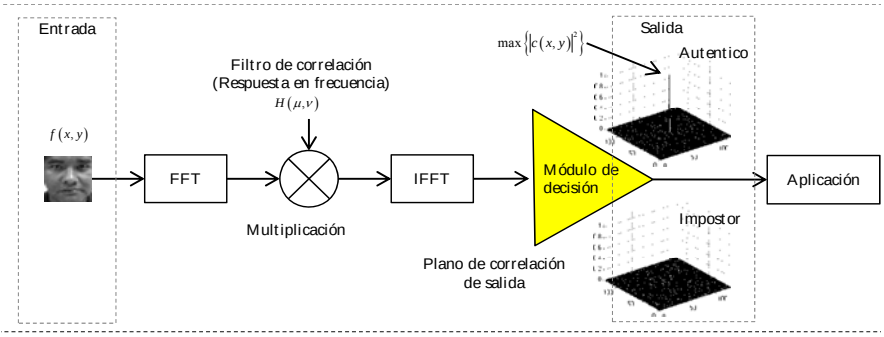


Fig. 1. Diagrama a bloques de un sistema básico de reconocimiento de objetos con filtros de correlación

2.1 Filtro de altura máxima de correlación promedio (MACH)

Un filtro MACH $\mathbf{h}_{mach}$ se diseña para maximizar la relación entre la intensidad de la altura de correlación promedio (ACH) a la salida y la medida de similitud promedio (ASM) a través de todas las imágenes de entrenamiento [5]. En otras palabras, un filtro MACH está diseñado para maximizar la función $J = |ACH|^2 / ASM$. Sean $\mathbf{X}_i$ y $\mathbf{M}$, ambas matrices diagonales de tamaño $d \times d$ que contienen los elementos de los vectores de entrenamiento $\mathbf{t}_i$, y el vector de entrenamiento promedio, dado por

$$\mathbf{m} = \frac{1}{\mathrm{N}} \sum_{i=1}^{\mathrm{N}} \mathbf{t}_i \ . \tag{1}$$

La ACH se define como el promedio de los valores de intensidad a la salida del filtro en respuesta a las imágenes de entrenamiento; esta métrica puede expresarse como

$$ACH = \frac{1}{\mathrm{N} \cdot d} \sum_{i=1}^{\mathrm{N}} \mathbf{t}_i^{+} \mathbf{h}_{mach} \ . \tag{2}$$

Adicionalmente, la ASM puede ser vista como el error promedio entre las respuestas de correlación producidas por las imágenes de entrenamiento $\mathbf{v}_i = \mathbf{X}_i^{*} \mathbf{h}_{mach}$, y la función de correlación producida por la imagen de entrenamiento promedio $\overline{\mathbf{v}} = \mathbf{M}^{*} \mathbf{h}_{mach}$, es decir

$$ASM = \frac{1}{\mathrm{N} \cdot d} \sum_{i=1}^{\mathrm{N}} |\mathbf{t}_i - \mathbf{m}|^2 \ . \tag{3}$$

Utilizando una notación matriz-vectorial, la ACH y ASM pueden ser reescritas como

$$ACH = \mathbf{m}^{+} \mathbf{h}_{mach} , \tag{4}$$

y

$$\begin{aligned} ASM &= \mathbf{h}^{+}\left[\frac{1}{\mathrm{N}\cdot d}\sum_{i=1}^{\mathrm{N}}(\mathbf{t}_i-\mathbf{m})(\mathbf{t}_i-\mathbf{m})^{*}\right]\mathbf{h}, \\ &= \mathbf{h}^{+}\mathbf{S}\mathbf{h}, \end{aligned} \tag{5}$$

donde

$$\mathbf{S}=\left[\frac{1}{\mathrm{N}\cdot d}\sum_{i=1}^{\mathrm{N}}(\mathbf{t}_i-\mathbf{m})(\mathbf{t}_i-\mathbf{m})^{*}\right]. \tag{6}$$

El filtro MACH $\mathbf{h}_{mach}$ se obtiene maximizando la función objetivo

$$J(\mathbf{h}_{mach})=\frac{|ACH|^2}{ASM}=\frac{\mathbf{h}_{mach}^{+}\mathbf{m}\mathbf{m}\mathbf{h}_{mach}}{\mathbf{h}_{mach}\mathbf{S}\mathbf{h}_{mach}}, \tag{7}$$

donde el filtro MACH resultante está dado por [5]

$$\mathbf{h}_{mach}=\mathbf{S}^{-1}\mathbf{m}. \tag{8}$$

2.2 Filtro SDF de compromiso óptimo (OTSDF)

El filtro MACH optimiza tanto la ASM como la ACH. No obstante, en ocasiones el filtro MACH presenta problemas en la aparición de lóbulos laterales en el plano de correlación de salida, los que pueden disminuir el desempeño del filtro. Este problema puede resolverse minimizando la energía de correlación promedio (ACE), lo que provoca la generación de valores cercanos a cero en todo el plano de correlación, excepto en los valores de correlación correspondientes a la localización de las imágenes de entrenamiento. Sea $\mathbf{D}$ una matriz diagonal de $d\times d$ elementos, en la que los valores de la diagonal principal están definidos por $E\left\{|\mathbf{t}_i|^2; i=1,\ldots,\mathrm{N}\right\}$; es decir, por el espectro de energía promedio de las imágenes de entrenamiento en la escena. Sea $\mathbf{h}$ cualquier filtro de correlación, por lo tanto, la ACE puede ser calculada como [6]

$$ACE=\mathbf{h}^{+}\mathbf{D}\mathbf{h}. \tag{9}$$

Es importante notar que las métricas ACE y ASM están en conflicto entre sí, por lo que debe establecerse un compromiso entre ellas. El filtro OTSDF está diseñado para realizar un compromiso entre varias métricas que están en conflicto. Un filtro OTSDF que establece un compromiso entre las métricas ACE, ASM y ACH puede obtenerse minimizando la siguiente función objetivo:

$$\begin{aligned} J\left(\mathbf{h}_{otsdf}\right) &= \omega_1 ACE+\omega_2 ASM-|ACH|, \\ &= \omega_1\mathbf{h}_{otsdf}^{+}\mathbf{D}\mathbf{h}_{otsdf}+\omega_2\mathbf{h}_{otsdf}^{+}\mathbf{S}\mathbf{h}_{otsdf}-\left|\mathbf{h}_{otsdf}^{+}\mathbf{m}\right|. \end{aligned} \tag{10}$$

En la Ec. (10), la ACE y ASM son funciones a ser minimizadas, ACH es una función a ser maximizada y $\omega_1+\omega_2=1$ son las constantes de compromiso. El filtro OTSDF resultante, está dado por [7]

$$\mathbf{h}_{otsdf}=\left(\omega_2\mathbf{S}+\omega_2\mathbf{D}\right)^{-1}\mathbf{m}. \tag{11}$$

3. Base de datos de los rostros

En esta sección se describe la base de datos de rostros utilizada en las pruebas realizadas con el algoritmo propuesto. Usaremos la base de datos de la CMU AMP Face Expression Database [8], que se compone de 13 diferentes sujetos. Por cada sujeto distinto existen 75 imágenes con diferentes expresiones faciales. Las imágenes son monocromáticas, de tamaño 64 x 64 pixeles y fueron tomadas bajo condiciones idénticas de iluminación, por lo que solo cambios de expresiones faciales son considerados. La Tabla 1 muestra ejemplos de diferentes imágenes de rostros por sujeto contenidas en la base de datos.

Tabla 1. Ejemplos de rostros contenidos en la CMU AMP Face Expression Database.

Sujeto:	1	2	3	4	5	6	7
Índices:	1-75	76-150	151-225	226-300	301-375	376-450	451-525
							
Sujeto:	8	9	10	11	12	13	
Índices:	526-600	601-675	676-750	751-825	826-900	901-975	
							

Sea $\{S_k\}$ una secuencia de $N_S = 75$ imágenes correspondientes al k-ésimo sujeto de la base de datos. Además, sea $\{U\} = \bigcup_{k=1}^{13} \{S_k\}$ la secuencia completa de todas las $N_U = 975$ imágenes de la base de datos. Para asegurar resultados estadísticos correctos en nuestros experimentos, hemos dividido cada uno de los conjuntos de imágenes correspondientes a cada sujeto de la base de datos en dos subconjuntos $\{S_k\} = \{A_k\} \cup \{T_k\}$ donde $\{A_k\}$ y $\{T_k\}$ son llamados el subconjunto de imágenes disponibles y el subconjunto de imágenes de prueba, respectivamente. Las imágenes disponibles $\{a_k^i(x,y) \in \{A_k\}; i = 1,\ldots,N_A\}$ se asumen como conocidas; por lo que todas pueden ser elegidas para el entrenamiento de los filtros. Las imágenes de prueba $\{t_k^i(x,y) \in \{T_k\}; i = 1,\ldots,N_T\}$ se asumen como desconocidas, por lo que sólo pueden ser usadas para probar el desempeño de los filtros. El subconjunto de imágenes disponibles se forma con $N_A = 35$ imágenes, que fueron elegidas de $\{S_k\}$ al azar, mediante el uso de una distribución uniforme. Las $N_F = N_S - N_A$ imágenes restantes se identifican como imágenes de entrenamiento. La Tabla 2 muestra la distribución de los índices de las imágenes de los rostros utilizadas en los subconjuntos $\{A_k\}$.

Tabla 2. Índice de las imágenes de la base de datos en los subconjuntos disponibles $\{A_k\}$.

Sujeto	Índices $\{U\}$	Índices de las imágenes $\{U\}$ utilizadas en $\{A_k\}$
$\{S_1\}$	1-75	$\{A_1\}$: índices {6 ,7 ,8 ,9 ,15 ,16 ,17 ,19 ,22 ,24 ,26 ,28 ,29 ,30 ,33 ,37 ,39 ,43 ,47 ,49 ,51 ,53 ,54 ,55 ,56 ,59 ,60 ,61 ,63 ,67 ,69 ,70 ,72 ,74 ,75}
$\{S_2\}$	76-150	$\{A_2\}$: índices {76 , 79 , 80 , 81 , 84 , 86 , 88 , 89 , 91 , 92 , 93 , 95 , 96 , 97 ,100 ,104 ,105 ,114 ,120 ,121 ,122 ,123 ,127 ,131 ,132 ,133 ,134 ,137 ,138 ,139 ,140 ,141 ,142 ,143 ,144}
$\{S_3\}$	151-225	$\{A_3\}$: índices {152 ,153 ,154 ,155 ,157 ,160 ,163 ,164 ,165 ,171 ,172 ,173 ,174 ,175 ,176 ,178 ,179 ,181 ,182 ,185 ,186 ,187 ,190 ,201 ,204 ,206 ,207 ,209 ,211 ,212 ,213 ,216 ,219 ,222 ,224}
$\{S_4\}$	226-300	$\{A_4\}$: índices {226 ,228 ,230 ,231 ,232 ,234 ,235 ,236 ,237 ,238 ,240 ,242 ,247 ,249 ,250 ,251 ,253 ,261 ,268 ,269 ,273 ,274 ,276 ,277 ,279 ,281 ,285 ,288 ,289 ,290 ,292 ,294 ,296 ,298 ,299}
$\{S_5\}$	301-375	$\{A_5\}$: índices {301 ,302 ,305 ,310 ,311 ,316 ,317 ,318 ,319 ,322 ,323 ,324 ,325 ,331 ,333 ,334 ,335 ,337 ,340 ,344 ,347 ,349 ,350 ,352 ,357 ,359 ,363 ,365 ,366 ,367 ,369 ,370 ,372 ,373 ,375}
$\{S_6\}$	376-450	$\{A_6\}$: índices {376 ,377 ,381 ,382 ,383 ,385 ,395 ,401 ,403 ,404 ,406 ,407 ,410 ,412 ,413 ,414 ,417 ,419 ,421 ,422 ,423 ,426 ,429 ,430 ,433 ,435 ,436 ,439 ,441 ,442 ,444 ,445 ,446 ,449 ,450}
$\{S_7\}$	451-525	$\{A_7\}$: índices {452 ,453 ,454 ,455 ,459 ,461 ,462 ,465 ,466 ,469 ,470 ,471 ,473 ,474 ,477 ,480 ,482 ,487 ,488 ,489 ,493 ,497 ,502 ,503 ,504 ,507 ,508 ,509 ,510 ,514 ,516 ,517 ,519 ,521 ,524}
$\{S_8\}$	526-600	$\{A_8\}$: índices {528 ,529 ,535 ,536 ,538 ,539 ,550 ,552 ,554 ,556 ,558 ,560 ,562 ,563 ,565 ,566 ,569 ,573 ,574 ,575 ,577 ,580 ,581 ,582 ,584 ,585 ,587 ,588 ,589 ,593 ,594 ,596 ,597 ,598 ,600}
$\{S_9\}$	601-675	$\{A_9\}$: índices {601 ,608 ,609 ,610 ,611 ,612 ,613 ,615 ,619 ,620 ,624 ,630 ,634 ,635 ,638 ,642 ,644 ,645 ,646 ,647 ,651 ,652 ,654 ,656 ,659 ,660 ,661 ,663 ,664 ,665 ,666 ,667 ,669 ,673 ,675}
$\{S_{10}\}$	676-750	$\{A_{10}\}$: índices {677 ,678 ,679 ,680 ,682 ,686 ,687 ,688 ,690 ,691 ,692 ,694 ,695 ,696 ,697 ,701 ,703 ,706 ,709 ,713 ,714 ,717 ,719 ,720 ,724 ,726 ,728 ,732 ,735 ,742 ,744 ,746 ,748 ,749 ,750}
$\{S_{11}\}$	751-825	$\{A_{11}\}$: índices {751 ,754 ,755 ,756 ,757 ,759 ,760 ,761 ,763 ,767 ,768 ,771 ,772 ,773 ,778 ,780 ,781 ,782 ,783 ,785 ,786 ,787 ,793 ,796 ,797 ,801 ,803 ,804 ,806 ,808 ,812 ,817 ,821 ,822 ,823}
$\{S_{12}\}$	826-900	$\{A_{12}\}$: índices {827 ,829 ,830 ,831 ,833 ,834 ,838 ,839 ,840 ,845 ,846 ,847 ,848 ,854 ,856 ,857 ,858 ,859 ,863 ,864 ,868 ,869 ,870 ,872 ,875 ,881 ,882 ,887 ,888 ,889 ,894 ,895 ,896 ,898 ,900}
$\{S_{13}\}$	901-975	$\{A_{13}\}$: índices {901 ,903 ,904 ,906 ,909 ,910 ,911 ,913 ,914 ,915 ,917 ,919 ,923 ,926 ,927 ,928 ,929 ,930 ,933 ,934 ,935 ,936 ,937 ,941 ,953 ,957 ,958 ,961 ,962 ,965 ,967 ,971 ,972 ,973 ,975}

4. Diseño de filtros compuestos de correlación utilizando optimización combinatoria

En esta sección se describe el algoritmo propuesto para la síntesis de filtros compuestos de correlación optimizados para el reconocimiento facial. Dado un conjunto $\{A_k\}$ de imágenes disponibles, el algoritmo busca un subconjunto óptimo $\{O_k\} \subseteq \{A_k\}$ representando las imágenes de entrenamiento elegidas. Es esperado que con las imágenes de rostros contenidas en $\{O_k\}$ se sintetice el filtro compuesto con el mejor desempeño en comparación con cualquier otra combinación de imágenes seleccionada. El desempeño del filtro puede caracterizarse en términos de una métrica de calidad específica. Para ilustrar la complejidad del problema de búsqueda, consideremos el caso de un conjunto de imágenes disponibles con $N_A = 35$ imágenes. En este caso, existen 34,359,738,367 posibles soluciones (combinaciones), por lo que una búsqueda exhaustiva seria intratable. El diagrama de bloques del algoritmo propuesto se presenta en la Fig. 2, y el procedimiento se explica de la siguiente manera:

- Paso 1: Lectura de todas las imágenes del subconjunto $\{A_k\}$.
- Paso 2: Generación de una población inicial de soluciones candidatas (combinaciones de imágenes a partir de $\{A_k\}$). Cada conjunto solución se codifica

como una cadena de longitud variable de números enteros generados al azar; de tamaño máximo N_A .

- Paso 3: Evaluar la aptitud de cada individuo de la población; esto se realiza de la siguiente manera: Sintetizar un filtro compuesto $\mathbf{h}_{current}$ utilizando las imágenes de rostros codificados en la cadena actual. A continuación, se evalúa el desempeño de $\mathbf{h}_{current}$ en el reconocimiento de rostros codificados en la cadena correspondiente. La función de aptitud busca maximizar la función

$$J(\mathbf{h}_{current}) = \frac{|ACH|^2}{\omega_1 ASM + \omega_2 ACE}. \tag{12}$$

- Paso 4: Evaluar la condición de parada, definida como un número máximo de generaciones. Si ésta condición no se ha alcanzado, continuar al paso 5.
- Paso 5: Marcar los mejores individuos dentro de la población actual, es decir, elegir los individuos con mayor aptitud (selección del 50%).
- Paso 6: Generar nuevos individuos utilizando el operador genético (cruza) basado en codificación de longitud variable [9].
- Paso 7: Remplazar los peores individuos por los nuevos individuos generados en el paso 6.
- Paso 8: Aplicar el operador de mutación a la población (5% a cada elemento de cada individuo). Regresar al paso 3.

Puede observarse que al utilizar el algoritmo propuesto es posible encontrar un conjunto de imágenes de entrenamiento que genere el mejor desempeño en términos de un compromiso entre las métricas ACH, ACE y ASM. Esto significa que el filtro resultante tendrá un desempeño balanceado entre tolerancia a distorsiones y capacidad de discriminar patrones desconocidos. Es importante observar que para que los sujetos de la base de datos puedan ser reconocidos y clasificados a través de filtros de correlación diseñados con el algoritmo propuesto, es necesario diseñar un sistema de reconocimiento de patrones basado en un banco de filtros.

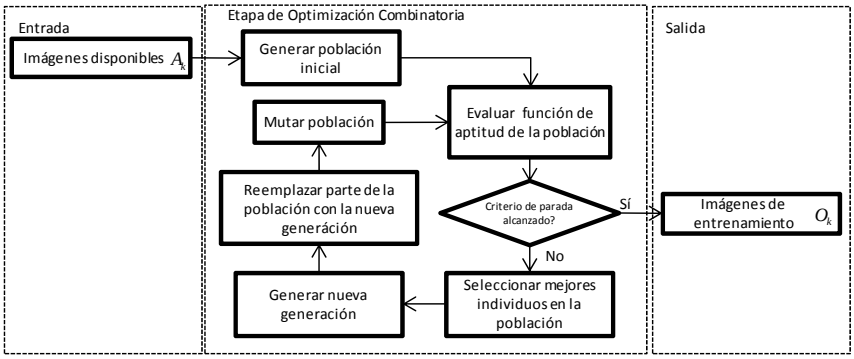


Fig. 2. Diagrama de bloques del algoritmo iterativo propuesto.

El sistema de reconocimiento debe ser capaz de validar la identidad de una imagen de entrada arbitraria perteneciente a la base de datos. El diagrama de bloques para construir un banco de filtros se presenta en la Fig. 3.

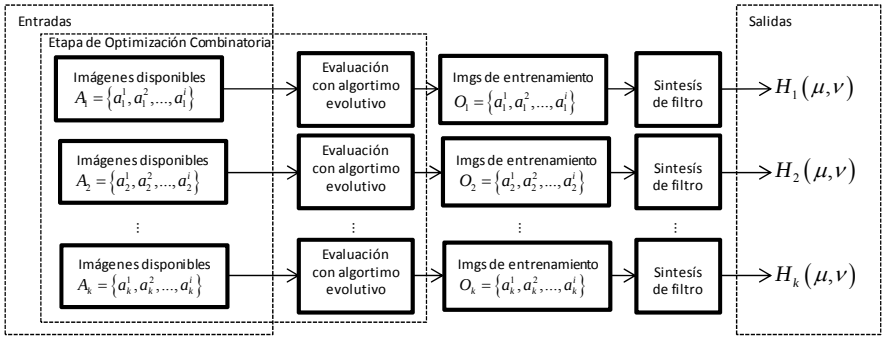


Fig. 3. Diagrama de bloques para la construcción de un banco de filtros compuestos entrenados con el algoritmo mostrado en la Fig. 2.

5. Resultados

En esta sección se presentan los resultados obtenidos con la metodología propuesta en la evaluación del reconocimiento facial con la base de datos. Los resultados obtenidos se presentan en términos del desempeño del reconocimiento facial evaluado por medio del porcentaje de eficiencia de clasificación correcta. Los resultados obtenidos con el método propuesto serán comparados con los resultados del sistema basado en características propuesto por Abusham et al [10]. Para evaluar el desempeño del algoritmo propuesto, se construyeron trece filtros de correlación (un filtro por cada sujeto) con la ayuda del algoritmo iterativo de la Fig. 2. Adicionalmente, el algoritmo se ha probado para diferentes valores de ω_1 y ω_2. Los filtros resultantes fueron entrenados usando únicamente imágenes del conjunto $\{A_k\}$. El sistema de reconocimiento facial utilizado para evaluar el desempeño de los filtros se presenta en la Fig. 4, y el procedimiento se explica a continuación. Sea $t_k^i(x,y)$ la *i*-ésima imagen de prueba a lo largo del *k*-ésimo sujeto; se busca que esta imagen sea identificada por el sistema de reconocimiento de patrones. Primero, $t_k^i(x,y)$ es transformada al dominio de Fourier y su espectro es multiplicado por la respuesta en frecuencia de cada filtro dentro del banco $\{H_k(\mu,\nu); k = 1,\ldots,13\}$. Los espectros resultantes son transformados al dominio espacial para obtener el conjunto de planos de correlación $\{cp_k(x,y); k = 1,\ldots,13\}$. Posteriormente, el sistema calcula la capacidad de discriminación (DC) para cada plano de correlación de la siguiente manera:

$$DC_k = 1 - \frac{MAX\{p_i; i = 1,\ldots,k = 13; (i \neq k)\}}{p_k}, \tag{13}$$

donde $p_k = MAX\{|cp_k(x,y)|^2\}$ es el valor de máxima intensidad en la función $cp_k(x,y)$. Nótese que el valor de DC para cada filtro es calculado respecto a los picos de correlación generados por los otros filtros dentro del banco. De esta manera, se espera que el filtro entrenado con las imágenes de los rostros correspondientes al sujeto de entrada mantenga un valor alto de DC. A continuación, el sistema localiza el filtro con el valor de DC más alto, y si este valor es DC $\geq$ 0.5 entonces la imagen de entrada se considera reconocida, y es consecuentemente asignada a la misma categoría del sujeto para el cual fue entrenado el filtro. De otra manera, la imagen de entrada se considera rechazada.

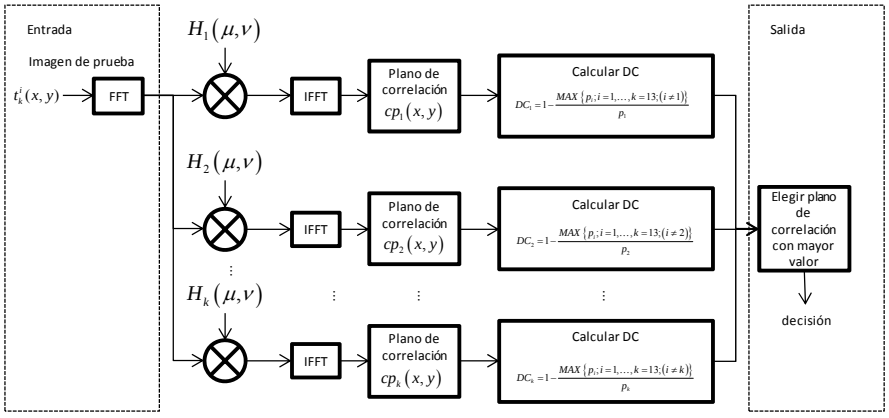


Fig. 4. Sistema de reconocimiento de patrones usado para evaluar el desempeño de los filtros compuestos de correlación entrenados con el algoritmo propuesto.

Para evaluar el desempeño del sistema propuesto, se realizaron experimentos para reconocer y clasificar las imágenes de los rostros de la base de datos. En nuestros experimentos, se utilizaron solamente las imágenes en los subconjuntos $\{T_k\}$ como las imágenes de prueba de entrada. Los resultados obtenidos se resumen en la Tabla 3, las entradas muestran el número de imágenes de entrada pertenecientes a cada sujeto (denominador) frente a las decisiones tomadas por el sistema de reconocimiento de patrones (numerador). La Tabla 3 muestra también el porcentaje de clasificación correcta obtenido en cada experimento. Los resultados obtenidos cuando $\omega_1 = 0.3$ y $\omega_1 = 0.5$ alcanzan un 100% de clasificación correcta. Para el caso de $\omega_1 = 0.7$ se ha clasificado una de las imágenes pertenecientes al sujeto 10 como si fuese el sujeto 8. Los experimentos realizados, también permiten observar el comportamiento del diseño adaptativo para los diferentes valores que puede tomar ω_1 y ω_2. Valores de ω_1 cercanos a la unidad caracterizan al filtro con más tolerancia a las distorsiones y produce que el número de imágenes de entrenamiento necesarias, disminuyan conforme se incrementa el valor de ω_1.

Tabla 3. Desempeño de clasificación del sistema de reconocimiento facial para valores de ω .

			Sujeto de entrada														
			$\{T_1\}$	$\{T_2\}$	$\{T_3\}$	$\{T_4\}$	$\{T_5\}$	$\{T_6\}$	$\{T_7\}$	$\{T_8\}$	$\{T_9\}$	$\{T_{10}\}$	$\{T_{11}\}$	$\{T_{12}\}$	$\{T_{13}\}$	Rechaza-dos	Correcta-mente Clasificados
Decisión	ω_1=0.3, ω_2=0.7	1	40/40	0	0	0	0	0	0	0	0	0	0	0	0	0	100
		2	0	40/40	0	0	0	0	0	0	0	0	0	0	0	0	100
		3	0	0	40/40	0	0	0	0	0	0	0	0	0	0	0	100
		4	0	0	0	40/40	0	0	0	0	0	0	0	0	0	0	100
		5	0	0	0	0	40/40	0	0	0	0	0	0	0	0	0	100
		6	0	0	0	0	0	40/40	0	0	0	0	0	0	0	0	100
		7	0	0	0	0	0	0	40/40	0	0	0	0	0	0	0	100
		8	0	0	0	0	0	0	0	40/40	0	0	0	0	0	0	100
		9	0	0	0	0	0	0	0	0	40/40	0	0	0	0	0	100
		10	0	0	0	0	0	0	0	0	0	40/40	0	0	0	0	100
		11	0	0	0	0	0	0	0	0	0	0	40/40	0	0	0	100
		12	0	0	0	0	0	0	0	0	0	0	0	40/40	0	0	100
		13	0	0	0	0	0	0	0	0	0	0	0	0	40/40	0	100
	ω_1=0.5, ω_2=0.5	1	40/40	0	0	0	0	0	0	0	0	0	0	0	0	0	100
		2	0	40/40	0	0	0	0	0	0	0	0	0	0	0	0	100
		3	0	0	40/40	0	0	0	0	0	0	0	0	0	0	0	100
		4	0	0	0	40/40	0	0	0	0	0	0	0	0	0	0	100
		5	0	0	0	0	40/40	0	0	0	0	0	0	0	0	0	100
		6	0	0	0	0	0	40/40	0	0	0	0	0	0	0	0	100
		7	0	0	0	0	0	0	40/40	0	0	0	0	0	0	0	100
		8	0	0	0	0	0	0	0	40/40	0	1	0	0	0	0	100
		9	0	0	0	0	0	0	0	0	40/40	0	0	0	0	0	100
		10	0	0	0	0	0	0	0	0	0	39/40	0	0	0	0	97.5
		11	0	0	0	0	0	0	0	0	0	0	40/40	0	0	0	100
		12	0	0	0	0	0	0	0	0	0	0	0	40/40	0	0	100
		13	0	0	0	0	0	0	0	0	0	0	0	0	40/40	0	100
	ω_1=0.7, ω_2=0.3	1	40/40	0	0	0	0	0	0	0	0	0	0	0	0	0	100
		2	0	40/40	0	0	0	0	0	0	0	0	0	0	0	0	100
		3	0	0	40/40	0	0	0	0	0	0	0	0	0	0	0	100
		4	0	0	0	40/40	0	0	0	0	0	0	0	0	0	0	100
		5	0	0	0	0	40/40	0	0	0	0	0	0	0	0	0	100
		6	0	0	0	0	0	40/40	0	0	0	0	0	0	0	0	100
		7	0	0	0	0	0	0	40/40	0	0	0	0	0	0	0	100
		8	0	0	0	0	0	0	0	40/40	0	1	0	0	0	0	100
		9	0	0	0	0	0	0	0	0	40/40	0	0	0	0	0	100
		10	0	0	0	0	0	0	0	0	0	39/40	0	0	0	0	97.5
		11	0	0	0	0	0	0	0	0	0	0	40/40	0	0	0	100
		12	0	0	0	0	0	0	0	0	0	0	0	40/40	0	0	100
		13	0	0	0	0	0	0	0	0	0	0	0	0	40/40	0	100

En base a los resultados obtenidos, podemos observar que los filtros de correlación diseñados a través de optimización combinatoria producen altos índices de clasificación en el reconocimiento facial. Notemos que el peor índice de clasificación correcta alcanzado con el enfoque propuesto es del 97.5%. Este índice de clasificación es superior al mejor caso del sistema basado en características propuesto por Abusham et al [10], donde el índice de clasificación correcta es del 90%. En base a estos resultados, podemos decir que el reconocimiento facial con filtros de correlación es una atractiva alternativa al los sistemas basados en características.

6. Conclusiones

Se ha presentado un algoritmo iterativo basado en optimización combinatoria para la síntesis de filtros de correlación para reconocimiento facial. Se ha mostrado que dado un conjunto de imágenes de rostros, el algoritmo propuesto es capaz de encontrar una combinación óptima de imágenes de entrenamiento para la síntesis de un filtro compuesto de correlación con un desempeño optimizado para reconocimiento facial en términos de las métricas ACH, ACE y ASM. Los resultados de simulaciones por computadora obtenidos con esta propuesta exhiben altas tasas de clasificación para distintas combinaciones de parámetros del algoritmo iterativo.

7. Referencias

1. Vijaya Kumar, B.V.K., "Efficient approach to designing linear combination filters," Appl. Opt., 22(10), 1445-1448 (1983).
2. Casasent, D. and Chang, W. T., "Correlation synthetic discriminant functions," *Appl. Opt.* 25(14), 2343-2350 (1986).
3. Diaz-Ramirez et al., "Multiclass pattern recognition using adaptive correlation filters with complex constraints," *Opt. Eng.* 51(3), (Marzo 2012).
4. Vijaya Kumar , B.V.K., and Pochapsky , E., "Signal-to-noise ratio consideration in modified matched spatial filters," *J. Opt. Soc. Am. A.* 3(6), 777-786 (1986).
5. Mahalanobis, A., Vijaya Kumar, B.V.K., Song, S., Sims, S.R.F., and Epperson, J.F., "Unconstrained correlation filters," *App. Opt.* 33(17), 3751–3759 (June 1994).
6. Refregier, P. "Optimal trade-off filters for noise robustness, sharpness of the correlation peak and horner efficiency". Opt. Lett. 16(11), 829-31 (1991).
7. Vijaya Kumar, B.V.K., Carlson, D., and Mahalanobis, A. "Optimal tradeoff synthetic discriminant function (OTSDF) filters for arbitrary devices," Opt. Lett. 19(19), 1556-1558 (1994).
8. “CMU AMP Face Expression Database”, <http://chenlab.ece.cornell.edu/downloads.html>(01 Octubre 2012)
9. Inman Harvey. The saga cross: The mechanics of recombination for species with variable-length genotypes. In R. Manner and B. Manderick, (Eds.), Parallel Problem, pp. 269-278. North-Holland, 1992.
10. E. E. Abusham, D. Ngo, A. Teoh, “Fusion of locally linear embedding and principal component analysis for face recognition (FLLEPCA)”, 3rd International Conference on Advances in Patten Recognition (ICAPR '05), vol. 3687, pp. 326–333, 2005.

Evidencia de mejora en los sistemas de reconocimiento basados en iris, utilizando esquemas adaptados de fusión de imágenes

Juan M. Colores-Vargas[1], Mireya S. García-Vázquez[1], Alejandro A. Ramírez-Acosta[2] y Héctor M. Pérez-Meana[3]

[1] Centro de Investigación y Desarrollo de Tecnología Digital (CITEDI-IPN)
Avenida del Parque 1310, Tijuana, B.C. México 22510
[2] MIRAL R&D, Palm Garden, Imperial Beach, USA 91932
[3] Sección de Graduados de Mecánica y Eléctrica (ESIME-IPN), DF., México

[1]{colores, mgarciav}@citedi.mx, [2]ramacos10@hotmail.com, [3]hmperezm@ipn.mx
Paper received on 22/09/12, Accepted on 23/10/12.

Resumen. Actualmente las investigaciones sobre el desarrollo de sistemas de reconocimiento basados en iris se enfocan en los sistemas que operan bajo esquemas de adquisición no cooperativos y/o controlados, los cuales debido a su naturaleza de adquisición son limitados en información biométrica y por ende necesitan de métodos muy sofisticados para generar altos índices de reconocimiento. En este artículo se presenta el trabajo donde se propone utilizar técnicas de fusión para captar mayor información biométrica a partir de video-iris, mediante la conformación de plantillas biométricas digitales que integren las características de un grupo de marcos del video, incrementando los índices de reconocimiento del iris en un sistema no-cooperativo. Se analizan siete técnicas de fusión sobre un subconjunto de marcos de la base de datos MBGC.v2, los resultados hacen evidente la utilidad de las técnicas de fusión al lograr una mayor extracción de información biométrica. El método de fusión por PCA, presenta el mejor desempeño al mejorar los valores de reconocimiento en función a las distancias Hamming en aproximadamente 83% de los experimentos.

Palabras Clave: Fusión, Iris, MBGC, PCA, Reconocimiento.

1 Introducción

Actualmente los sistemas comerciales de reconocimiento funcionan bajo esquemas y principios propuestos en las decadas de los 80s y 90s ([1-2]) basicamente operan; adquiriendo imágenes o video del ojo de la persona, posteriormente dicha informacion es procesada para acceder a la textura del iris, la cual se utiliza para

Mireya García-Vázquez, Grigori Sidorov (Eds.).
Advances in Computing Science and Control.
Research in Computing Science 59, 2012, pp. 145-156.

generar un codigo digital asociado que es empleado como identificador. Los sistemas biometricos basados en iris, se diseñan para operar particularmente en entornos controlados y bajo condiciones de adquisición con participación activa del usuario. Esta participación en algunos casos puede ser indeseable pero necesaria para lograr adquirir información con calidad idónea para el reconocimiento, ya que influye directamente en el desempeño de los sistemas [3].

De manera general, los sistemas de reconocimiento biométricos basados en el iris, se contituyen de cuatro etapas principales mostradas en la figura 1 [1-2]. En la primera etapa (*adquisición de imagen*), se adquiere la imagen o video del ojo de la persona a ser reconocida. Es importante adquirir información que presente un buen contraste y enfoque que permita distinguir los detalles en la región del iris. Una vez adquirida la imagen digital, se envia a la siguiente etapa (*pre-procesamiento*). En la etapa de pre-procesamiento se realizan dos tareas: la segmentación y la normalización. La segmentación consiste en calcular los parámetros que determinan la ubicación y tamaño de la región del iris dentro de la imagen digital. La normalización tiene como objetivo, transformar la textura segmentada en una plantilla digital con dimensiones estandarizadas; ésta permitirá invarianza frente a variaciones en el proceso de adquisición.

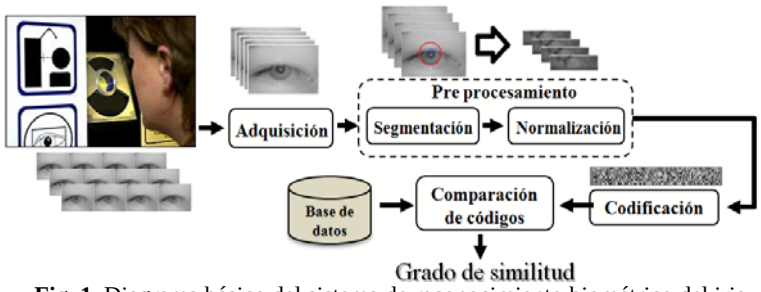


Fig. 1. Diagrama básico del sistema de reconocimiento biométrico del iris.

En la siguiente etapa (*codificación*), la textura normalizada del iris es procesada para extraer la información biométrica presente en la plantilla digital de base, permitiendo reconocer una persona de otra. El resultado es un código asociado "único" llamado plantilla digital biométrica. Cuando una identificación o verificación es requerida, se repite cada una de las etapas, comparando la plantilla digital biometrica contra las plantillas digitales biométricas almacenadas en la base de datos (*comparación de codigo*). Una decisión exitosa sera definida mediante una medida de similitud de códigos que establece la correlacion entre ambas plantillas comparadas, usualmente se basa en el porcentaje de similitud obtenido, si las plantillas digitales biométricas comparadas fueron generadas por el mismo iris, o sea, por la misma persona.

Actualmente las áreas más activas de investigación se han enfocado hacia el desarrollo de sistemas biométricos "no-cooperativos", es decir, que limitan el tiempo de exposición y el comportamiento participativo del usuario en la etapa de adquisición. La naturaleza de la etapa de adquisición en estos sistemas induce ruido, afectando de manera severa la calidad de las imágenes o video del iris [4]. Para lograr un sistema de tales magnitudes, es necesario el re-diseño de las etapas críticas del sistema de reconocimiento:

- *Sistema de adquisición:* un sistema no-cooperativo deberá permitir incrementar la distancia de adquisición y disminuir el tiempo de exposición entre el usuario y la cámara, es decir, realizar la adquisición de la información biométrica del usuario, estando éste en movimiento. Este tipo de sistema es más susceptible a los efectos de desenfoque y a las distorsiones proyectivas; afectando la textura del iris a un grado tal que no es posible analizar la textura del iris para identificar a una persona (ver Fig. 2a).
- *Pre-procesamiento:* En un sistema no-cooperativo, es realista suponer que debido a las condiciones variantes del sistema de adquisición se van a presentar una serie de características como obstrucciones, desenfoque, reflexiones o distorsiones proyectivas que causaran el incorrecto funcionamiento de los algoritmos de segmentación hechos para sistemas cooperativos (ver Fig. 2b).

Fig. 2. Factores de error presentes en un sistema de reconocimiento no-cooperativo. (a) Imágenes de baja calidad en la etapa de adquisición, (b) Resultados fallidos del algoritmo de segmentación de un sistema cooperativo.

Colores et al. [5] exponen una propuesta de arquitectura, donde se logra obtener un sistema de reconocimiento basado en iris con un error del 2.48%, el cual es un valor muy bueno en este tipo de sistemas. La propuesta consiste en la introducción de 2 nuevas etapas en el sistema de reconocimiento: una etapa de evaluación de calidad para eliminar aquellos marcos distorsionados o carentes de información biométrica y una etapa de evaluación de la segmentación la cual es útil a fin de validar el proceso o retroalimentar información que optimice los algoritmos de segmentación.

En este artículo, se plantea el uso del video-iris y la explotación de la información en los marcos que le componen, utilizando técnicas de fusión de imágenes. La fusión servirá para formar una plantilla digital representativa que reúna todas las características de la región del iris de un conjunto de marcos del video iris, adquiridos de la misma escena pero en diferentes instantes de tiempo. La plantilla digital teóricamente debe contener mayor información biométrica que permita reconocer mejor a un usuario, reduciendo las tasas de error para sistemas no-cooperativos. Aunque existen diversos métodos de fusión [6], el objetivo principal de este artículo es experimentar y analizar métodos de fusión a fin de determinar el más idóneo para ser incluido como etapa dentro del sistema de reconocimiento no-cooperativo. El artículo está organizado de la siguiente manera: en la sección 2 se explica detalladamente el principio de operación de los métodos de fusión de imágenes evaluados. La sección 3 presenta la metodología de evaluación y los resultados de la evaluación de 7 métodos de fusión, y finalmente en la sección 4 se exponen las conclusiones y el trabajo futuro.

2 Métodos de fusión de imágenes

Un método de fusión de imágenes es un procedimiento a partir del cual, un conjunto de imágenes dadas son mezcladas para obtener una nueva imagen que contiene la máxima información posible de todas ellas. Los métodos de fusión pueden dividirse en aquellos que actúan en las imágenes a nivel de los pixeles (pixel-pixel) y los métodos basados en el análisis multi-resolución.

- Métodos pixel-pixel: las técnicas incluyen las operaciones aritméticas básicas, las operaciones lógicas y las operaciones probabilísticas.
- Métodos multi-resolución: consisten en representar diferentes niveles de detalles de una imagen, conformando una pirámide de copias filtradas de una original, en cada escala de la pirámide la imagen es reducida en dimensión para obtener diferente tipo de información en la imagen. La idea fundamental de estos sistemas consiste en obtener una representación más conveniente de la señal original sin pérdida de la información, de forma que pueda posteriormente reconstruirse.

2.1 Fusión por media ponderada

El método de fusión por media ponderada se basa en una combinación lineal de los pixeles, asignando más peso al pixel que tenga un nivel de detalle mayor, es decir, al pixel más nítido [7]. De manera que se definen dos matrices de pesos W_1 y W_2, donde $0 \leq W_1, W_2 \leq 1$ y $W_1(x,y) + W_2(x,y) = 1$, obteniendo así, una imagen resultante dada por la siguiente ecuación 1.

$$I(x,y) = W_1(x,y)I_1(x,y) + W_2(x,y)I_2(x,y) \tag{1}$$

Para determinar las matrices de pesos, se usa la información procedente de los bordes de las imágenes, obtenidos mediante la aplicación de filtros pasa-altas que reflejan los cambios abruptos en las intensidades de los pixeles respecto a su entorno (*bordes*). Por otra parte, las bajas frecuencias de una imagen son aquellos pixeles con poca variación en su intensidad respecto a su entorno, es decir, las zonas homogéneas.

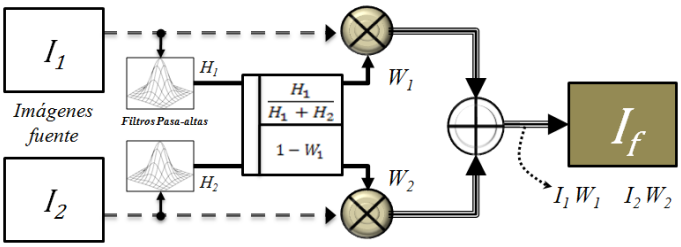


Fig. 3. Esquema de fusión de imágenes I_1 *e* I_2 , mediante la fusión por media ponderada [7].

El filtrado se ejecuta a través de un filtro Kernel pasa-altas Gaussiano procedente de la aplicación del concepto de la segunda derivada, dicho filtro es aplicado a las imágenes de entrada mediante la operación de convolución definida por la ecuación 2.

$$H(x,y) = |h * I(x,y)| \rightarrow h = \begin{bmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{bmatrix} \quad (2)$$

$$W_1(x,y) = \frac{H_1(x,y)}{H_1(x,y) + H_2(x,y)} \quad (3)$$

$$W_2(x,y) = 1 - W_1(x,y) \quad (4)$$

Así, si H_1 contiene la información de bordes de la imagen I_1 y H_2 la de la imagen I_2, en ambos casos por aplicación de la máscara definida previamente, podemos definir los pesos mediante las ecuaciones 3 y 4 (ver Fig. 3).

2.2 Fusión por análisis de componentes principales

El método por análisis de las componentes principales (PCA), permite reducir la dimensión del espacio donde se procesan las imágenes, reduciendo así la carga computacional [8]. El método consiste en transformar un conjunto de datos X de dimensión nxm en otro conjunto Y con menor dimensión nxl con la menor pérdida de información útil posible, es decir, $l \leq min\{n,m\}$. En la figura 4 se muestra el esquema básico donde se fusionan dos imágenes $I_1\ y\ I_2$ y se definen las matrices de pesos $W_1\ y\ W_2$, para obtener una imagen I_f resultante dada por la ecuación 5.

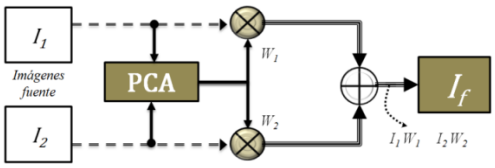


Fig. 4. Esquema de fusión de $I_1\ e\ I_2$, mediante fusión por PCA [8].

Para determinar las matrices de pesos, las matrices $I_1\ e\ I_2$ de tamaño $M \times N$ son reacomodadas como vectores columna $I_{k1}\ e\ I_{k2}$, el promedio entre las dos imágenes es un vector columna definido por Ψ utilizado para normalizar cada imagen I_k y obtener así $X_k = \{X_1, X_2\}$ definida por la ecuación 6.

$$I_f(x,y) = W_1(x,y)I_1(x,y) + W_2(x,y)I_2(x,y) \quad (5)$$

$$X_k = I_k - \Psi \quad (6)$$

$$C = \frac{1}{2}\left({X_1}^T X_1 + {X_2}^T X_2\right) \quad (7)$$

Posteriormente se obtiene un conjunto de vectores ortogonales U que describen las imágenes de entrada y que son los eigenvectores de la matriz de covarianza C de X_k (ec. 7). El número de eigenvectores λ_i de U pueden reducirse si existen eigenvalores asociados cuyo valor sea despreciable, de manera que sólo se toman los más significativos para obtener una matriz que represente cualquier imagen en un espacio de dimensión menor a la original. La matriz reducida de eigenvectores U se utiliza para calcular los coeficientes W_1 y W_2 que se modelan en las ecuaciones 8 y 9.

$$W_1 = U^T X_1 \qquad W_2 = U^T X_2 \quad (8)(9)$$

2.3 Fusión por descomposición en pirámide Laplaciana

Se conoce también como pirámide de diferencias pasa-bajas [9], el concepto es similar al de fusión por media ponderada, se procesan los detalles de la imagen obtenidos al aplicar filtros pasa-altas (filtro Laplaciano) sobre la imagen. En adición el método utiliza filtros pasa-bajas (filtro Gaussiano) sobre diferentes escalas de la imagen, para obtener cada vez menor cantidad de detalles denominados coeficientes de aproximación (pirámide Gaussiana). Posteriormente de calculan los coeficientes de detalle (pirámide Laplaciana) obteniendo la diferencia entre los coeficientes de detalle y la imagen original. Sea una imagen I, el primer nivel de su pirámide Gaussiana se define como una copia de la imagen original, mientras que el nivel *k-esimo* se define por la ecuación 10 y el nivel *k-esimo* de la pirámide Laplaciana es definido en la ecuación 11, la notación $\downarrow 2$ y $\uparrow 2$ denota un sub-muestreo y sobre-muestreo de la imagen en un factor de 2. La matriz w, define un filtro Kernel pasa-bajas definido por la ecuación 12, aplicado mediante la operación de convolución.

$$G_k(x,y) = [w * G_{k-1}(x,y)]_{\downarrow 2} \tag{10}$$

$$\tilde{L}_k(x,y) = G_k(x,y) - 4w * [G_{k-1}(x,y)]_{\uparrow 2} \tag{11}$$

$$w = \frac{1}{256}\begin{bmatrix} 1 & 4 & 6 & 4 & 1 \\ 4 & 16 & 24 & 16 & 4 \\ 6 & 24 & 36 & 24 & 6 \\ 4 & 16 & 24 & 16 & 4 \\ 1 & 4 & 6 & 4 & 1 \end{bmatrix} \tag{12}$$

Para la reconstrucción a partir de ambas pirámides, se utiliza la ecuación 13, donde $\widehat{\mathrm{G}}_0$ es la reconstrucción de la imagen original I. Para fusionar dos imágenes, se efectúa todo el procedimiento descrito sobre cada imagen, la fusión consiste en mezclar los coeficientes de aproximación y detalle de diversas maneras. Los coeficientes de detalle $\mathrm{G_k}$ se fusionan en todos los niveles, mientras que los de aproximación $\tilde{\mathrm{L}}_\mathrm{k}$ solo se combinan en el último nivel.

$$\hat{G}_k(x,y) = \tilde{L}_k(x,y) + 4w * [\hat{G}_{k+1}(x,y)]_{\uparrow 2} \tag{13}$$

Zhang y Blum [10] propusieron un método de combinación denominado "selección por medida de actividad", donde los coeficientes son considerados de manera separada; los de aproximación se combinan utilizando la media aritmética y los coeficientes de detalle eligiendo el mayor valor absoluto (ver ec. 14 y 15). Los coeficientes de detalle proporcionan información más relevante en las imágenes; bordes, líneas o limites de regiones. De manera que al elegir el mayor valor absoluto estamos seleccionando el coeficiente con mayor actividad, es decir, información.

$$G_k^C(x,y) = \frac{G_k^A(x,y) + G_k^B(x,y)}{2} \tag{14}$$

$$\tilde{L}_k^C(x,y) = \begin{cases} \tilde{L}_k^A(x,y) & si \quad \left|\tilde{L}_k^A(x,y)\right| > \left|\tilde{L}_k^B(x,y)\right| \\ \tilde{L}_k^B(x,y) & en\ otro\ caso \end{cases} \tag{15}$$

2.4 Fusión por descomposición en pirámide FSD

Este método definido por sus siglas en ingles FSD que describen los procesos que se ejecutan [11]; filtrado, resta y reducción de tamaño (*filter*, *subtract*, *decimate*). Estas operaciones se ejecutan también en el método de fusión por pirámide Laplaciana, pues la pirámide FSD es una variación de la misma.

$$L_k(x,y) = G_k(x,y) - w * G_k(x,y) \tag{16}$$

$$\hat{G}_k(x,y) = L_k(x,y) + w * (L_k(x,y) + \left[4\hat{G}_{k+1}(x,y)\right]_{\uparrow 2}) \tag{17}$$

La variación se encuentra en el cálculo del nivel *k-esimo* de la pirámide Laplaciana, que es computacionalmente más eficiente (ec.16), sin embargo esta variación restringe el uso de la ecuación 13 utilizada en la fusión por pirámide Laplaciana, pues produce perdidas en el detalle de la imagen. De tal manera que resulta necesario definir una reconstrucción diferente la cual es definida por la ecuación 17. La fusión de imágenes emplea la misma metodología utilizada en la fusión por pirámide Laplaciana (ecuaciones 14 y 15), la media aritmética combina los coeficientes de aproximación y el mayor valor del absoluto combina los coeficientes de detalle.

2.5 Fusión por descomposición en pirámide de contraste

Se conoce también como la pirámide de razón en paso bajo [12], se basa en la modificación del método por pirámide Laplaciana. La variación permite fusionar imágenes y representar con menor cantidad de anomalías perceptibles al ojo humano. Se sabe que el ojo humano es más sensible al contraste (detecta cambios en orden del 5%) que al brillo (detecta cambios en orden del 50%). Por ello, el método utiliza razones de cambio en lugar de las diferencias. De tal modo que la pirámide Gaussiana se re-define mediante la ecuación 18, la cual describe el nivel *k-esimo* de la pirámide y la reconstrucción se define por la ecuación 19.

$$R_k(x,y) = G_k(x,y)/4w * [G_{k-1}(x,y)]_{\uparrow 2} \tag{18}$$

$$\hat{G}_k(x,y) = R_k(x,y) * 4w * [\hat{G}_{k+1}(x,y)]_{\uparrow 2} \tag{19}$$

La fusión de coeficientes emplea una metodología similar al método por pirámide Laplaciana en la mezcla de coeficientes de aproximación (ver ecuación 14). Sin embargo, para la mezcla de coeficientes de detalle no se toma el máximo absoluto si no que se aplica un criterio descrito por la ecuación 20,

$$R_k^C(x,y) = \begin{cases} R_k^A(x,y) & si \quad \left|R_k^A(x,y) - 1\right| > \left|R_k^B(x,y) - 1\right| \\ R_k^B(x,y) & en\ otro\ caso \end{cases} \tag{20}$$

2.6 Fusión por descomposición en pirámide de gradiente

La pirámide de gradiente surge también como una variación a la pirámide Laplaciana. En particular, este método se basa en formar no solo una sino cuatro pirámides

Gaussianas obtenidas mediante el filtrado a diferentes orientaciones (horizontal, vertical y diagonales). El nivel *k-esimo* con orientación l de la pirámide del gradiente se define mediante la ecuación 21

$$D_{k,l}(x,y) = d_l * [G_k(x,y) + \dot{w}G_k(x,y)] \tag{21}$$

G_k es el nivel *k-esimo* de la pirámide Gaussiana, d_l es el filtro gradiente para la orientación l y w es el filtro tipo Kernel Gaussiano descrito por la ecuación 22, y los filtros gradiente están definidos mediante las ecuaciones (23-26).

$$\dot{w} = \frac{1}{16}\begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix} \tag{22}$$

$$d_1 = [1 \quad -1] \tag{23}$$

$$d_2 = \frac{1}{\sqrt{2}}\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \tag{24}$$

$$d_3 = \begin{bmatrix} -1 \\ 1 \end{bmatrix} \tag{25}$$

$$d_4 = \frac{1}{\sqrt{2}}\begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} \tag{26}$$

Para la reconstrucción de la imagen a partir de las pirámides del gradiente se utiliza la ecuación 27. La fusión de coeficientes se lleva a cabo empleando la misma metodología descrita por las ecuaciones 15 y 16, utilizadas en el método de fusión por pirámide Laplaciana.

$$\tilde{L}_k(x,y) = [1+w] * \sum_{l=1}^{4}(-\frac{1}{8}d_l * D_{k,l}(x,y)) \tag{27}$$

2.7 Fusión por descomposición en pirámide DWT

El método de fusión por transformada discreta de Wavelets (DWT) propuesto por Mallat et al.[13], se basa en la teoría de la pirámide del Gradiente. Para este método solo es necesario almacenar una pirámide Gaussiana y generar solo cuatro pirámides correspondientes a los coeficientes de aproximación y detalle en orientación horizontal y vertical. La representación por Wavelets tiene la ventaja de no generar información redundante dado que las funciones Wavelets son ortogonales y la señal original se puede reconstruir a partir de la descomposición de Wavelets con un algoritmo inverso.

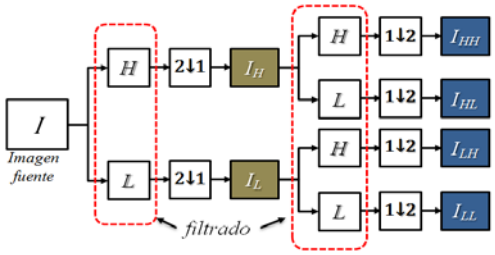


Fig. 5. Esquema de descomposición de una imagen mediante Transformada Discreta Wavelets.

Podemos definir el nivel *k-esimo* de las pirámides de Wavelets mediante las ecuaciones (28-31). Donde la notación $(1 \downarrow 2)$ se refiere a eliminar la mitad de las filas de la imagen y la notación $(2 \downarrow 1)$ se refiere a eliminar la mitad de las columnas de la imagen. Los filtros Kernel pasa-altas y pasa-bajas son definidos sobre las ecuaciones (29-32). La pirámide se forma aplicando esta descomposición de modo recursivo sobre los coeficientes de aproximación. La figura 5 muestra un nivel de descomposición de la pirámide.

$$LL_k(x,y) = [w_L * [w_L * I_k(x,y)]_{2\downarrow1}]_{1\downarrow2} | W_L = \frac{1}{\sqrt{2}}[1 \quad 1] \tag{28}$$

$$LH_k(x,y) = [w_H * [w_L * I_k(x,y)]_{2\downarrow1}]_{1\downarrow2} | W_H = \frac{1}{\sqrt{2}}[-1 \quad 1] \tag{29}$$

$$HL_k(x,y) = [w_L * [w_H * I_k(x,y)]_{2\downarrow1}]_{1\downarrow2} | W_L = \frac{1}{\sqrt{2}}[1 \quad 1] \tag{30}$$

$$HH_k(x,y) = [w_L * [w_H * I_k(x,y)]_{2\downarrow1}]_{1\downarrow2} | W_H = \frac{1}{\sqrt{2}}[-1 \quad 1] \tag{31}$$

Para realizar la reconstrucción se aplican las mismas transformaciones sobre las cuatro sub-imágenes, ecuaciones 32-35.

$$LL_k^{-1}(x,y) = w_{L^{-1}}^T * [w_{L^{-1}} * LL_k(x,y)_{1\uparrow2}]_{2\uparrow1} \,|W_{L^{-1}} = \frac{1}{\sqrt{2}}[1 \quad 1] \tag{32}$$

$$LH_k^{-1}(x,y) = w_{H^{-1}}^T * [w_{L^{-1}} * LH_k(x,y)_{1\uparrow2}]_{2\uparrow1} | W_{H^{-1}} = \frac{1}{\sqrt{2}}[-1 \quad 1] \tag{33}$$

$$HL_k^{-1}(x,y) = w_{L^{-1}}^T * [w_{H^{-1}} * HL_k(x,y)_{1\uparrow2}]_{2\uparrow1} | W_{L^{-1}} = \frac{1}{\sqrt{2}}[1 \quad 1] \tag{34}$$

$$HH_k^{-1}(x,y) = w_{H^{-1}}^T * [w_{H^{-1}} * HH_k(x,y)_{1\uparrow2}]_{2\uparrow1} | W_{H^{-1}} = \frac{1}{\sqrt{2}}[-1 \quad 1] \tag{35}$$

Por último, la imagen reconstruida se obtiene a partir de la ecuación 36.

$$\hat{I} = LL_k^{-1}(x,y) + LH_k^{-1}(x,y) + HL_k^{-1}(x,y) + HH_k^{-1}(x,y) \tag{36}$$

3 Resultados experimentales

3.1 Base de datos y experimentos

Con el propósito de evaluar el rendimiento de los siete métodos de fusión, la base de datos MBGC.v2 [14] fue seleccionada pues presenta diversos factores de ruido, los cuales se presentan en los sistemas no-cooperativos que son parte del objeto de estudio en este artículo. La base se conforma por 986 videos del ojo capturados con una cámara LG2200 EOU con iluminación cercana infra-roja (NIR). Algunas de las características principales de los videos son: formato MPEG-4, una resolución de los marcos de 480x640 pixeles y una profundidad de 8 bits en escala de grises (valores de intensidad entre 0-255).

A partir de la base de datos se genero un pequeño subconjunto de 100 marcos provenientes de 10 videos (*10 marcos de cada video seleccionados de manera aleatoria*). Es importante señalar que los marcos seleccionados fueron analizados para verificar que cumplieran con los parámetros de calidad de imagen y calidad segmentación expuestas por Colores et al. [5]. La segmentación del iris en los marcos seleccionados se llevo a cabo mediante los algoritmos de segmentación de Libor Masek [16]. Además de los 100 marcos, fueron elegidos 2 marcos adicionales de cada video para fines de comparación biométrica, estos fueron elegidos respecto al criterio de mayor energía [15] (referencia 1) y mejor calidad subjetiva (referencia 2).

3.2 Resultados

Como se menciona en la primera sección, el proceso de reconocimiento se basa en el valor de la distancia Hamming, dicho valor refleja la correlación entre plantillas digitales biométricas. Es decir, la distancia Hamming tendrá un valor pequeño cuando se comparan plantillas digitales generadas a partir del mismo iris (*comparación Intra-clase*) o de otro modo tendera a un valor cercano a 1 (*comparación Inter-clase*).

Dado que el objetivo en este artículo trata la reducción de la distancia Hamming, se realizaron 200 comparaciones Intra-clase (*100 comparaciones para cada marco de referencia*) y se implementaron y aplicaron los métodos de fusión descritos en la sección 2 sobre combinaciones entre pares de imágenes de la base de datos. De modo que para cada método se re-calcularon las comparaciones Intra-Clase con el propósito de determinar el método que pudiera incrementar los índices de reconocimiento gracias a la reducción de los valores de la distancia Hamming.

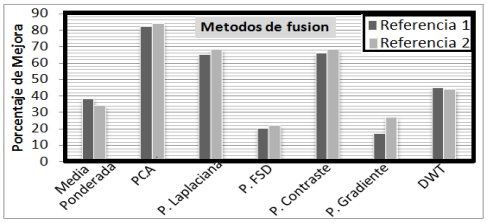


Fig. 6. Porcentaje de reducción de la Distancia Hamming al implementar métodos de fusión.

La figura 6 muestra los resultados porcentuales de mejora respecto a las distancias Hamming (reducción) para cada método analizado, es claramente apreciable que el método de fusión basado en el análisis de componentes principales PCA tiene un mejor desempeño respecto a los otros métodos, este logra reducir los valores de distancia Hamming en aproximadamente el 83% de los experimentos.

En la figura 7, se representan de manera detallada la reducción en los valores de las distancias Hamming al implementar el método de fusión por PCA. En la gran mayoría de experimentos se redujo considerablemente el valor de la distancia Hamming, lo cual proyecta una posible reducción en las tasas de error de reconocimiento al implementarse como un nuevo modulo en un sistema de reconocimiento del iris bajo un esquema de adquisición no-cooperativo.

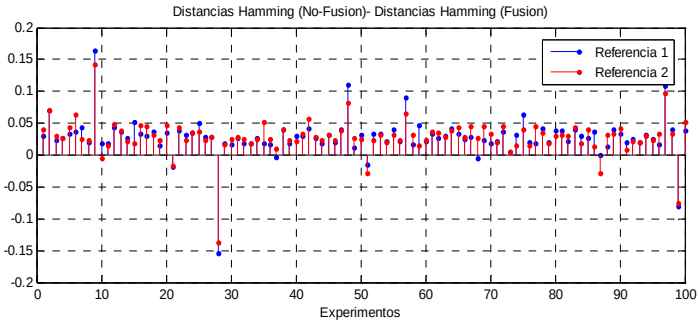


Fig. 7. Variación entre las distancias Hamming obtenidas de las comparaciones Intra-Clase, antes y después de implementar la fusión de imágenes por PCA.

4 Conclusiones

En este artículo se evalúa de manera experimental el rendimiento de los diferentes métodos de fusión de imágenes sobre una aplicación para fusionar plantillas digitales biométricas del iris. El objetivo de la fusión es conjuntar información de múltiples marcos provenientes de un video adquiridos bajo un esquema de adquisición no-cooperativo. En este artículo se ha constatado que los métodos de fusión pueden emplearse para reducir las distancias de comparación Hamming, la disminución de estos valores está relacionada a la disminución de las tasas de error del sistema no-cooperativo de reconocimiento del iris completo. Los experimentos muestran que al utilizar la fusión de imágenes por PCA, se logra reducir las distancias Hamming en el 83% de los casos. De manera que, los resultados sugieren que al agregar un modulo de fusión a la arquitectura del sistema no-cooperativo de reconocimiento del iris, se podría aumentar el rendimiento del sistema. Por lo anterior podemos decir que nuestra propuesta puede formar parte de lo proporcionado por García et al [17], para aplicaciones de reconocimiento del iris a distancia y en ambientes no controlados.

Referencias

1. Daugman, J.: How Iris Recognition Works. *IEEE Transactions on Circuits and Systems for Video Technology*, 14, pp.21-30 (2004)
2. Wildes, R.: Iris Recognition: An Emerging Biometric Technology. *Proceedings of the IEEE*, 85-9, pp.1348-1363 (1997)
3. Gamassi, M.; Lazzaroni, M.; Misino, M.; Piuri, V.: Quality assessment of biometric systems: a comprehensive perspective based on accuracy and performance measurement. *IEEE Transactions on Instrumentation and Measurement*, 54, pp.1489–1496 (2005)
4. Vatsa, M.; Singh, R.; Gupta, P.: Comparison of Iris Recognition. *International Conference on Intelligent Sensing and Information Processing*, pp.354–358 (2004)
5. Colores-Vargas, J.; García-Vázquez, M.; Ramírez-Acosta, A.; Nakano, M; Pérez-Meana, H.: Iris recognition system based on video for unconstrained environments. *Scientific Research and Essays*.7-35, pp. 3114-3127 (2012)
6. Mitchell, H.; Singh, R.; Gupta, P.: Multifocus Method for Controlling Depth of Field. *Grafica Obscura*. (1994)

7. Haeberli, P.; Singh, R.; Gupta, P.: Image Fusion: Theories, Techniques and Applications. *Springer-Verlag Berlin Heidelberg* 2010.
8. Pajares, G.; De la Cruz, J.: Visión por Computador: Imágenes Digitales y Aplicaciones. *Madrid: RA-MA*. (2001)
9. Burt, P.; Kolczynski, R.: Enhanced image capture through fusion, *Proc. Fourth Int. Conf. on Computer Vision*, pp. 173–182. (1993)
10. Zhang, Z.; Blum, R.: A categorization of Multiscale-Decomposition-Based Image Fusion Schemes with a Performance Study for a Digital Camera Application. *Proc. IEEE* , 87-8, 1315–1326.(1999)
11. Anderson, H.: A filter-subtract-decimate hierarchical pyramid signal analyzing and synthesizing technique. *U.S. Patent* 4.718 104. (1987)
12. Toet, A.; van Ruyven, J.; Valeton, J.: Merging termal and visual images by a contrast pyramid. *Optical Engineering* , 28-7, 789-792.(1989)
13. Mallat, S.: A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Trans. Pattern Anal. Mach. Intell*., 11, 674–693. (1989)
14. Multiple Biometric Grand Challenge. face.nist.gov/mbgc/.
15. Colores-Vargas, J.; García, M.; Ramírez, A.: Measurement of defocus level in iris images using convolution kernel method. *Lect. Notes Comput*. 6256:164-170.(2010)
16. Libor Masek, (2003). Recognition of human iris patterns for biometric identification. Master's thesis, University of Western Australia.
17. García-Vázquez, M.; Ramírez-Acosta, A.: Avances en el Reconocimiento del Iris: Perspectivas y Oportunidades en la Investigación de Algoritmos Biométricos. *Computación y Sistemas Vol. 16 No. 3, ISSN 1405-5546*.(2012)

A comparison of predictive measures of problem difficulty for classification with Genetic Programming

Yuliana Martínez[1], Leonardo Trujillo[1*], Edgar Galván-López[2], and Pierrick Legrand[3]

[1] Doctorado en Ciencias de la Ingeniería, Departamento de Ingeniería Eléctrica y Electrónica, Instituto Tecnológico de Tijuana, Tijuana BC, México
[2] Distributed Systems Group, School of Computer Science and Statistics, Trinity College Dublin, Ireland
[3] Université Victor Segalen, Bordeaux, France
IMB, Institut de Mathématiques de Bordeaux, UMR CNRS 5251, France
ALEA Team, INRIA Bordeaux Sud-Ouest, France
{ysaraimr,leonardo.trujillo.ttl,}@gmail.com,
edgar.galvan@scss.tcd.ie,
pierrick.legrand@u-bordeaux2.fr
Paper received on 22/09/12, Accepted on 24/10/12.

Abstract. In the field of Genetic Programming (GP) a question exists that is difficult to solve; how can problem difficulty be determined? In this paper the overall goal is to develop predictive tools that estimate how difficult a problem is for GP to solve. Here we analyse two groups of methods. We call the first group Evolvability Indicators (EI), measures that capture how amendable the fitness landscape is to a GP search. The second are Predictors of Expected Performance (PEP), models that take as input a set of descriptive attributes of a problem and predict the expected performance of a GP system. These predictive variables are domain specific thus problems are described in the context of the problem domain. This paper compares an EI, the Negative Slope Coefficient, and a PEP model for a GP classifier. Results suggest that the EI does not correlate with the performance of GP classifiers. Conversely, the PEP models show a high correlation with GP performance. It appears that while an EI estimates the difficulty of a search, it does not necessarily capture the difficulty of the underlying problem. However, while PEP models treat GP as a computational black-box, they can produce accurate performance predictions.

Keywords: Genetic Programming, Performance prediction, Classification

1 Introduction

Genetic Programming (GP) deals with the development of evolutionary algorithms (EA's) for automatic program induction [1], However, as for every EA, GP systems are stochastic search process, with many degrees of freedom and heuristic components. Therefore, as of yet, it is not possible to derive, from first principles, weather GP can solve a particular problem or task. A current goal within the GP community is to estimate how *hard* a problem instance might be for a specific GP. Such a measure could

* Corresponding author.

Mireya García-Vázquez, Grigori Sidorov (Eds.).
Advances in Computing Science and Control.
Research in Computing Science 59, 2012, pp. 159- 170

allow researchers to correctly choose and tune a GP search without the need of actually executing the code [2], which usually is computationally expensive.

If we want to measure the difficulty of a problem in GP, we can consider at least two different frames of reference. The first is the problem domain, which is independent of the method used to solve the problem [3]. The second frame of reference is to consider a perspective directly related to the process used to find a solution; in the case of GP this frame of reference corresponds with the search space and fitness landscape [4]. Let us first describe the latter.

The concept of a fitness landscape has dominated the way geneticists think about biological evolution and has been adopted by the EA community as a way to visualize evolution dynamics. Formally, a fitness landscape, as specified in [5], can be defined as a triplet (x, χ, f): (a) a set x of configurations, (b) a notion χ of neighbourhood, distance or accessibility on x, and finally, (c) a fitness function f. The local and global structure of the fitness landscape describes the underlying difficulty of a search. In general, most meta-heuristics work under the assumption that the fitness of a candidate solution, a point on the fitness landscape, is positively correlated with the fitness of (some) of its neighbours. Such a property can be defined as the *evolvability* of a landscape [6, 7]. Hence, some researchers have proposed measures that characterize the evolvability of a fitness landscape, what are here referred to as evolvability indicators (EI).

Another approach is to use the problem domain as the frame of reference, and characterize the difficulty of a problem based on the expected performance the GP search, a quantity that is derived from domain specific features of each problem instance [8–10]. This is a more pragmatic approach, the evolutionary search is taken as a black-box process and the performance of GP on a set of training problems is used to build predictors of the expected performance on unseen problems, following a machine learning methodology. In what follows, we refer to such measures of problem difficulty as Predictors of Expected Performance (PEPs).

The remainder of this paper proceeds as follows. Section 2 gives an overview of related work. Then, Section 3 describes how GP can be used for classification and presents the GP classifier used to perform the comparative analysis. The experimental results and analysis are given in Section 4. Finally, Section 5 contains a summary and conclusions.

2 Background

Landscapes and problem difficulty have been the subject of a good deal of research regarding EA's. For instance, researchers have developed work on landscape correlation [11], autocorrelation [12], epistasis [13] and monotonicity [14]. For GP, locality has been used to measure problem difficulty based on genotype to phenotype mappings [15–17]. However, this paper focuses on one of the most successful EI in GP literature, the Negative Slope Coefficient (NSC). The following discussion presents the NSC and reviews its predecessor, Fitness Distance Correlation (FDC).

2.1 Evolvability Indicators

FDC is a measure for problem difficulty originally proposed for genetic algorithms (GA's) [18] and later extended to GP [19]. The logic behind FDC proceeds as follows. Assume that we can compute the genotypic distance between each valid individual and the (global) optimum to a problem. If this distance is negatively correlated with the fitness of each individual then the search problem should be characterized as *easy*, and it should be characterized as difficult if no correlation is detected. Moreover, a problem should be considered to be deceptive if the correlation is positive. While FDC has shown to be reliable in many test cases, its more glaring weakness is that the optimal solution must be known *a priori*, somewhat not realistic for real-world problems.

Following the same general assumptions of FDC, Vanneschi et al [20] propose the NSC. In the case of NSC, knowledge about the global optimum is not required. Instead, NSC relies on the concept of *fitness clouds*, a scatter plot where for each genotype x a point is plotted on a 2-D plane, where the horizontal axis corresponds with the fitness of x given by $f(x)$, and the vertical axis represents the fitness $f(y)$ of a neighbouring genotype y. The hypothesis behind NSC is that the *fitness cloud shape* provides a meaningful description of the evolvability of a problem for GP-based search. The NSC is computed by assuming a piecewise linear relationship between $f(x)$ and $f(y)$ for a sample of M individual genotypes and computing the slope of the scatter points within a set of equally spaced segments of the $f(x)$ axis. In the original implementation, individuals are sampled using the Metropolis-Hastings algorithm, neighbours are generated using standard sub-tree mutation, and the representative neighbour y for each genotype x is chosen using tournament selection. The NSC is given in the range of $(-\infty, 0]$, where a value of 0 represents a highly evolvable (assumed to be easy) problem, and a negative NSC indicates a less evolvable (more difficult) problem.

2.2 Predictors of Expected Performance

Another way to characterize problem difficulty is to attempt to predict the expected performance that a GP search will achieve on a given problem instance, a more direct approach. Following this line of thought, two approaches have been proposed in GP literature. First, consider the work of [21], where the frame of reference of the problem and of the search method are combined to derive Predictors of Expected Performance (PEP's) for GP. In [21], the authors propose linear predictive models based on a sampling of the fitness landscape, given by

$$P(\mathbf{t}) \approx a_0 + \sum_{\mathbf{p} \in \mathcal{S}} a_{\mathbf{p}} \cdot d(\mathbf{p}, \mathbf{t}) \, , \tag{1}$$

where $P(\mathbf{t})$ is the predicted performance, $\mathbf{t}$ is the target functionality, $d(\mathbf{p}, \mathbf{t})$ is a distance measure (such a distance measure is a common fitness function for many application domains of GP), $\mathcal{S}$ is the set of all possible program behaviours, and where each behaviour $\mathbf{p}$ represents the set of program outputs obtained from the set of fitness cases for a particular problem. Hence, PEP's are derived by sampling the space $\mathcal{S}$ of possible program behaviours, which can also be seen as the space of possible programs. These

models were tested on symbolic regression problems and 4-input Boolean problems for GP, achieving good results.

The second, more recent work, is more tightly centered within the problem frame of reference [8–10], and proceeds as follows. Given a problem p, for which we want to compute a performance prediction, extract a feature vector $\boldsymbol{\beta} = (\beta_1, \beta_2, ...\beta_N)$ of N distinct features that *describe the properties of* p. Then, a PEP P is given by a kernel function K,

$$P(\boldsymbol{\beta}) \approx K(\boldsymbol{\beta}) \, . \tag{2}$$

Notice that the form of K is not restricted in any way. For instance, [8] uses a linear function similar to the one proposed in [21]. However, [9] tests more complex linear models and also non-linear models. Moreover, using this approach the feature vector $\boldsymbol{\beta}$ should be designed specifically for the domain of p. For example, [8] propose problem features for symbolic regression and Boolean problems, and their results show that the predictive accuracy surpasses that of the fitness-based models of [21]. In the case of [9], the authors predict the performance of a GP-classifier and use descriptive features that describe the geometry of the class samples. In that work, a quadratic linear model and symbolic regression models achieve the best performance.

In both approaches described above, the task of deriving the predictive model K is solved using a machine learning strategy that proceeds as follows, First, generate a set $\mathcal{Q}$ of problems (for instance, symbolic regression or classification problems), such that each sample $p \in \mathcal{Q}$ represents a distinct problem instance. The problems in $\mathcal{Q}$ can be real world problems, however its more practical to generate synthetic problems using a predefined probabilistic model. Second, from each $p \in \mathcal{Q}$ extract a vector of descriptive features $\boldsymbol{\beta}$. Third, solve each problem in $\mathcal{Q}$ using a specific GP system, this generates a performance estimate ϵ for each problem. In this formulation, ϵ could be the performance of a single GP execution or a statistic computed from m independent runs. Then, the problem is to find an optimal predictive model K^o of GP performance, such that

$$K^o = \underset{K}{argmin} \{Err[K(\beta_i), \epsilon_i]\} \, \forall p \in Q, \tag{3}$$

where $Err[,]$ represents an error measure, such as the root-mean-square error (RMSE). In practice, in order to derive K, the set $\mathcal{Q}$ is divided into a training set $\mathcal{T}$, a validation set $\mathcal{V}$ and a testing set $\mathcal{U}$. The learning problem can then be solved with standard regression techniques [8–10] or with non-linear symbolic regression [9, 10]. An important final observation regarding PEP's is that they can generalize quite easily, and can used to predict the performance of other stochastic or black-box algorithms such as neural networks [21, 10].

2.3 Discussion and Limitations

After reviewing the basic methodology of EI's and PEP's, one practical and computational difference stands out. On the one hand, EI's use a very large sampling of the search space to derive an accurate measure. In effect, this means that for every new problem instance it is necessary to perform this costly computational step. However,

executing the actual GP search might in fact be faster. Therefore, the best use of EI's would be to characterize a whole class of problems, where the estimate of search difficulty provided by the EI could generalize to the entire class of problems.

On the other hand, a PEP model is used quite easily and directly for each new problem instance. Practically, the only possible bottleneck would be the computational cost of calculating the set of descriptive features for each problem. However, in order to learn a new PEP a very large number of experimental runs must be carried to derive the training data. Moreover, each PEP is strongly linked to a specific GP system and implementation, and even small deviations from the configuration of the GP system might cause the predictive model to break-down.

3 Classification with GP

In supervised classification a pattern $\mathbf{x} \in \mathbb{R}^P$ has to be classified as belonging to one of M distinct classes $\omega_1, ..., \omega_M$ using a training set $\mathcal{T}$ of P-dimensional patterns with a known classification. The idea is to build a mapping $g(\mathbf{t}) : \mathbb{R}^P \rightarrow M$, that assigns each pattern $\mathbf{t}$ to a corresponding class ω_i, where g is derived based on evidence provided by $\mathcal{T}$. GP can be used in different ways to solve such classification tasks [1, 22]. However, this work uses the approach proposed in [23], denoted as the Probabilistic GP Classifier (PGPC).

3.1 PGPC Classifier

In PGPC, GP is used to evolve a mapping $h(\mathbf{x}) : \mathbb{R}^P \rightarrow \mathbb{R}$ that transforms each input pattern $\mathbf{x}$ into a point on the real line. Moreover, it is assumed that the behaviour of h can be modeled using multiple Gaussian distributions, each corresponding to a single class [23]. The distribution of each class $\mathcal{N}(\mu, \sigma)$ is derived from the examples provided for it in set $\mathcal{T}$, by computing the mean μ and standard deviation σ of the outputs obtained from h on these patterns. Then, from the distribution $\mathcal{N}$ of each class a fitness measure can be derived using Fisher's linear discriminant; for a two class problem it proceeds as follows. After the Gaussian distribution $\mathcal{N}$ for each class are derived, a distance is required. In [23], Zhang and Smart propose a distance measure between both classes as

$$d = \frac{|\mu_1 - \mu_2|}{\sigma_1 + \sigma_2} \,, \tag{4}$$

where μ_1 and μ_2 are the means of the Gaussian distribution of each class, and σ_1 and σ_2 their standard deviations. When this measure tends to 0, it is the worst case scenario because the mapping of both classes overlap completely, and when it tends to ∞ it represents the optimal separation. To normalize the above measure, the fitness for an individual mapping h is given by

$$f_d = \frac{1}{1 + d} \,. \tag{5}$$

After executing the GP, the best individual found determines the parameters for the Gaussian distribution $\mathcal{N}_i$ associated to each class. Then, a new test pattern $\mathbf{x}$ is assigned to class i when $\mathcal{N}_i$ gives the maximum probability.

Table 1. Parameters for the PGPC system used in the experimental tests.

Parameter	**Description**
Population size	200 individuals.
Generations	200 generations.
Initialization	*Ramped Half-and-Half*, with 6 levels of maximum depth.
Operator probabilities	Crossover $p_c = 0.8$; Mutation $p_\mu = 0.2$.
Function set	$\left\{+, -, *, /, \sqrt{}, sin, cos, log, x^y, \lvert\cdot\rvert, if\right\}$
Terminal set	$\{x_1, ..., x_i, ..., x_P\}$ Where each x_i is a dimension of the data patterns $\mathbf{x} \in \mathbb{R}^P$
Bloat control	Dynamic depth control.
Initial dynamic depth	6 levels.
Hard maximum depth	20 levels.
Selection	Lexicographic parsimony tournament
Survival	Keep best elitism

4 Comparative Analysis

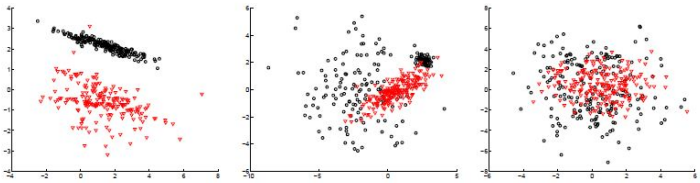
The main goal of the experimental work is to evaluate and compare the predictive accuracy of a state-of-the-art EI (NSC) and a PEP model for a GP-based classifier PGPC. To this end, a large set of synthetic classification problems are generated and solved with PGPC, executing 30 independent runs on each problem and computing the average classification error as the estimate of the expected performance of PGPC.

4.1 Classification Performance

Table 1 presents the setup for the PGPC system. A GP with Koza style crossover and mutation and dynamic depth control to minimize bloat [24] was used. The PGPC classifier is implemented using Matlab 2011a and the GPLAB toolbox [25].

To evaluate the performance of PGPC, 300 two-class classification problems are randomly generated using Gaussian mixture models (GMM's), these conform set $\mathcal{Q}$. Examples of the classification problems generated are shown in Figure 1, depicting sample points of two different classes(circles and triangles), scattered over the $\mathbb{R}^2$ plane. The use of randomly generated GMM's allows us to generate either unimodal or multimodal classes, with different amounts of class overlap. All class samples lie within the closed 2-D interval $x, y \in [-10, 10]$, and 200 sample points were randomly generated for each class.

Then, for every problem $p \in \mathcal{Q}$ the average test error of PGPC is computed from 30 independent runs, where the training (70% of the 200 samples) and testing (30%) sets where randomly determined at the start of each run. Figure 1 also specifies the average classification error ϵ achieved by PGPC on each problem and the corresponding standard deviation σ. In all three cases, and for all problems, the average provides a useful performance estimate given the small standard deviation.



(a) $\epsilon = 0.01$, $\sigma = 0.01$ (b) $\epsilon = 0.15$, $\sigma = 0.03$ (c) $\epsilon = 0.33$, $\sigma = 0.03$

Fig. 1. Three classification problems and the average classification error ϵ achieved by PGPC and standard deviation σ of 30 independent runs.

4.2 Evolvability for Classification Problems

This section presents the results of computing the NSC on each of the classification problems in set $\mathcal{Q}$. The approach is quite straightforward, since it is possible to directly apply the NSC algorithm to every problem. The algorithm described in [20] is used her with the same parameters except for the total amount of sampled individuals M. Whereas in [20] $M = 40,000$, here $M = 10,000$, a practical choice to reduce computation time; however, some informal tests showed that the results are consistently similar for both values in the group of experiments reported here. Figure 2 presents a scatter plot where the horizontal axis is the average classification error and the vertical axis is the NSC.

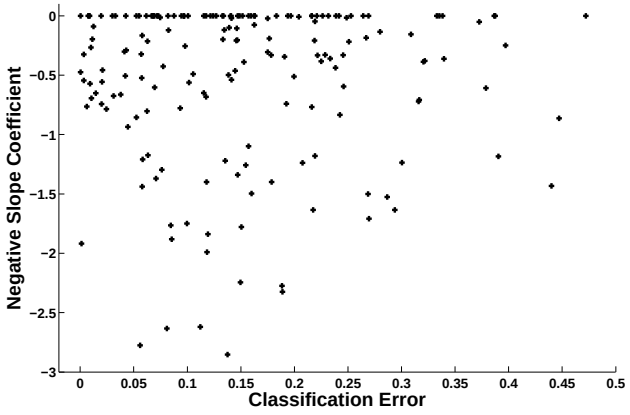


Fig. 2. Scatter plot of the average classification error achieved by PGPC on each problem and the corresponding NSC value. Pearson's correlation coefficient $\rho = 0.02$.

The results clearly suggest that the NSC does not correlate with PGPC performance, in particular we can see how many problems are characterized as easy (with NSC equal

or close to zero) even when the performance achieved by PGPC is quite poor. This suggests that an EI such as the NSC is limited as a predictor of GP performance since it only considers the frame of reference of the search process; i.e., it can only provide an approximate measure of the difficulty of the search but tells us very little regarding the difficulty of the underlying problem that the GP is intended to solve.

4.3 Prediction of Classification Performance

This section, we show how a PEP estimates the performance of PGPC on the set of classification problems presented above. The PEP is derived following the approach described in [9]. Therefore, the feature vector for each problem β is composed of the following problem descriptors.

Problem Descriptors

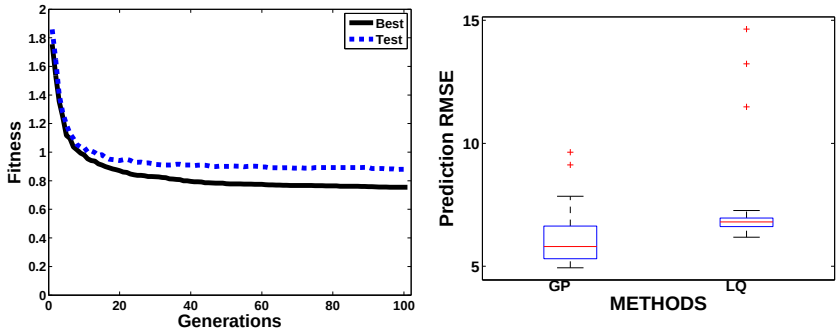
- The geometric mean ratio of the pooled standard deviations to standard deviations of the individual populations (SD), often used as part of a homogeneity test [26].
- Volume of Overlap Region (VOR) provides an estimate of the amount of overlap between both classes in feature space [3]. This measure is computed by finding, for each feature, the maximum and minimum value of each class and then calculating the length of the overlap region. The length obtained from each feature can then be multiplied in order to obtain a measure of volume overlap. VOR is zero when there is at least one dimension in which the two classes do not overlap.
- Feature efficiency (FE), measures the amount by which each feature dimension contributes to the separation of both classes. When there is a region of overlap between two classes on a feature dimension, then data is considered ambiguous over that region along that dimension. However, it is possible to progressively remove the ambiguity between both classes by separating those points that lie inside the overlapping region. The efficiency of each feature is defined as the fraction of all remaining points separable by that feature, and the maximum feature efficiency (FE) is taken as the representative value for a two-class problem.
- The Class Distance Ratio (CDR) compares the dispersion within the classes to the gap between the classes [3]. It is computed as follows: for each data sample the Euclidean distance to its nearest-neighbour is computed, within and outside its own class. Then, the CDR is the ratio of the averages of all intraclass and interclass nearest-neighbour distances.

Predictors of Expected Performance: Two different PEP models are tested, a linear model with quadratic terms (LQ-PEP) and a symbolic regression models derived with GP (GP-PEP), reported to achieve the best results in [9]. In the case of the GP-PEP models, the problem descriptors are used as terminal elements
$T = \{SD, VOR, FE, CDR\}$, while the function set is defined as
$F = \left\{+, -, *, /, \sqrt{}, sin, cos, log, x^y, |\cdot|, if\right\}$. Moreover, fitness is computed by the

RMSE calculated on a set of n training problems, given by

$$f(K) = \sqrt{\frac{\sum_{i=1}^{n}(K(\beta_i) - \epsilon_i)^2}{n}}. \quad (6)$$

where β_i is the vector of descriptive features and ϵ_i is the performance estimate on a training problem i. Finally, the GP is executed with the following parameters: 200 individuals, 100 generations, ramped half-and-half initialization, 0.8 crossover probability, 0.2 mutation probability, dynamic depth bloat control with a maximum depth of 12 levels, and lexicographic tournament selection.



(a) (b)

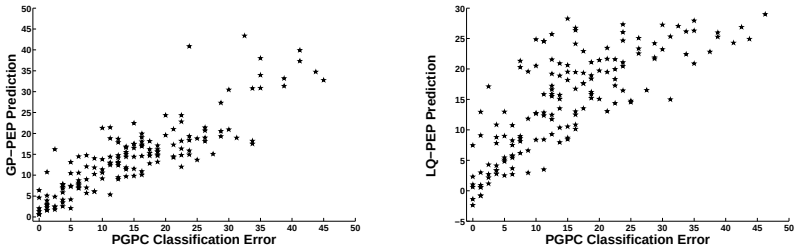
Fig. 3. Evolution of best fitness and fitness computed on the testing set, where both plots show the median over 30 independent runs and Box plot comparison of both PEP models.

For both types of PEP models the set $\mathcal{Q}$ of classification problems is divided into a training set and a testing set, each with 50% of the problems, and 30 runs are executed with different random partitions.

For the GP-PEP models the evolution of best fitness and test fitness is presented in Figure 3(a), which shows that the learning process is not over-fitted based on the similarity, with only a marginal difference, of both curves, which represent the median value over all runs.

Figure 3(b) presents a box plot comparison of the 30 executions of the GP-PEP and LQ-PEP models, based on the predictive accuracy achieved on the test set of each run. In general, both PEP models exhibit a very good predictions, with the median error around 5 and 7 percentage points of classification accuracy. However, the figure also shows that the symbolic regression models achieve a better performance, and a statistical t-test confirms this at the 1% significance level. Another look at the predictive accuracy of the PEP models is shown in the scatter plots of Figure 4. In these plots the predictive classification error is plotted against the average error of PGPC on each problem, using the best LQ-PEP and GP-PEP models. It is clear that the prediction of

the PEP models is strongly correlated with the average performance of PGPC, and the Pearson's correlation coefficient presented with each plot confirms this observation.



(a) GP-PEP $\rho = 0.86$ (b) LQ-PEP $\rho = 0.77$

Fig. 4. Scatter plots that show the average performance of PGPC (x-axis) and the predicted performance of each PEP model (y-axis). The legend specifies Pearson's correlation coefficient ρ.

5 Concluding Remarks

Is it possible to predict how *hard* a problem is for a Genetic Programming system without actually running the search process? This has been questioned at considerable length within the GP community over the last twenty years, where some good work has been done, with the development of promising proposals and perspectives.

In this paper, two groups of problem difficulty prediction tools are analyzed, named Evolvability Indicators and Predictors of Expected Performance. The former group of measures attempt to capture how amenable the fitness landscape is to a GP, whereas the latter groups takes as input a set of descriptive features of a problem and produces as output an estimate of the expected performance of the GP search.

The key lessons of this study are the following. Firstly, while EIs (in this work the Negative Slope Coefficient is considered) can give a good estimation of the difficulty of the search problem, they are not necessarily correlated with expected performance; Secondly, the results suggest that PEPs achieve a highly accurate prediction of GP performance.

References

1. Koza, J.R.: Genetic programming II: automatic discovery of reusable programs. MIT Press, Cambridge, MA, USA (1994)
2. Wedge, D.C., Kell, D.B.: Rapid prediction of optimum population size in genetic programming using a novel genotype -: fitness correlation. In: Proceedings of the 10th annual conference on Genetic and evolutionary computation. GECCO '08, New York, NY, USA, ACM (2008) 1315–1322
3. Ho, T.K., Basu, M.: Complexity measures of supervised classification problems. IEEE Trans. Pattern Anal. Mach. Intell. **24**(3) (March 2002) 289–300

4. Kinnear, K.E.: Fitness landscapes and difficulty in genetic programming. In: Proceedings of the First IEEE Conference on Evolutionary Computing, Piscataway, NY, IEEE Press (1994) 142–147
5. Stadler, P.F., Stephens, C.R.: Landscapes and Effective Fitness. Comments on Theoretical Biology **8**(4) (2003) 389–431
6. Altenberg, L. In: The evolution of evolvability in genetic programming. MIT Press, Cambridge, MA, USA (1994) 47–74
7. O'Neill, M., Vanneschi, L., Gustafson, S., Banzhaf, W.: Open issues in genetic programming. Genetic Programming and Evolvable Machines **11**(3-4) (September 2010) 339–363
8. Graff, M., Poli, R.: Performance models for evolutionary program induction algorithms based on problem difficulty indicators. In: Proceedings of the 14th European conference on Genetic programming. EuroGP'11, Berlin, Heidelberg, Springer-Verlag (2011) 118–129
9. Trujillo, L., Martínez, Y., Galván-López, E., Legrand, P.: Predicting problem difficulty for genetic programming applied to data classification. In: Proceedings of the 13th annual conference on Genetic and evolutionary computation. GECCO '11, New York, NY, USA, ACM (2011) 1355–1362
10. Trujillo, L., Martínez, Y., Melin, P.: Estimating classifier performance with genetic programming. In: Proceedings of the 14th European conference on Genetic programming. EuroGP'11, Berlin, Heidelberg, Springer-Verlag (2011) 274–285
11. Weinberger, E.: Correlated and uncorrelated fitness landscapes and how to tell the difference. Biological Cybernetics **63**(5) (1990) 325–336
12. Manderick, B., de Weger, M.K., Spiessens, P.: The genetic algorithm and the structure of the fitness landscape. In Belew, R.K., Booker, L.B., eds.: ICGA, Morgan Kaufmann (1991) 143–150
13. Goldberg, D.E., Deb, K., Horn, J.: Massive Multimodality, Deception, and Genetic Algorithms. In Männer, R., Manderick, B., eds.: Parallel Problem Solving from Nature, 2, Amsterdam, Elsevier Science Publishers, B. V. (1992)
14. Naudts, B., Kallel, L.: A comparison of predictive measures of problem difficulty in evolutionary algorithms. IEEE Transactions on Evolutionary Computation **4**(1) (2000) 1–15
15. Galván-López, E., McDermott, J., O'Neill, M., Brabazon, A.: Defining locality in genetic programming to predict performance. In: IEEE Congress on Evolutionary Computation, IEEE (2010) 1–8
16. Galván-López, E., McDermott, J., O'Neill, M., Brabazon, A.: Towards an understanding of locality in genetic programming. In Pelikan, M., Branke, J., eds.: GECCO, ACM (2010) 901–908
17. Galván-López, E., McDermott, J., O'Neill, M., Brabazon, A.: Defining locality as a problem difficulty measure in genetic programming. Genetic Programming and Evolvable Machines (accepted) **12**(4) (2011) 365–401
18. Jones, T., Forrest, S.: Fitness distance correlation as a measure of problem difficulty for genetic algorithms. In: Proceedings of the 6th International Conference on Genetic Algorithms, San Francisco, CA, USA, Morgan Kaufmann Publishers Inc. (1995) 184–192
19. Tomassini, M., Vanneschi, L., Collard, P., Clergue, M.: A study of fitness distance correlation as a difficulty measure in genetic programming. Evol. Comput. **13**(2) (June 2005) 213–239
20. Vanneschi, L., Clergue, M., Collard, P., Tomassini, M., Vérel, S.: Fitness clouds and problem hardness in genetic programming. In: GECCO (2). (2004) 690–701
21. Graff, M., Poli, R.: Practical performance models of algorithms in evolutionary program induction and other domains. Artif. Intell. **174**(15) (October 2010) 1254–1276
22. Eggermont, J., Kok, J.N., Kosters, W.A.: Genetic programming for data classification: partitioning the search space. In: Proceedings of the 2004 ACM symposium on Applied computing. SAC '04, New York, NY, USA, ACM (2004) 1001–1005

23. Zhang, M., Smart, W.: Using gaussian distribution to construct fitness functions in genetic programming for multiclass object classification. Pattern Recogn. Lett. **27**(11) (August 2006) 1266–1274
24. Silva, S., Costa, E.: Dynamic limits for bloat control in genetic programming and a review of past and current bloat theories. Genetic Programming and Evolvable Machines **10**(2) (2009) 141–179
25. Silva, S., Almeida, J.: Gplab–a genetic programming toolbox for matlab. In Gregersen, L., ed.: Proceedings of the Nordic MATLAB conference. (2003) 273–278
26. Michie, D., Spiegelhalter, D.J., Taylor, C.C., Campbell, J., eds.: Machine learning, neural and statistical classification. Ellis Horwood, Upper Saddle River, NJ, USA (1994)

Algoritmos PSO y DE aplicados al problema de inestabilidad en sistemas multiagentes nómadas

Alejandro Sosa, Víctor Zamudio, Rosario Baltazar, Carlos Lino, Miguel Angel Casillas y Marco Sotelo

División de Estudios de Posgrado e Investigación, Instituto Tecnológico de León, Av. Tecnológico S/N, 37290 Guanajuato, México
alejandro_sosa@ieee.org, vic.zamudio@ieee.org, r.baltazar@ieee.org, carloslino@itleon.edu.mx, miguel.casillas@ieee.org, masotelof@ieee.org
http://posgrado.itleon.edu.mx/
Paper received on 22/09/12, Accepted on 21/10/12.

Resumen El problema de inestabilidad cíclica en un sistema multi-agente nómada basado en reglas, se origina cuando un agente externo se incorpora al sistema, provocando así ajustes en las topologías de interconexión. Dichos ajustes pueden ocasionar comportamientos oscilatorios en el sistema, lo cual no es deseable. Se ha demostrado que dichas oscilaciones puede ser minimizadas mediante el bloqueo de uno o varios agentes. Los algoritmos evolutivos, como el PSO y el DE, pueden ser utilizados para la optimización numérica siempre y cuando se tenga una función objetivo. En el presente artículo se muestran los resultados de aplicar estos algoritmos al problema de inestabilidad cíclica, tratando de minimizar la función definida como el promedio de cambios en el sistema. Para dichas pruebas se generaron diversas instancias a las cuales se les aplicaron los algoritmos evolutivos antes mencionados. Los resultados obtenidos por dichos algoritmos se contrastaron mediante la prueba de Wilcoxon para poder determinar cual algoritmo tuvo mejor desempeño.

Palabras claves: PSO, DE, Inestabilidad Cíclica, Agentes Nomadas,Topologías de Interconexión

1. Introducción

Los ambientes inteligentes[1] se basan en el concepto de la computación ubicua, es decir, la integración de la informática en nuestro entorno de forma que los ordenadores y otros sistemas de procesamiento no sean elementos diferenciables, sino que formen parte natural de esos entornos. Basándose en ello, tenemos los entornos inteligentes, que son entornos digitales formados por redes de dispositivos inteligentes, que detectan la presencia y acciones del usuario y actuar en consecuencia (es decir, de manera reactiva), e incluso se pueden anticipar y adaptar a las necesidades y preferencias del usuario. Unos de los retos de los entornos inteligentes es minimizar o eliminar la inestabilidad que se pudiera generar en ellos, además de ser un problema poco abordado en

Mireya García-Vázquez, Grigori Sidorov (Eds.).
Advances in Computing Science and Control.
Research in Computing Science 59, 2012, pp. 171- 182

la literatura[2].
Una de las opciones propuestas es el algoritmo INPRES, basado en el uso de locking (que consiste en no permitir a un agente cambiar de estado) y la detencion de ciclos en la Interaction Network asociada. Con INPRES, se pretende obtener un sistema estable, con la restricción de que la cantidad de agentes bloqueados sea el mínimo para evitar que este bloqueo no sea prioridad dentro del sistema sino una herramienta para tener un control del sistema; este sistema es llamado c-INPRES[3] es la nueva versión del algoritmo antes mencionado donde se pretende que los agentes bloqueados sea el mínimo posible dentro del sistema. Uno de los algoritmos que se ha probado de manera exitosa usando el locking en ambientes estáticos es la optimización mediante cúmulo de partículas (Particle Swarm Optmization, PSO)[4] pues ha permitido reducir al mínimo las oscilaciones del sistema. Otro de los algoritmos usados en ambientes estáticos es el de DE[5] [6] el cual mantiene una población de soluciones candidatas, las cuales se recombinan y mutan para producir nuevos individuos los cuales serán elegidos de acuerdo al valor de su función de desempeño. Debido a lo anterior, en este trabajo representamos los resultados de haber aplicado los algoritmos PSO y DE a sistemas multiagentes dinámicos; es decir, donde los agentes presentan un comportamiento nómada.

2. Algoritmo de Evolución Diferencial

La evolución diferencial(Differential Evolution, DE)[5] [6] es una rama de la computación evolutiva desarrollada por Rainer Storn y Kenneth Price para optimización en espacios continuos, aplicado en la resolución de problemas complejos. Al igual que otros algoritmos de esta categoría, la DE mantiene una población de soluciones candidatas, las cuales se recombinan y mutan para producir nuevos individuos los cuales serán elegidos de acuerdo al valor de su función de desempeño. Lo que caracteriza a la DE es el uso de vectores de prueba, los cuales compiten con los individuos de la población actual a fin de sobrevivir. La DE es un método de búsqueda que utiliza N vectores:

$$x_{i,G}; i = 0, 1, 2, ..., N - 1 \tag{1}$$

como la población de cada generación (G). El valor de N no cambia durante el proceso de optimización. La población inicial se elige de manera aleatoria si no se conoce nada acerca del problema. Como regla, se asume una distribución uniforme para las decisiones aleatorias a tomar. La idea principal detrás de la DE es un nuevo esquema para generar vectores. La DE genera estos nuevos vectores cuando se suma la diferencia de pesos entre dos vectores miembros de la población a un tercer vector miembro. Si la aptitud del vector resultante es menor que el miembro de la población elegido entonces el nuevo vector reemplaza al vector con el cual fue comparado. Este vector a comparar puede ser (aunque no necesariamente lo es) parte del proceso de generación arriba mencionado. Además, el mejor vector $x_{mejor,G}$ se evalúa en cada generación G para no perder de vista el progreso durante el cual se hace la minimización.[7]

El algoritmo asume que las variables del problema a optimizar están codificadas como un vector de números reales. La longitud de estos vectores (n) es igual al número de variables del problema, y la población esta compuesta de np (número de padres)

vectores. Se define un vector x_p^g, en donde p es el índice del individuo en la población $(p = 1...np)$ y g es la generación correspondiente. Cada vector está a su vez compuesto de las variables del problema $x_{p,m}^g$, en donde m es el índice de la variable en el individuo $(m = 1...n)$. Se asume que el dominio de las variables del problema está restringido entre valores mínimos y máximos x_m^{min} y x_m^{max}, respectivamente. DE se compone básicamente de 4 pasos:

- Inicialización: Se genera una población inicial aleatoria con una distribución uniforme[8].
- Mutación: Se selecciona aleaoriamente tres vectores diferentes entre sí, se restan dos de ellos y a la diferencia se aplica un peso dado a ellos por un factor y por último se suma la diferencia la diferencia al tercer vector[9].
- Recombinación: Se realiza la recombinación, tomando a cada uno de los individuos de la población como el padre principal y otros 3 padres se seleccionan aleatoriamente generando un hijo. Si el hijo generado tiene un mejor valor de la función objetivo que el padre principal, entonces lo reemplaza[8].
- Selección: Todos los vectores son seleccionados una sola vez como padre principal sin depender de la función objetivo, se comprueba si el padre seleccionado resulta mejor que el hijo generado este conserva su valor de lo contrario se sustituye por el hijo[9].

Estos 4 procesos se realizan hasta que se cumpla el criterio de paró definido por el usuario, este proceso se muestra en el Algoritmo 1.

Algoritmo 1. Evolución diferencial.

```
f función objetivo a minimizar.
límites valores máximos y mínimos que pueden tomar las variables.
|P| número de individuos en la población  mayor o igual a 4.
N número de llamadas a función.
n dimensión.
F parámetro de entrada entre 0 y 2.
CR parámetro de entrada entre 0 y 1.
Return Pbest tal que f(Pbest) es el mejor fitness.
P = Generar una población aleatoria.
f(P) Se calcula el fitness de cada individuo de la población.
While {no se cumpla al numero máximo de llamadas a función}
    tempXi = Xj + F(Xk - Xl)
    for cada coordenada j del individuo
        r = numero aleatorio uniformemente distribuido _
             entre 0 y 1.
        if r <= CR
            tempXi[j]=  temXi[j] si r <= CR; en otro caso Xi[j]
        endif
    endfor
    if f(tempXi) mejor f(Xi)
        Xi =  tempXi
        f(Xi) = f(tempXi).
    endif
endWhile
```

3. Algoritmo de Optimización mediante Cúmulo de Partículas

El algoritmo de Optimización mediante Cúmulo de Partículas (Particle Swarm Optimization, PSO)[4], originalmente propuesto por Kennedy and Eberhart, en 1995 es un algoritmo bioinspirado, que se basa en una analogía con el comportamiento social del vuelo de las aves o del cúmulo de peces. Se fundamenta en el enfoque conocido como la "metáfora social", la cual puede resumirse de la siguiente forma: los individuos que conviven en una sociedad tienen una opinión que es parte de un "conjunto de creencias" compartido por todos los posibles individuos. Cada individuo, puede modificar su propia opinión basándose en tres factores:

- Conocimiento sobre el entorno (desempeño).
- Conocimiento histórico sobre experiencias personales anteriores (componente cognitivo).
- Conocimiento histórico sobre experiencias anteriores de los individuos situados en su vecindario (componente social).

Este algoritmo usa dos ecuaciones: la ecuación (2) es usada para encontrar la velocidad, describe la magnitud y la dirección del paso que será tomado por las partículas y está basado en el comportamiento adquirido hasta este momento.

$$v_{id} = wv_{id} + c_1(lBest_{id} - x_{id}) + c_2(gBest_{id} - x_{id}) \tag{2}$$

Donde:
v_{id}: es la velocidad de la i-ésima partícula.
$d = 1, 2, ..., N$, siendo N el tamaño de la dimensión.
$i = 1, 2, ..., N$, siendo N el tamaño de la población.
w es el coeficiente de adaptación al entorno.
c_i: es el coeficiente de memoria del vecindario.
c_2: es el coeficiente de memoria.
$lBest$ es el mejor partícula local encontrada por la i-ésima partícula.
$gBest$ es la mejor partícula encontrada hasta ese momento por las partículas.
x_i representa la i-ésima partícula.

La ecuación (3) actualiza la posición actual de la i-ésima partícula a la nueva posición utilizando el resultado obtenido del calculo de la velocidad.

$$x_{id} = x_{id} + v_{id} \tag{3}$$

PSO-Binario

El PSO Binario[10] fue diseñado para trabajar en espacios binarios. El PSO Binario selecciona las partículas lBest y gBest de la misma forma que el PSO. La principal diferencia con el PSO radica en la ecuación utilizada para actualizar la velocidad y la posición de la partícula. La ecuación utilizada para actualizar la velocidad se basa en probabilidades en el rango [0,1]. Para ello establece un mapeo para todos los valores

reales de la velocidad a el rango [0,1]. La ecuación 3 permite convertir el valor en binario 0 o 1.

$$v_{ij}(t) = sig(v_{ij}(t)) = \frac{1}{1 + e^{-v_{ij}(t)}} \tag{4}$$

Donde:
v_{ij} es la velocidad de la i-ésima partícula
$i = 1, 2, ...N$, siendo N el tamaño de la población
$j = 1, 2, ...D$, siendo D el numero de dimensiones

Mientras tanto la ecuación (5) es usada para actualizar la posición de la partícula a la nueva posición.

$$x_{ij}(t+1) = \begin{cases} 1 & \text{si } r_{ij} < sig(v_{ij}(t+1)) \\ 0 & \text{en otro caso} \end{cases} \tag{5}$$

Donde:
r_{ij} es un número aleatorio uniforme en el rango [0,1]
$i = 1, 2, ...N$, siendo N el tamaño de la población
$j = 1, 2, ...D$, siendo D el número de dimensiones

A continuación se muestra el Algoritmo 2, donde se observa a detalle la optimización por cúmulos de partículas.

Algoritmo 2. Optimización por Cúmulo de Partículas.

```
n: El número de partículas de la población.
LI: Límite Inferior.
LS: Límite Superior.
C1: Coeficiente de la memoria.
c2: Coeficiente de la memoria del vecindario.
w: Coeficiente de adaptación al entorno.
for cada partícula i
   Inicializar la población y buscar el mejor local
   de la población inicial.
   X= x1, x2, x3,...,xd |xi que pertenece [LI,LS]
   V= v2, v2, v3,...,vd |vi que pertenece [0,1]
   LBest =X
endfor
Buscar el mejor Global
if LBestx es mejor que GBestx
   GBest = LBest
endif
while se cumpla la condición de paro
   for cada partícula i
      Vi = wVi y c1(GBest - Xi) + c2(LBest - Xi)
      Xi = Xi + Vi
   if f(Xi) es menor que f(LBest)
      f(LBest) = f(Xi)
   endif
   endif
Buscar el mejor Global
for cada partícula i
```

```
   if LBest es mejor que GBest
      GBest = LBest
   endif
endfor
endwhile
Return GBest
```

4. Cálculo de Inestabilidad Dinámica

Uno de los retos que presenta los ambientes inteligentes es la inestabilidad ciclica.Un ejemplo es cuando un dispositivo A esta esperando una acción de un dispositivo B, sin la cual no puede continuar trabajando sin embargo el dispositivo B, requiere de información o ejecución de una tarea dada por A, generando un ciclo entre ambos dispositivos provocando que ninguno termine la tarea que se les ha asignado. Para entender mejor el fenómeno debemos retomar el hecho que dentro de los Ambientes Inteligentes existen múltiples agentes interconectados entre si, dichos agentes son gobernados por diversas reglas que afectan su comportamiento y pueden llevar a los agentes a cambiar de estado múltiples veces en un periodo de tiempo dado. De lo anterior suelen suceder múltiples casos ya documentados en donde se presenta interferencia o comportamiento no deseado entre los dispositivos[11],[12]. Al efectuar una evaluación del sistema todos los dispositivos y/o agentes que se encuentren en conflicto provocan que en general el sistema se vuelva inestable, obteniendo a su vez comportamientos erráticos o simplemente que los dispositivos no lleven a cabo tarea para la que fueron programados.

Para medir la oscilación de un sistema se han propuesto diferentes medidas de esta. En primera instancia se tiene la Oscilación Acumulativa Promedio[14] la cual trata de medir la relación de cambio entre el estado actual del sistema y el estado siguiente. En este trabajo se utiliza el Promedio de Cambios en el Sistema, ésta medida evalúa la estabilidad en base a la cantidad de cambios en los estados del sistema. [13]

Cálculo del Promedio de Cambios en el Sistema

El valor considerado para medir la inestabilidad es el Promedio de Cambios en el Sistema[14]. Para calcular este valor se utiliza la ecuación 5. En dicha ecuación se considera solo los cambios entre el estado actual del sistema con respecto al estado siguiente (sin considerar la magnitud del cambio):

$$O = \frac{\Sigma_{i=1}^{n-1} x_i}{n-1} \begin{cases} 1 \text{ si } S_i \neq S_{i+1} \\ 0 \text{ en otro caso} \end{cases} \tag{6}$$

Donde:
O: Promedio de cambios en el sistema.
n: Generaciones del ciclo de vida con S_i siendo el estado del sistema en el tiempo *i* y S_{i+1} siendo el estado del sistema en el tiempo *i+1*.

5. Minimizando la Inestabilidad en Ambientes Inteligentes

Los experimentos se realizaron bajo diferentes instancias de prueba donde cada instancia tuvo las siguientes características: se inicio con una matriz de 30x2 agentes, posteriormente se fue incrementando el tamaño y la dimensión, terminando con una matriz de 30x30 agentes. Los agentes entrantes se fueron agregando en diferentes periodos, usando la misma topología(Figura 1). La función objetivo es el cálculo del promedio de cambios en el sistema(Ecuación 6) donde se evaluan y conseervan los estados de los agentes que presentan menos oscilaciones. Con ello se pretende que el sistema no oscile, que afecten lo menos posible al sistema, para asegurar que conserve su propia integridad y funcionalidad para el usuario final. En la topología usada en el sistema se muestra el siguiente comportamiento: se está iniciando con un número definido de nodos o agentes, que en este caso son 2, con dependencias del anterior estado para su interacción. Dado que el sistema se va incrementando, los nodos se van insertando de forma sucesiva; de igual manera se pueden formar diferentes topologías de conexión, es decir, de un nodo pueden depender uno o más nodos con lo que se puede formar diferentes topologías con el mismo número de nodos, como se muestra en la Figura 1:

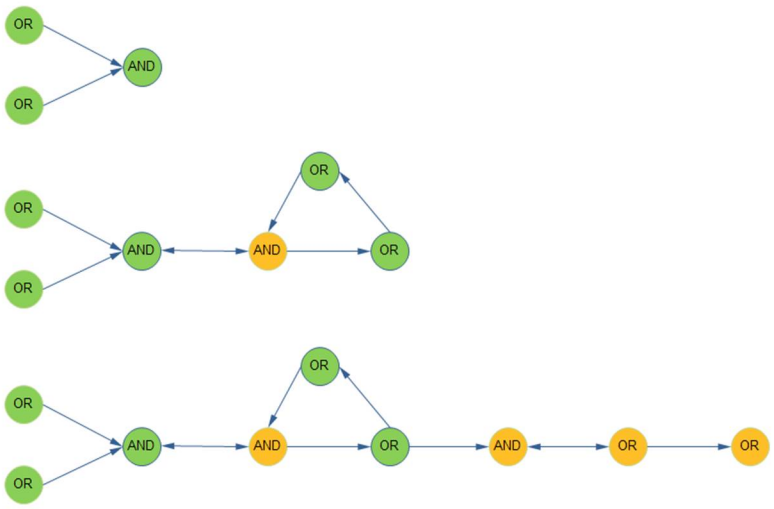


Figura 1. Topología usada en los experimentos realizados

La Figura 1 muestra el crecimiento del sistema iniciando con 2 agentes y terminando con 9, para este caso se llegará a una dimensión mayor.Las instancias de pruebas implementadas en los algoritmos fueron diseñados para incrementar el tamaño de manera uniforme iniciando un periodo del ciclo establecido, como por ejemplo en la primera instancia inicio con 2 agente y fue incrementando en diferentes periodos fijos del algoritmo terminando con 30 agentes en total,donde cada agente entrante posee sus propia

regla de interacción con los agentes ya establecidos en el sistema y con lo cual se pretende emular la entrada de agentes en un ambiente inteligente. Un escenario donde se ha dejado evolucionar al sistema se muestra en al Figura 2, utilizando la misma topología mencionada en la Figura 1 y sin aplicar ninguna técnica que permita minimizar las oscilaciones en el sistema.

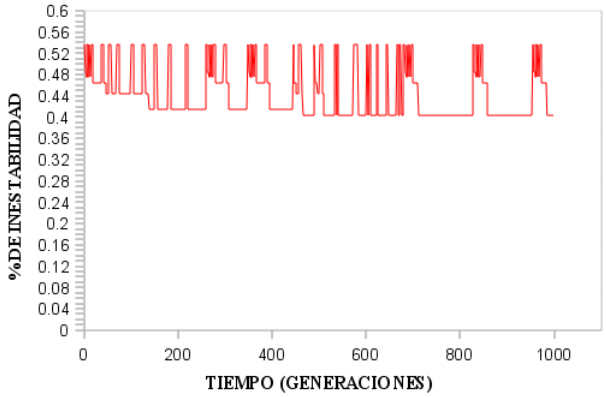


Figura 2. Gráfica de Inestabilidad

En la Figura 1 se muestra la topología de interconexión que tendría el sistema. El estado de cada agente es representado por un 0 o 1 en una cadena binaria, la cual representa su estado en un determinado tiempo y se utilizan las compuertas lógicas AND y OR como reglas de interacción respectivamente[15], formando una cadena de AND y OR representado todas las reglas en el sistema. Los parámetros utilizados para los algoritmos DE y PSO se muestran en la Tabla 1.

Tabla 1. Parámetros del algoritmo ED y PSO.

Parámetros	Valores PSO	Valores ED
Individuos	30	30
Dimensión inicial	2	2
Dimensión final	30	30
Llamadas a función	1000	1000
% de agentes bloqueados	0.2	0.2
W	1	//
C1	0.4	//
C2	0.6	//
F	//	0.9
CR	//	0.8

6. Resultados

Los resultados son mostrados en la Tabla 1 y Figura 3 y 4, los cuales permiten decir que el ambiente inteligente dinámico puede ser estabilizado exitosamente. En los experimentos realizados, se inició el sistema con 2 agentes, incrementandose hasta llegar a 30. Los algoritmos PSO y DE implementados permitiran encontrar combinaciones de agentes bloqueados, que minimizan las oscilaciones del sistema. En la Tabla 2 se muestran los resultados obtenidos en varias pruebas realizadas aplicadas a los algoritmos antes mencionados.

Tabla 2. Resultados.

Escenario	DE	PSO
Instancias 1	0.3831	0.3941
Instancias 2	0.3840	0.3985
Instancias 3	0.3737	0.4040
Instancias 4	0.3636	0.3989
Instancias 5	0.3787	0.4040
Instancias 6	0.3777	0.4144
Instancias 7	0.3643	0.4033

Los anteriores resultados fueron sometidos a la Prueba No Paramétrica de Rangos con Signo de Wilcoxon[16] para analizar la identidad estadística con respecto a los resultados del Cálculo del Promedio de Cambios en el Sistema de los algoritmos, de donde se obtuvieron T+ = 0,T− = 28, con una T0 = 4 para una muestra de 7 instancias de prueba con un nivel de significancia del 0.10. Los resultados indican que hay suficiente evidencia estadística para establecer que el algoritmo muestra un mejor desempeño, por lo cual podemos decir que el algoritmo de Evolución Diferencial se encuentra más a la izquierda obteniendo mejores resultados.

7. Conclusiones y Trabajos Futuros

En el presente trabajo se estudio el problema de inestabilidad cíclica en sistemas multiagentes. Se aplicaron y compararon los algoritmos PSO y DE bajo diferentes instancias de prueba. Con los resultados obtenidos se concluye que la inestabilidad que se genera en un escenario dinámico puede ser controlada usando alguno de estos algoritmos. Al aplicar la prueba de Wilconxon (debido a que existe suficiente evidencia estadística) se concluye que el algoritmo de Evolución Diferencial tuvo mejores resultados que el algoritmo de Optimización mediante Cúmulos de Partículas. Como trabajo futuro se plantea implementar otras técnicas que permitan mejorar los resultado obtenidos.

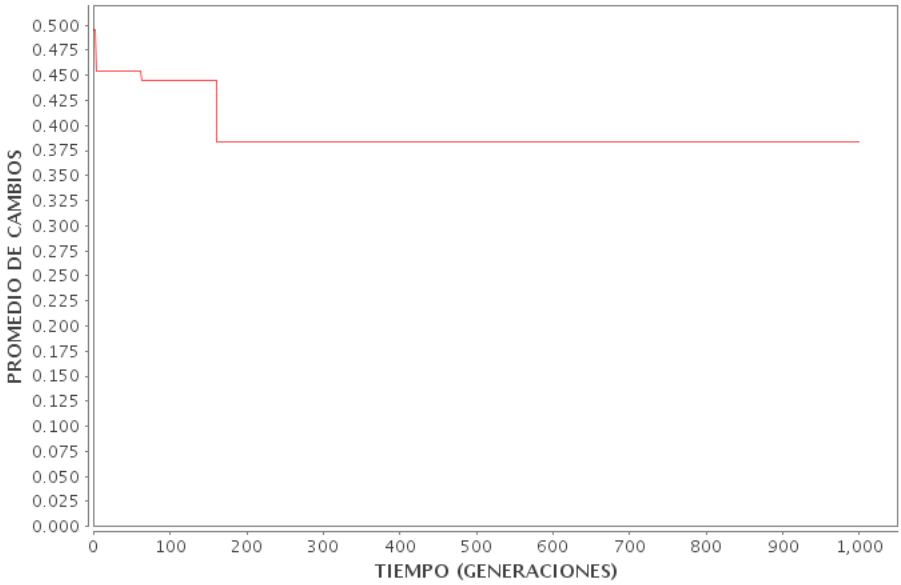


Figura 3. Promedio de Cambios en el Sistema al aplicar DE

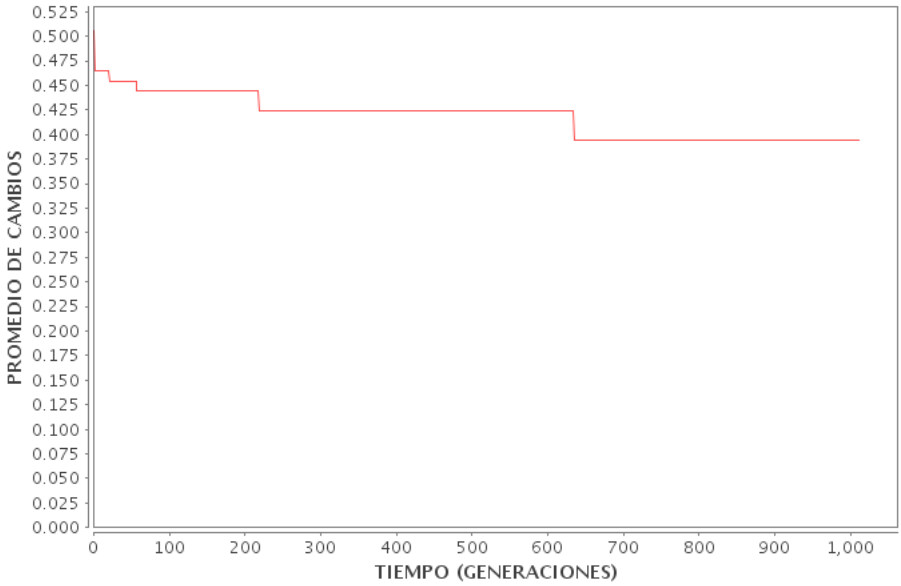


Figura 4. Promedio de Cambios en el Sistema al aplicar PSO

Agradecimientos

Alejandro Sosa Sales agradece el apoyo del Consejo Nacional de Ciencia y Tecnología (CONACyT), por el apoyo recibido para la realización de este trabajo de investigación y a la DGEST por el apoyo del proyecto 4311.11P

Referencias

1. Stuart Russell & Peter Norvig; Inteligencia Artificial, un enfoque moderno, Prentice Hall
2. Victor Zamudio, Vic Callaghan and Jeannette Chin "A Multi-dimensional Model for Task Representation and Allocation in Intelligent Environments", (2006), Volume: 3823, Publisher: Springer Berlin Heidelberg, Pages: 345-354
3. Victor Zamudio, Rosario Baltazar, Miguel Angel Casillas,Vic Callaghan "c-INPRES: Coupling Analysis Towards Locking Optimization in Ambient Intelligence". The 6th International Conference on Intelligent Environments IE10. 19-21 Julio 2010, Monash University (Sunway campus), Kuala Lumpur, Malaysia
4. Kennedy, J. and Eberhart, R. (1995). Particle swarm Optimization. In Proceedings of IEEE International Conference on Neural Networks, 1995., volume 4, pages 19421948 vol.4.
5. Rainer Storn and Kenneth Price. Differential evolution -a simple and efficient adaptative scheme for global optimization over continuous spaces. Technical Report TR-95- 12, International Computer Science, Berkeley, California, March 1995.
6. Rainer Storn and Kenneth Price. Differential evolution - a fast and efficient heuristic for global optimization over continuous spaces. Journal of Global Optimization, (11):341–359, 1997.
7. Dr. Carlos A. Coello Coello, Luis Vicente Santana Quintero. Un Algoritmo Basado en Evolución Diferencial para Resolver Problemas Multiobjetivo,Tesis Maestría, México DF, 2004.
8. Luis Vicente Santana Quintero y Carlos A. Coello Coello. Un Algoritmo Basado en Evolución Diferencial para Resolver Problemas Multiobjetivo.Tesis, CINVESTAV-IPN(Grupo de Computación Evolutiva).
9. Rubí del Carmen Gómez Ramón. Estudio empírico de variantes de Evolución Diferencial en optimización con restricciones. Tesis, Laboratorio Nacional de Informática Avanzada.
10. J. Kennedy, R. Eberhart, and Y. Shi. Swarm Intelligence. San Francisco: Morgan Kaufmann Publishers, 2001.
11. Victor Zamudio and V. Callaghan. Facilitating the ambient intelligent vision: A theorem, representation and solution for instability in rule-based multi-agent systems. Special Section on Agent Based System Challenges for Ubiquitous and Pervasive Computing. International Transactions on Systems Science and Applications., 4(2):108–121, May 2008.
12. Víctor Manuel Zamudio and Vic Callaghan. Understanding and Avoiding Interaction Based Instability in pervasive Computing Environmentes. Journal: International Journal of Pervasive Computing and Communications , vol. 5, no. 2, pp. 163-186, 2009
13. Leoncio Alberto Romero, Victor Zamudio, Rosario Baltazar, Aplicacion de Locking por Medio de Tecnicas de Inteligencia Artificial en Ambientes de Computo Pervasivo con Alta Inestabilidad, Tesis, Instituo Tecnologico de León 2012.
14. Leoncio Alberto Romero, Victor Zamudio, Rosario Baltazar and Marco Sotelo, "A Comparison Between PSO and MIMIC as Strategies for Minimizing Cyclic Instabilities in Ambient Intelligent", in the 5th International Symposium on Ubiquitous Computing and Ambient Intelligence (UCAmI´11), Rivera Maya, México, December 5-9, 2011.
15. V. Zamudio, V. Callaghan, "Preventing Instability in Rule-Based Multi-agent Systems; A Challenge to the Ambient Intelligence Vision". In workshop on Multi-agent Systems Challenges for Ubiquitous and Pervasive Computing MASUPC07 held at First International Conference on New Technologies, Mobility and Security (NTMS'2007), Telecom Paris, France, 2 to 4 May, 2007.

16. Wilcoxon, F.: Individual comparisons by ranking methods. Biometrics Bulletin 1(6) (1945) 80–83

Three metaheuristics solving a scheduling problem in a RIA environment

Adriana Perez-Lopez[1*], Rosario Baltazar[1], Martín Carpio[1], Arnulfo Alanis[2]

[1]División de Estudios de Posgrado e Investigación, Instituto Tecnológico de León, AV. Tecnológico S/N, 37290 Guanajuato, México
adriana_perez@ieee.com, charobamx1@yahoo.com.mx, jmcarpio61@hotmail.com
[2]Instituto Tecnológico de Tijuana, Baja California, Mexico
alanis@tectijuana.edu.mx
Paper received on 22/09/12, Accepted on 18/10/12.

Abstract. Scheduling decides how to commit resources between a variety of tasks and it is used in different areas as schools, factories, even hospitals; in surgery area of an hospital their main resources are surgery rooms and surgeries, where our objective is maximize number surgeries with respective emergency through metaheuristics like Ant Colony System, Genetic Algorithm, Memetic Algorithm; however to user is not enough the optimize resources in scheduling, also the user wish to get a fast solution and from anyplace, the option is the use of Rich Internet Applications, where the user can see his results of the surgery rooms scheduling in smartphones, tablets and laptops.

Keywords: Scheduling, Rich Internet Application, Ant Colony System, Genetic Algorithm, Memetic Algorithm.

1 Introduction

Scheduling is a process decision-making to use manufacture and services industry, bussiness, and university environment [1] [2]. Between the typical goals of scheduling problems, maximizing resource utilization is not only a measure of academic interest, but also useful and important in practice.

Job shop [3] is combination problem and this is NP-Hard problem. Actually the metaheuristics has been a good way to solve this kind of problems, as Ant Colony System (ACS), Particle Swarm Optimization (PSO), Genetic Algorithms (GA), Memetic Algorithm (MA) etc. [3] [1] [4].

Scheduling problems plays a central role in Ant System System (ACS) research, and many different types of scheduling problems have been attacked with ACS algorithms, one them is Job Scheduling [1], [5]. ACS is inspired by the trail following behavior of ant colonies. Ants, when moving along path to find food, leave along their path a chemical called pheromone as a signal for other ants to follow. Each ant build a solution, the best solutions is marked with more attenuacion from the colony to a food source. [6].

However GA have been applied successfully in many optimization problems. When applied GA to scheduling, view sequences or schedules as individuals or member of a

Mireya García-Vázquez, Grigori Sidorov (Eds.).
Advances in Computing Science and Control.
Research in Computing Science 59, 2012, pp. 183- 193

population can be observed. Each individual has fitness, and each iteration best individuals survive to next iteration (generation), also commonly a individue is mutated, this way the population preserve genetic diversity, avoiding a local optimum [2].

Memetic Algorithm has many research about scheduling problem where its results are acceptable, as [7], [8], due MA worked combinatorial optimization enviroments, as GA, although GA and MA have many similar processes, MA represent a algorithm with desirable characteristics [8].

Floren Devin [9] use Rich Internet Applications (RIA) to show results from timetabling, says Rich Internet Applications (RIA) can address many users without any requirement because they are run on a web browser, also applications that RIAs' popularity owes much to their powerful presentation and interaction capabilities.

Actually at Social Security Institute by its acronym in spanish (IMSS) does scheduling of surgeries manually through a excel program, they have problems, sometimes, the surgeries are overlap in a surgery room and other surgery room isn't used at the same time. In this work, a program to scheduling the surgery, maximizing the use surgery rooms and number emergency surgeries is proposed.

The analized algorithms find an optimal solution, but this solution must be obtained as soon as posible and in any place, where the user remains. To solve this problem the RIA are used in that way that the surgeon can reach the information if a change in the schedule has occurred.

2 Scheduling of Surgeries

Scheduling is concerned with allocating limited resources to tasks to optimize the performance, such as completion time or production cost.[3]

The models studied in schedule are known as static where the activities, resources, process times are predefined, in other words this are not modified during the process. The automation of the timetable hospital in medical has a database with information about medical, surgical rooms, and patients [10] [11]. In this work, we are going to optimize the use of surgery rooms with scheduling.

To schedule the surgeries in medical unit should observe the following:

- Service hours.
- Duration of service.
- Patient's importance criterion.

It's important to say that every surgery has medical area, every medical area have it surgery hours with itself surgery rooms, time for morning service is *360 min* according to data of Instituto Mexicano del Seguro Social.

In figure 1 shows the representation elements of surgeries, where e is the medical area, C is scheduling's surgery, in this case, it is considered that all patients are equally important. The user will give to system a file with surgeries to scheduling, the structure of file cointains a id surgery, the medical area, time slot and grade emergency

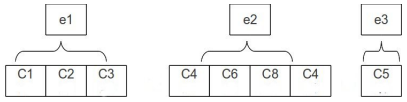


Fig. 1. Representation elements of surgeries

Formally, given *n* of independent surgeries $S = \{s_1, s_2, ..., s_n\}$, *l* surgeries rooms $G = \{g_1, g_2, ..., g_l\}$.

And each $s = \{e_i, ts_j, u_k\}$ where ts_j is time slot and u_k is emergency $u_k = \{1, 2\}$, e_i is the medical area where $e_i \in E$ being E the set of medical areas, then $E = \{e_1, e_2, ..., e_m\}$ and each e contains $\overline{r}$ an a $\overline{d}$

$$e = \{\overline{g}, \overline{d}\} \tag{1}$$

where

- $\overline{g}$ = vector assigned rooms
- $\overline{d}$ = vector assigned days
- $\overline{r} \subseteq R$
- $|\overline{r}| \leq m$
- m = number of rooms
- $|\overline{r}| \leq 7$ (days of the week)

The objective function 2 is maximize the number surgeries to schedule taking in account the emergency grade. Where the grade one is the most importance.

$$\sum_{i=0}^{n} w(ts_i) \leq \texttt{service time} \tag{2}$$

if $w = 1$
where

$$w = \begin{cases} -1 & \texttt{not scheduled} \\ 1 & \texttt{scheduled} \end{cases} \tag{3}$$

The objective function have account that really every day there are patients that should have surgery as soon as possible, and if we consider only time slot surgeries, it should not schedule surgeries with emergency grade one.

When we have more pending scheduling surgeries, the objetive function will be negative number. Also in case we have a few surgeries but with time-slot more high than scheduling surgeries maybe the objetive function will be negative too.

3 Ant Colony System

Ant Colony System (ACS) is an algorithm inspired from the foraging behavior of ant species. The ants deposit pheromone on the ground in order to mark some favorable path that should be followed by other ants.[6]

To find a solution, the surgeries are organizated, given a path, it means a sequence of task to do, then the goal is to travel greatest number surgeries, with time restriction (See figure 2).

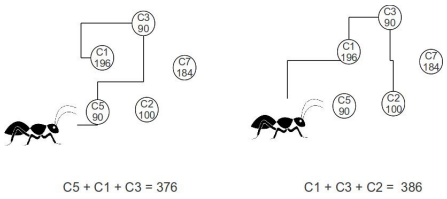


Fig. 2. Representation the ants search the solution

In Ant System each (artificial) ant is placed on a randomly chosen surgery, Setting off from its starting surgery an ant builds a complete tour by probabilistically selecting surgery to move to next until all surgeries have been visited. There are two way to select the next surgery

If *q (random number [0,1]) $\leq q_0$ (probability of exploitation)* then (see equation 4)

$$j = \arg\max_{j \in LT_k} = [\tau_{ij}]^{\alpha}[\eta_{ij}]^{\beta} \tag{4}$$

otherwise do equation 5

$$p_{ij}^{k} = \begin{cases} \frac{[\tau_{ij}]^{\alpha}[\eta_{ij}]^{\beta}}{\sum_{l \varepsilon N_i^k}[\tau_{ij}]^{\alpha}[\eta_{ij}]^{\beta}} \\ 0 \text{ otherwise} \end{cases} \tag{5}$$

where α is the constant importance pheromone, β is the constant importance distance, η is $\frac{1}{d(i,j)}$, τ the matrix of values asociate to pheromone and LT_k ant's (k) list tabu.

The value's matrix pheromones(τ_{ij}) continue preserve it concept according to [2][5], [12]; where there is more pheromone, there will are probability to better solution, however, η_{ij} changes it concept to η_i, because there aren't distance matrix ($d(i,j)$), this variable is calculate heuristic by authors as $\eta_i = \frac{1}{time-slot}$ [13].

After tour, each ant update matrix asociate to pheromone (τ). [14]. That is to say they search a local solution through the equation 6.

$$\tau(i,j) = (1-\rho)[\tau(i,j)] + \tau_0 \tag{6}$$

where ρ is constant evaporation, τ is the matrix of values asociates to pheromone and τ_0 = pheromone.
When get tour every ant, the next step is update pheromone (global search) as show in equation 7.

$$\tau(i,j) = (1-\sigma)[\tau(i,j)] + \Delta\tau_{i,j} \tag{7}$$

where σ is global search's factor, τ is a value's matrix asociate to pheromone, and $\Delta\tau_{i,j}$ is represented by equation 8.

$$\Delta\tau_{i,j} \begin{cases} \frac{1}{d_{i,j}} \text{if it is the best} \\ 0 \text{ otherwise} \end{cases} \tag{8}$$

4 Genetic Algorithm

GA are search algorithm, where main objective is find a parameter set that maximize the function 2, through the algorithm 1

Algorithm 1 Genetic Algorithm

Require: (numPopulation, totalCall, percentageElitism, percentageMutation)
1: Choose the initial population of individuals
2: **repeat**
3: Evaluate the fitness of each individual in that population
4: Do process elitism
5: Selec individuals for 3 tournament
6: Do cross of individual selected
7: Do mutation
8: Update population
9: Do process intensifier
10: Add countcall
11: **until** totalCall $\leq$ countCall

The population is represented id surgery, where it is sequence surgeries. Initial population is generated randomly, in select better individuals, ensure don't lose the the best. Through tournament selection where subgroups of individuals are chosen from larger population, possible schedule, the best of subgroup is chosen, then the parents are reproduced, the method used is annular cross where the parents do interchange of information like as figure 3. In case mutation is select a bit, a id surgery, and it is moved from its place.

During reproduction process, sometimes offspring are like as parents, then individual hasn't changes when the new individuals do reproduction process again, so solution is local optimum, then there is percentage of clones and percentage of scouts, where if

population exceed the percentage of scouts, the population get n worst individual and regenerate them (See figure 3).

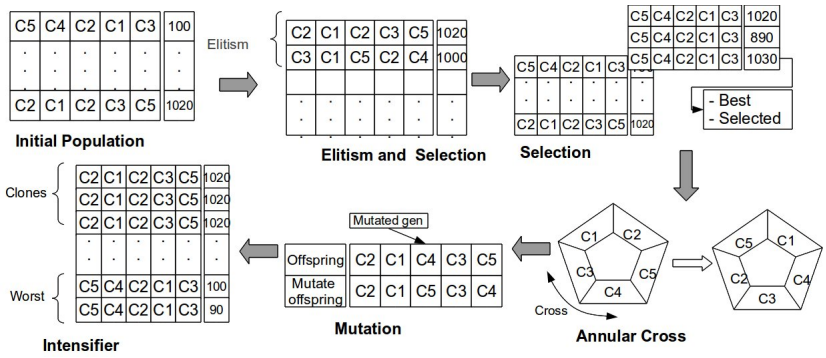


Fig. 3. Genetics algorithm's process

5 Memetic Algorithm

Dawkins in his book propound the idea called *meme* [15], which it means unit of cultural transmission or a unit of imitation. This idea, meme, was accepted and tranforming itself [8], [4].

A memetic algorithm have agent wich is each posible solution. Making a analogy in genetic algorithm is like individual. The memetic algorithm begins like as genetic algorithm (see algorithm 2), building a population of a sequences surgeries, get a subset of better population, and select agents to cross. After creating a new population, we obtain an agent to mutate, like as the mutation in section 4. Is easy to see memetic algorithm is like genetic algorithm, but it has small changes to help it, the transformation and finally the intesifier.

The transformation updates population, makes a small mutation in which a new solution can potentially replace any existing solution [4], that to say, the algorithm select a small subset of population of surgeries, make mutation moving from place an id surgery, and get a new fitness, it is compared with previous fitness to select the best between the new agent and the selected.

6 Rich Internet Applications

Interactive, dynamic interfaces are produced by Rich Internet Applications (RIAs), wich can be used like traditional desktop application, enable moving part of the computation to the client. This mean, every investigation has applications, then, the purpose is to apply the above algorithms to a hospital, were given a list of surgeries with their

Algorithm 2 Memetic Algorithm

Require: (numPopulation, totalCall, percentageElitism, percentageMutation)

```
Choose the initial population of individuals
repeat
    Evaluate the fitness of each individual in that population
    Do process elitism
    Do the method of selection call three tournament
    Do cross of individual selected
    Do mutation
    Do transformation
    Update population
    Do process intensifier
    Add countcall
until totalCall ≤ countCall
```

propieties are scheduled through ACS, GA or MA; it shows surgeries scheduled through interfaz friendly.

Interface user must be clear, that is to say, the user don't know if ACS or GA is used, but the answer must be fast and efficient. When used Rich Internet Applications, we talk about ajax. because Ajax applications can add or retrieve new data for a page it working, therefore the page will be update inmediatly.

Framework ZK is an event-driven, component-based framework to enable rich user interfaces for web applications [16]. The main mechanism of ZK is Ajax, however, the framework ZK is diferent for other frameworks because not require you to have any knowledge of JavaScript to develop Ajax-based web applications. There are three important parts in architecture framework Zk, the ZK loader, ZK AU (asynchronous update) engine, and ZK client engine (see Figure 4).

When an user make an activity (*click*), this event, "upload" for example, is bubbled up to the ZK Client Engine; ZK Client Engine decides whether and when to send the event back to the server in an Ajax request to the ZK Update Engine on the server. ZK Update Engine will invoke for handling an AU request and send a collection of commands back to the client; the ZK Client Engine evaluates each of these commands to update the widgets accordingly. Each activity is represented in figure 4

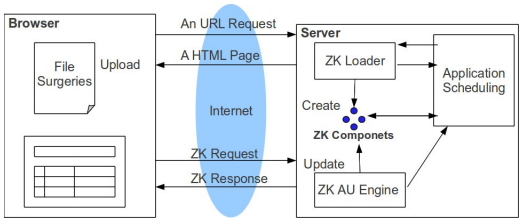


Fig. 4. Architecture of Framework ZK applied to the scheduling of surgery rooms.

7 Results

The instances used for test mentioned metaheuristics was generated initially real data from a series of surgeries done during six weeks in Instituto Mexicano del Seguro Social (IMSS), however the data showed are random continued variables by Poisson process. Those data being sorted by creating two kind of files:

.dat file : This file contains surgeries will be to schedule
.cld file : This file contains the medical area with each days and assigned surgery rooms.

The initial data for methods of ACS, GA and MA can show in table 1 where we have ten thousand function call, and diferent kind population, because each algorithm has its characteristics, strengths and weaknesses, in case ACS used few population, because if we use a lot population there aren't enought iterations to share it better local solution. In case GA and MA the mutation is least to hasn't a lot change in poputation.

Table 1. Initial data for metaheuristics Anto Colony System (ACS), Genetic Algorithm (GA), Memetic Algorithm (MA)

Item	**ACS**	**Item**	**GA**	**Item**	**MA**
Population	10	Population	50	Population	100
Function calls	10,000	Function calls	10,000	Function calls	10,000
α	5	Elitism	0.3	Elitism	0.20
β	2	Mutation	0.2	Mutation	0.01
ρ	0.9	scouts	0.4	scouts	0.10
σ	0.9				
q_0	0.5				

The table 2 show the results according objetive function, where we can observe, six intances was evaluated with ACS, GA and MA, a standart desviation mixed because in case instance textit1-5 ACS and MA show less dispersion than GA, however in case instance *6* ACS shows less dispersion than MA, a low standard deviation indicates that the data points tend to be very close to the mean, then we try to find algorithms that show results acceptables with standard deviation low.

Table 2. Instances results of metaheuristics ACS and GA

Instance	**Median**			**Standar Deviation**			**Best**		
-	**ACS**	**GA**	**MA**	**ACS**	**GA**	**MA**	**ACS**	**GA**	**MA**
1	10,136.01	9,542.82	9,414.91	287.40	346.15	184.34	10,629.50	10,126.75	9,834.50
2	3,300.00	-667.50	-1,342.50	119.21	310.81	290.62	3,420.00	-255.00	-1,800.00
3	17,100.00	17,820.00	17,520.00	310.42	345.24	173.95	17,580.00	18,375.00	17,280.00
4	14,295.00	14,857.50	14,505.00	280.37	236.88	177.63	14,805.00	15,195.00	14,250.00
5	12,450.00	12,375.00	12,195.00	179.00	404.00	236.09	12,810.00	13,365.00	11,700.00
6	15,225.00	16,785.00	16,230.00	439.95	330.23	296.77	15,810.00	17,460.00	15,630.00

The test instances selected have a set of medical areas, surgeries, times. this items are organized through the metaheuristics before mentioned obtained finally the table 3, it shows total number surgeries has each instance, and the median of surgeries scheduled; it should be emphasized every surgery is different size, therefore in time t is possible to have two surgeries inside this time t or only surgery inside this time t.

Table 3. Median of surgeries scheduled

Instance	ACS	GA	MA	Total
1	106.5	106	105	158
2	61	69	67	144
3	116	122	118.5	151
4	118.5	121	112.5	162
5	117	121.5	118.5	175
6	137	142	133	187

The user need know when the patient will be make the surgery, then the Rich Internet Applications begins to work, the next figure show a result after run anything metaheuritics before cited.

In figure 5 the results of the RIA and application of the algorithms are showed, the first combobox is medical area, and two buttons to upload or schedule but it doesn't show which algorithm is used, because this process is clear for user, then the user get a viable solution, and fast answer, recharge only the table when is changed the medical area.

When the user touch the button *calendarizar* the datas are collected, and send to metaheuristics, the user doesn't know the algorithm used, but he wait a valid response, surgeries with it day and surgery room appropiate with a scheduling that optimize the surgery rooms. Then the user can see the id surgery when and where will be. In case *N/A* means unallocated, there isn't surgery rooms available.

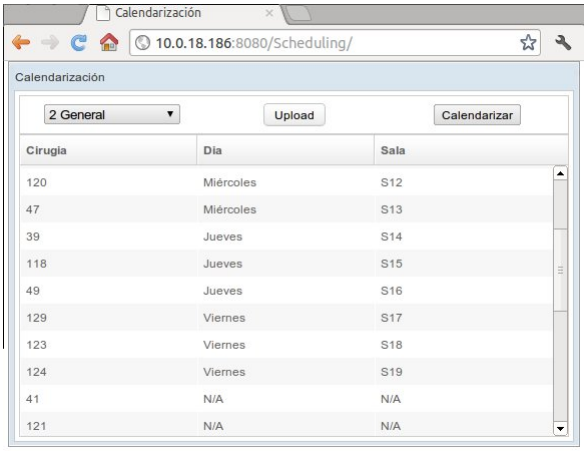


Fig. 5. GUI of scheduling for user

8 Conclusion

The scheduling of surgeries can reduce the effort of scheduling surgeries and avoid overlapping them, besides that can be optimize the resources available to the hospital, because it reduces dead-time, so adding the easy access to schedule representing the surgery.

The objetive function presents maximize the time, however sometimes the fitness is different but the result is the same, the best solution is just as equals as the worst solution, but both solutions are differents itself, because each solution has a sequence different.

Ant Colony Optimization has presented acceptable results to solution, because this algorithm build a solution itself. Additionally other advantage is to use resouces computer during it execution, because Ant Colony Optimization is provided to use threads, take some resorces togther.

However Genetic Algorithm is a algorithm make to combinatorial problems, because their results are viables how solution to problem of scheduling, also, GA is easy to understand and transfer for his simulation.

Both algorithms were comparated, however their performace was similar, through Wilcoxons Signed Sum Ranking Test can see there isn't enough statistical evidence to say which has a better performance.

On the other hand Memetic Algorithm has had a good performance, but comparing through Wilcoxons Signed Sum Ranking Test, is possible to observe that the Genetic Algorithm have a better performance than Memetic Algorithm, however comparing ACS with MA, we can observe that there isn't enough statistical evidence to say which has a better performance.

Moreover, the RIA is a good tool to show result to the user, because is better responsiveness and information flow, it has better interactivity than the tradidional web pages also this system can be seen online from any smartphone in any place.

How future job is to compare with other algorithms. To make a most appropiate interface with their respective data base. On the other hand generate new instances that cause stress in this algorithms.

Acknowledgment

Thanks to Consejo Nacional de Ciencia y Tecnología (CONACyT), for their support to carry out this research through a scolarship to Adriana Rubí Pérez López, and thanks to the project 4310.11P of DGEST.

References

1. Djamarus, D., Ku-Mahamud, K.R.: Ant system algorithm with negative pheromone for course scheduling problem. In: Eighth International Conference on Intelligent Systems Design and Applications. (2008)
2. Wu Zheng-jia, Zhang Li-ping, W.W.W.K.: Research on job-shop scheduling problem based on genetic ant colony algorithms. In: International Conference on Computational Intelligence and Security. (2009)
3. Zhixiong, L.: Investigation of particle swarm optimization for job shop scheduling. In: Third International Conference on Natural Computation Problem (ICNC2007). (2007)
4. Cotta, C., Ferndez, A.J.: Memetic Algorithms in Planning, Scheduling, and Timetabling, Evolutionary scheduling. Springer (2007)
5. Merkle Daniel, M.M., Schmeck, H.: Ant colony optimization for resource-constrained project scheduling. IEEE Transactions on Evolutionary Computation **6** (2002) 333–346
6. Dorigo Marco, S.T.: Ant Colony Optimization. A Bradford book (2004)
7. Garcia Vinícius Jacques Garcia, Morelato Franga Paulo, M.A.d.S.M.P.: A parallel memetic algorithm applied to the total tardiness machine scheduling problem. In: IPDPS'06. (2006)
8. Moscato Pablo, C.C.: An introduction to memetic algorithms. Inteligencia Artificial, Revista Iberoamericana de Inteligencia Artificial. **19** (2003) 131–148
9. Florent Devin, Y.L.N.: Timetabling RIA in action. In: Association for the Advancement of Artificial Intelligence. (2010)
10. Dolz Abadía, C.: Ventajas de la gestin informatizada en una unidad de endoscopía digestiva. Gastroenterologa y Hepatologa **28** (2005) 2005
11. De San Pedro M., Pandolfi D., L.M.V.A.: Metaheurística ACO aplicada a problemas de planificación en entornos dinámicos. In: IX Workshop de Investigadores en Ciencias de la Computacin. (2007)
12. Lutuksin, T., Pongcharoen, P.: Best worst ant colony system parameter investigation by using experimental design and analysis for course timetabling problem. In: Second International Conference on Computer and Network Technology. (2010)
13. Ritchie, G.: Static multi processor scheduling with ant colony optimisation and local search. Technical report, University of Edinburgh (2003)
14. Dorigo Marco, G.L.M.: Solving symmetric and asymmetric tsps by ant colonies. In: IEEE Conference on Evolutionary Computation. (1996)
15. Dawkins, R.: The Selfish Gene. Oxford University Press (1976)
16. Chen, H., Cheng, R.: ZK ajax without javascript framework. First press (2007)

Emotions characterization over EEG analysis : a survey

Adrian R. Aguiñaga[1], Miguel Ángel López Ramírez[1], Arnulfo Alanís Garza[1], Rosario Baltazar[2]

[1] Instituto Tecnológico de Tijuana
Blvd. Industrial y Av. ITR Tijuana S/N, 22500, Mesa Otay, Tijuana, B.C., México
[2] Instituto Tecnológico de León
Avenida Tecnológico S/N, 37000, Fracc. Industrial Julián de Obregn León, Gto., México
nefer@live.com, danym23@aol.com, alanis@tectijuana.edu.mx, charobalmx@yahoo.com.mx
Paper received on 22/09/12, Accepted on 16/10/12.

Abstract. This paper presents a review of the emotion characterization, through the analysis of electroencephalogram (EEG) signals and intelligent classifier algorithms. The set of emotions to study are those that are based on basic human survival motivation: *anger, disgust, fear, happiness, surprise and sadness*. In order to achieve the overall vision, topics are addressed as follows. First, an overview of brain computer interfaces (BCI), and a review of previous studies on the evaluation of emotions and their interactions with the environment. Second, a EEG measurement and feature extraction methods based on the implementation of the Wavelet Transform (WT), to rationalize the efficacy of the EEG data to classify the emotions. Finally, the artificial classifiers that allow to perform a characterization of the emotions over a computational models (i.e. Fuzzy C-Means, Neuronal Networks, Support Vector Machines).
Keywords: Emotions, Electroencephalogram (EEG), Wavelets, Artificial classifiers.

1 Introduction

The research of the Brain Computer Interfaces (BCI), like the opportunity to establish a connection between the users and their environment are a very active research area, due the necessity of generate models that can allow to people with some physical illness or mental disorder to interact with their environment more naturally. The main subject on this paper are focused on the fact that the emotions are one of the most important features of the humans, on recent years the research effort in the Human Computer Interaction (HCI), have a particular interest on the humans emotions and the possibility of create a computational model capable to construe human emotions, some of this works are based on the physically reactions generated by the emotions [1] [2], and psychological signal analysis [3].

In the literature, we can find different techniques that allow us to perform assessments to the human emotions activity : Electromyogram (EMG), Electrocardiogram (ECG), Skin Conductive Resistance (SCR),Respiration Rate (RR), Blood volume pressure (BVP),Heart Rate (HR) [3] [5] [8], all this techniques allows to the researches to obtain data based on the physically reaction to the emotions, some of this methods

Mireya García-Vázquez, Grigori Sidorov (Eds.).
Advances in Computing Science and Control.
Research in Computing Science 59, 2012, pp. 194- 204

had show results that achieve up to 70-98% of emotion recognition [9], however these results need a high level of environmental control.

Another methods that allow us to monitor the status of the brain activity while the subject are exposed to a emotion, are that ones that monitored the brain activity which are divided in two methods : unicellular register and brain images. The first one uses invasive techniques, which make it a not feasible due the implications that this produces. Otherwise the brain images have a good speed test and gives a good approach to the real data, some of this techniques are: Positron Emission Tomography (PET), Magnetic Resonance (MRI), Functional MRI (fMRI), Electroencephalogram (EEG), Magnetic Electroencephalogram (MEG) and simulation of Transcranial Magnetic Stimulation (TMS). Product of this previous research and methods a relation between the brain activity that are bound to the emotions has been established, however the main difficulty lies in the fact that, it is very hard to uniquely map physiological patterns onto emotion types and the physiological data are very sensitive to artifacts and noises [10].

Exist a great interest on the EEG technique due the feasibility and low cost of his implementation, compared with the others, this technique measured the activity of a group of neuronal cells of the cerebral cortex or scalp, this technique provides physical, physiological and pathological information, which can be analyzed for a medical diagnostic or the research of cognitive processes, with the advantage of not perturb the environment of the test subjects, due that the implementation of the EEG are minimally invasive.

2 Emotions

The emotions are a simple expression that consist on feelings and thoughts but at the same time are subconscious internal processes [11], this processes encompass the daily life and play a key role on it, however the emotions cannot be objectively observed and cannot be measured [10].

As result of the daily experiences we could suggest the emotions as the visible product of another way of thinking, which complement to reason on the daily's decision maker process, based on the perceptions related to the knowledge and the goals, characterized by a trigger and a value [12]. One of the main functions of the emotions are select or provide weight to one or several features from the attention center, and bring a comparative of alternatives in decisions which facilitate the communication.

Emotions involve multiples areas of the brain and shows a complex activation sequences in time. Trough signals measurements the central nervous system provides a relation between physiological changes of emotions and the brain activity [13]. In robotic and virtual agents areas to analyze this situations with a computational approach, exist an agreement of which are the basic emotions, this are the emotions related to survival human motivations, which are listed below: i) Anger: This emotion allow us to prevent that some unwanted situation continues (Right frontal cortex activation). ii)Disgust: Allows us to make a change/or correct something that not contribute to the satisfaction of the needs (Right Prefrontal Cortex Area). iii) Fear: is a defense mechanism to discover the threats (Bilateral Temporal Activation). iv) Happiness: Is a reward that motivates the search (Right Prefrontal Cortex Area). vi) Sadness: This are the manifest of that

there's something that need to be satisfied (Left Temporal Areas). The sadness and the happiness involves all most all the brain areas, and also all emotions shares pre-frontal cortex, cingulate gyrus, and temporal cortex areas [11].

Several schemes of emotions classification are defined, all of them shows that a emotion reflects an unique motivational tendency and behavior. This provides the idea of emotions as a very significant part of a system, since they represent unique forms of action related to a physiological patterns [15] [16]. The researchers use this physiological patterns to classify the emotions into three types: (i) Distress (ii) Interest and (iii) Pleasure [17]. Exist many other sets of emotions accepted as the basic sets of basic emotions, i.e. virtual emotions manage like basic emotions, seven basic emotions: Fear, Anger, Joy, Disgust, Acceptance, Anticipation, and Surprise [18], for pattern recognition (linear discriminant) manage four: emotions happiness, sadness, anger, fear and for physiological patterns a set of six emotions defined : happy, sad, disgust, fear, joy and anger and many others[9] [21].

3 EEG Data Acquisitions

The electroencephalogram (EEG) is a recording of the electrical activity of the brain from the scalp. The recorded wave forms reflect the cortical electrical activity, this provides a frequency that in some literature's are refereed as the rhythmic repetitive activity (in Hz). Related to this the frequency of EEG activity can be classified for it different properties: Rhythmic (EEG activity consisting in waves of approximately constant frequency), Arrhythmic (EEG activity in which no stable rhythms are present), Dysrhythmic (Rhythms and/or patterns of EEG activity that characteristically some disorder).

The frequency content of the EEG signals are the fundamental information to appraising it and is on the 10 Hz where the most significant information are contained [19]. The most common brain signals provides five EEG rhythms that are classified on on table 1, under the premise of a human presents changes of emotions, the frequencies of the activity the brain also present a variations. In fact a individual band of frequency can yield information located on subtle change of emotion [24].

Table 1. Rhythms classifications [24].

$Band$	$Frecuency(Hz)$	$Range(Hz)$
$Delta(\delta)$	.5-4	3.5
$Theta(\theta)$	4-8	4
$Alpha(\alpha)$	8-12	4
$Beta(\beta)$	12-30	18
$Gamma(\gamma)$	> 35	–

Exist also an alpha-like variant called mu (μ) and can be found over the motor cortex (central scalp) that is reduced with movement, or the intention to move for this reason this band are not included in table 1. In the study of the emotions of based on the EEG

signal, the alpha (8 Hz 12 Hz) and beta (12 Hz to 30 Hz) bands were highlighted as particular rhythms for recognition.

The arrangement of the placement for scalp electrodes for a EEG plays a important role of recording of the signals, on the recent researches the adoption of the 10-10 international standard system are placed ad the most common electrode placement used by the researchers but there's not a benchmark which indicates a electrodes placement. This standard are a extended version of the 10-20 system of electrode placement as is shown on the Figure 1, also known as the 10% system.

The placement of electrodes is based on landmarks on the skull, namely the nasion (Nz), the inion (Iz), and the left and right pre-auricular points (LPA and RPA). The first step is to form the line from Nz to Iz, approximately over the vertex. To determine the location of the vertex, the contour from LPA to RPA is also passed over the vertex. These two contours should intersect at 50% of their lengths and the point thus obtained is the exact vertex, also exist other proposed the standard for placing the electrodes in various locations at the scalp, and called as 5% system (proportional distances 5% of total length along the contours between skull landmarks), however the use of larger number of electrodes may increase the recognition rate of EEG signal with the cost of compromising physical burden to the subjects and longer time consumption [15].

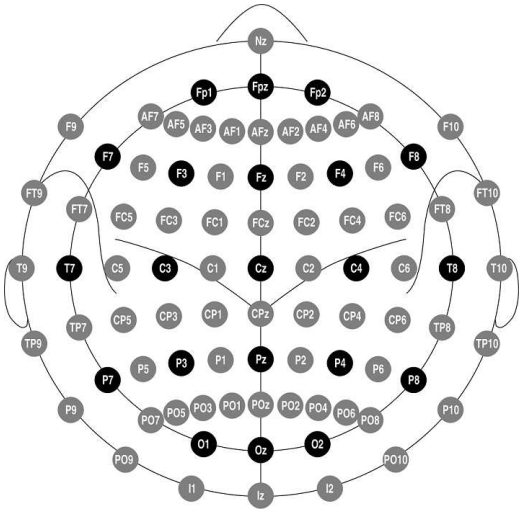


Fig. 1. Electrode positions based on the 10-10 extension of the 10-20 standard [25].

4 Feature Extraction

The electric activity obtained by the EEG, are contaminated with noises generated by the power line and external interfaces, in addition with other kind of signals that aren't

product of a brain activity which are known as artifacts (effects produced by the ocular movement and the relation with the brain activity, muscular movements, vascular movements and gloss kinetic artifacts), i.e. previous contributions to identify regular oscillations around 10 Hz to 12 Hz shows that both the alpha (α) rhythm and theta (θ) rhythm are significantly affected by the subject's eye blinking [23] .

The emotion recognition research based on EEG signals, the implementations of non-parametric methods of feature extraction that are based on multi-resolution analysis, on of this method are the Wavelet Transform (WT) [24]. The joint time-frequency resolution obtained by WT makes it as a good candidate for the extraction of details as well as approximations of the signal, which cannot be obtained either by Fast Fourier Transform (FFT) or by Short Time Fourier Transform (STFT), because even though Fourier allow us to obtain a representation of the signal on the frequency, do not provides a time resolution, this means we could know the main frequencies but not in which moment occur, in other case the STFT could solve this problem but is only viable to stationary periodic signals, and like all the most of the biological systems the emotions are non-stationary signals.

Wavelet is a wave limited by a time, which has an average value of zero that allows to describe anomalies, pulses and other events that start and finish within signal that allows analysis time located in a large signal, providing the possibility of encountering discontinuities or peaks of short duration that would otherwise, it would be difficult to detect and treat and this technique are a based on spectral estimation where we can express any general function that can be expressed on a infinite series of wavelets. The idea of this function is represented as a linear combination of a particular set of functions, obtained by translating and scaling of a basic function called mother wavelet ($\Psi_{a,b}$).

For this kind of signals the $\Psi_{a,b}$ is given as

$$\Psi_{a,b}(t) = \frac{1}{\sqrt{a}}\Psi\frac{t-b}{a} \tag{1}$$

Where $_{a,b} \in R$,$a > 0$, R is the wavelet space.

The parameters "a" are the scaling factor and "b" are the shifting factor, by selecting a suitable scaling values and a time offset for a wavelet it can be established like a effective method to analyze a non stationary bio-metrical signals [22]. The WT are then obtained by the internal multiplication of *f(t)* with the wavelet function.

$$W_f(a,b) = \{f(t), \Psi_{a,b}(t)\} \tag{2}$$

This transform reflects the state of the function on *f(t)* on the scale (frequency) and the position (time). The only limitation for chose a prototype as a mother wavelet is to satisfy the admissibility condition.

$$C_\Psi = \int_{-\infty}^{\infty} \frac{|\Psi(\omega)|^2}{\omega} d(\omega) < \infty \tag{3}$$

Where the $\Psi(\omega)$ is the Fourier transform of $\Psi_{a,b}(t)$.

The time-frequency representation is performed by repeatedly filtering the signal with a pair of filters that cut the frequency domain in the middle, this means that WT

decomposes the signal into approximation coefficients (CA) and detailed coefficients (CD). The approximation coefficient is subsequently divided into new approximation and detailed coefficients. This process is carried out iterates and producing a set of approximation coefficients with detail coefficients at different levels or scales as we can appreciate on table 2 [24] [38].

Table 2. Decomposition of EEG signals into different frequency [24].

$Band$	$Frecuency\ Range(Hz)$	$Decomposition\ Level$
$Delta(\delta)$	0-4	D6
$Theta(\theta)$	4-8	D5
$Alpha(\alpha)$	8-14	D4
$Beta(\beta)$	14-32	D3
$Gamma(\gamma)$	32-64	D2
$Noises)$	64-128	D1

There's another method to filter the signal based on the of Surface Laplacian (SL) filter for removing the noises and artifacts (Eq. 4). The SL filter is used to emphasize the electric activities and filtering out those that might have an origin outside the skull, however this method could lose various spatial frequencies from the middle frequencies which could be a potentially useful information.

$$X_{new} = X(t) - \frac{1}{N_E}\sum_{i=1}^{N_E} X_i(t) \tag{4}$$

By now we can divide the feature extraction techniques as: Time Domain Analysis, Frequency Domain Analysis and Time-Frequency Analysis, however also exist two other methods for feature extraction: Fractal Analysis and Interval Analysis, the first are a new scientific paradigm that has been successfully used in quantifying the complexity of dynamical signals in biology and medicine[26]. The intervals analysis is also widely accepted due to its simplicity and usefulness, but is not sensitive to the noises and other artifacts [27].

5 Emotion Classification

Recognize a emotional state based on analyze the features from inputs with a good accuracy, are one of the main goals of the emotions recognition, to achieve this the time taken for training through an intelligent method are a important factor for emotions classification. Many techniques to perform the classification can be used, i.e. Support Vector Machines (SVM), Neural Networks (NN), Linear Discriminant Analysis (LDA), Multi Layer Perceptron Network (MLPN), Naïve Bayes Classifier (NBC), Fisher Discriminant Analysis(FDA), Binary Fisher Discriminant Analysis (BFDA), Transferable Belief Model (TBM) and many others. The table 3, shows the previous researches made

to achieve the emotions recognition by the analysis of the brain signals and intelligent classifiers.

The actual researches are focused to provide a better interpretation from the correlation between emotions and brainwaves are growing with the main goal of create BCI applications, nevertheless only few results add conclusions which correlate the EEG with particular emotion. In correlation with the filters and the classifiers, the use of visual, audio or audio-visual stimuli to provoke emotions, plays a very important role. Which are used to intentionally create lab settings based on exposing to the subject to a selected emotional images from the international standard data bases to obtain brainwaves of a desired emotions.

The intensity of the emotion experienced by the subjects during the experimental analysis with audio-visual stimuli provoke much better results compared the other two stimulus methods (audio or visual)[10]. In the same panorama the consideration of use statistical features to classify emotions are implemented too, such as mean, standard deviation, power, variance and Analysis of Variance (ANOVA).

Other thing to note in table 3, is that the valence-arousal based on two dimensional analysis of emotion classification is discussed in majority of works, and that the neural network and LDA methods have higher classification rate than other classification methods such as SVM, Linear Mapping (LM), and LBC classifiers as can be observed on table 3. A recognition accuracy of over 90% of average seems to be acceptable for realistic applications and actual research findings suggest this are possible and brainwaves can be successfully identified using statistical features [15]. In general, the assessment of human emotions with greater accuracy is depends on number of electrodes, placement of electrodes, method of pre-processing and feature extraction used for emotion detection.

Since the classification methods depends on grouping the unknown data by previous learning process[15]. A large amount of researchers tend to use lesser number of electrodes for recording EEG signals, most of the electrodes are placed on the region of frontal lobe and parietal lobes are considered for acquiring the EEG. In addition, there is no standard benchmark available to select the number of electrodes and location of electrodes in human scalp for emotion recognition applications [24].

6 Conclusion

A significant amount of investigators have dealt researches for determining the best methodology to obtain a acceptable classification accuracy. However since,the EEG patterns are different for each person, a set of interdisciplinary and collaborative works are required for concluding some significant results. Other aspect to note is, that only a few researchers have discussed about the effects of multiple emotions on given stimuli, generating on this is one of the important area to research since, as we know multiple emotions under natural conditions could occur and are not stationary which mean that a state of a emotion may vary in time.

To reach the goal of classify emotions on efficient way, a several further studies are required to analyze the relation between brain waves and the effect of multiple emotions,coupled with this a development of more efficient pre-processing and feature ex-

Table 3. Review of the implementation of artificial classifiers for emotion recognition

Classification Method	Summary
Support Vector Machines (SVM)	*The classification of EEG signals for BCI applications through adaptive learning and co-variance matrices with adaptive learning and Common Spatial Patterns (CSP). Classification Average : 70-76%* [28].
SVM	*Recognition of five emotions (joy, anger, sadness, fear, and relax) through multi-modal bio-potential Signals. Classification Average : 41.7 %* [3].
MLPN	*Recognition of four emotions (angry, sadness, pleasure, and joy) have been classified using MLP network. Classification Average : 69.69%* [29].
BFDA	*PCA is applied for feature selection and emotion is classified in two dimensional axis that means alpha and beta band power and power ratio between alpha and beta band. Classification Average: 90%* [30].
SVM + NN	*Estimation of two emotions (Pleasure and unpleasure) by the classification using NN and SVM. Classification Average : 62.3 %* [31].
LDA + NN	*Estimation of five human emotions (joy, anger, relaxation, sadness, and worry) using fractal- dimension of EEG is classified using NN and Linear mapping network* [32].
FDA + NBC	*Estimation of emotions in two dimensional space is detected, with the power at six frequency bands. Classification Average : 54 - 55 %* [33].
NBC + SVM + NN	*Emotions are classified in valence-arousal space using three classifiers, using the frequency band power, Cross correlation coefficients, Peak frequency in alpha and beta band, and Hjorth parameters. Classification Average : 29% - 35%* [34].
Transferable Belief Model(TBM)	*Estimation of two emotions (Positive -happy- and negative emotions-sad-) are detected, by using the patterns of the energies at various frequency bands of EEG signal, and power at selected frequency bands.* [35].
Mean, Analysis of Variance, Standard deviation	*Estimation of three emotions (pleasant, aversive, and neutral) are classified according features such as mean, standard deviation, power, variance and Analysis of Variance* [36].
Multivariate Analysis of Variance (MANOVA)	*Recognition of (depressed or Non-depressed) emotions derived through MANOVA* [37].
Fuzzy C-Means Clustering (FCM)	*Measures Emotions using EEG signals on classifying human emotions using fuzzy c-means clustering for Wavelets, based on minimizing : Fuzziness Performance Index (FPI), Modified Partition Entropy (MPE) and Separable Distance (SD). In addition, the Objective Function Value (OFV) is also considered as a measure for classifying emotions* [10].
LDA + K Nearest Neighbor (KNN)	*Classification of emotions divided on subsets (disgust, happy, surprice, fear and neutral), whit a maximum subsets of emotions classification rate of 91.67% for disgust, 81.67% for happy and surprise, 81.25% for fear and 93.75% for neutral is achieved using 62 channels EEG signals* [24].

traction methods are required to. Also, further studies are required to select the number of electrodes and the placement of electrodes on brain region due to these are essential to obtain valuable information in the emotion recognition. Furthermore there is no comparative study has been possible by determining the statistical correlation between different emotions and EEG signals. Also a factor to empathize are the applications of this methods to help people with some illness, by the development on systems that can help them to interact them on a more natural way.

References

1. Bung, H. and Furui, S. : Automatic recognition and understanding of spoken languages-a first step toward natural human machine communication: Proceedings of IEEE, 88, 1142-1165 (2000).
2. Cowie, R., Douglas, E., Tsapatsoulis, N., Votsis, G., Kollias, G., Fellenz, W. and Taylor, J.G.: Emotion Recognition in human-computer interaction: IEEE Signal Processing, 1, 3773-3776 (2001)
3. Takahashi, K.: Remarks on emotion recognition from bio-potential signals : The Second International Conference on Autonomous Robots and Agents, 186-191 (2004).
4. Picard R.W, Vyzas E., and Healey J : Towards Machine Emotional Intelligence: Analysis of Affective Physiological State: IEEE Transactions on Pattern Analysis and MachineIntelligence, Vol 23(10), pp, 1175- 1191 (2001).
5. Wagner J, Kim J., and Andre E : From Physiological Signals to Emotions: Implementing and Comparing Selected Methods for Feature Extraction and Classification: IEEE Proceedings on Multimedia and Expo, pp, 940-943 (2005).
6. Nasoz F, Lisetti C L, Alvarez K., and Finkelstein N : Emotion Recognition from Physiological Signals for User Modeling of Affect: Proceedings UM'2003, 9th International Conference on User Model, Pittsburg, USA, June 22-26, pp, 1-9 (2003).
7. Egon L, Broek V D, Schutt M. H, Westerink J H D M, Herk J V., and Tuinenbreijer K : Computing Emotion Awareness through Facial Electromyography: Human Computer Interaction 2006, pp, 52- 63 (2006).
8. Kim K.H, Bang S W, and Kim S. R,: Development of Person Independent Emotion Recognition System based on Multiple Physiological Signals: IEEE proceedings on EMBS, pp, 50-51(2002).
9. Picard R.W, Vyzas E., and Healey J : Towards Machine Emotional Intelligence: Analysis of Affective Physiological State. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol 23(10), pp, 1175- 1191 (2001).
10. M. Murugappan, M. Rizon, R Nagarajan,S. Yaacob, : Inferring of Human Emotional States using Multichannel EEG. European Journal of Scientific Research ISSN 1450-216X Vol.48 No.2, pp.281-299 (2010).
11. AlMejrad, A.S.: Human emotions detection using brain wave signals: A challenging. European Journal of Scientific Research, 44(4), 640-659 (2010).
12. L.D. Canamero. Designing Emotions for Activity Selection in Autonomous Agents, in R. Trappl, P. Petta, S.Payr, eds., Emotions in Humans and artifacts, cambridge, MA: the MIT Press, pp, 115-148 (2003).
13. Winton WM, Putnam L., and Krauss R : Facial and Autonomic Manifestations of the Dimensional Structure of Emotion: Journal of Experimental Social Psychology, pp, 195-216 (1984).
14. Cacioppo C.J and Tassinary LG : Inferring Physiological Significance from Physiological Signals : American Psychologist, Vol 45(1), pp, 16-18 (1990).

15. M. Murugappan, R. Nagarajan and S. Yaacob: Discrete Wavelet Transform Based Selection of Salient EEG Frequency Band for Assessing Human Emotions. J. Biomedical Science and Engineering, 3, 390-396 (2010).
16. Garcia O, Favela J., and Machorro R : Emotional Awareness in Collaborative Systems: IEEE Proceedings on Signal Processing and Information Retrieval Symposium, pp, 296-303 (1999).
17. Levis M : Self-Conscious emotions: American Scientist, Vol 83, pp, 68-78 (1995).
18. Plutchik R and Kellerman H. A General Psychoevolutionary Theory of Emotion. Emotion Theory, Research, and Experience. Vol 1, Academic Press, pp, 99-100 (1996).
19. Berger H : Uber das Elecktroenzephalogramm des Menschen: Arch Psychiat nervenkrankheitten, Vol 87, pp, 527-570 (1929).
20. Gonzalo Ulloa, Camilo E. Valderrama: Spectral analysis of physiological parameters for consumers emotion detection: Revista S&T, 10(20), 27-49. (2012).
21. Flidlund A.J and Izard E.Z : Electromyographic Studies of Facial Expressions of Emotions and Patterns of Emotions. Social Psychophysiology: A Source Book (1983).
22. Pardue, J.H., Landry, J.P., & Clark, T.D.,Jr. : A soft systems approach to input distribution estimation for a non-stationary demand process: En WSC '95 Proceedings of the 27th conference on Winter simulation, pp. 982-987. Washington, DC: IEEE Computer Society. doi: 10.1145/224401.224761 (1995)
23. Adrian ED . J Physiol 61, 4972 (1926).
24. M. Murugappan, Nagarajan ,R Nagarajan,S. Yaacob: Classification of human emotion from EEG using discrete wavelet transform : Chaos, SJ. Biomedical Science and Engineering, 3, 390-396 (2010).
25. Robert Oostenveld,Peter Praamstra: The five percent electrode system for high-resolution EEG and ERP measurements: Clinical Neurophysiology 112 713-719 (2001).
26. Adlakha A: Single trial EEG classification, Technical Report: Swiss Federal Institute of Technology (2002).
27. Saltzberg B : A New Approach to Signal Analysis in Electroencephalography: IRE Transactions on Medical Electronics, Vol 8, pp, 24-30 (1957).
28. Sun S and Zhang C : Adaptive Feature Extraction for EEG Signal Classification:IEEE Transactions on Medical and Biological Engineering and Computing, Vol 44 (10), pp, 931-935 (2006).
29. Lin Y P, Wang C H, Wu T L, Jeng S K, and Chen J H : Multilayer Perceptron for EEG Signal Classification during Listening to Emotional Music : Proceedings of TENCON 2007, pp, 1-3 (2007).
30. Danny O B : EEG-based Emotion Recognition The Influence of Visual and Auditory Stimuli: Technical Report, pp, 1-16 (2008).
31. Takahashi K and Tsukaguchi A : Remarks on Emotion Recognition from Multi-Modal Bio-Potential Signals: IEEE Transactions on Industrial Technology, Vol 3, pp, 1654-1659 (2003).
32. Iizuka T and Nakawa M : Emotion Analysis with Fractal-Dimension of EEG Signals : IEIC Technical Report, Vol 102 (534), pp, 13-18 (2005).
33. Chanel G, Kronegg J, Granjean D., and Pun T : Emotion Assessment: Arousal Evaluation using EEGs and Peripheral Physiological Signals: Lecturer Notes in Computer Science, Springer, pp, 530 537 (2006).
34. Horlings R, Datcu D, and Rothkrantz L J M : Emotion Recognition using Brain Activity: Proceedings on Int Conference on Computer Systems and Technologies, pp, II-1-1 II-1-5 (2008).
35. Savran A, Ciftci K, Chanel G, Javier Cruz Mota, Luong Hong Viet, Sankur B, Akarun L, Caplier A., and Rombaut M : Emotion Detection in the Loop from Brain Signals and Facial Images: Final Project Report, eNTERFACE06 (2006).

36. Oya H, Kawasaki H, Mathhew A, Howard III. and Adolphs R : Electrophysiological Responses in the Human Amygdala Discriminate Emotion Categories of Complex Visual Stimuli: The Journal of Neuroscience, Vol 22(21), pp, 9502-9512 (2002).
37. Jones N A, Field T, Fox N A M, Davalos., and Gomez C : EEG during Different Emotions in 10-month-old infants of Depressed Mothers : Journal of Reproductive and Infant Psychology, Vol 19, pp, 296- 312 (2001).
38. Chethan, P. and Cox, M. : Frequency characteristics of wavelets : IEEE Transactions on Power Delivery, 17(3), 800-804 (2002).

Series de Tiempo Difusas aplicadas al pronóstico de la remuneración por la fabricación del calzado

Marisol Gutiérrez[1], Luis Ernesto Mancilla Espinoza[1], Alfonso Gutiérrez Lugo[2] y Marco A. Gutiérrez[2]

[1] División de Estudios de Posgrado e Investigación, Instituto Tecnológico de León, Av. Tecnológico S/N, 37290 Guanajuato, México
marisol.gutierrez@ieee.org, lmancilla01@hotmail.com
http://posgrado.itleon.edu.mx/
[2] Ingeniería Industrial, Instituto Tecnológico Superior de Lagos de Moreno Libramiento Tecnológico No. 5000, 47480 Jalisco, México
http://www.teclagos.edu.mx/
Paper received on 04/10/12, Accepted on 25/10/12.

Resumen En el presente trabajo, se utiliza el método llamado Series de Tiempo Difusas en tiempo variante, para realizar el pronóstico de la remuneración por la fabricación del calzado (cuyos datos son obtenidos por medio de la Encuesta Mensual de la Industria Manufacturera, en el ciclo de Enero-2008 a Julio-2011), donde los conjuntos manejados son difusos y basado en la diferencia de segundo orden para obtener la tendencia. Al final los resultados obtenidos con el método de series de tiempo difusas son comparados con el método de regresión lineal, respecto al error obtenido por ambos.

Palabras claves: Conjuntos Difusos, Series de Tiempo Difusas, Pronóstico.

1. Introducción

La toma de decisiones depende del nivel de exactitud de la predicción que se desee (por lo cual es un proceso muy complejo, ya que interviene el futuro). En la vida diaria se desempeña el proceso de predicción muy cotidianamente, como por ejemplo el más popular el pronóstico del clima por mencionar alguno.

En algunos casos los modelos clásicos no pueden pronosticar (predecir el futuro) si los valores están representados en términos lingüísticos, es decir, en conjuntos difusos (es un conjunto que puede contener elementos de forma parcial [1] [2], es decir, en los conjuntos clásicos, no pertenece = 0 y pertenece = 1, pero los conjuntos difusos permiten grados de pertenencia entre 0 y 1) introducidos por Zadeh en 1965 [3]. Song y Chissom [4] [5] [6] presentan la teoría de Series de Tiempo Difusas para mejorar los inconvenientes que presentan los métodos de Series de Tiempo Clásicas.

La predicción con Series de Tiempo Difusas tiene una amplia aplicación la cual resulta muy importante, por ejemplo: planificación económica y de negocios, inventarios y control de producción, etcétera.

Mireya García-Vázquez, Grigori Sidorov (Eds.).
Advances in Computing Science and Control.
Research in Computing Science 59, 2012, pp. 205- 216

Song y Chissom pronostican el número de matrículas en la Universidad de Alabama, basados en la teoría de Series de Tiempo Difusas. Chen [7] continuó con la investigación de Song y Chissom, el cual redujo el tiempo y proceso de cálculo. Huang [8] usó la diferencia de matrículas, para predecir y posteriormente en el artículo [9] extiende el trabajo [7] con la adición de reglas heurísticas. Chen [10] utilizó un método para ajustar la longitud de cada intervalo con Algoritmos Genéticos. Jilani [11] presenta un método propuesto que pertenece al orden k-ésimo. Pathak y Singh [12] continúan con el pronóstico del número de matrículas en la Universidad de Alabama, introducen el concepto de intervalo de banda ancha de $(4/3)\sigma$ para el pronóstico. Jasim [13] pronostica el número de matriculas, en base al método de primer orden en tiempo variante.

Chen y Hwang [14] presentan un método para predecir la temperatura, basado en Series de Tiempo Difusas. Shah [15] pronóstica el capital de la India con uso de Series de Tiempo Difusa, donde calcula la tendencia con la diferencia de segundo orden para el pronóstico. Ecerkale [16] empleó un periodo de tres años de los datos de producción de combustibles de aviación de Turquía como datos experimentales para el pronóstico.

Song [5] utilizó el siguiente modelo para el pronóstico de la matrícula universitaria:

$$A_i = A_{i-1} \, {}^{O}R \tag{1}$$

Donde A_i, se refiere a las inscripciones difusas del año i, el símbolo o denota el operador de composición Max-Min, y R es la relación difusa formada por las inscripciones difusas, extraídas de las Series de Tiempo Difusas. Este método tiene algunas desventajas: requiere una gran cantidad de cálculos para obtener la relación difusa R. La operación de composición de máximos y mínimos de la ecuación (1) realiza una gran cantidad de cálculos, cuando la relación difusa R es muy grande.

Esta investigación, es estructurada como a continuación se muestra. En la sección 2, se explican los conceptos básicos necesarios para comprender las Series de Tiempo Difusas. En la sección 3, se presentan los objetivos de utilizar Series de Tiempo Difusas en esta investigación. En la sección 4, se describe detalladamente la metodología utilizada. En la sección 5, se muestran los resultados obtenidos con las Series de Tiempo Difusas y son comparados con los resultados obtenidos con Regresión Lineal por medio de sus errores. Finalmente en la sección 5, se exponen las conclusiones.

2. Conceptos básicos de Series de Tiempo Difusas

Song [5] propone la definición de series de tiempo difusas, basadas en conjuntos difusos [1] [2] [3].

Sea U el universo en discurso. $U = \{u_1, u_2, ..., u_n\}$, y sea A el conjunto difuso del universo en discurso U definido como a continuación:

$$A = \mu_A(u_1)/u_1 + \mu_A(u_2)/u_2 + ... + \mu_A(u_n)/u_n \tag{2}$$

Donde μ_A es la función de pertenencia de A, $\mu_A : U \rightarrow [0, 1]$, $\mu_A(u_i)$ indica el grado de pertenencia de u_i en el conjunto difuso A. $\mu_A(u_i)\varepsilon[0, 1]$ para $1 \leq i \leq n$.

$X(t)$ (para $t = ..., 1, 2, ...$) se encuentra dentro del universo en discurso y es un subconjunto de R, y el conjunto difuso $f_i(t)$ (para $i = 1, 2, ...$) está definido en $X(t)$.

Sea $F(t)$ la colección de $f_i(t)$ (para $i = 1, 2, ...$). Entonces $F(t)$ es llamada Serie de Tiempo Difusa de $X(t)$ (para $t = ..., 1, 2, ...$).

Si $F(t)$ es producida por $F(t-1)$, que se denota por $F(t-1) \rightarrow F(t)$, Esta relación puede representarse por $F(t) = F(t-1) \, {}^{O}R(t, t-1)$, donde el símbolo o denota el operador de composición Max-Min, $R(t, t-1)$ es la relación difusa que hay entre $F(t)$ y $F(t-1)$ y es llamado modelo de primer orden de $F(t)$.

Sea $F(t)$ una Serie de Tiempo Difusa y sea $R(t, t-1)$ la relación difusa de un modelo de primer orden de $F(t)$. Si $R(t, t-1)$ es igual a $R(t-1, t, 2)$ para cualquier tiempo t, entonces $F(t)$ se llama Serie de Tiempo Difusa en tiempo invariante. Si $R(r, t-1)$ es dependiente del tiempo t, es decir, $R(t, t-1)$ puede ser diferente de $R(t-1, t-2)$ para cualquier t, entonces $F(t)$ se llama Serie de Tiempo Difusa en tiempo variante.

Song [6] propuso un modelo de Series de Tiempo Difusas, para tiempo variable y pronosticar el número de matrículas de la Universidad de Alabama. A continuación se muestra una breve reseña del método [6]:

Paso 1: Definir el universo en discurso U donde están definidos los conjuntos difusos.

Paso 2: Dividir el universo en discurso U en intervalos de longitudes iguales.

Paso 3: Determinar los valores lingüísticos representados por los intervalos de los conjuntos difusos del universo en discurso.

Paso 4: Cambiar a conjuntos difusos los datos históricos de las matrículas.

Paso 5: Elegir un adecuado parámetro w, donde $w > 1$, para calcular $R^w(t, t-1)$ y el pronóstico de las matrículas como $F(t) = F(t-1) \, {}^{O}R^w(t, t-1)$.

Donde $F(t)$ denota el pronóstico difuso de la matrícula del año t, $F(t-1)$ denota el dato de la matrícula difusa del año $t-1$.

$$R^w(t, t-1) = F^T(t-2) \times F(t-1) \cup F^T(t-3) \times F(t-3) \cup ... \cup F^T(t-w) \times F(t-w+1) \tag{3}$$

Donde w es llamado el modelo base que indica el número de años antes de t, $\times$ es el operador de producto cartesiano, y T es el operador de transposición.

Paso 6: Obtiene el pronóstico de la matrícula por medio de Redes Neuronales.

3. Objetivo

Los modelos de series de tiempo clásicos no pueden predecir los periodos próximos, si las series de tiempo están expresadas lingüísticamente. El objetivo de la investigación es pronosticar la Remuneración por la fabricación del calzado, de los periodos siguientes dados los periodos anteriores.

La Serie de Tiempo de la Remuneración, se tomó del sitio oficial en Internet de PROSPECTA [17] a su vez obtenidos del Instituto Nacional de Estadística y Geografía (INEGI) por medio de Encuesta mensual de la industria manufacturera (EMIM), con una muestra de 603 empresas de Fabricación de calzado, en el ciclo de Enero-2008 a Julio-2011 (comprendido por 43 periodos cada uno con duración de un mes).

4. Metodología

A continuación se describe la metodología utilizada para el pronóstico por medio Series de Tiempo Difusas.

Este método está compuesto por 5 pasos principales, el primero paso, donde se define el universo en discurso y es dividido en intervalos de igual tamaño, el segundo paso, el universo es re-dividido en sub-intervalos de tamaños difusos según la distribución estadística de los datos, el tercer paso, se define para cada conjunto difuso A un valor lingüístico y un grado de pertenencia para cada intervalo del conjunto, el cuarto paso, se establece la relación lógica difusa entre los datos, el quinto paso, con el fin de predecir el dato del próximo periodo se utiliza una serie de reglas para encontrar si la tendencia es a la alza, la baja o tiene un ritmo constante, el método se basa en diferencias de segundo orden para encontrar el valor del próximo periodo.

PASO 1:

Primero se define el universo en discurso de acuerdo con los datos históricos (ver Tabla 1). Se encuentra el valor máximo (356,000) y mínimo (225,360), de los datos históricos, después se busca un número redondeado mayor que el máximo (360,000), así como un número redondeado menor que el mínimo (220,000). Con el objetivo de poder dividir el universo en n intervalos, para que los límites de dichos intervalos sean enteros. Entonces se tiene el universo que se denota con la letra U.

$$U = [220000, 360000] \tag{4}$$

Se divide el universo en discurso U (ecuación 4), en varios intervalos (n) de igual longitud $u_1, u_2, ..., u_n$. Para este trabajo el universo se divide en siete intervalos ($n = 7$) de la misma longitud (tamaño de la longitud es 20,000). Por lo tanto los intervalos quedan de la siguiente manera: $u_1 = [220000, 240000]$, $u_2 = [240000, 260000]$, $u_3 = [260000, 280000]$, $u_4 = [280000, 300000]$, $u_5 = [300000, 320000]$,
$u_6 = [320000, 340000]$ y $u_7 = [340000, 360000]$.

PASO 2:

Obtener la distribución estadística de los datos históricos en cada intervalo (ver Tabla 2). Se contabilizan todos los datos que se encuentran en cada intervalo, es decir, la frecuencia en la que cada intervalo se repite en los datos históricos (ver Fig .1).

Una vez que se tiene la distribución estadística de los datos, los intervalos obtenidos en el paso 1 se dividen nuevamente en sub-intervalos de tamaño difuso. El intervalo que tiene menor frecuencia de datos se divide solamente entre la unidad, así, sucesivamente se incrementa el valor en que se dividen los intervalos, hasta llegar al intervalo con mayor frecuencia de datos, se toma en cuenta que si dos intervalos tienen la misma frecuencia de datos ambos se dividen entre el mismo valor.

Para este trabajo se cuenta con siete datos distintos de frecuencias, es decir, empezamos el intervalo u_5 con menor frecuencia de datos solo es dividido entre la unidad (es decir, el intervalo queda igual), el intervalo que continua con menor frecuencia de datos

Tabla 1. Total de remuneraciones por la fabricación del calzado [17].

Mes/año	Moneda Nacional	Mes/año	Moneda Nacional
1/2008	256,330	*1/2010*	246,000
2/2008	260,250	*2/2010*	245,540
3/2008	267,510	*3/2010*	281,110
4/2008	265,020	*4/2010*	254,300
5/2008	265,190	*5/2010*	264,800
6/2008	247,560	*6/2010*	268,450
7/2008	273,690	*7/2010*	286,980
8/2008	258,710	*8/2010*	280,900
9/2008	255,750	*9/2010*	283,950
10/2008	282,930	*10/2010*	289,050
11/2008	261,890	*11/2010*	270,490
12/2008	324,880	*12/2010*	356,000
1/2009	226,700	*1/2011*	259,850
2/2009	225,360	*2/2011*	254,910
3/2009	241,450	*3/2011*	280,850
4/2009	236,900	*4/2011*	268,360
5/2009	233,780	*5/2011*	270,200
6/2009	230,200	*6/2011*	277,750
7/2009	258,900	*7/2011*	280,130
8/2009	243,320		
9/2009	250,650		
10/2009	273,390		
11/2009	252,600		
12/2009	324,930		

Tabla 2. Distribución de los datos históricos.

Intervalos	u_1	u_2	u_3	u_4	u_5	u_6	u_7
Frecuencia	5	14	13	8	0	2	1

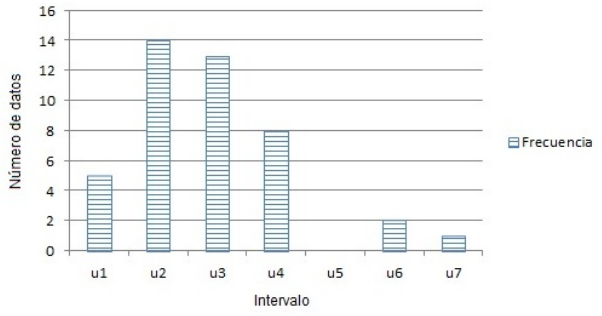


Figura 1. Histograma de la distribución de los datos históricos.

Tabla 3. Intervalos difusos.

$u_l = [valor_{l1}, valor_{l2}]$	
$u_{1,1} = [220000, 225000]$	$u_{3,4} = [270000, 273333.33]$
$u_{1,2} = [225000, 230000]$	$u_{3,5} = [273333.33, 276666.66]$
$u_{1,3} = [230000, 235000]$	$u_{3,6} = [276666.66, 280000]$
$u_{1,4} = [235000, 240000]$	$u_{4,1} = [280000, 284000]$
$u_{2,1} = [240000, 242857.14]$	$u_{4,2} = [284000, 288000]$
$u_{2,2} = [242857.14, 245714.28]$	$u_{4,3} = [288000, 292000]$
$u_{2,3} = [245714.28, 248571.42]$	$u_{4,4} = [292000, 296000]$
$u_{2,4} = [248571.42, 251428.57]$	$u_{4,5} = [296000, 300000]$
$u_{2,5} = [251428.57, 254285.71]$	$u_{5,1} = [300000, 320000]$
$u_{2,6} = [254285.71, 257142.85]$	$u_{6,1} = [320000, 326666.66]$
$u_{2,7} = [257142.85, 260000]$	$u_{6,2} = [326666.66, 333333.33]$
$u_{3,1} = [260000, 263333.33]$	$u_{6,3} = [333333.33, 340000]$
$u_{3,2} = [263333.33, 266666.66]$	$u_{7,1} = [340000, 350000]$
$u_{3,3} = [266666.66, 270000]$	$u_{7,2} = [350000, 360000]$

se divide entre dos sub-intervalos iguales, así sucesivamente, hasta llegar al intervalo u_2 con mayor frecuencia de datos es dividido en siete sub-intervalos iguales.

NOTA: Si algún intervalo no tiene ningún dato histórico, dicho intervalo puede ser eliminado ya que no tiene uso en el resto de la investigación.

PASO 3:

Se define cada conjunto difuso A_i basándose en el número total de sub-intervalos obtenidos en el paso anterior (28 sub-intervalos en total), se denota cada conjunto difuso A_i con un valor lingüístico, como se muestra a continuación:

$A_1 =$ Muy^{10} bajo, $A_2 =$ Muy^9 bajo, $A_3 =$ Muy^8 bajo, $A_4 =$ Muy^7 bajo, $A_5 =$ Muy^6 bajo, $A_6 =$ Muy^5 bajo, $A_7 =$ Muy^4 bajo, $A_8 =$ Muy^3 bajo, $A_9 =$ Muy^2 bajo, $A_{10} =$ Muy bajo, $A_{11} =$ Bajo, $A_{12} =$ Por debajo de la media, $A_{13} =$ Poco abajo de la media, $A_{14} =$ Media, $A_{15} =$ Poco arriba de la media, $A_{16} =$ Por arriba de la media, $A_{17} =$ Alto, $A_{18} =$ Muy^1 alto, $A_{19} =$ Muy^2 alto, $A_{20} =$ Muy^3 alto, $A_{21} =$ Muy^4 alto, $A_{22} =$ Muy^5 alto, $A_{23} =$ Muy^6 alto, $A_{24} =$ Muy^7 alto, $A_{25} =$ Muy^8 alto, $A_{26} =$ Muy^9 alto, $A_{27} =$ Muy^{10} alto y $A_{28} =$ Muy^{11} alto.

Los conjuntos difusos A, tienen un grado de pertenencia para cada intervalo de longitud u_i. Por simplicidad el grado de pertenencia de los conjuntos difusos son 0, 0.5 y 1. El objetivo de representar los datos históricos (conjuntos clásicos), en datos difusos (conjuntos difusos), es obtener una serie de tiempo difusa. Para evitar acumular demasiados datos, si el grado de pertenencia es $(0/u_i)$ se omitirá escribirlo.

$A_1 = 1/u_{1,1} + 0{,}5/u_{1,2}$
$A_2 = 0{,}5/u_{1,1} + 1/u_{1,2} + 0{,}5/u_{1,3}$
$A_3 = 0{,}5/u_{1,2} + 1/u_{1,3} + 0{,}5/u_{1,4}$
$A_4 = 0{,}5/u_{1,3} + 1/u_{1,4} + 0{,}5/u_{2,1}$
$A_5 = 0{,}5/u_{1,4} + 1/u_{2,1} + 0{,}5/u_{2,2}$
$A_6 = 0{,}5/u_{2,1} + 1/u_{2,2} + 0{,}5/u_{2,3}$

$A_7 = 0{,}5/u_{2,2} + 1/u_{2,3} + 0{,}5/u_{2,4}$
$A_8 = 0{,}5/u_{2,3} + 1/u_{2,4} + 0{,}5/u_{2,5}$
$A_9 = 0{,}5/u_{2,4} + 1/u_{2,5} + 0{,}5/u_{2,6}$
$A_{10} = 0{,}5/u_{2,5} + 1/u_{2,6} + 0{,}5/u_{2,7}$
$A_{11} = 0{,}5/u_{2,6} + 1/u_{2,7} + 0{,}5/u_{3,1}$
$A_{12} = 0{,}5/u_{2,7} + 1/u_{3,1} + 0{,}5/u_{3,2}$
$A_{13} = 0{,}5/u_{3,1} + 1/u_{3,2} + 0{,}5/u_{3,3}$
$A_{14} = 0{,}5/u_{3,2} + 1/u_{3,3} + 0{,}5/u_{3,4}$
$A_{15} = 0{,}5/u_{3,3} + 1/u_{3,4} + 0{,}5/u_{3,5}$
$A_{16} = 0{,}5/u_{3,3} + 1/u_{3,4} + 0{,}5/u_{3,5}$
$A_{17} = 0{,}5/u_{3,4} + 1/u_{3,5} + 0{,}5/u_{3,6}$
$A_{18} = 0{,}5/u_{3,5} + 1/u_{3,6} + 0{,}5/u_{4,1}$
$A_{18} = 0{,}5/u_{3,6} + 1/u_{4,1} + 0{,}5/u_{4,2}$
$A_{19} = 0{,}5/u_{4,1} + 1/u_{4,2} + 0{,}5/u_{4,3}$
$A_{20} = 0{,}5/u_{4,2} + 1/u_{4,3} + 0{,}5/u_{4,4}$
$A_{21} = 0{,}5/u_{4,3} + 1/u_{4,4} + 0{,}5/u_{4,5}$
$A_{22} = 0{,}5/u_{4,4} + 1/u_{4,5} + 0{,}5/u_{5,1}$
$A_{23} = 0{,}5/u_{4,5} + 1/u_{5,1} + 0{,}5/u_{6,1}$
$A_{24} = 0{,}5/u_{5,1} + 1/u_{6,1} + 0{,}5/u_{6,2}$
$A_{25} = 0{,}5/u_{6,1} + 1/u_{6,2} + 0{,}5/u_{6,3}$
$A_{26} = 0{,}5/u_{6,2} + 1/u_{6,3} + 0{,}5/u_{7,1}$
$A_{27} = 0{,}5/u_{6,3} + 1/u_{7,1} + 0{,}5/u_{7,2}$
$A_{28} = 0{,}5/u_{7,1} + 1/u_{7,2}$

PASO 4:

Establecer la relación difusa formada por los datos históricos, de la siguiente manera:
$A_i \rightarrow A_j$
$A_j \rightarrow A_k$
$A_k \rightarrow ...$
$... \rightarrow A_m$

Donde A_i es el dato difuso el periodo $n-1$ y A_j el dato difuso del periodo n, por lo tanto, la relación difusa es $A_i \rightarrow A_j$. El significado de esta relación lógica difusa este definida como: **SI** los datos difusos del periodo $n-1$ es A_i **ENTONCES** para el periodo n sera A_j.

Se basa en los datos históricos de la tabla 1, para obtener la relación lógica difusa es la siguiente:

PASO 5:

Para averiguar si la tendencia es creciente, decreciente o casi lo mismo, se dividen los intervalos difusos (obtenidos en el paso 2), en cuatro sub-intervalos de igual longitud, donde el primer cuartil (o el punto 0.25 del intervalo) es tomado como el valor de

Tabla 4. Relación lógica difusa.

$r : A_i \rightarrow A_j$			
$r_1 : A_{10} \rightarrow A_{12}$	$r_{12} : A_{24} \rightarrow A_2$	$r_{23} : A_9 \rightarrow A_{24}$	$r_{34} : A_{20} \rightarrow A_{15}$
$r_2 : A_{12} \rightarrow A_{14}$	$r_{13} : A_2 \rightarrow A_2$	$r_{24} : A_{24} \rightarrow A_7$	$r_{35} : A_{15} \rightarrow A_{28}$
$r_3 : A_{14} \rightarrow A_{13}$	$r_{14} : A_2 \rightarrow A_5$	$r_{25} : A_7 \rightarrow A_6$	$r_{36} : A_{28} \rightarrow A_{11}$
$r_4 : A_{13} \rightarrow A_{13}$	$r_{15} : A_5 \rightarrow A_4$	$r_{26} : A_6 \rightarrow A_{18}$	$r_{37} : A_{11} \rightarrow A_{10}$
$r_5 : A_{13} \rightarrow A_7$	$r_{16} : A_4 \rightarrow A_3$	$r_{27} : A_{18} \rightarrow A_{10}$	$r_{38} : A_{10} \rightarrow A_{18}$
$r_6 : A_7 \rightarrow A_{16}$	$r_{17} : A_3 \rightarrow A_3$	$r_{28} : A_{10} \rightarrow A_{13}$	$r_{39} : A_{18} \rightarrow A_{14}$
$r_7 : A_{16} \rightarrow A_{11}$	$r_{18} : A_3 \rightarrow A_{11}$	$r_{29} : A_{13} \rightarrow A_{14}$	$r_{40} : A_{14} \rightarrow A_{15}$
$r_8 : A_{11} \rightarrow A_{10}$	$r_{19} : A_{11} \rightarrow A_6$	$r_{30} : A_{14} \rightarrow A_{19}$	$r_{41} : A_{15} \rightarrow A_{17}$
$r_9 : A_{10} \rightarrow A_{18}$	$r_{20} : A_6 \rightarrow A_8$	$r_{31} : A_{19} \rightarrow A_{18}$	$r_{42} : A_{17} \rightarrow A_{18}$
$r_{10} : A_{18} \rightarrow A_{12}$	$r_{21} : A_8 \rightarrow A_{16}$	$r_{32} : A_{18} \rightarrow A_{18}$	
$r_{11} : A_{12} \rightarrow A_{24}$	$r_{22} : A_{16} \rightarrow A_9$	$r_{33} : A_{18} \rightarrow A_{20}$	

tendencia a disminuir, y el tercer cuartil (o el punto 0.75 del intervalo) es tomado como el valor de tendencia al alza. Para predecir el valor del próximo periodo se utiliza el siguiente procedimiento:

Este procedimiento se basa en la diferencia de segundo orden. Donde la diferencia de segundo orden está dada por dS como a continuación se muestra:

$$dS = (S_{n-1} - S_{n-2}) - (S_{n-2} - S_{n-3}) \tag{5}$$

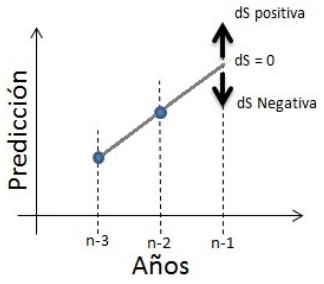


Figura 2. Si la proyección de una línea recta con los dos primeros datos ($n-3$ y $n-2$) es igual al valor de $n-1$, se dice que la tendencia tiene un crecimiento constante ($dS = 0$), mientras que si es mayor (dS positivo), se dice que la tendencia es a la alza y de lo contrario, si es menor (dS negativo) la tendencia es a la baja.

Para predecir el valor futuro se aplicaran las siguientes 4 reglas:

Regla 1: Dado que se tiene una diferencia de segundo orden dS (ver ecuación 5), el cual es imposible calcular para los dos primeros datos históricos (ver Tabla 1), en el tercer dato, donde no está disponible el dato del periodo $n-3$, pero si se encuentran disponibles los datos de los periodos $n-2$ y $n-1$, se utilizará:

Si $\mid S_{n-1} - S_{n-2} \mid >$ que la mitad de la longitud del intervalo correspondiente al conjunto difuso A_j, con valor de pertenencia igual a 1, entonces la tendencia del pronóstico es a la baja y el valor esperado sera el punto ubicado en 1/4 (0.25) del intervalo.

Si $\mid S_{n-1} - S_{n-2} \mid <$ que la mitad de la longitud del intervalo correspondiente al conjunto difuso A_j, con valor de pertenencia igual a 1, entonces la tendencia del pronóstico es a la alza y sera el punto ubicado en 3/4 (0.75) del intervalo.

Si $\mid S_{n-1} - S_{n-2} \mid =$ que la mitad de la longitud del intervalo correspondiente al conjunto difuso A_j, con valor de pertenencia igual a 1, entonces el valor esperado sera el punto medio (0.5) del intervalo.

Regla 2: Si la $S_{n-1} + (\mid (S_{n-1} - S_{n-2}) - (S_{n-2} - S_{n-3}) \mid \times 2)$ o $S_{n-1} - (\mid (S_{n-1} - S_{n-2}) - (S_{n-2} - S_{n-3}) \mid \times 2)$ cae dentro del intervalo correspondiente al conjunto difuso A_j, con valor de pertenencia igual a 1, entonces la tendencia es a la alza y el valor esperado sera el punto ubicado en 3/4 (0.75) del intervalo.

Regla 3: Si no se cumple la regla 2, se continua con la siguiente condición. Si la $S_{n-1} + (\mid (S_{n-1} - S_{n-2}) - (S_{n-2} - S_{n-3}) \mid \div 2)$ o $S_{n-1} - (\mid (S_{n-1} - S_{n-2}) - (S_{n-2} - S_{n-3}) \mid \div 2)$ cae dentro del intervalo correspondiente al conjunto difuso A_j, con valor de pertenencia igual a 1, entonces la tendencia es a la baja y el valor esperado sera el punto ubicado en 1/4 (0.25) del intervalo.

Regla 4: Si los datos no siguen la regla 2 o la regla 3, el pronóstico tomara el valor del punto medio (0.5) del intervalo correspondiente al conjunto difuso A_j, con valor de pertenencia igual a 1.

5. Resultados

Al usar la metodología mencionada anteriormente, la tendencia de la Serie de Tiempo se predice como se muestra en la tabla 5. El gráfico muestra la compración entre los datos históricos, regresión lineal (dada por la ecuación 6) y la predicción de la serie de tiempo difusa.

$$y_i = 626.22x_i + 252,212 \tag{6}$$

Para los periodos x_j y los pronósticos resultantes y_j. Donde $j = 1, 2, ...$
El error promedio se define como:

$$Error\ promedio = \frac{\left[\sum_{i=1}^{m} (\frac{\lfloor valor\ pronosticado_i - valor\ histrico_i \rfloor}{valor\ histrico_i}) \times 100\right]}{m} \tag{7}$$

El error cuadrático medio (ECM) para comparar los resultados, el ECM es calculado de la siguiente manera:

Tabla 5. Resultados de la predicción con Series de Tiempo Difusa comparados con los datos históricos y la regresión lineal.

Mes/Año	Remuneración histórica	Tendencia	Predicción	Regresión lineal
1/2008	256,330			
2/2008	260,250			
3/2008	267,510	DOWN	267, 500	254, 090.66
4/2008	265,020	DOWN	264166.66	254, 726.88
5/2008	265,190	MIDDLE	265, 000	255, 343.1
6/2008	247,560	MIDDLE	247142.85	255, 969.32
7/2008	273,690	MIDDLE	275, 000	256, 595.54
8/2008	258,710	MIDDLE	258, 571.42	257, 221.76
9/2008	255,750	MIDDLE	255, 714.28	257, 847.98
10/2008	282,930	MIDDLE	282, 000	258, 474.2
11/2008	261,890	MIDDLE	261, 666.66	259, 100.42
12/2008	324,880	MIDDLE	323, 333.33	259, 726.64
1/2009	226,700	MIDDLE	227, 500	260, 352.86
2/2009	225,360	MIDDLE	227, 500	260, 979.08
3/2009	241,450	MIDDLE	241, 428.57	261, 605.3
4/2009	236,900	MIDDLE	237, 500	262, 231.52
5/2009	233,780	MIDDLE	232, 500	262, 857,74
6/2009	230,200	UP	233, 750	263, 483.96
7/2009	258,900	MIDDLE	258, 571.42	264, 110.18
8/2009	243,320	MIDDLE	244, 285.71	264, 736.4
9/2009	250,650	MIDDLE	250, 000	265, 362.62
10/2009	273,390	MIDDLE	275, 000	265, 988.84
11/2009	252,600	MIDDLE	252, 857.14	266, 615.06
12/2009	324,930	MIDDLE	323, 333.33	267, 241.28
1/2010	246,000	MIDDLE	247, 142.85	267, 867.5
2/2010	245,540	MIDDLE	244, 285.71	268, 493.72
3/2010	281,110	MIDDLE	282, 000	269, 119.94
4/2010	254,300	MIDDLE	255, 714.28	269, 746.16
5/2010	264,800	MIDDLE	265, 000	270, 372.38
6/2010	268,450	MIDDLE	268, 333.33	270, 998.6
7/2010	286,980	MIDDLE	286, 000	271, 624.82
8/2010	280,900	MIDDLE	282, 000	272, 251.04
9/2010	283,950	DOWN	282, 000	272, 877.26
10/2010	289,050	MIDDLE	289, 000	273, 503.48
11/2010	270,490	MIDDLE	271, 666.66	274, 129.7
12/2010	356,000	MIDDLE	355, 000	274, 755.92
1/2011	259,850	MIDDLE	258, 571.42	275, 382.14
2/2011	254,910	MIDDLE	255, 714.28	276, 008.36
3/2011	280,850	MIDDLE	282, 000	276, 634.58
4/2011	268,360	MIDDLE	268, 333.33	277, 260.8
5/2011	270,200	MIDDLE	271, 666.66	277, 887.02
6/2011	277,750	DOWN	277, 500	278, 513.24
7/2011	280,130	DOWN	281, 000	279, 139.46
Error promedio (ecuación 7)			0.33	6.55
Error cuadrático medio (ecuación 8)			1,138.39	24, 616.69

$$ECM = \sqrt{\frac{\sum_{i=1}^{m}(valor\ histrico_i - valor\ pronosticado_i)^2}{m}} \tag{8}$$

Donde m es el número de datos pronosticados.

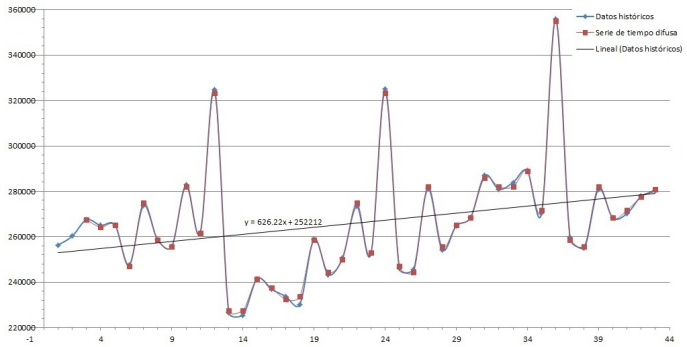


Figura 3. Predicción con Series de Tiempo Difusas.

6. Conclusiones

El uso de técnicas difusas, proporciona mejoras para la predicción de series de tiempo, ya que se puede trabajar el lenguaje natural (o lingüístico), por medio de conjuntos difusos.

Se puede observar que la Serie de Tiempo Difusa, tiene mejores resultados para el pronóstico de la remuneración por la fabricación del calzado, que la Regresión Lineal, ya que ambos errores son inferiores. También se puede observar en la gráfica, la Serie de Tiempo Difusa tiene una mayor aproximación a los datos históricos que la Regresión Lineal.

Por lo tanto, se concluye que la técnica de Series de Tiempo Difusas es capaz de predecir la remuneración por la fabricación del calzado, para los próximos periodo, así como identificar la tendencia.

Agradecimientos Al Consejo Nacional de Ciencia y Tecnología (CONACYT).

Referencias

1. Klir, G.J., Yuan, B.: Fuzzy Sets and Fuzzy Logic: Theory and Applications. Prentice Hall, New Jersey (1995).
2. Kaufmann, A.: Introducción a la teoría de los subconjuntos borrosos. Para el uso de ingenieros. Continental, México (1982).
3. Zadeh, L.A.: Fuzzy Sets. Information and Control 8, 338–353 (1965).

4. Song, Q.: Fuzzy Time Series and Its Models. Fuzzy Sets ans System 54, 269–277 (1993).
5. Song, Q., Chissom, B.S.: Forecasting enrollments with fuzzy time series - Part I. Fuzzy Sets and System 54, 1–9 (1993).
6. Song, Q., Chissom, B.S.: Forecasting enrollments with fuzzy time series - Part II. Fuzzy Sets and System 62, 1–8 (1994).
7. Chen, S.M., Hsu, C.C.: A New Method to Forecast Enrollments Using Fuzzy Time Series. Inter. Journal of Applied Science and Engineering 2004. 2, 3: 234–244 (2004).
8. Hwang, J.R., Chen, S.M., Lee, C.H.: Handling forecasting problems using fuzzy time series. Elsevier 100, 217–228 (1998).
9. Huarng, K.: Heuristic models of fuzzy time series for forecasting. Fuzzy Sets and Systems 123, 369–386 (2001).
10. Chen, S.M., Chung, N.Y.: Forecasting enrollments of students by using fuzzy time series and genetic algorithms. International journal of information and management sciences 17, 1–17 (2006).
11. Jilani, T.A., Burney, S.M.A., Ardil, C.: Fuzzy metric approach for fuzzy time series forecasting based on frequency density based partitioning. Proceedings of World Academy of Science, Engineering and Technology 23, 333–338 (2007).
12. Pathak, H.K., Singh, P.: A New Bandwidth Interval Based Forecasting Method for Enrollment Using Fuzzy Time Series. Applied Mathematics 2011. 2, 504–507 (2011).
13. Jasim, H.T., Salim, A.G.J., Ibraheem, K.I.: A Novel Algorithm to Forecast Enrollment Based on Fuzzy Time Series. Applications and Applied Matematics: An international Journal (AAM) 7, 385–395 (2012).
14. Chen, S.M., Hwang, J.R.: Temperature prediction using fuzzy time series. IEEE Trans. on Systems, Man, and Cybernetics-Part B: Cybernetics 30, 263–275 (2000).
15. Shah, M.: Fuzzy Time Series: A Realistic Method to Forecast Gross Domestic Capital of India. Analysis and Design of Intelligent Systems using Soft Computing Techniques, 255–264 (2007).
16. Ecerkale, K., Küçükdeniz, T., Esnaf, Ş.: Comparison of fuzzy time series based on difference parameters and two-factor time-variant fuzzy time series models for aviation fuel production forecasting. Journal of aeronautics and space technologies 4, 57–63 (2010).
17. PROSPECTA, Centro De Innovación y Competitividad, `http://www.prospecta.org.mx`

Consideraciones para el Control de congestión en redes inalámbricas de sensores utilizando la optimización crosslayer

Raymundo Buenrostro-Mariscal, Juan Iván Nieto-Hipólito, María Cosio-León, Mabel Vazquez-Briseno, and Juan de Dios Sánchez-López

Facultad de Ingeniería, Arquitectura y Diseño, UABC
Carr. Tijuana-Ensenada Km. 103, 22860 Ensenada, BC., México
{rbuenrostro,jnieto,maria.cosio,mabel.vazquez,jddios}@uabc.edu.mx
http://telematica.uabc.mx
Paper received on 03/10/12, Accepted on 24/10/12.

Resumen La utilización de las redes inalámbricas de sensores (WSN) ha evolucionado rápidamente. Hoy se proponen para la recolección de información de naturaleza crítica. Estas aplicaciones tienen requerimientos muy específicos en la transmisión de datos, como la tasa de paquetes perdidos y el retardo. La congestión en la red es un problema que compromete el logro de los requerimientos; por ello, requiere una atención especial. Existen mecanismos para controlar la congestión, como los protocolos de transporte, que incluyen funciones para detectarla, notificarla y resolverla. En el estado del arte para el control de congestión en WSN, las propuestas son diseñadas para aplicaciones con requerimientos más simples y de forma independiente de otras capas involucradas en la comunicación (Red y Enlace de Datos). En este trabajo se presentan las consideraciones y requerimientos para un protocolo de Control de Congestión crosslayer consciente de las aplicaciones, la subcapa MAC y la capa Física.

Keywords: Congestión, Crosslayer, QoS, WSN

1. Introducción

Las Redes Inalámbricas de Sensores (WSN, del inglés Wireless Sensor Networks) son redes compuestas por dispositivos interconectados entre sí, que trabajan de forma cooperativa para recolectar datos de interés y transmitirlos de forma inalámbrica a otro punto. En los últimos años, se ha incrementado los trabajos que exploran su uso para aplicaciones críticas, como la supervisión de la salud de pacientes de forma remota [1], [2]. Estas aplicaciones son consideradas como heterogéneas, ya que generan distintos tipos y volúmenes de tráfico; que a su vez tienen diferentes requerimientos de transmisión en términos de velocidad de datos, confiabilidad y prioridad en la entrega. Las WSN tienen limitaciones importantes para cumplir con estos requerimientos [3], debido a su naturaleza inalámbrica y a su medio compartido, donde todos los nodos contienden por el acceso al medio. Además, la mayor parte del tiempo la red mantiene flujos de datos desde cada nodo sensor hacia el nodo final (Sink), tráfico muchos-a-uno

Mireya García-Vázquez, Grigori Sidorov (Eds.).
Advances in Computing Science and Control.
Research in Computing Science 59, 2012, pp. 219- 230

(*many-to-one*); provocando que lo nodos intermedios manejen una mayor cantidad de tráfico y se saturen rápidamente. Este problema se conoce como *efecto embudo o funneling* [4]. Aunado a lo anterior, los dispositivos que integran la red tienen capacidades limitadas de hardware, especialmente de procesamiento, capacidad de almacenamiento y suministro de energía; que incrementan los retos de la transmisión. Bajo este contexto, la *congestión* es un problema que puede presentarse frecuentemente tanto en los nodos como en el propio enlace inalámbrico. La congestión tiene un impacto directo en el desempeño global de la red y en la calidad de servicio (QoS, del inglés Quality of Service) entregada a las aplicaciones; ya que incrementa la cantidad de paquetes perdidos, retrasa la entrega de datos y aumenta el gasto de energía [5]. Por lo tanto, solucionar la congestión en la redes WSN es un reto obligado para cumplir con el objetivo primario de la recolección remota de datos. Las soluciones deben incluir tres funciones principales: *detección de congestión, notificación y resolución*. Las propuestas existentes para las WSN implementan las funciones de forma distinta y con objetivos diferentes, pero en su mayoría están diseñadas para aplicaciones homogéneas con requerimientos más sencillos y trabajando de forma independiente de otros mecanismos claves en la trasmisión de datos. De lo anterior, crear un protocolo de control de congestión para aplicaciones críticas y heterogéneas es necesario. Que además, optimice el uso de los recursos y alcance el mayor desempeño. El diseño *crosslayer* se propone como una de las mejores alternativas para este escenario, ya que obtiene mejores rendimientos que las soluciones tradicionales que trabajan de forma independiente [6]. El diseño crosslayer propone aprovechar las funciones de los protocolo de diferentes capas, por ejemplo transporte, red y subcapa MAC (del inglés Medium Access Control) para trabajar en conjunto. En las siguientes secciones se presentan las principales consideraciones para el Control de Congestión en WSN, para guiar el diseño de nuevos protocolos, lo cual es la principal contribución de éste trabajo.

2. Clasificación de las principales líneas de diseño para el control de congestión

A partir de los trabajos consultados se realiza una propuesta de clasificación de las soluciones existentes para el control de congestión; específicamente los mecanismos utilizados para la detección, notificación y resolución de congestión. El cuadro 1 resume esta clasificación, resalta la preferencia por la notificación implícita y el ajuste de la tasa de transmisión en los nodos, como mecanismos para mitigar la congestión. De igual forma se aprecia que sólo tres propuestas utilizan el diseño crosslayer. A continuación se presenta el trabajo relacionado, divido por las funciones propuestas del control de congestión.

2.1. Detección de congestión

STCP [7] utiliza el grado de utilización del búfer en cada nodo intermedio involucrado en una transmisión de datos, para detectar congestión de forma local; mientras que SenTCP [8] utiliza la relación del tiempo promedio de arribo de paquetes contra

Cuadro 1. Clasificación de los protocolos analizados

Propuesta	Detección	Notificación	Resolución	Diseño
STCP	Nivel de búfer	Implícita	Ajusta velocidad Re-dirige tráfico	Tradicional
SenTCP	Nivel de búfer T. servicio p. T. inter-arribo p.	Implícita	Ajusta velocidad	Tradicional
TRCCIT	Tasa de datos	Implícita	Ajusta velocidad Re-dirige tráfico	Tradicional
CODA	Carga del canal Nivel de búfer	Explícita	Descarta paquetes Ajusta velocidad	Tradicional
CTCP	Tasa de error T. Nivel de búfer	Explícita	Ajusta velocidad	Tradicional
RT^2	Retardo del nodo Nivel de búfer	Implícita	Ajusta velocidad	Crosslayer
LCART	Carga de canal T. servicio p. T. inter-arribo p. Nivel de búfer	Implícita	Ajusta velocidad	Crosslayer
PCCP	T. inter-arrib p. T. servicio p.	Implícita	Ajusta velocidad	Crosslayer
CRRT	Nivel de búfer Tasa de datos	Implícita	Ajusta velocidad	Tradicional

el tiempo promedio de servicio de paquetes en combinación con la cantidad empleada del búfer para detectar el grado de congestión local en cada nodo. De forma similar TRCCIT [9] utiliza la tasa de datos, donde compara la tasa de arribos de paquetes contra la tasa de envío de paquetes, para declarar congestión en el nodo. TRCCIT asume un control de congestión pro-activo, al monitorear la carga local de la red y actuar en consecuencia. En CODA [10], además del nivel de ocupación del búfer, se utiliza una combinación de las condiciones de carga actual y pasada del canal inalámbrico para inferir una detección más precisa de congestión en cada receptor. CTCP [11] determina la presencia de congestión cuando se rebasa un umbral de búfer y una tasa de error de bit o pérdida de paquetes, mediante el mecanismo de reconocimiento de paquetes (Ack, del inglés Acknowledgment) en modo *salto-por-salto* (HbH, del inglés Hop-by-Hop); CTCP debido a éstos mecanismos, puede distinguir si la causa de congestión es por desbordamiento de búfer o por errores de transmisión. RT^2 [12] define un umbral de retardo promedio del nodo, con el cual, el nodo tiene una idea acerca del nivel de contención o saturación del medio en su entorno; y lo combina con un nivel de búfer para detectar de forma precisa la congestión. PCCP [13] utiliza el tiempo promedio de inter-arribo de paquetes t_a^i y el tiempo promedio de servicio de paquete t_s^i en la subcapa MAC para determinar el *grado de congestión* $d(i)$ actual en el nodo o del enlace; este nuevo índice de congestión es definido como la proporción del tiempo de servicio sobre

el tiempo de arribo de paquetes $d(i) = t_s^i/t_a^i$; entonces cuando $d(i) > 1$ asume que el nodo experimenta congestión e informa a los nodos vecinos. PCCP nombra a ésta forma de detección como *Detección de Congestión Inteligente*. LCART [14] resuelve la detección de congestión de la red mediante el uso simultáneo de varios eventos como el t_s, t_a, nivel de uso del búfer y un umbral límite de carga del canal.

2.2. Notificación de congestión

La mayoría de las propuestas de notificación de congestión utiliza el concepto de piggyback para enviar de forma implícita el aviso de congestión a los nodos de la red. STCP [7] incorpora esta forma de notificación y lo realiza dentro del paquete de datos y en el paquete de reconocimiento Ack, fijando un bit en el campo de *notificación de congestión* (CN, del inglés Congestion Notification) del encabezado del paquete. En STCP cada nodo de la red tiene la posibilidad de detectar la congestión y generar la notificación hacia el Sink para informarle del estado de la red, a su vez el Sink genera el aviso de congestión utilizando el paquete Ack hacia los nodos fuente. TRCCIT utiliza la notificación implícita HbH, pero a diferencia de STCP, éste avisa de la congestión local de forma inmediata a los nodos vecinos, mediante la activación de un bit en cada paquete que sale del nodo. PCCP utiliza el mismo tipo de notificación implícita que STCP y TRCCIT pero no sólo cambia el estado de un bit, si no que agrega información útil al mensaje de notificación: nivel de prioridad global, t_s^i, t_a^i y cantidad de nodos hijos; para que los nodos que reciben el mensaje realicen acciones diferenciadas en base a la información. Además, PCCP tiene dos eventos que pueden provocar el envío del mensaje de notificación, la primera ocurre cuando el número de paquetes reenviados por el nodo rebasa un umbral predefinido y la segunda cuando el nodo recibe el mensaje de notificación de congestión enviado por otros nodos. CODA utiliza notificación explícita para difundir el aviso de congestión hacia los nodos fuente en un modo de comunicación HbH; el nodo al recibir este mensaje reduce la tasa de envío y retransmite el aviso de congestión a los nodos vecinos, esta forma es conocida como *Backpressure*. CTCP utiliza la notificación explícita mediante la cual, el nodo que detecta la congestión genera un mensaje a todos los nodos de su vecindario, para indicarles que no puede recibir más paquetes. De la misma forma, cuando se resuelve la congestión, los nodos son informados para re-establecer las tasas de transmisión.

2.3. Resolución de congestión

CRRT [15] implementa un mecanismo centralizado en el Sink para eliminar la congestión de la red en modo *extremo-a-extremo* (E2E, del inglés End-to-End); esta forma contribuye a una decisión imparcial ya que el Sink tiene una visión global de la red y puede controlar la velocidad de todos los nodos en la red. El mecanismo base para resolver la congestión en CRRT es el esquema ”Incremento Aditivo y Decremento Aditivo” (AIAD, del inglés Additive Increase Additive Decrease), el cual ajusta gradualmente la tasa de transmisión en lo nodos fuente evitando una reducción agresiva. ART [16] utiliza un esquema distribuido para manejar la congestión, en los nodos marcados como

esenciales *E-Nodes*, con el cual ajustan la velocidad de transmisión en la red; sin embargo, sólo aplican este ajuste a un conjunto de nodos etiquetados como *no esenciales* para detener temporalmente su tráfico. Los nodos no esenciales podrán reactivar el envió de datos hasta que reciban de la red el aviso de operación normal. La clasificación de los nodos es realizada por un algoritmo que se basa en la energía residual de cada nodo en cada ejecución. En STCP cuando un nodo recibe la notificación de congestión, éste puede en-rutar los paquetes sucesivos del flujo de datos por una ruta diferente, siempre y cuando se tenga un algoritmo en la capa de red que permita este proceso, o disminuir la velocidad de transmisión de los nodos fuente. CODA utiliza dos mecanismos para resolver la congestión: un esquema de lazo abierto HbH y un esquema de lazo cerrado E2E. En el primer esquema los nodos utilizan el método Backpressure para propagar el aviso e indicarles a los nodos vecinos reducir la tasa de envío, esquema conocido como auto-regulación de nodo fuente. En el esquema de lazo cerrado, el que controla la tasa de transmisión de los nodos fuente es el Sink, que emite un mensaje de regulación hacia todos los nodos cuando existe congestión persistente, esquema llamado regulación múlti-fuente. CODA decide que esquema utilizar de acuerdo al grado de congestión de la red, si la tasa de eventos de los nodos fuente está por debajo de cierta fracción de la capacidad teórica del canal utiliza auto-regulación de nodo, de lo contrario se utiliza regulación desde el Sink. PCCP implementa un control de congestión HbH de acuerdo al *grado de congestión* $d(i)$ *y a los índices de prioridad del nodo*; con lo cual garantiza que el nodo que tiene un mayor índice de prioridad obtenga mas ancho de banda de la red. Cada nodo cuenta con dos "colas", una para el tráfico propio y otra para el tráfico en tránsito que recibe; entonces, de acuerdo a la prioridad definida para cada tráfico el nodo ajusta y programa el envió de los datos. Bajo esta base, PCCP ofrece una resolución de congestión flexible, diferenciada y distribuida para ofrecer *equidad ponderada* a los nodos de la red. TRCCIT resuelve la congestión de la red de forma pro-activa mediante la selección de varias rutas para redirigir el tráfico de los nodos fuente. La definición de congestión pro-activa se refiere a la capacidad de seleccionar sobre la marcha una ruta múltiple o individual para controlar la congestión. TRCCIT evita la congestión en la red, esto lo fundamenta al resolver rápidamente la *congestión transitoria* en los nodos. Entonces, en la medida que un nodo se adapta o resuelve la congestión transitoria evita el incremento y una condición prolongada de congestión, en una trayectoria dada.

3. Control de congestión en WSN

Existen principalmente dos causas que provocan *congestión* en las WSN. La primera se presenta a nivel de nodo, cuando la velocidad de arribo de paquetes excede la tasa de servicio de paquetes. Esto es, el tiempo para procesar los paquetes que se reciben en un nodo es mayor a los tiempos de inter-arribo de paquetes; ocasionando que el búfer del nodo se sature y *tire* los nuevos paquetes que arriban. Esta pérdida se clasifica como un *paquete descartado o packet dropped*. Los paquetes descartados provocan un gran número de retransmisiones en la red, incrementando considerablemente el retraso en la entrega de los paquetes y el consumo de energía [17]. Considerando la dirección del flujo de datos en las WSN, tráfico many-to-one, se espera que este problema se agudicé en los nodos intermedios cercanos al Sink; ya que ellos manejan un mayor tráfico com-

binado de otros nodos [18]. La segunda causa de congestión es responsabilidad de los problemas del enlace inalámbrico, donde altas tasas o ráfagas de datos pueden exceder la capacidad límite del enlace y provocar congestión. Además, problemas relacionados con el método de acceso por contienda, la interferencia de las señales y las colisiones son otra causa a *nivel enlace* (subcapa MAC y capa Física). Los paquetes dañados por estas causas son considerados como *paquetes perdidos o packet loss*.

Los paquetes descartados y perdidos afectan fuertemente la *confiabilidad de la red* (capacidad de la red para transmitir paquetes con éxito), la *utilización del enlace* inalámbrico y el consumo de energía del nodo; ya que independientemente de la causa es necesario retransmitir los paquetes. Bajo este contexto, la congestión impacta negativamente el rendimiento global de la WSN y el cumplimiento de QoS de las aplicaciones, de ahí la importancia de incorporar mecanismos que controlen sus efectos o impidan que se presente.

3.1. Modelo funcional propuesto para el protocolo de control de congestión

Un protocolo de control de congestión bajo nuestra perspectiva debe incluir tres módulos funcionales: *detección de congestión, módulo de notificación y resolución de congestión*. La figura 1 muestra los tres módulos funcionales y propone los mecanismos que pueden ser utilizados para lograr el control de congestión en las WSN. El protocolo de control de congestión debería ser diseñados para trabajar, tanto en modo *salto-por-salto*, para ejecutar los procesos entre nodos vecinos a un salto; como en modo *extremo-a-extremo*, que considera la ejecución de los procesos entre nodos finales (nodo sensor y Sink). Con la capacidad para seleccionar el modo de acuerdo a los procesos que realice; con la intención de optimizar los tiempos de respuesta y el consumo de recursos de la red. Además, es importante que operé en ambas direcciones de la red; privilegiando el flujo en dirección hacia el Sink que tiene mayor presencia en la red.

El modulo de *detección* identifica la presencia de congestión mediante eventos definidos en los nodos de la red, principalmente en los nodos involucrados en una transmisión de datos. Existen mecanismo que pueden utilizar diferentes eventos para alcanzar una detección eficiente; por ejemplo, revisar el nivel de uso de los búfer de cada nodo, monitorear el tiempo de arribo de paquetes contra el tiempo de servicio de paquetes, el número de paquetes retransmitidos, entre otros.

El módulo de *notificación* debe difundir el aviso de congestión, desde el nodo que la detecta, a los nodos de la red o idealmente a los nodos involucrados en la congestión. Esta notificación puede ser *explícita* o *implícita*, en el primer caso se debe crear un paquete de control diseñado exclusivamente para este fin. En la notificación implícita el nodo utiliza el paquete de datos para incluir la información de control necesaria para informar a los nodos de la presencia de congestión; esta forma de empaquetado se conoce como ”Piggybacks”. La información de control que se incluya, en cualquiera de los dos casos, puede ser tan solo un cambio de bit de algún campo de control o información adicional para indicar la velocidad de transmisión permitida, nivel de prioridad o grado de congestión.

El modulo de *resolución de congestión* debe utilizar los medios posibles para eliminar la congestión o disminuirla; el principal mecanismo, y el más utilizado, es ajustar la velocidad de transmisión de paquetes en los nodos de la red a un valor límite, llamado

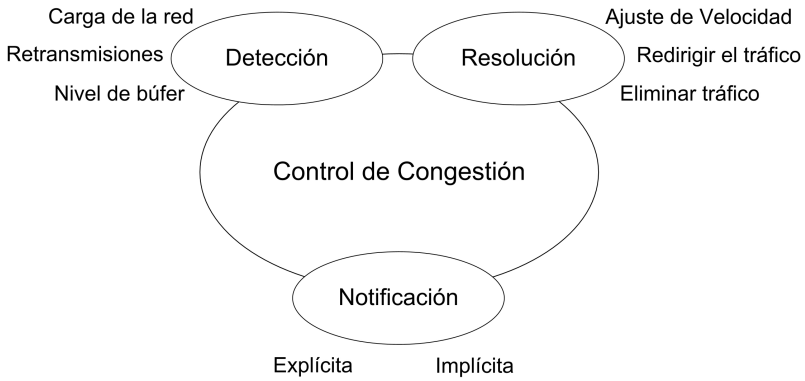


Figura 1. Modelo funcional del control de congestión para WSN

control de flujo de datos. Este mecanismo funciona para los problemas de desbordamiento de búfer o retardo de procesamiento del nodo. Sin embargo, en la pérdida de paquetes debido a los errores en la transmisión, una mejor alternativa es re-dirigir el tráfico a otro canal para reducir la congestión presente; si la topología y los protocolos de comunicación lo permiten.

Por lo general los tres módulos se consideran procesos exclusivos de la capa de transporte; sin embargo, es necesario un diseño crosslayer, con el cual la información y las capacidades de cada capa se comparten para el logro del objetivo del control de congestión.

4. Consideraciones de diseño del modelo funcional propuesto para el control de congestión

En esta sección se exponen las *consideraciones de diseño* claves para un protocolo de control de congestión crosslayer en WSN. Explicando en una primera fase aquellas que impactan en su operación de forma global. Para después introducir las consideraciones relevantes a cada uno de los módulos funcionales de la Figura 1; concluyendo con la heurística que guiara el diseño del protocolo crosslayer de control de congestión en WSN:

1. El *consumo de energía* tiene que estar presente en todas las etapas y procesos del control de congestión, con el fin de extender la vida útil de la red. Por lo tanto, es necesario reducir la pérdida de paquetes y mantener al mínimo la carga de tráfico de paquetes de control.
2. Habilidad para decidir cuándo operar en modo *HbH y/o E2E* para agilizar el tiempo de respuesta de los procesos de detección, notificación y resolución.
3. *Control de congestión dinámico* que le permita adaptarse a las prioridades de la aplicación y al estado actual de la red; a través del cual pueda asignar diferentes tasas de transmisión a los diferentes niveles de prioridad.

4. Ofrecer *equidad ponderada* a los nodos de la red en el uso del canal de comunicaciones, que se adapta a las prioridades del tráfico o del nodo.
5. Centrar el diseño en *evitar congestión*, más que pensar en mitigarla; lo cual necesita de mecanismos rápidos y efectivos de detección y notificación.
6. Con la intención de optimizar la notificación de congestión es necesario que el mecanismo de detección distinga casos de *congestión transitoria*; que pueden ser resueltos de forma local con el objetivo de evitar que la congestión evolucione a niveles más severos.
7. Es importante retroalimentarse de los parámetros de confiabilidad, tales como porcentaje de paquetes perdidos, retardo en la entrega de los paquetes, tasa de error de bit del canal, entre otros; para contar con mayor información que optimice el proceso de evaluación del módulo de detección. Lograr la interacción con el canal de comunicaciones es fundamental para este fin.
8. Diferenciar el origen de la pérdida del paquete, por desbordamiento del búfer o por problemas del enlace inalámbrico, es vital para las decisiones del módulo de resolución de congestión.

El conjunto de estas consideraciones involucran diferentes capas del modelo *OSI* (del inglés Open System Interconnection) para su desarrollo, lo cual requiere del paradigma de diseño crosslayer. En la literaturas encontramos propuestas que incluyen al menos dos capas [14] o algunas tan extensas como en [6].

4.1. Consideraciones para el módulo de detección de congestión

Se propone utilizar el esquema HbH para la detección de congestión por su rapidez y por la posibilidad de identificar el sitio de congestión dentro de la red. Sólo hay que tener en cuenta, en el diseño, que incrementa los procesos en cada nodo y en algunos casos el uso del búfer. Las consideraciones más importantes para este módulo son:

- Medir constantemente el *nivel del búfer* en los nodo. Es un método sencillo y rápido que se implementa tanto en modo HbH como E2E. Sin embargo, esta medida por si sola no es efectiva, ya que sería difícil establecer con precisión el grado de congestión. Ya que puede estar con una nivel bajo del nivel del búfer, mientras experimenta congestión por el trafico de otros nodos cercanos a él.
- Adicionar al nodo la habilidad para conocer la *carga de tráfico actual* del canal. Aprovechando la naturaleza compartida del medio inalámbrico, un nodo puede escuchar el canal para medir la actividad o el tráfico local de la red. Sin embargo, activar el dispositivo de radio del nodo es costoso, en términos de energía. Por lo cual, debe hacerse eficientemente, por ejemplo, a intervalos regulares y en cortos periodos de tiempo.
- Medir el *tiempo de arribo de paquetes y el tiempo de servicio de paquetes* en cada nodo; con lo cual no sólo mejora las medidas del grado de congestión del nodo y la carga local de la red; si no que además, puede disminuir la frecuencia de escucha del canal y ahorrar energía.

4.2. Consideraciones para el módulo de notificación de congestión

La notificación debe ser enviada lo más rápido posible para que el mecanismo de resolución de congestión inicie su proceso y elimine la congestión de la red. Se propone el método implícito como base para el mecanismo de notificación de congestión; que incluya como información de control: niveles de prioridad para nodos y tráfico, tasa máxima de transmisión y grado de congestión. Con la restricción de ocupar el menor espacio posible del paquete de datos:

- El *método explícito* genera un paquete especial de control. Sus ventajas son, el uso del espacio total del paquete, para incluir la información de control, y que puede ser enviado en cualquier momento. La principal desventaja es el aumento del tráfico en la red, contribuyendo a la congestión, al consumo de energía del nodo y al deterioro global de la utilización del enlace.
- El *método implícito* utiliza el concepto de piggyback; con lo cual resuelve el problema anterior. Sus desventajas son: el uso del espacio útil del paquete de datos, para incorporar los datos de control; y la espera para notificar la congestión hasta que se genere un paquete de datos.
- Por último, un evento importante del mecanismo de notificación, es generar un mensaje de *aviso de congestión terminada* a los nodos de la red, con la intención de recuperar la operación normal de la red lo más pronto posible.

4.3. Consideraciones para el módulo de resolución de congestión

La resolución de congestión debe ser dinámica, para fijar diferentes tasas de transmisión a los nodos y/o flujos de datos, de acuerdo a los requerimientos de la aplicación y a los recursos del nodo. Para evitar que la utilización del canal se degrade y se incumpla con los requisitos de QoS:

- El *ajuste de velocidad* puede ser E2E centralizado en el Sink o distribuido en cada nodo en modo HbH. El primero ofrece un ajuste imparcial y global a todos los nodos. Sin embargo, el esquema distribuido puede resolver la congestión en menor tiempo, a costa de mayor complejidad. Se propone éste último, ya que mitigar rápidamente la congestión es el objetivo principal.
- *Diferenciar el tráfico del nodo*, en tráfico propio y de transito, ayuda aplicar políticas de reducción diferencial de la tasa de transmisión, dándole mayor prioridad a sus datos o incluso el descarte del tráfico de transito.
- *Re-dirigir el tráfico* en rutas alternas cuando se tiene una ruta saturada es una opción. Sin embargo, se incrementa la complejidad de la capa de ruteo y la necesidad de vigilar el orden de los paquetes que llegan al Sink.

4.4. Heurística del protocolo crosslayer para el control de congestión

Derivado del análisis anterior y a las necesidades de los nuevos entornos de aplicación de las WSN proponemos una heurística para un protocolo crosslayer de control

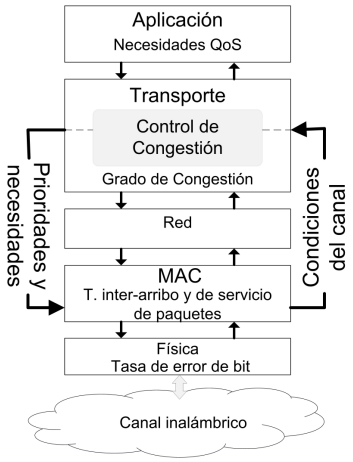


Figura 2. Esquema de la heurística del protocolo crosslayer para control de congestión

de congestión pro-activo, que ajuste la tasa de transmisión de forma dinámica, diferenciada y exacta; para indicar a los nodos que tanto aumentar o disminuir sus tasas de transmisión en relación al grado de congestión experimentado, a la ubicación del nodo dentro de la red y al tipo de tráfico de la aplicación. En nuestra propuesta la definición de optimización crosslayer se entiende como: la capacidad del protocolo de control de congestión para acceder a la información de parámetros de dos o más capas. Información que puede utilizar o modificar con el objetivo de controlar la congestión (Figura 2). Se propone utilizar la subcapa MAC y las capas Física y Transporte para la operación de los diferentes procesos. La subcapa MAC y la capa física proporcionan las condiciones actuales del canal y del comportamiento del tráfico para estimar el grado de congestión mediante los parámetros: *tasa de error de bit, el tiempo de servicio y de inter-arribo de paquetes*. En la capa de transporte se ubica la base del control de congestión, que proporciona información a la MAC para informarle de las *necesidades de operación* del protocolo y las *prioridades de las aplicaciones* soportadas en la capa de aplicación. La Figura 2 muestra la interacción de las capas, la información y parámetros que comparten. La heurística propone dos estrategias: *ajustar las tasas de transmisión* de cada nodo, de acuerdo al grado de congestión presente en el nodo y en la red. Y *gestionar el acceso al medio*, ajustando en el nodo la prioridad de servicio del búfer de salida y del mecanismo de control de acceso al medio, en aquellos casos que el nodo este en congestión o necesite mayor prioridad. Esta sinergia debería mejora los resultados de operación de los módulos de detección, notificación y resolución del protocolo.

5. Conclusiones

En este trabajo se expusieron los principales mecanismos que pueden ser utilizados para el control de congestión, bajo el nuevo paradigma crosslayer; utilizado para el diseño de protocolos de comunicación en redes de conmutación de paquetes, con mayor énfasis en las WSN. Se mostró en la sección 3 que la congestión es uno de los problemas que afectan fuertemente el rendimiento global de la red y los niveles de QoS ofrecidos. Además, que para abordar los problemas de congestión en las WSN es necesario que el protocolo integre las funciones de detección, notificación y resolución de congestión; y que estén diseñados bajo optimización crosslayer. Lo presentado puede servir de guía de diseño para protocolos de transporte que traten de mitigar el problema de congestión de la red; lo cual haría un uso más eficiente del ancho de banda y los recursos de los dispositivos de la red. Que en el caso de las WSN es un recurso muy limitado si se compara con tecnologías inalámbricas como Bluetooth, Wi-Fi o 3G. Como trabajo futuro de investigación se propone: Determinar las funciones y parámetros que deben tomarse en cuenta para la optimización crosslayer y evaluar la eficiencia de los mecanismos propuestos.

Referencias

1. M.A. Ameen, A. Nessa, and Kyung Sup Kwak. QoS issues with focus on wireless body area networks. In *Convergence and Hybrid Information Technology, 2008. ICCIT '08. Third International Conference on*, volume 1, pages 801 –807, November 2008.
2. Benoît Latré, Bart Braem, Ingrid Moerman, Chris Blondia, and Piet Demeester. A survey on wireless body area networks. *Wireless Networks*, 17(1):1–18, 2011.
3. M. Aykut Yigitel, Ozlem Durmaz Incel, and Cem Ersoy. QoS-aware MAC protocols for wireless sensor networks: A survey. *Comput. Netw.*, 55(8):1982–2004, June 2011.
4. Chieh-Yih Wan, Shane B. Eisenman, Andrew T. Campbell, and Jon Crowcroft. Siphon: overload traffic management using multi-radio virtual sinks in sensor networks. In *Proceedings of the 3rd international conference on Embedded networked sensor systems*, SenSys '05, page 116–129, New York, NY, USA, 2005. ACM.
5. Phumzile Malindi. QoS in telemedicine. In Georgi Graschew, editor, *Telemedicine Techniques and Applications*. InTech, June 2011.
6. M.C. Vuran and I.F. Akyildiz. XLP: a cross-layer protocol for efficient communication in wireless sensor networks. *IEEE Transactions on Mobile Computing*, 9(11):1578 –1591, November 2010.
7. Y.G. Iyer, S. Gandham, and S. Venkatesan. STCP: a generic transport layer protocol for wireless sensor networks. In *14th International Conference on Computer Communications and Networks, 2005. ICCCN 2005. Proceedings*, pages 449 – 454, October 2005.
8. Md. Abdur Rahman, Abdulmotaleb El Saddik, and Wail Gueaieb. Wireless sensor network transport layer: State of the art. In S.C. Mukhopadhyay and R.Y.M. Huang, editors, *Sensors*, volume 21 of *Lecture Notes in Electrical Engineering*, pages 221–245. Springer Berlin Heidelberg, 2008.
9. F.K. Shaikh, A. Khelil, A. Ali, and N. Suri. TRCCIT: tunable reliability with congestion control for information transport in wireless sensor networks. In *Wireless Internet Conference (WICON), 2010 The 5th Annual ICST*, pages 1 –9, March 2010.

10. Chieh-Yih Wan, Shane B. Eisenman, and Andrew T. Campbell. CODA: congestion detection and avoidance in sensor networks. In *Proceedings of the 1st international conference on Embedded networked sensor systems*, SenSys '03, page 266–279, New York, NY, USA, 2003. ACM.
11. E. Giancoli, F. Jabour, and A. Pedroza. CTCP: reliable transport control protocol for sensor networks. In *International Conference on Intelligent Sensors, Sensor Networks and Information Processing, 2008. ISSNIP 2008*, pages 493 –498, December 2008.
12. Vehbi Cagri Gungor, Özgür B. Akan, and Ian F. Akyildiz. A real-time and reliable transport (RT) ^{2} protocol for wireless sensor and actor networks. *IEEE/ACM Trans. Netw.*, 16(2):359–370, April 2008.
13. C. Wang, B. Li, K. Sohraby, M. Daneshmand, and Y. Hu. Upstream congestion control in wireless sensor networks through cross-layer optimization. *IEEE Journal on Selected Areas in Communications*, 25(4):786 –795, May 2007.
14. A. Sharif, V.M. Potdar, and A.J.D. Rathnayaka. LCART: a cross-layered transport protocol for heterogeneous WSN. In *2010 IEEE Sensors*, pages 793 –796, November 2010.
15. Muhammad Mahbub Alam and Choong Seon Hong. CRRT: congestion-aware and rate-controlled reliable transport in wireless sensor networks. *IEICE TRANSACTIONS on Communications*, E92-B(1):184–199, January 2009.
16. Nurcan Tezcan and Wenye Wang. ART: an asymmetric and reliable transport mechanism for wireless sensor networks. *Int. J. Sen. Netw.*, 2(3/4):188–200, April 2007.
17. Chonggang Wang, K. Sohraby, Bo Li, M. Daneshmand, and Yueming Hu. A survey of transport protocols for wireless sensor networks. *IEEE Network*, 20(3):34 – 40, June 2006.
18. M.O. Rahman, M.M. Monowar, and Choong Seon Hong. A QoS adaptive congestion control in wireless sensor network. In *Advanced Communication Technology, 2008. ICACT 2008. 10th International Conference on*, volume 2, pages 941 –946, February 2008.

Diseño de un Permutador para un Decodificador Turbo 3GPP LTE de Tasa Variable

Rodríguez Aguiñaga Adrian, Sánchez Adame Moisés, Calvillo Téllez Andrés

Centro de Investigación y Desarrollo de Tecnología Digital, CITEDI - IPN, Tijuana B. C.
TEL: 01(664)6231344,
correo-e: arodriguez@citedi.mx, msanchez@citedi.mx; calvillo@citedi.mx
Paper received on 01/10/12, Accepted on 25/10/12.

Abstract. Se presenta el análisis y modelación en Matlab de un turbo codificador con permutaciones variables. En la evaluación de la estimación se consideran las especificaciones técnicas de 3GPP LTE, para sistemas de comunicación de última milla sometidas a un ambiente ruidoso caracterizado por AWGN. La estructura del turbo codificación presento un mejor desempeño para una tasa de error de $1x10^{-7}$, mejorando la relación BER vs Eb/N_0, para comunicaciones inalámbricas con estándares UMTS.

Keywords: permutadores dou-binarios, turbo códigos, 3GPP, Tasa de error.

1 Introducción

La turbo codificación presenta múltiples trabajos en la comunidad internacional de las ciencias de la comunicación y continúa siendo un tema de investigación actual, en 1993, Claude Berrou, Alain Glavieux y Punya Thitimajshima, presentaron los Turbo códigos. Ellos centraban su interés en la codificación para corrección de errores, demostrando un excelente rendimiento y resultados que concordaban con la teoría desarrollada por el estadounidense Claude Shannon. Con ello nace una nueva era en el campo de la codificación y más específicamente, para los códigos decodificados iterativamente. Hoy día la demanda de servicios de telecomunicaciones se centra en tecnologías de cuarta generación (4G) donde las tasas máximas previstas son de 100 Mbps en enlace descendente y 50 Mbps en ascendente (con un ancho de banda en ambos sentidos de 20MHz).

Las comunicaciones inalámbricas se desarrollan constantemente, provocando la evolución de dispositivos móviles. La propuesta de este trabajo es crear un esquema de permutado basado en la arquitectura de la 3GGP, que resulte totalmente genérico, con esto se pretende que este modelo resulte aplicable a cualquier otro modelo de turbo codificación con el cual se quiera trabajar, con tan solo modificar las características y/o técnicas implementadas en la construcción de codificador o el decodificador y a su vez que el modelo permita la interacción con los bloques de bits y control se la energía simulada.

Mireya García-Vázquez, Grigori Sidorov (Eds.).
Advances in Computing Science and Control.
Research in Computing Science 59, 2012, pp. 231-240.

En la Sección dos se describe el desarrollo del algoritmo permutador, las problemáticas inmediatas, consideraciones, soluciones para la adaptación, el desarrollo e implementaciones del algoritmo, mostrando en la sección tres los resultados y finalmente en la sección cuatro las conclusiones

2 Desarrollo del Algoritmo Permutador

Los algoritmos de permutación, definen el gran parte el desempeño del turbo codificador, por lo cual tomamos como base los estándares para áreas locales y metropolitanas [1, 2], y en base a ellos se construyó el algoritmo que representara el funcionamiento del permutador conforme a los parámetros muestreados en el apéndice A.

- *N*: tamaño del bloque. El codificador/decodificar es alimentado por k bits ($k = 2 \cdot N\ bits$).

- *P0, P1, P2* y *P3*: parámetros del algoritmo permutador o estados de transición según sea el tamaño de *N*, *Pi*, i=1, 2, 3…

Esta arquitectura de algoritmo (apéndice B) de doble paso permite construir y determinar el tamaño del bloque, por lo que debe de ser capaz de trabajar con bloques de datos de distintos tamaños, leer los bits de forma ordenada o reordenada según sea su caso, debido que cada paso emplea una técnica de asignación diferente.

De igual manera los valores de P, como se observa en la en el apéndice a, dependen de N, por lo que un ciclo recorre todo el rango de valores para N, desde 0 hasta N-1 y también a su vez conforme se ejecuta el algoritmo, los bloques entrantes a la cadena se pueden expresar como sub ciclos o procesos adyacentes, lo cual expone la obtener un grado de ejecución simultanea para los procesos, lo cual hace factible la implementación de la técnica de trabajo divide y vencerás, con lo que N, cambiara a n, siendo $n < N$, $n \subset N$ y $n = \frac{N}{Pe}$, se asigna Pe como el grado o nivel de bloques para efectuar el proceso en sub bloques.

2.1 Problemáticas Inmediatas

Normalmente los valores de Pe, son expresados en potencias de 2, sin embargo estos valores resultan no ser eficientes cuando presentan denominaciones elevadas, además de resultar en casos en los cuales los valores de Pe, podrían no ser divisibles a N, estos casos se tienes que definir y se atacaran con una técnica distinta [2], para toda Pi, donde Pe puede tomar estos valores por los niveles de conveniencia definidos para evitar la complejidad [3], [4].

```
Para j= rango desde 0 hasta N-1
Subdivisión en procesos:
Proceso 1: desde 0 hasta n-1.
```

```
Proceso 2: desde n hasta 2·n-1.
Proceso n: desde (Pe-1)·n hasta N-1.
(k-1)·n hasta k·n-1.
```

Esta arquitectura presenta problemáticas para los casos de Pe = 8 y N= 36,108 y 180 y son tratados por la función fix, que se encarga de eliminar el fraccionario del resultante de la operación.

$$n = fix\left(\frac{n}{Pe}\right) \tag{1}$$

Otro punto a tener en cuenta es que la creación de rejillas o sub bloques creará solapamiento realizado por el decodificador. Con un solapamiento suficiente de muestras la decodificación será realizada óptimamente y a una alta tasa de bits [5] .Cada sub bloque n es dividido en rejillas, y cada rejillas debe de estar compuesta por un número óptimo de muestras, para evitar conflictos de direccionamiento de memoria, ya que se disponen N bloques y con Pe procesos, a cuales se necesitará almacenar los resultados en memoria y los datos serán almacenados y referenciados para n posiciones, es decir desde 0 hasta N-1, como se muestra en la Fig. 1.

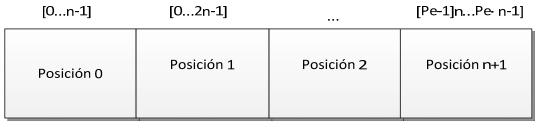


Fig. 1. Acceso a memoria evitando traslapes en la información de cada bloque.

Así la posición 0 de la memoria corresponderá con el índice 0, generado por el permutador y la posición n-1 corresponderá al índice N-1, así cuando el permutador genera secuencias de índices ordenados, cada uno de los procesos alternos también generara índices.

$$[(k-1)\cdot n, \ldots, k\cdot n-1] \tag{2}$$

Las entradas y salidas “suaves” deben de leer los datos entregados por el permutador en forma sincrónica, aun cuando haya existido algún error o colisión, para evitar la implementación de sincronizadores muy complejos [6 - 8].

2.2 Consideraciones y Soluciones para Adaptación del Algoritmo

Se tomaran en cuentas las consideraciones propuestas para resolver errores o colisiones, que son las siguientes:

1) Detectar los casos específicos de mayor ocurrencia error= 38, 108,180.

2) La proporción de recursiones con errores frente a las correctas, cuando se dan valores de *N* y *Pe* que producen un error es muy alta, se normaliza en todos los casos.
3) Cuando ocurre colisión, son únicamente dos los productores que colisionan es decir, una petición puede ser procesada al instante mientras la otra ha de esperar al siguiente ciclo para ser procesada.

Cuando la tercera consideración sea efectiva se detiene el permutador para meter el error a un ciclo nulo, el cual pude ayudar al sistema sincronizar, mediante la activación de un control de flujo de bits determinado por un diagrama de estados el cual auxiliara a mantener la sincronía y evitar errores de entrada y salida de datos, para el codificador o el decodificador, como se muestra en la Fig. 2.

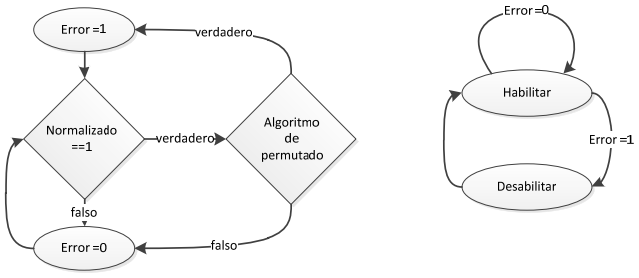


Fig. 2. Momentos en cual el permutador permitirá sincronización para evitar errores.

2.3 Desarrollo del Algoritmo

Los bloques son implementados en base a los algoritmos MAP, estos algoritmos realizan recorridos hacia adelante y hacia atrás, por lo cual, otra de las características a tomar en cuenta es que el permutador debe de operar estos bloques en la misma dirección que los decodificadores ver Fig.3. Para ello se hace uso del algoritmo LIFO (Last-Input, First Output), en el algoritmo permutador se generan únicamente índices, no direcciones y referencias por lo cual resulta importante implementar el sistema LIFO, para disponer de la lista direcciones cuando se requieran.

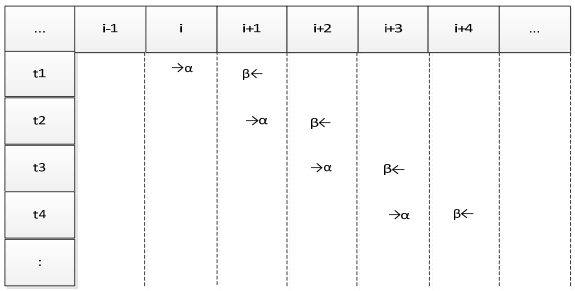


Fig. 3. Sub bloques para un algoritmo MAP, para el permutado.

Si observamos en la ilustración 3, $i+1$ y $i+2$, en $t2$ alfa y beta ocurren al mismo tiempo, esta situación introduce la necesidad de crear un arreglo bidimensional o una pila doble para evitar conflictos y proveer direcciones así como referencias en conjunto con el generador de direcciones donde ya han sido reordenadas para conservar el orden obedeciendo el conmutado del algoritmo permutador.

2.4 Implementación del Algoritmo

Es posible manipular el algoritmo para lograr implementaciones de menor complejidad, tal es el caso de la división o inclusión de parámetros como sub procesos, estos permiten simplificar el orden del algoritmo y su ejecución, por lo que nos conviene establecer puntos de referencia en el algoritmo donde esto resulte valido (Tabla 2),a través del empleo de pocos recursos para calcular del módulo donde únicamente es necesario un registro como pila, un sustractor y un comparador de magnitud, esto permite mejorare la métrica aplicada a los cálculos en el modulo-N, esto significa que bastara con los dos bits menos significativos (BMS), para establecer su módulo.

Existe un incremento de ciclos en los bloques mayores, pero esto no necesariamente indica mayor complejidad dado que al ser un proceso de sumas recursivas y sub procesos mantiene su simplicidad, a diferencia de utilizar quizá menos ciclos de una arquitectura más compleja, por lo que el algoritmo queda descrito como se muestra a continuación:

Tabla 1. Algoritmo

```
for j = 0 … N-1
      if j = 0  { pila = 0}
          else  { pila = pila + P0}
end if
      if pila >= N { pila = pila – N}
end if
aux = 2BMS( binario( j ))
if aux = 00 { i = pila + 1}
         else if aux = 01
         { i = pila + mod(1+N/2+P1, N)}
```

Comportamiento

	j	P0·j
Proceso 1	0	0
	:	:
	n-1	n-1·P0
Proceso 2	n	n·P0
	:	:
	2·n-1	(2·n-1)·P0

```
else if aux = 10
        { i = pila + mod(1+P2, N)}
else if aux = 11
       { i = pila + mod(1+N/2+P3, N)}
end if
if i >= N
       { i = i – N}
end if
```

Proceso Pe	(Pe-1)·n	((Pe-1)·n)·P0
	:	:
	n·Pe-1	(n·Pe-1)·P0

Con esta implementación se consigue el mismo comportamiento así como la reducción de recursos empleados para ello.

Debido a que no todos los casos iniciales son cero, esto porque *n* no es en todos los casos múltiplo de N, el algoritmo implementa un análisis de los dos bits menos significativos del tamaño de sub-bloque *n* que son la compensación entre los casos del proceso *k* y *k+1*, así que en lugar de iniciar en cero para cada proceso se inicializa con una $variable = P0 \cdot k \cdot n$.

Para asegurar que no se realice ninguna recursión extra en la normalización del registro pila, el nuevo parámetro será normalizado antes de ser almacenado temporalmente, entonces la $variable = mod\,(P0 \cdot k \cdot n, N)$.

3 Resultados

Los experimentos se realizaron con bloques de datos desde 24 hasta 2400 bits (Tabla 2), con un código 1/3, para las simulaciones se empleó matlab, aplicando algoritmos de turbo codificadores 3GPP, trabajando en conjunto con las mejoras descritas en la sección 2 de este artículo, relacionadas al permutador de datos. Se establece un rango de pasos mayor a 0.2 dB e inferior a 2.4 dB para la relación BER vs Eb/N0, en la caracterización mostrada en las gráficas.

Para las cadenas de bloques mayores a mil bits, se observa un desempeño superior contra las de menor dimensión, como podemos observar en la ilustración 4 y 5, esto debido a que se observa como conforme aumenta el tamaño del bloque, así mismo como el grado información provista por los procesos es mayor y por tanto el rendimiento del permutador es mayor.

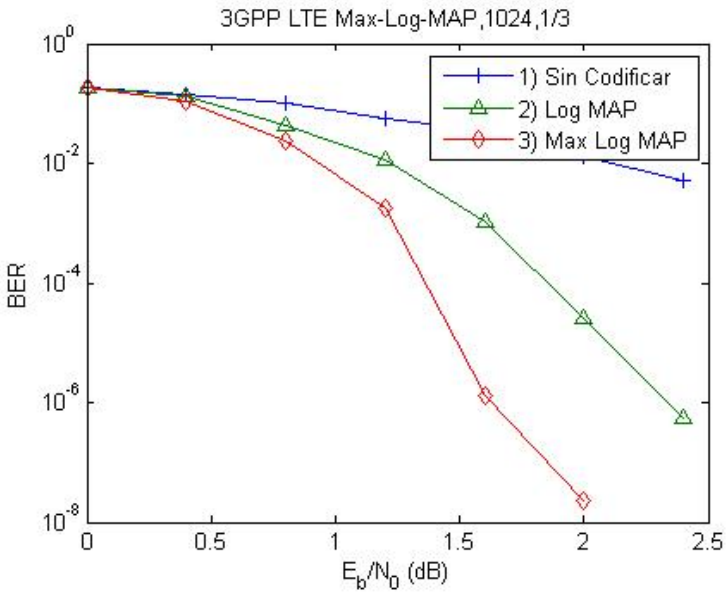


Fig. 4. Comportamiento de la tasa de error de bits, con bloques de permutación superiores a mil bits.

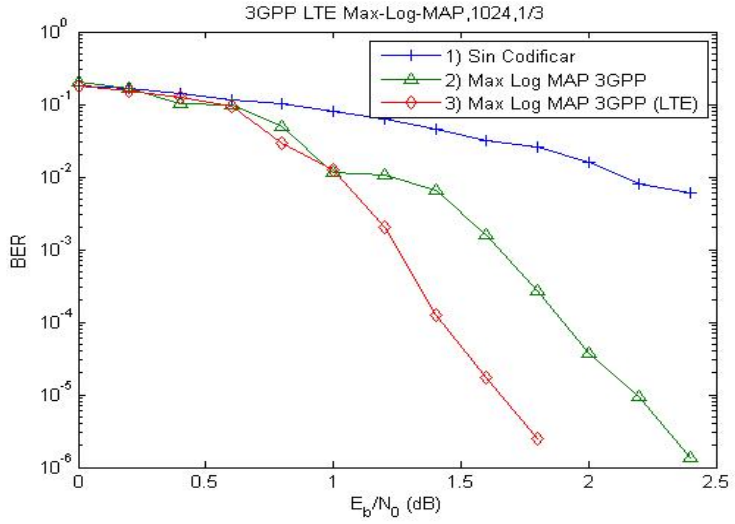


Fig. 5. Comportamiento de la tasa de error de bits, con bloques de permutación menores a mil bits.

Sin embargo debido a la necesidad de entregar los datos contemporáneamente y que en los bloques de bits de menor tamaño se encuentran los casos de error de N, resulta no ser divisible entero de un byte, el desempeño sigue siendo bajo.

4 Conclusiones

La implementación de las técnicas de permutado descritas en este artículo de logran tasas de error de bits, que superan la tasa de un error por cada millón de bits transmitidos, en base a las simulaciones realizadas con la sincronización de procesos de hasta $9.53x10^{-7}$, logrando esto por la retroalimentación de ambos CRS en forma sincrónica y con entradas controladas, en el cual se observa que el sacrifico de energía es de 0.2 dB, con un esquema de modulación BPSK y a través de un canal AWGN.

También se observa que a partir de la cuarta iteración los parámetros de BER, no sufren mayores ganancias, con lo cual se elimina tiempo de cómputo que ya no aportara mayores ganancias, al desempeño.

Al hacer estas adaptaciones a los permutadores, obtenemos una mejora en la tasa de bits, a diferencia de los modelos basados en cadenas de pseudo aleatorias de bits en un permutador. Además de disminuir el tiempo de procesamiento, definiendo el rango de iteraciones útiles para un proceso de codificación, y no alcanzar procesamientos en los cuales la información de ganancia ya no aporta datos relevantes para las decisiones suaves y duras del sistema de turbo codificación.

5 Referencias

1. IEEE 802.16-2004. IEEE Standard for Local and Metropolitan area Networks. Part 16: Air Interface for Fixed Broadband Wireless Access Systems.
2. IEEE 802.16e2005. IEEE Standard for Local and Metropolitan area Networks. Part 16: Air Interface for Fixed Broadband Wireless Access Systems.
3. S. Yoon, Y. Bar-Ness, "A Parallel MAP Algorithm for Low Latency Turbo Decoding",IEEE Communications Letters, Vol. 6, Nº 7, July 2002.
4. F. Speziali, J. Zory, "Scalable and Area Efficient Concurrent Permutador for HighThroughput Turbo-Decoders", IEEE Computer Society.
5. R. Dobkin, M. Peleg, R. Ginosar, “Parallel Permutador Design and VLSI Architecturefor Low-Latency MAP Turbo Decoders”, IEEE Transactions on VLSI Systems, Vol. 13, Nº 4, April 2005.
6. Corneliussen, A, Silpakar, P. Hardware-Accelerated NIOS-II Implementation of a Turbo Decoder. Computer and Electrical Engineering, 2009. ICCEE '09. Second International Conference on Date of Conference: 28-30 Dec. 2009
7. Han, J. Erdogan, A. Arslan, T. Power and Area Efficient Turbo Decoder Implementation for Mobile Wireless Systems. Signal Processing Systems Design and Implementation, 2005. IEEE Workshop on Digital Object Identifier. 2005, Page(s): 705 – 709.
8. Han, J.H.; Erdogan, A.T.; Arslan, T. Implementation of an efficient two-step SOVA turbo decoder for wireless communication systems. Global Telecommunications Conference, 2005. GLOBECOM '05. IEEE Volume: 4 2005, Page(s): 5 pp. - 2433

Tabla 2. Valores de P0, P1, P2 y P3 según N.

N	P0	P1	P2	P3
24	5	0	0	0
36	11	18	0	18
48	13	24	0	24
72	11	6	0	6
96	7	48	24	72
108	11	54	56	2
120	13	60	0	60
144	17	74	72	2
180	11	90	0	90
192	11	96	48	144
216	13	108	0	108
240	13	120	60	180
480	13	240	120	360
960	13	480	240	720
1440	17	720	360	540
1920	17	960	480	1440
2400	17	1200	600	1800

Tabla 3. Algoritmo binario de reordenamiento de datos para sistemas 3GPP.

Paso 1: alternar los pares de bits.

Asignar *u0 = [(A0, B0), (A1, B1), (A2, B2), (A3, B3),... (AN-1, BN-1)]* , como entrada del codificador 1.
Para i = 0 hasta *N*-1
Si (i mod 2 == 1) intercambiar los pares *(Ai,Bi)* • *(Bi,Ai)*

Con lo cual obtendremos una secuencia similar a esta:
u1 = [(A0,B0), (B1,A1), (A2,B2), (B3,A3), ... (BN-1,AN-1)] = [u1(0), u1(1), u1(2), u1(3), ... , u1(N-1)]

Paso 2: P(j), la función *P(j), proporciona las direcciones en la secuencia* u1, que deberán ser asignada en direcciones "*j*" ,de la secuencia permutada • *u2(i) = u1(P(j))*

Para *j* = 0 hasta N-1
Conmutar *j* mod 4:
Caso 0: $P(j) = (P0 \cdot j + 1) \bmod N$
caso 1: $P(j) = (P0 \cdot j + 1 + N/2 + P1) \bmod N$
caso 2: $P(j) = (P0 \cdot j + 1 + P2) \bmod N$

caso 3: $P(j) = (P0 \cdot j + 1 + N/2 + P3) \bmod N$
la secuencia obtenida por el paso 2 es la siguiente:
u2 = [u1(P(0)), u1(P(1)), u1(P(2)), u1(P(3)), ... ,
u1(P(N-1))] =
= [(BP(0), AP(0)), (AP(1), BP(1)), (BP(2), AP(2)),
(AP(3), BP(3)), ... , (AP(N-1), BP(N-1))].
u2 es la secuencia de entrada para el codificador 2.

Análisis comparativo de las características de radiación de antenas monopolo de microcinta para dispositivos móviles basados en tecnología 4G - LTE

Rainer García Aldama, José Luis Medina Monroy, Ricardo Arturo Chávez Pérez

Centro de Investigación Científica y de Educación Superior de Ensenada; Carretera Ensenada-Tijuana No. 3918, Zona Playitas, C.P. 22860, Ensenada, B. C. México.
rainer@cicese.edu.mx, jmedina@cicese.mx, chavez@cicese.mx
Paper received on 01/10/12, Accepted on 25/10/12.

Abstract. En este trabajo se presenta el análisis comparativo de tres estructuras diferentes de antenas impresas tipo monopolo, diseñadas para dispositivos móviles 4G basados en tecnología LTE. Estas estructuras deberán cubrir las bandas de microondas de ultra-ancho de banda definidas para LTE, manteniendo un tamaño reducido adecuado para dispositivos móviles. Se realiza un análisis comparativo de resultados del análisis electromagnético de las tres estructuras diseñadas con respecto al ancho de banda de impedancias, el patrón de radiación, la ganancia y el tamaño de los mismos. Entre los resultados, se han logrado antenas pequeñas con ganancia, patrones de radiación omnidireccionales y anchos de banda fraccionales superiores al 93%, entre otros.

Keywords: antena de microcinta, antena monopolo, banda ultra-ancha (UWB), comunicaciones inalámbricas 4G.

1 Introducción

Los requerimientos de servicios cada vez mayores de las redes de comunicaciones móviles, han llevado consigo un desarrollo vertiginoso de las tecnologías móviles existentes, en busca de ofrecer un mayor ancho de banda y de incrementar la eficiencia espectral de los sistemas de comunicaciones existentes. La tecnología móvil de 4ta generación LTE (Long Term Evolution) aparece como una solución viable a los requerimientos establecidos.

LTE brinda la posibilidad de seleccionar entre un gran número de canales y bandas de frecuencia de microondas, en el momento de realizar su implementación y posterior explotación. Estas bandas de frecuencia de microondas comienzan desde los 1428MHz y se extienden hasta los 2690MHz, lo cual constituye el único requerimiento de un sistema 4G – LTE relativo a las antenas. El resto de los requerimientos de la interfaz aérea 4G – LTE están relacionados con el aprovechamien-

Mireya García-Vázquez, Grigori Sidorov (Eds.).
Advances in Computing Science and Control.
Research in Computing Science 59, 2012, pp. 241-252.

to óptimo de las capacidades del canal, el traspaso suave entre celdas y el control de potencia de la señal, por lo que no son de interés para el tema abordado.

En aras de permitir que los dispositivos móviles 4G sean capaces de operar en diferentes bandas, ampliando su mercado y su posible área de cobertura, se han desarrollado estructuras radiantes capaces de cubrir bandas específicas del espectro donde se han detectado solicitudes de autorización de licencias de explotación. Estas estructuras generalmente se presentan como sistemas radiantes del tipo multi-banda, capaces de abarcar regiones discretas dentro de determinado rango de frecuencias, a través de múltiples resonancias, mayormente de banda estrecha. Estructuras como las descritas pueden encontrarse en la literatura [1–5].

Sin embargo, la norma LTE define bandas de comunicaciones que se presentan de forma continua en el espectro. Esto ocasiona que en las regiones del espectro que se encuentran entre las resonancias multi-banda anteriormente descritas, las estructuras radiantes posean valores de pérdidas por regreso tales que no puedan ser consideradas aceptables para ser utilizadas, debido al desacoplamiento de la antena con el receptor y transmisor del equipo y ocasiona una pérdida de cobertura para el dispositivo móvil.

Recientemente se han hecho esfuerzos de desarrollar estructuras de antenas de banda ultra-ancha diseñadas para cubrir regiones amplias del espectro. A pesar de ello, las estructuras encontradas en la literatura [6–10] presentan un límite inferior de su banda de trabajo, superior al requerido para cubrir las primeras bandas de frecuencia de microondas autorizadas para LTE cercanas a los 1428 MHz. Una reducción de la frecuencia mínima de operación de estas estructuras, ocasiona un aumento de sus dimensiones físicas, lo cual dificulta su implementación o instalación dentro de un dispositivo móvil 4G - LTE.

En este trabajo, se presentan resultados del diseño y análisis electromagnético de tres estructuras geométricas diferentes de antenas monopolo diseñadas para: cubrir el rango de frecuencias requerido para la aplicación 4G – LTE, tener un patrón de radiación omnidireccional que permite incrementar la cobertura espacial, y mantener un tamaño reducido. Asimismo, se presenta una comparación de los resultados de las tres estructuras estudiadas, incluyendo sus patrones de radiación, anchos de banda, ganancias y pérdidas por regreso en toda la banda de trabajo.

2 Análisis de las Antenas Monopolo

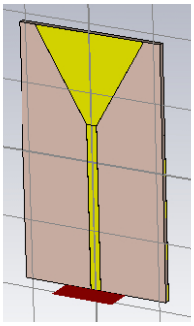
Se efectúa el análisis electromagnético de tres estructuras de antenas monopolo: la triangular, la rectangular y la circular en el intervalo de frecuencias de 500MHz a 5GHz mediante el software CST Microwave Studio®. Las tres estructuras se diseñaron empleando un substrato FR-4 con espesor h =1.5778mm, permitividad ε_r = 4.08 y tangente de pérdidas $Tan\delta = 0.018$, lo cual permite disminuir los costos de fabricación; sin embargo pueden obtenerse estructuras más pequeñas utilizando substratos con mayor permitividad, pero a un costo mayor.

Para definir si el tamaño de la antena es adecuado para un dispositivo móvil 4G se consideraron los dispositivos de última generación desarrollados por las empresas líderes que dominan el mercado actual en esta materia: Samsung, Apple, Blackberry y Nokia. El equipo con menor ancho (We=58.6mm) es el de la gama iP-

hone (versiones 4, 4S y 5) y el de menor altura (Le=105mm) es el Blackberry Bold. Por lo tanto estas dimensiones establecen el tamaño máximo que deberán cumplir las antenas monopolo a diseñar.

2.1 Antena Monopolo Triangular

La geometría del monopolo triangular propuesto se muestra en la Fig. 1, conjuntamente con el valor y descripción de cada una de las variables de diseño. La antena ocupa un área total de 74.5 X 44 mm, lo cual constituye una opción viable tomando en consideración el requerimiento definido.



Name	Value	Description
Lf	48	Lenght of the feed line
Lg	48	Lenght of the ground plane
Lx	a1+10	Width of the structure
Ly	Lg+b1	Lenght of the structure
W	3.12	Width of the feed line
a1	34	Upper monopole width
b1	26.5	Monopole height
h	1.5778	Substrate height (orig 60mils)
t	0.0298	Copper layer height

Fig. 1. Antena monopolo triangular y sus variables de diseño

Es importante mencionar que la longitud de la línea de alimentación Lf y la altura del monopolo b1 establecen la frecuencia mínima de la banda de trabajo.

Los resultados del análisis de la antena monopolo triangular presentan dos frecuencias de resonancia. Al disminuir el parámetro a1 se provoca un acercamiento entre ambas frecuencias de resonancia, así como una disminución del ancho de banda de impedancias. Por el contrario si se aumenta el parámetro a1, las frecuencias de resonancia se separan, a expensas de aumentar las pérdidas en la región intermedia entre ambas resonancias, llegando a exceder el límite de los -10dB.

En la Fig. 2 se muestra como el ancho banda de impedancias para esta estructura queda definido entre 1.41GHz y 2.87GHz, obteniéndose un ancho de banda fraccional de 68.9%. Este rango de frecuencias cumple con el requisito de diseño de ancho de banda establecido. Se puede observar que la reflexión mínima se localiza en 2.54GHz, donde las pérdidas de retorno alcanzan los -20.3dB.

El comportamiento de ganancia de la antena triangular se muestra en la Fig. 3, donde los marcadores 1 y 2 representan los límites del ancho de banda de impedancias definido, mientras el marcador 3 señaliza la frecuencia a la cual se tiene la ganancia máxima, manteniéndose dicha ganancia en el rango de 2.05dB a 4.35dB.

El patrón de radiación de la antena triangular en el plano H se muestra en la Fig. 4 para las frecuencias límites de la banda de trabajo, ya que es omnidireccional en el plano E. A la frecuencia mínima, el patrón de radiación presenta un comportamiento omnidireccional, con máximos aproximadamente perpendiculares al plano de la antena. En el límite superior de la banda se observa un corrimiento de 25º en la dirección de máxima radiación, acompañado de un estrechamiento de 30.6º en

el ancho del haz de mediana potencia (HPBW). El comportamiento simétrico del patrón y su buen ancho del haz de mediana potencia, permiten al dispositivo móvil mejorar su rango espacial de cobertura.

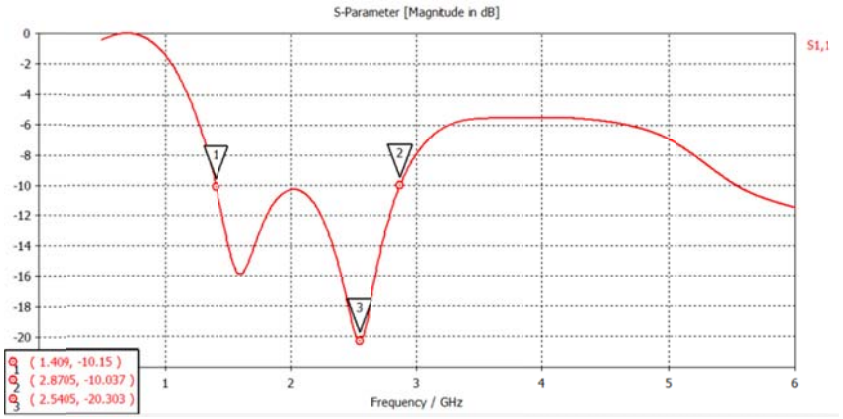


Fig. 2. Pérdidas de retorno de la antena monopolo triangular

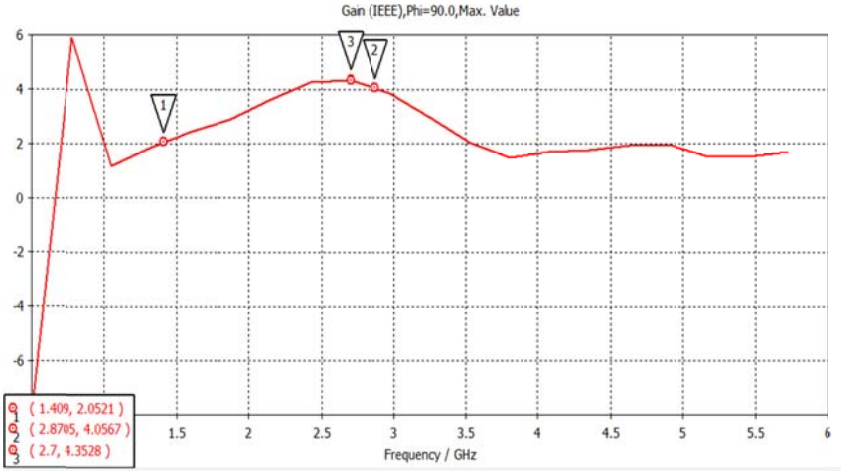


Fig. 3. Comportamiento de la ganancia en dB de la antena monopolo triangular

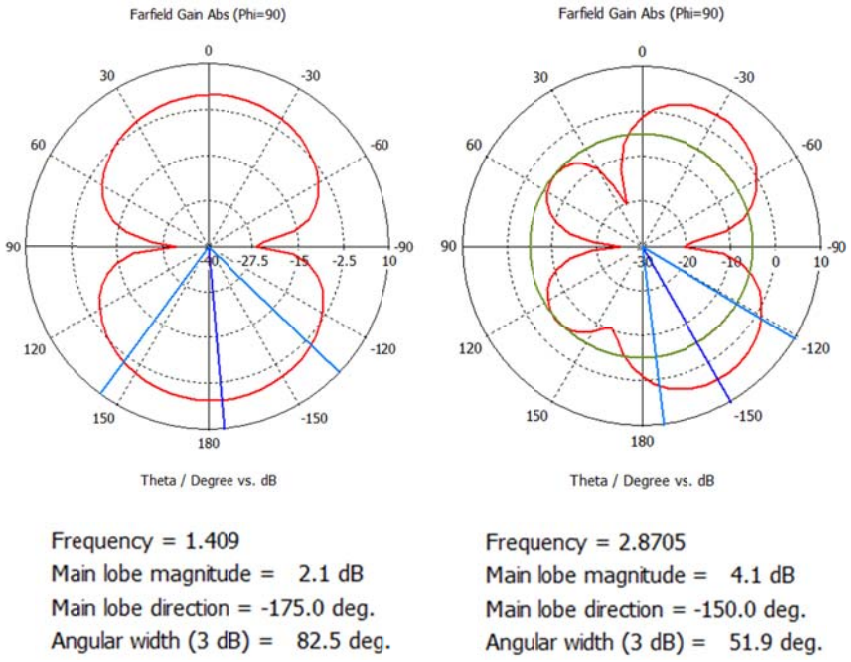
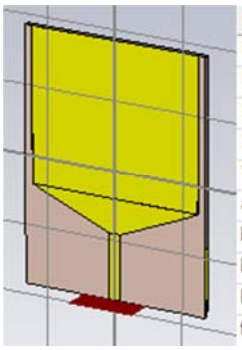


Fig. 4. Patrones de radiación de la antena triangular en 1.409GHz y 2.8705GHz

2.2 Antena Monopolo Rectangular

En la Fig. 5 se muestra la antena monopolo rectangular propuesta, así como los valores de cada variable de diseño. Esta antena presenta, al igual que la antena triangular, dos frecuencias de resonancia. La altura b1 del acoplador triangular determina si el monopolo se comporta como una antena de ultra-ancho de banda o del tipo multi-banda. La antena presenta un área de 74 X 54 mm, acorde con los requerimientos necesarios.



Name	Value	Description
Lf	19	Lenght of the feed line
Lg	19.2	Lenght of the ground plane
Lx	a1+4	Width of the structure
Ly	Lf+b1+b2	Lenght of the structure
W	3.12	Width of the feed line
a1	50	Upper monopole width
b1	10	Triangular zone height
b2	45	Rectangular zone height
h	1.5778	Substrate height (orig 60mils)
t	0.0298	Copper layer height

Fig. 5. Antena monopolo rectangular y sus dimensiones físicas

El comportamiento de las pérdidas por retorno se muestra en la Fig. 6, donde se puede apreciar su ancho de banda de impedancias desde 1.4GHz hasta 3.2GHz, con un ancho de banda fraccional de 78.6%, lo cual cumple con la especificación de diseño. Las pérdidas de inserción menores (-17.2dB) están localizadas en 1.667GHz.

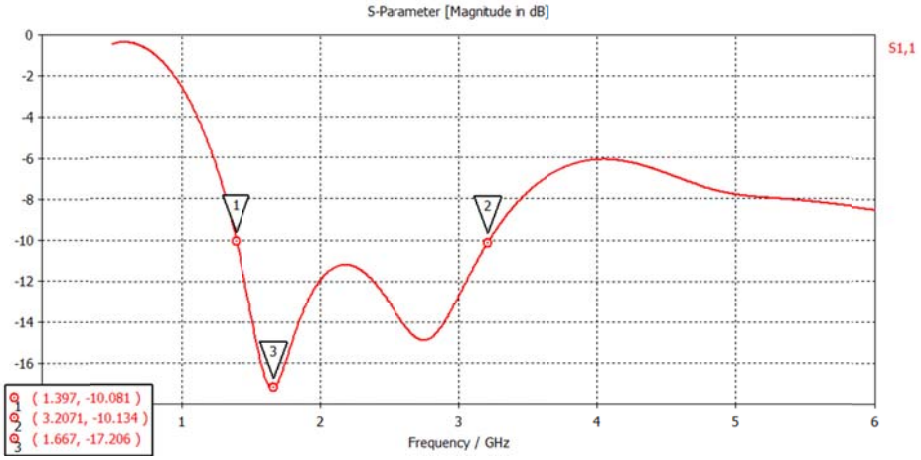


Fig. 6. Pérdidas de retorno de la antena monopolo rectangular

En la Fig. 7 se muestra el comportamiento de ganancia la cual está acotada entre 2.415dB y 4.9264dB dentro del mismo ancho de banda de impedancias. Los marcadores 1 y 2 representan los límites de la banda de trabajo, mientras el marcador 3 señala la frecuencia a la cual la ganancia es máxima.

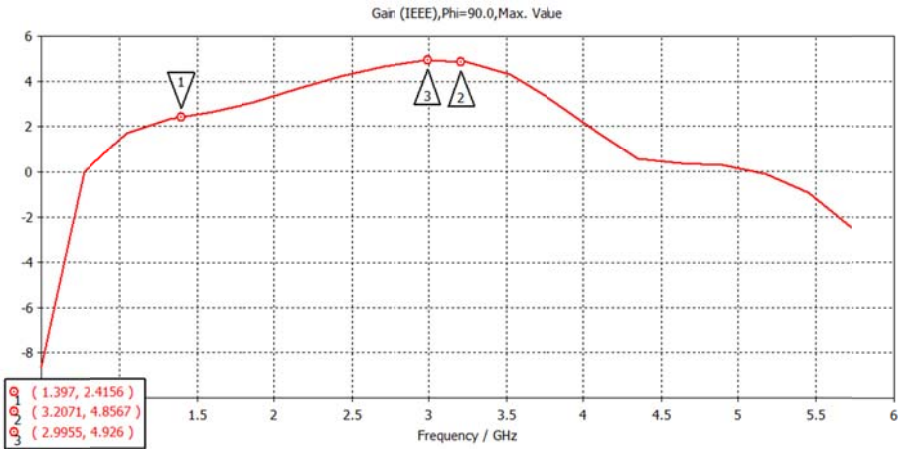


Fig. 7. Comportamiento de la ganancia en dB de la antena monopolo rectangular

En la Fig. 8 se muestra el comportamiento del patrón de radiación en el plano H, ya que en el plano E es omnidireccional. Se puede observar que en 1.397GHz la radiación es máxima en la dirección perpendicular al plano de la antena, mientras

que a 3.2071GHz se presenta una reducción de 33.7º en el ancho del haz de mediana potencia así como un corrimiento de 20º en la dirección de radiación máxima.

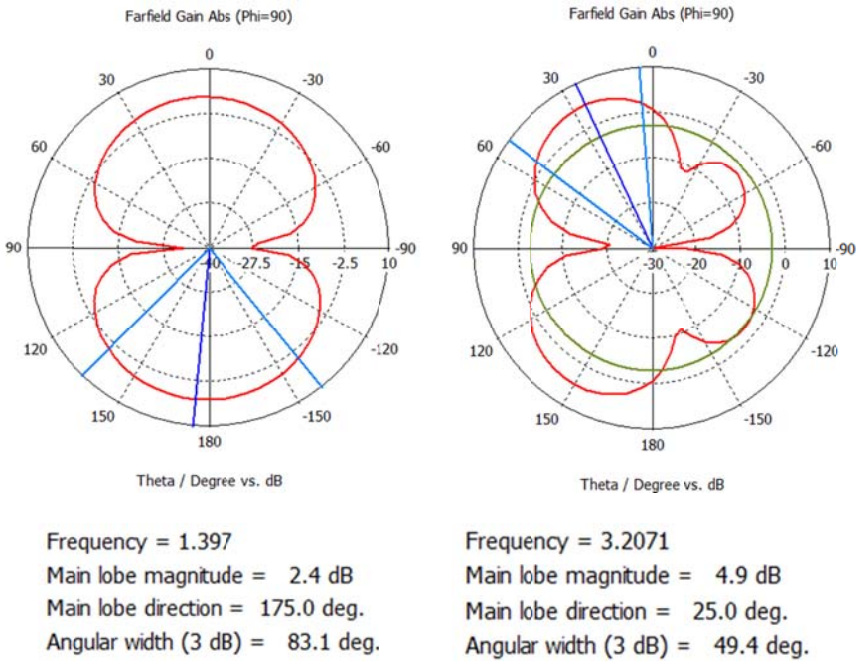
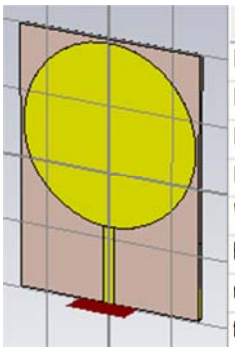


Fig. 8. Patrones de radiación de la antena rectangular a 1.397GHz y 3.2071GHz

2.3 Antena Monopolo Circular

El tercer monopolo presenta una geometría circular y se muestra en la Fig. 9 con los valores correspondientes a todas sus variables de diseño. Esta antena presenta un área de 79 X 57 mm, lo cual cumple con las especificaciones de tamaño.



Name	Value	Description
Lf	Lg	Lenght of the feed line
Lg	23	Lenght of the ground plane
Lx	2*r+2	Width of the structure
Ly	Lf+2*r+1	Lenght of the structure
W	3.12	Width of the feed line
h	1.5778	Substrate height (orig 60mils)
r	27.5	Monopole radio
t	0.0298	Copper layer height

Fig. 9. Geometría y dimensiones físicas de la antena monopolo circular

A diferencia de los monopolos anteriores, donde cada una de las variables de diseño afectaba solo una de las dimensiones del plano, en este tipo de estructura la variación del radio del círculo afecta a ambas dimensiones espaciales. Esta condición impone una limitación de diseño.

En la Fig. 10 se muestra el comportamiento de las pérdidas por retorno, donde su ancho de banda de impedancias presenta como frecuencia límite inferior 1.383GHz y se extiende más allá de los 6GHz. El mejor valor de acoplamiento (-24.984dB) se encuentra en 3.1235GHz.

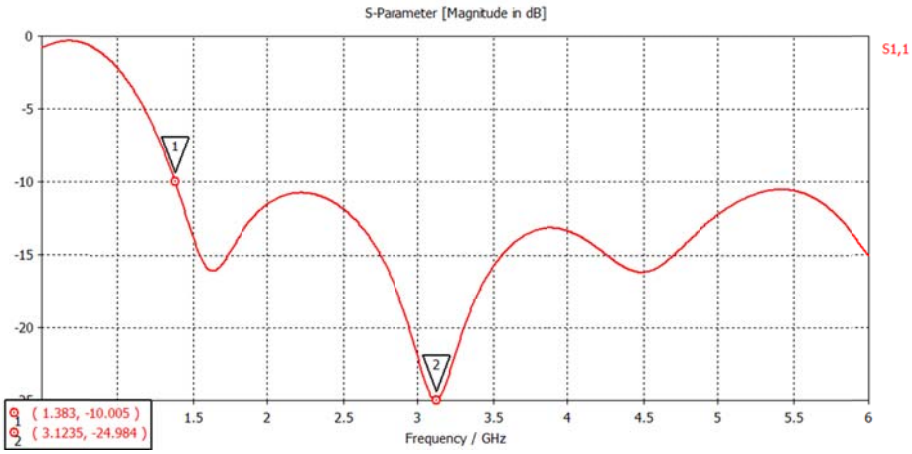


Fig. 10. Pérdidas de retorno de la antena monopolo circular

La Fig. 11 muestra el comportamiento de la ganancia en función de la frecuencia. A partir del inicio de la banda de trabajo de impedancias la antena posee ganancia hasta los 5.1GHz. Aunque el rango de frecuencias de interés para 4G – LTE culmina en los 2.69GHZ, dado el amplio ancho de banda de impedancias mostrado en la Fig. 10, se analiza su comportamiento hasta 3.8GHz ya que la estructura puede ser utilizada para dispositivos que implementen otras tecnologías móviles 4G, como WIMAX.

Se puede observar en la Fig. 11 que en el rango de frecuencias definido, la ganancia está acotada entre 2.176dB y 4.671dB. Se señalizan los límites de esta banda 4G por medio de los marcadores 1 y 3, mientras el marcador 2 muestra la frecuencia a la que se alcanza la mayor ganancia y el marcador 4 la frecuencia a la cual la antena monopolo circular propuesta deja de tener ganancia.

En la Fig. 12 se muestra el patrón de radiación en el plano H. La dirección de máxima radiación, para la frecuencia de 1.383GHz, es perpendicular al plano del monopolo. Mientras que para 3.8GHz existe un corrimiento de 30º en la dirección de radiación máxima, acompañado de una disminución de 34º en el ancho del haz de mediana potencia. Este comportamiento omnidireccional del patrón de radiación unido a una ganancia adecuada en toda la banda de trabajo, favorece la cobertura espacial del dispositivo móvil.

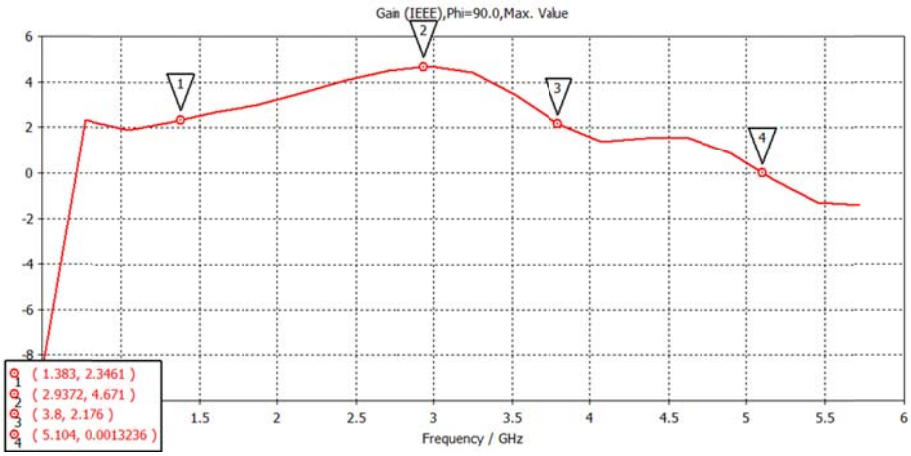


Fig. 11. Comportamiento de la ganancia en dB de la antena monopolo circular

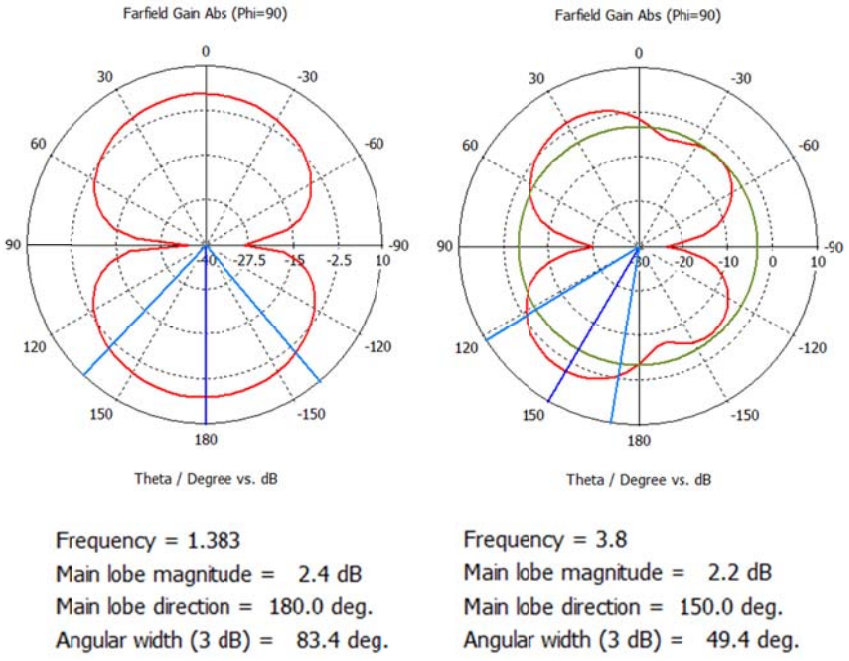


Fig. 12. Patrones de radiación de la antena circular para 1.383GHz y 3.8GHz

Tomando en consideración todas las figuras de mérito analizadas, se puede decir que la antena circular tiene un ancho de banda entre 1.383GHz y 3.8GHz, resultando un ancho de banda fraccional del 93.26%, el cual supera al ancho de banda requerido para 4G – LTE.

3 Resultados Experimentales

Las tres estructuras diseñadas y analizadas, se construyeron en un substrato FR-4 y se añadieron conectores SMA. La Fig. 13 muestra las antenas monopolo construidas, en donde sus dimensiones físicas coinciden con los valores obtenidos del diseño.

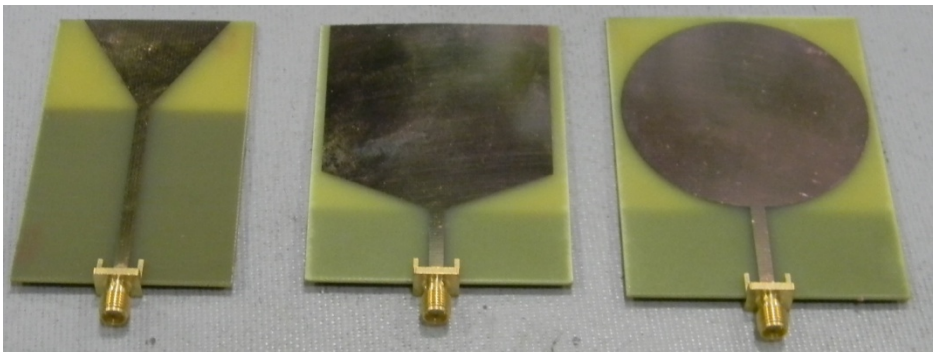


Fig. 13. Antenas monopolo construidas

Las estructuras se caracterizaron mediante un analizador de redes vectorial en el intervalo de frecuencias de 1 a 3 GHz. En la Fig. 14 se muestran los resultados de la medición de las pérdidas de retorno, en donde para todas las estructuras se muestra un corrimiento de la frecuencia mínima de trabajo, llevándola hasta las 1.45GHz para la antena circular y hasta los 1.48GHz para el resto de las antenas. En todos los casos la frecuencia máxima se extiende más allá de 2.89GHz, cumpliendo exitosamente el requerimiento necesario en este aspecto.

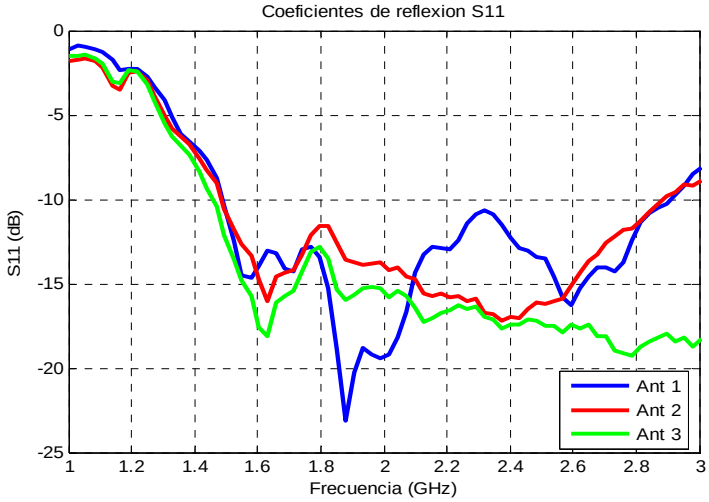


Fig. 14. Resultados de la medición de las pérdidas de retorno para las antenas monopolos triangular (Ant 1), rectangular (Ant 2) y circular (Ant 3)

Es importante resaltar que la antena circular, extiende su frecuencia máxima más allá de los 7GHz, constituyendo la estructura con mejor ancho de banda.

La mayor variación en la frecuencia mínima de trabajo es inferior a los 0.08GHz y se atribuye a pequeños cambios en los valores de la permitividad eléctrica relativa y la tangente de pérdidas del substrato FR-4.

Una comparación de las tres estructuras presentadas anteriormente se muestra en la Tabla 1 la cual resume los valores de las figuras de mérito analizadas para cada antena.

Table 1. Valores de las figuras de mérito para las tres antenas monopolo propuestas

Antena Monopolo	Dimensiones físicas [mm]	Ancho de banda [GHz]	Ancho de banda fraccional [%]	Rango de ganancia [dB]	Variación de la dirección de radiación máxima
Triangular	74.5 X 44	1.46	68.9	2.05 – 4.35	25º
Rectangular	74 X 54	1.8	78.6	2.41 – 4.92	20º
Circular	79 X 57	2.417	93.26	2.17 – 4.67	30º

Se puede apreciar que las dimensiones resultaron inferiores a 79 X 57 mm, el ancho de banda fraccional fue superior al 68.9% y sus ganancias superan los 2.05dB en todos los casos.

4 Conclusiones

Se proponen y analizan tres antenas monopolo para dispositivos móviles 4G – LTE.

Se estudiaron las geometrías triangular, rectangular y circular, en donde el análisis electromagnético demostró que estas estructuras son capaces de cumplir con los requerimientos de diseño.

La elección de la mejor antena monopolo entre las propuestas depende de la figura de mérito que se considere más importante. El monopolo triangular resulta el de menores dimensiones físicas, la estructura rectangular con tapper presenta la mayor ganancia y la antena monopolo circular muestra el mejor ancho de banda. En general las tres antenas construidas mostraron un comportamiento de ancho de banda desplazado 80MHz pero que excede lo especificado por la tecnología 4G – LTE.

5 Referencias

1. Liang, Z. et al: Simulation and design of multi-band planar meandered monopole antenna for mobile phone application. In: 2012 International Conference on Microwave and Millimeter Wave Technology (ICMMT). IEEE Conference Publications. 1-4.
2. Mehdipour, A. et al: Compact Multiband Planar Antenna for 2.4/3.5/5.2/5.8-GHz Wireless Applications. In: Antennas and Wireless Propagation Letters, IEEE. (2012) 144-147.
3. Cihangir, A. et al: A novel multi-band antenna design with matching network for use in mobile terminals. In: 2012 6th European Conference on Antennas and Propagation (EUCAP). IEEE Conference Publications. 1667-1671.
4. Wen Tao, Li et al: Novel design of printed multiband antenna for wireless applications. In: 2012 International Conference on Microwave and Millimeter Wave Technology (ICMMT). IEEE Conference Publications. 1-3.
5. Cho-Kang, Hsu et al: Low Profile Multi-band Antenna for Mobile Phones. In: 2012 15th International Symposium on Antenna Technology and Applied Electromagnetics (ANTEM). IEEE Conference Publications. 1-4.
6. Zhang, Y. et al: Miniaturized planar UWB antenna on a trapezoidal ground fed with a tapered microstrip line. In: Microwave and Optical Technology Letters. Volume 54. No. 11 (2012). 2468-2472.
7. Salmani, Z. et al: Design of log-periodic dipole array inspired antenna with very wide bandwidth. In: Microwave and Optical Technology Letters. Volume 54. No. 11 (2012) 2446-2450.
8. Moradhesari, A. et al: Band-notched UWB planar monopole antenna using slotted conductor-backed plane. In: Microwave and Optical Technology Letters. Volume 54. No. 10 (2012) 2237-2241.
9. Siahkal-Mahalle, B.H. et al: A new design of small square monopole antenna with enhanced BW by using cross-shaped slot and conductor-backed plane. In: Microwave and Optical Technology Letters. Volume 54. No. 11 (2012) 2656-2659.
10. Lin, S. et al: A UWB printed dipole antenna and its radiation characteristic analysis. In: Progress In Electromagnetics Research C. Volume. 31 (2012) 83-96.

Privacy threat warning in eHealth events wireless sensor networks transmissions

Maria de los Angeles Cosio León, Juan Iván Nieto Hipólito, Raymundo Buenrostro Mariscal, and Mabel Vazquez Briseno

Autonomous University of Baja California, Faculty of Engineering Architecture and Design, Telematic Research Group,
Carr. Tijuana-Ensenada km 103, 22860 Ensenada, Baja California, México
{maria.cosio, jnieto, rbuenrostro, mabel.vazquez}@uabc.edu.mx
http://telematica.uabc.mx
Paper received on 22/09/12, Accepted on 15/10/12.

Abstract. New technological paradigms as Wireless Sensors Networks (WSNs) are proposed to a broad application spectrum to support sensing infrastructures. These kinds of networks gather, process and transmit data with restrictions about energy, computing capabilities and transmission rates. A field where WSNs have found a comfortable application area is in eHealth. Patients wear devices with mobile capabilities sending by consistent time intervals an invariant amount of bits about physiological parameters. Due to data transmission patterns on WSNs, as well as sensed data and considering medical restrictions, privacy issues emerge. In this paper we describe a privacy threat about trajectories built while eHealth events are transmitted. Once, we do a qualitative analysis about the threat, we model it from an adversarial viewpoint by proposing an attack model whose aim is to uncover identifiable personal information and other highly sensitive information.

1 Introduction.

WSNs are a conjoined set of devices used to collect, process and transmit events gathered from their context. In this work, we analyzed a public WSN containing fixed nodes to bring support for transmissions of mobile nodes' events. The latter set of nodes are worn by patients getting in and out from WSN' coverage. The aforementioned interactions leave trails containing rich information about the mobile actors in the scenario. A trail is a trajectory described by a mobile device after an event transmission —referring event as a set of bits produced by sensing devices about physiological parameters—. Event transmissions spread information about the data's sources in near real-time such as a time when a transmission starts, type and length of the event and their trajectory. Interactions among the WSN devices and mobile devices can allow for the building of trajectory history and deduction of personal information from the connection between the mobile device and the person wearing it. Information is strongly related to four basic *Context's* elements: location; identity; activity and time [1].

The aforementioned scenario open an opportunity for a myriad of services based on location as well as activities performed by patients. Hence, unauthorized people could

Mireya García-Vázquez, Grigori Sidorov (Eds.).
Advances in Computing Science and Control.
Research in Computing Science 59, 2012, pp. 253- 263

be interested in gaining access to private information. The scenario imposes challenges to protect people's privacy and further security issues. In order to show potential damage that an adversary could cause after gathering enough information of physiological events [2], we do a qualitative analysis about the threat related to primary indices uncovered by the trajectory history. We model it from an adversarial viewpoint, therefore, proposing an attack model whose aim is to uncover identifiable personal information and other highly sensitive information.

Organization Section 2 introduces works resolving privacy threats on transmissions events and the proposal about trajectory reduction. The System Model is described in Section 3, including the network and the data model. Section 4 describes a privacy threat of interest being our main aim in this paper, adversaries and concluding with the Attack model. In section 5 we outline our approach to resolve the threat previously described. Section 6 has a discussion about the privacy issue and our approach; finally, we close this paper in section 7 with Conclusions and future work.

2 Related Work.

Thwarting privacy damage could be accomplished through privacy enhancing techniques (PETs). Cipher is a PET to provide content privacy; in [3] is proposed an instance of these sort of mechanisms. Although it is not enough to provide protection against contextual privacy threats [4] in WSN, as transmissions are an immediately sensed event, they follow routes built by routing algorithms and consider specifically a known restrictions up to sink node (see Figure 2). So an adversary could gather information about existing transmission.

Other kinds of mechanisms achieve privacy through policies that restrict access to data. However, policies are vulnerable to inadvertent or malicious disclosure of private information as described in [5]. With the same aim, there exists anonymity techniques. Their aim is making it impossible to connect data with the data's owner. However, they have limitations; specifically in spatial application domains. Person's identity can be inferred from his or her location. Pseudonymity maintains information restricted for a subset of users. But, anonymity as well as pseudonymity are vulnerable to data mining. Further, anonymity presents a barrier to authentication and personalization, which are a paramount requirement on eHealth services.

A new branch to protect privacy in WSNs is proposed in [2]. In order to protect events observed at different instants that could be and related among them to conclude information about the data's source, (e.g. events caused by the same entity); authors propose a solution considering events with the following characteristics: (i) mobile; (ii) Start from the outermost part of the network —the perimeter—; and (iii) after a given (non predictable) interval of time expires in some place within the network. To measure solution's performance, authors used *communication overhead measure*. Also, communication and computational cost incurred through: (i) the amount of messages produced by the protocol; and (ii) the amount of messages used to forward an event itself (real or dummy) to the base station (BS). The Protocol correctness was measured using Pearson X^2 with the aim to know the association grade between two sets and them the un-observability provided. In [6] authors address the problem of protecting

query privacy (e.g., hiding which node matches the query) and data privacy (e.g., hiding sensed data). They introduce a realistic network model and two novel, adversarial models; resident and non-resident adversaries. For each of them propose a distributed privacy-preserving technique and evaluate its effectiveness via analysis and simulation.

In [5], authors use obfuscation techniques, degrading data quality about location in order to protect person's location privacy while he uses location base services (LBS). To prove their proposal, obfuscating data feed a location-based service of acceptable quality. Their experiments show there exists a trade-off between quality of service offered and the privacy level. Key information is the size of the initial obfuscation set; although their results show that obfuscation techniques are adequate to resolve the person's location privacy problem. Authors conclude that there is needed a deepest research about some setup parameters to achieve a good trade-off between privacy and quality of service.

Besides the above mentioned proposals, there exist WSNs applied to know animal habits; they are constantly sending information about animals location. Considering this scenario, authors of [7] propose the question: *How many transmissions are possible to reduce trajectories without losing our solutions intention?*, in reference to reducing energy consumption by these sort of application. They answer the question, tackling the problem through a distributed variant of the Douglas-Peuker heuristic for polyline reduction, from *Computational Geometric literature*, augmented with temporal awareness. To control the quality of their solution, authors settled error boundaries about location of Object of Interest.

Above mentioned works provided us insights to resolve the problem proposed in this paper; though they are not aware of satellite information spread by data transmissions from observed entities (e.g. trajectories performed by low mobility devices or eHealth events appearing in fixed intervals can be related to physiological measures by basic mathematical operations over transmitted packets, including events that expire at any node in the WSNs). Being aware of above mentioned characteristics, an entity has capabilities to deduce personally identifiable information (PII) as well as sensitive information. Therefore, our aim is to show a contextual privacy threat produced by additional information spread by eHealth event transmissions on WSNs.

3 The System Model.

The system model has four basic actors: (i) a mobile node; (ii) a public WSN; (iii) a sink node; (iv) an eHealth system in Internet, besides of interactions allowed among them (see Figure 1 for details). The first actor could have sensing capabilities, as sensing devices installed on it. The mobile device's profile allows to configure sensing intervals, as well as quality (maximal bits by sample). Both features are carefully monitored by the eHealth system. Hence, if an event is not received or accomplished with characteristics aforementioned, an alarm is activated to notify potential emergency or device failure. Also there exists a mechanism to control packet loss parameter up to where there is no information damage. Public WSN is supporting eHealth events transmissions, described in detail at Network Model below. A sink node extends WSN coverage to Internet making a vertical hand-off.

The eHealth system includes a set of control mechanisms to safely interact with the actors in the network. The first barriers are authentication and authorization subsystem managing access to upper layer services such as data storage, data processing, routing and others.

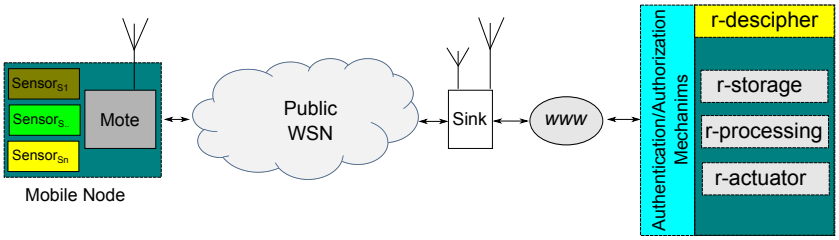


Fig. 1. System model

3.1 The Network Model.

There exists a public WSN that supports eHealth events transmissions from patients wearing devices with mobile capabilities. It is deployed over a finite area forming a grid. So patients could get in or out of the coverage as they perform their activities. We denote the WSN as G. It is defined by the pair $< N, E >$; where N is the set of nodes in G containing four nodes' subsets; $N = \{I, P, M, S\}$. The subset of *inside nodes* defines a *connected dominating subset* in the WSN denoted as I. Another subgroup of nodes is P, corresponding to *perimetrical nodes* in the WSN (graph search and the graph frontier problem). Third subset is M, the group of mobile devices denoted as $M = m_1, \ldots, m_{|M|}$. These sort of devices are $M \notin N$, when there are no transmissions requested. The subset's size M is related to t (a time instant), hence $M_t = |M|$, the number of mobiles nodes connected through some node of $(N - (M + S))$. Offering a broad vision about the network, Figure 2, shows nodes in the scenario as well as the trajectory followed by a mobile device performing hand-off processes to transmit an eHealth event, and the set of routes built by a routing protocol to drive packets up to sink node.

Without restrictions in low layers (MAC and physical) to get into the network a mobile device can start to have interactions with nodes in the subset $N - M$. However, restrictions exist in the upper layer to provide services such as routing R. Therefore, an authorization and authentication mechanism should be successfully accomplished by the mobile node $m_i \in M$ as Figure 1 shows. Once m_i fulfils the entire security requirements, the first interaction between $m_i \; and \; r\_Services$ is routing to build and activated routes, getting them ready for transmissions. Lastly, S is the subset of sink nodes, following an uniform distribution in the WSN. The subset S has a size $|S| \geq 1$.

Besides the aforementioned nodes subsets, in G exists E, the set of sensed events. They are produced by M and transmitted by the subset of nodes $(N - (M + S))$.

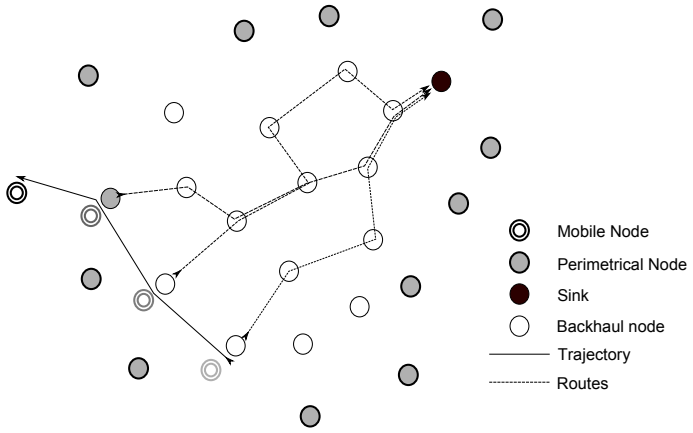


Fig. 2. Trajectory followed by a mobile node, as well as routes carrying packets up to sink node

3.2 The Data Model.

Previous considerations: Mobile devices turn-off the ratio when there is no necessity to transmit data and turn on at fixed time intervals to request possible queries on the eHealth system. About the type of events; we are focusing on ones which: (i) initiate on a random within the WSN (perimetrical or inside nodes); (ii) after a known amount of bits transmitted finishes within the network. It is possible that a mobile device does not finished to send the event due to transmission starts near perimetrical nodes and it losses WSN coverage; (iii) a trajectory in network nodes are a projection of patients' trajectory in physical space.

Data interactions as well as nodes characteristics could be modelled as follows. Mobiles nodes M have at least one sensing device, hence a set of sensing devices are defined as $< st_1, st_2, \ldots, st_n >$; where n is the number of sensing devices allowed in m_i. Sensing tasks are configured off-line, considering medical standards to measure physiological parameters. Sensing interval features allow to define the time interval between two sensed events. Hence this feature related to entire sensing device by node could be defined as follows: $< st_{1t_1}, st_{2t_2}, \ldots, st_{nt_n} >$, and a sensing device $< st_1 >$ will produce an event once a time t_1 is elapsed. Event length is related to a kind of sensing device, so a sample is defined by a triad $< St, T, \mathbb{L} >$ where St is a class of sensing device, T is the interval between samples and $\mathbb{L}$ is event's length in bits. Figure 3 describes a set of events, as well as packets required by the event.

4 Threat: Trajectories.

Based on Kerckhoff Principles [8], adversaries have knowledge about network protocols and privacy mechanisms. Besides the characteristics about people wearing mobile nodes (e.g. old or sick people, walking slow). The adversary performing a passive attack could gather information about hardware and data. Through samples resolution and

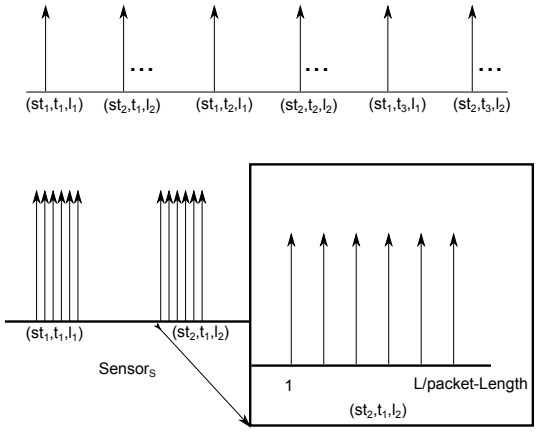


Fig. 3. Samples transmissions

type of event analysis. The same attack class could evolve into active attacks in the network, also in the real world considering that monitoring tasks are carefully configured and monitored. So changes in transmissions patterns could be detected as emergency events or node failures.

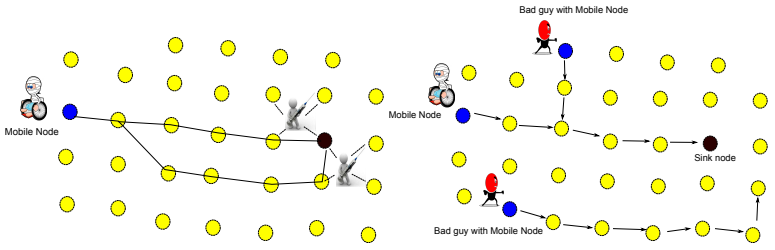
Patterns of eHealth events in WSN transmissions could be classified as follows: transmission domains (private or clear); source node's trajectory (trajectory on a set of nodes or in a simple node); time (periodic or single events); and lastly, size of event (time intervals to transmit data or bit numbers). Complete information is a valuable asses for business people. On the other hand, location is inextricably linked to personal safety, so unrestricted access to information about patient location could potentially lead to harmful encounters (for example stalking or physical attacks). We consider a sequence of recorded locations for a person constitutes a quasi-identifier that can be used in combination to identify the user [9]. So trajectories can be used to infer much about a person, even without a name attached to the data. We are aware about it, so our goal is to show a strategy to perform attacks using a trajectory's history and event characteristics with intention to deduce PII (personal identifiable information). In intention to build mechanism to thwart possible privacy damage.

4.1 Attack Model.

In this subsection we define the types of adversaries and kind of damage that could be inferred by them, as a result of related adversarial behaviour and events along the WSN. To uncover vulnerabilities on trajectories, we model events considering them in order to make relations between trajectories and the adversaries.

Adversaries. We are considering two types of adversaries. The first one is an external, passive and global coverage adversary. This means that the adversary is no part of WSN, she has her own device to sniff traffic and to gather all information about events

in the network (e.g. node, laptop with capabilities to store trajectory records). Her aim is to know identifiable personal information and other highly sensitive information. To achieve her aim, the adversary uses traffic analysis techniques, such as temporarily on transmission and their length. The second one is an active, external and short coverage adversary. It modifies routes in the scenario by making them unfeasible near the sink devices, through *traffic injection*, in order to modify parameters related to routing algorithm (see Figure 4.a). One strategy of this attacker is to drive the traffic through specific nodes to perform a *sink-hole* attack or accessing to data using less resources. The second strategy is to request by routing services an emergency transmission with the intention to reduce emergency service effectiveness. Figure 4.b shows this strategy. Once the adversary knows how many patients and what class of disease exists, it can perform a *denials of service* attack and reduce service confidence by the users.



(a) Exhausting routes attack (b) Requesting for Routing Services

Fig. 4. Active adversary strategies.

Modelling Events. An event can be described by its length as:

$$e_j = \{< t_x, l_j(t_x) > \ldots < t_{x+\Delta x}, l_j(x+\Delta x) >\}, \tag{1}$$

where

- l_j is St related and l_j has a maximum length defined by sensor type;
- $l_j : \mathbb{T} \rightarrow \mathbb{N}$ is a crescent monotonic function, where its upper threshold is defined $< St, t_{\Delta x}, \mathbb{L} >$;
- t_x is distinct time instant up to $t_{x+\Delta x}$;
- Δx is the elapse time to perform a sensing task.

The trajectory built by a mobile node, whit intention to transmit an event is related to the subset nodes in $(N - (M + S))$ through m_i sending partial or full event's bits. The trajectory along the nodes could be described as:

$$traj_e = \{< n_i, t_x >, \ldots, < n_{i+h}, t_{x+\Delta x} >\}, \tag{2}$$

where n_i is a node from a mobile device starts to send an event up to n_{i+h}, h hops away from n_i in mobile node's trajectory. It is important to denote that if the node

$n_{i+h} \in P$ and $l_j \notin Ls$, where Ls is the set of event size allowed. Then is possible to conclude that the mobile node lost network coverage. Therefore, trajectories related to unfinished transmissions are not adequate to define the event type. On the contrary, it is useful information to relate physical locations in the scenario at t_x along of events with $l_j \in Ls$. So, a trajectory describes a set of physical locations from a mobile node in a period $T_{0,x}$ while it is transmitting a class of physiological lectures defined through Ls.

The relation ship of equations 1 and 2 can offer valuable information about events, allowing to evolve into natural language expressions as (e.g. Patient is walking in the park while sensing heart rate). Continuous tracking services allow to related location with activities of daily living as an opportunist for sensing relevant information by medical personnel and caregivers or adversaries described above.

Uncovering Information from Trajectories. Authors in [2] propose Equation 3 to acquire events with a length of interest; we will use it with the same goal. Now, given a time instant $t_i \in \mathbb{T}$ and a subset of events E, the function $length_E : \mathbb{N} \times \mathbb{T} \rightarrow \mathbb{N}$ will provide the number of events with ls_i at the t_i:

$$length_E(l_i, t_i) = \sum_{e_j \in E} \epsilon_{e_j}(t_i, l_i), \tag{3}$$

$$where\ \epsilon_{e_j}(t_i, l_i) = \begin{cases} 1 \text{ if } < t_i, l_i(t_j) > \in e_j \\ 0 \text{ otherwise} \end{cases}$$

Selecting a point in the network time scale (elapsing time from the network start-up up to stop) of one event type, we will model this relation to show information discover.

A mobile node m_i will send information to the sink node through the subset of nodes $(N - (M + S))$. Each node in $(N - (M + S))$ have a single network interphase and the communication channel is shared by clients along the network time scale, t_i, so interactions among nodes are restricted to time slots.

The history of event transmissions performed by node m_i is a set of ordered pairs with $m_i = \{< t_0, e_0(l_j(t_t)) > \ldots, < t_n, e_n(l_j(t_t)) >\}$. The history of events in nodes $(N-(M+S))$ is defined by a set of ordered pairs $< t_i, e_x(l_j(t_t)) >$, where t_i is a point in the network time scale, e_x, an event x attended by $node_i$ and $l_j(t_t)$ is the $event_x$'s length on the event time scale (time elapsed by event transmission). The trajectory of an event through nodes $< n_i, n_{i+1}, \ldots, n_{i+h} >$ is defined in Equation 4.

$$te_j = \{< n_i, t_i, e_x(l_j(t_t)) >, \ldots, < n_{i+h}, t_{i+\Delta i}, e_x(l_j(t_t)) >\} \tag{4}$$

Where e_x is an event sensed by a mobile device and transmitted through a subset of nodes, $l_j(t_t)$ is the number of bits transmitted by nodes along the mobile node's trajectory in the time interval $T_{i,i+\Delta i}$. Figure 5 illustrates a trajectory followed by a patient while wearing a device that is transmitting an event.

Considering that each node in N has an unique ID, m_i requires to send its unique ID along with the routing query in intention to identify owner request. Here, the system has enough information to detect who is requesting by a service [10]. On the other hand, the restricted service routing, R, once it is authorized, offers a short path between source and destination nodes. At this point, an adversary could get enough information about

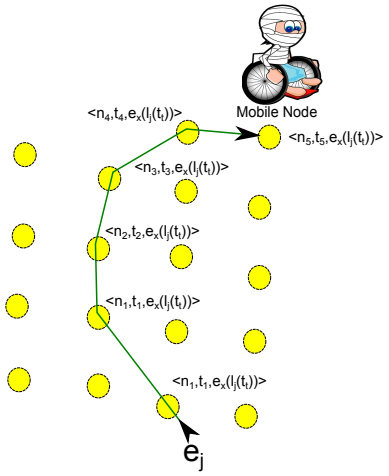


Fig. 5. Event network and physical trajectory

each event and the trial could be defined as $< m_i, St_j, \mathbb{L}, T >$; where $0 < j \leq |St|$. These proposed models allow to relate nodes with events, to filter by event type and to denote m_i locations.

5 Outlining Our Solution

Our system model has security considerations on upper layers outside of the WSN. To control access to remote services (r-services), there exists a r-service provider it includes subservices of: authorization; authentication; decipher; storage; processing and actuator subsystem shown in Figure 1. Data transmissions along the network are performed following routes that build a bioinspired algorithm as AntNET [11]. Due to routes been activated by forwards ants (ants tracking back to path to source node) instead of explorer ants (ants searching the sink node), allows the sink node to have control over the activation process. Routing process activation requires non-reusable tokens from mobile nodes, so an initial token is stored in the mobile device. When a mobile node requests a routing process, it sends hashed token, $h^y(token)$. Once the sink node receives the token, it sends it up to internet system. After this, the system provides a token, $h^x(token)$, and sink nodes wait for the same token from the mobile node to activate a routing process as Figure 6 shows. In other cases, nodes around the source are notified about a possible attack.

We assumed that nodes form a grid. this grid is divided by set regions, A. For each region of a_i, there exists a node called a sink. This node has capabilities to extend Internet services into the WSN. This device has a coverage area defined as: $sink_i =< area_i, s_i >$, where $area_i$ border is CA hops away from $sink_i$. Further, CA defines the maximum coverage area of $sink_i$, and $m_i t$ denotes people in the area at t using n_i sensor node to transmit its data along $area_i$.

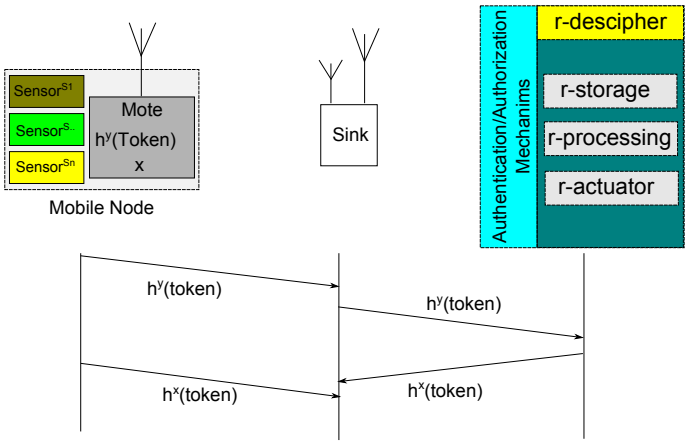


Fig. 6. Chain hash token

Afterwards an event starts and m_i is losing n_i's coverage. It starts a hand-off request, instead to build a new route. Current route is extended using previous knowledge provided by the hand-off mechanism about destination node. However, making a temporal analysis on transmissions, it is possible to know about this change. Thus we propose to introduce a random delay on transmissions in order to prevent adversary knowledge about the change.

6 Discussion.

Information by it self has no privacy requirements, although when it is possible to make a relation between information and the owner, a privacy threat appears. Trajectories allow to build aforementioned relation ship, although providing privacy on trajectories is a paramount task, because additional considerations are necessary to protect them. So the satellite mechanism, in the application layer must be aware of privacy and security requirements; in intention to enhance performance on the lower layer privacy mechanisms. This work proposes aforementioned consideration. Although, it is necesary to define the scope of each mechanism besides of their relation ship. Single solutions cut a path, but a broad opportunity to perform an attack persist. This work has challenges to resolve about the privacy mechanism implementation, metrics to quantify damage and the privacy threat about information spread by transmission length. However this research in progress offers information to resolve them, through mathematical models proposed.

7 Conclusions.

This proposal shows a privacy threat related to eHealth events through WSNs transmissions, adversaries as well as their strategies. The scenario above describes that without

privacy considerations, a broad opportunities for unauthorized personnel to gain access to PII is very real. Considering events' length; time; location an adversary has information from at lease two primary indices from context. This increases the likelihood for successful attacks to gather sensitive information by crossing efforts of active and passive adversaries as defined in section 4. Protecting this scenario is a challenging task as once the adversary knows two types of the most important elements of patients' context, the rest can be deduced through them [12]. Our initial effort tries to cover time elapsed by hand-off mechanism. Although events length requires additional techniques that must be thoroughly analyzed for its performance. The future work in this project includes simulation, hand-off delay considerations, as well as metrics to quantify damage.

References

1. B. Schilit, N. Adams, and R. Want. Context-aware computing applications. In *Proceedings of the 1994 First Workshop on Mobile Computing Systems and Applications*, WMCSA '94, pages 85–90, Washington, DC, USA, 1994. IEEE Computer Society.
2. S. Ortolani, M. Conti, B. Crispo, and R. Di Pietro. Events privacy in wsns: A new model and its application. In *World of Wireless, Mobile and Multimedia Networks (WoWMoM), 2011 IEEE International Symposium on a*, pages 1 –9, june 2011.
3. David L. Chaum. Untraceable electronic mail, return addresses, and digital pseudonyms. *Commun. ACM*, 24(2):84–90, February 1981.
4. Trevor Darrell & Daniel J. Weitzner Mark Ackerman. Privacy in context. *HumanComputer Interaction.*, 16(2-4):167–176, 2001.
5. Matt Duckham and Lars Kulik. Simulation of obfuscation and negotiation for location privacy. In Anthony G. Cohn and David M. Mark, editors, *COSIT*, volume 3693 of *Lecture Notes in Computer Science*, pages 31–48. Springer, 2005.
6. Emiliano De Cristofaro and Roberto Di Pietro. Preserving query privacy in urban sensing systems. In *Proceedings of the 13th international conference on Distributed Computing and Networking*, ICDCN'12, pages 218–233, Berlin, Heidelberg, 2012. Springer-Verlag.
7. Goce Trajcevski, Oliviu C. Ghica, and Peter Scheuermann. Tracking-based trajectory data reduction in wireless sensor networks. In *Proceedings of the 2010 IEEE International Conference on Sensor Networks, Ubiquitous, and Trustworthy Computing*, SUTC '10, pages 99–106, Washington, DC, USA, 2010. IEEE Computer Society.
8. W. Trappe and L.C. Washington. *Introduction to cryptography: with coding theory*. Prentice Hall, 2002.
9. Claudio Bettini, Xiaoyang Sean Wang, and Sushil Jajodia. Protecting privacy against location-based personal identification. In Willem Jonker and Milan Petkovic, editors, *Secure Data Management*, volume 3674 of *Lecture Notes in Computer Science*, pages 185–199. Springer, 2005.
10. Naveen Sastry and David Wagner. Security considerations for ieee 802.15.4 networks. In *Proceedings of the 3rd ACM workshop on Wireless security*, WiSe '04, pages 32–42, New York, NY, USA, 2004. ACM.
11. G. A. Di Caro. *Ant Colony Optimization and its application to adaptive routing in telecommunication networks*. PhD thesis, Faculté des Sciences Appliquées, Université Libre de Bruxelles, Brussels, Belgium, November 2004.
12. Anind Dey, Jeffrey Hightower, Eyal de Lara, and Nigel Davies. Location-based services. *Pervasive Computing, IEEE*, 9(1):11 –12, jan.-march 2010.

Identificación de riesgos de sequías y sismos al noroeste de Baja California utilizando Sistemas de Información Geográfica

Michelle Hallack-Alegría[1], Mario González-Durán[1], Mauricio Peregrina-Llanes[1]

[1]Centro de Ingeniería y Tecnología, Universidad Autónoma de Baja California, Valle de las Palmas, Tijuana, México

{[1]mhallack, [1]gonzalezduranmario, [1]mperegrina} @uabc.edu.mx
Paper received on 02/10/12, Accepted on 25/10/12.

Resumen. Los sistemas de información geográfica (SIG) son una tecnología de la información y comunicación en la que se despliega información geo-estadística sobre un mapa, así como el procesamiento de imágenes digitales obtenidas por percepción remota. El empleo de los SIG, en este estudio, tiene como objetivo principal utilizarlos como herramienta auxiliar en la detección e identificación de factores relacionados con la manifestación de eventos extremos relacionados a sequías, y sismos, así como visualizar en espacio y tiempo, la magnitud y la frecuencia de tales eventos. Con los resultados obtenidos se han inferido zonas de riesgo en la zona Este del Municipio de Tijuana, Baja California, como preámbulo para la microzonificación y además se han utilizado en la construcción de Atlas de riesgos para Latinoamérica. Finalmente, la generación de conocimiento aplicado a la protección de la población civil ante los diversos riesgos naturales a los que estamos expuestos fortalecen las bases de datos y la toma de decisiones en los organismos operadores de los recursos hídricos y del Centro Nacional de Prevención de Desastres.

Palabras clave: Sistemas de información geográfica, eventos extremos, sequías, sismos, riesgo.

1 Introducción

En el extremo noroeste de México confluyen dos riesgos ambientales que consideramos necesario abordarlos desde una perspectiva espacial, con el fin de visualizar las magnitudes y también la extensión superficial en la cual la población es vulnerable; estos son riesgos hidrometeorológicos del tipo sequías y los riesgos de carácter geológico, particularmente los sismos. La motivación de este trabajo es hacer una valoración de las herramientas con las que se capturan y manipulan datos como partida para el análisis de los elementos que determinan la probabili-

Mireya García-Vázquez, Grigori Sidorov (Eds.).
Advances in Computing Science and Control.
Research in Computing Science 59, 2012, pp. 267-276.

dad de riesgo en una zona o región. Uno de los principios fundamentales en el estudio de los riesgos naturales es que el aumento de la población intensifica el efecto del riesgo [1]. Se entiende por riesgo la probabilidad de que se presente un nivel de consecuencias económicas, sociales o ambientales en un sitio en particular y durante un periodo determinado, resultado de interacciones entre amenazas y condiciones de vulnerabilidad. A su vez, la amenaza son los eventos probables con capacidad de generar daño sobre unos elementos con limitación o incapaces de soportar, adaptarse o resistir a dichas amenazas (vulnerabilidad) [2]. Por las condiciones del relieve accidentado, clima árido y semiárido [3], como también aspectos geológicos que enmarcan la región noroeste hacia el municipio de Tijuana, existe la ocurrencia de diferentes fenómenos físicos como sismos, deslizamientos, inundaciones y sequías; que sumados a las condiciones de crecimiento poblacional sostenidos, migración, degradación ambiental, rápida urbanización y cambio climático, ofrecen un caldo de cultivo para escenarios de riesgo complejos y desastres naturales inminentes ante lo cual es importante producir información que primero: identifique factores de riesgo; posteriormente analice causas, consecuencias y se realicen estimaciones de la magnitud de los riesgos, y finalmente se produzca una comunicación del riesgo ante autoridades, academia, población y público en general.

1.1 Riesgos hidrometeorológicos

Hoy en día, la escasez de agua es una realidad que afecta la mayor parte del mundo. Mientras algunas regiones del Planeta históricamente se han enfrentado a sequías constantes, otras, como un efecto del cambio climático en los últimos años, las han enfrentado de manera irregular. La cada vez más frecuente incidencia de períodos secos, aunado a la creciente demanda de agua debida al crecimiento poblacional, la contaminación indiscriminada de aguas superficiales y subterráneas, entre otros aspectos, obligan a la formulación de planes de manejo sustentables de este recurso vital. En regiones con climas semidesérticos que cuentan con un intenso crecimiento económico y poblacional, así como una floreciente actividad agrícola la planeación de recursos hídricos se vuelve indispensable para fomentar el crecimiento y mantener la vida a largo plazo [4].

En el caso específico de México, los recursos hídricos son generalmente abundantes, se tienen un aproximado de 1511 kilómetros cúbicos de agua cada año, de los cuales aproximadamente el 72%, de esa agua de lluvia se evapora. La mitad de las precipitaciones que ocurren en el país se concentran en la parte sur, que representa una quinta parte del área superficial total. En el extremo opuesto, solamente el 25% de esta precipitación ocurre en la parte norte del país, la que representa el 50% del área superficial [5]. El estado de Baja California se clasifica dentro de los más secos con un promedio aproximado de 202 mm de precipitación anual [6], lo que lo ubica en una situación vulnerable ante eventos de sequías.

1.2 Riesgos Geológicos

De acuerdo con CENAPRED [7] este riesgo se presenta en función de: peligro, vulnerabilidad, y exposición ante la manifestación de movimientos de masas de terreno y acciones sísmicas. El Municipio de Tijuana, B.C., se localiza en una zona geomorfológicamente condicionada por fallas y lineamientos geológicos, que a su vez representan áreas de liberación de energía en la presencia de temblores.

La zonificación de riesgos sísmicos en las zonas urbanas, denominados microzonificación sísmica, es el primer paso y el más importante así como arduo hacia un análisis de riesgo sísmico, y es una estrategia de mitigación en las regiones densamente pobladas. En zonificación es posible observar cuantificadamente, la variación espacial de la respuesta del subsuelo a un terremoto típico que se puede esperar en la zona [8].

En el zona este del Municipio, se localiza el desarrollo habitacional Valle San Pedro, que de acuerdo con el Programa de Desarrollo Urbano del Centro de Población de Tijuana, 2010 – 2030, en los próximos veinte años albergará una población aproximada de 680,000 habitantes, con los usos de suelo habitacional, industrial, educativo y salud entre otros, en una zona donde la probabilidad recurrente en la manifestación de sismos esta presente; por lo que es necesario validar el nivel de riesgo al que se encuentra sometida la región y la exposición de la población que se espera se albergue en la zona habitacional con proceso de construcción desarrollado a base de mampostería y concreto reforzado. Creemos necesario hacer investigación sobre zonas urbanas desarrolladas y nuevos desarrollos urbanos para que sea posible extender los resultados a la aplicación de análisis antisísmico de edificaciones para estudios presentes, y futuros cambios en las normas y reglamentaciones del estado de Baja California.

2 Materiales y métodos

2.1 Contexto espacial

Al definir el área de estudio se presentan ambos trabajos de identificación de factores de riesgo en diferentes escalas. Por un lado se utilizó un análisis para definir la frecuencia del riesgo de sequías en la región a través de un periodo de retorno de lluvia de 10, 50 y 100 años. Se consideró la extensión aérea estatal de Baja California y una porción de Sonora para tomar en cuenta mayor número de datos en estaciones climatológicas. Por otra parte los elementos del análisis de las zonas con periodos de vibración utiliza una extensión local con el fin de distinguir anomalías en una serie de datos sobre un mismo tipo de suelo o formación edafológica por lo que se hace mención de una escala regional que abarca los estados del noroeste de País, para el análisis del fenómeno de sequías y una escala local para el fenómeno de la sismicidad (Fig. 1).

2.2 Datos

Sequías. El estado de Baja California, situado en el extremo noroeste de México, se encuentra en un área semidesértica por lo que se caracteriza por ser una zona donde las sequías pueden ser un fenómeno normal y esperado.

Se analizó la precipitación media anual para 68 estaciones con registros en el periodo de 1950 – 2008, proponiendo la producción de mapas de precipitación que auxilien en la detección de sequías o inundaciones a partir de los datos obtenidos de un análisis estadístico que permitió definir la frecuencia con que se presentan las sequías para la región noroeste de México que incluyó los estados de Baja California y Sonora.

Para realizar el análisis el primer paso es obtener la base de datos de las variables a utilizar. Para el caso particular del estudio de sequías, se incorpora las precipitaciones de los sitios ubicados en la región de estudio como medio para obtener indicadores de aridez al utilizar valores mensuales o anuales para calcular el déficit anual mensual pluviométrico [3]. En este estudio los datos utilizados son de precipitación media anual y han sido extraídos del Extractor Rápido de Información Climatológica III.

Se utilizó un Sistema de Información Geográfica (SIG) para crear mapas de precipitación media para el área de estudio. Se eligieron dos métodos de interpolación espacial como herramientas geoestadísticas, los cuales presentan una metodología adecuada para esta zona en particular. El primer método, Inverso de la distancia ponderada (Inverse Distance Weighted, IDW), es un método determinístico que mantiene que un valor estimado de un punto es influenciado por los puntos conocidos mas cercanos que por aquellos mas alejados. El segundo, Kriging, es una técnica que asume que la variación espacial de un atributo no es ni determinístico ni estocástico en su totalidad. En su lugar, considera que la variación espacial consiste en tres componentes principales: una tendencia espacial, representando la variación de la variable a regionalizar, una estructura, y un término aleatorio de error. La presencia o ausencia de la estructura y la interpretación de la variable aregionalizar han guiado al desarrollo de diferentes métodos de Kriging para inter polación espacial: Ordinary Kriging, Simple Kriging, y Universal Kriging. Para

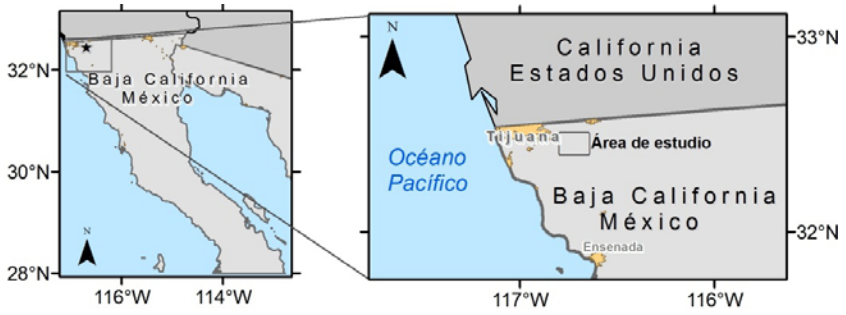


Fig. 1. Contexto espacial de la escala regional para la identificación de periodos de retorno de precipitación y representación de la escala local en el área de estudio de periodos fundamentales de vibración

este estudio, se eligió el mapa de precipitación arrojado por el método Simple Kriging, el cual se considera muy útil cuando se dispone de pocos datos de muestreo [9].

Sismicidad. Al representar geo-espacialmente los datos espaciales que tienen que ver con los elementos de riesgo sísmico a identificar, se llevo a cabo un levantamiento de los periodos de vibrar en diferentes puntos sobre el terreno capturando su información geográfica con un navegador GPS dentro del desarrollo inmobiliario "Valle de San Pedro" ubicado al suroeste de la ciudad de Tijuana.

Se investigó la variación en la respuesta sísmica de la superficie, para determinar donde el suelo se amplifica a un nivel que pueda causar daño estructural añadiendo atributos a los sitios puntuales como preámbulo a la obtención de mapas de microzonificación del periodo fundamental de vibrar en el suelo, parámetro que se amplifica cuando ocurre un sismo y esta vinculado al comportamiento estructural de las construcciones lo cual puede reflejar daños, perdida de rigidez y aparición de grietas.

La distribución espacial de los registros hechos fueron plasmadas en un mapa utilizando un SIG, que permitiera resaltar los períodos de acuerdo un rango, lo que permitió visualizar tendencias utilizando diferente simbología de acuerdo al campo de clasificación de la variable representada y la diferenciación de ésta medida en campo.

2.3 Métodos

Los diferentes aspectos de la representación geográfica que son aplicables a los SIG combinan formas geométricas definidas: puntos, líneas y polígonos con datos estadísticos asociados que determinan una forma y atributos deseados. En los dos casos trabajaremos con una capa de puntos que corresponden a las estaciones climatológicas y los sitios de muestreo de vibraciones. Se tomaran dos capas como base cartográfica en donde se distinguen: polígonos municipales, polígonos estatales, y polígonos de regiones fueron usadas y producidos para la producción de cartografía temática.

Periodos de retorno. El archivo de forma o *shapefile* que contiene la información de las estaciones ordenadas de acuerdo a las regiones identificadas por su espacio geográfico con características climatológicas y topográficas similares.

El análisis regional identifica la distribución de probabilidad que presente el mejor ajuste a los datos de precipitación media anual de diversas estaciones en una región, para estimar cuantiles (los valores de precipitación anual asociados a un determinado valor de probabilidad de ocurrencia) [10], con el fin de representar periodos de retorno de 10, 50 y 100 años, para lo cual representamos la cantidad de lluvia esperada por región en porcentaje del promedio de precipitación anual. La metodología utilizada maneja los estadísticos de orden, L-momentos. Utilizan-

do algoritmos Fortran proporcionados por Hosking y Wallis [11] y el software L-Rap de MGS Engineering Consultants.

Periodos fundamentales de vibrar. Se caracterizó la respuesta sísmica sobre la superficie del nuevo asentamiento en edificación, Valle de San Pedro, en cuanto a los periodos dominantes del movimiento del suelo en treinta y nueve puntos separados entre sí aproximadamente 100 metros. Se obtuvieron registros de vibración ambiental con un acelerómetro, compuesto por una grabadora y un sensor triaxial el cual se colocó directamente sobre el terreno orientando al norte la dirección X del sensor en todos los puntos. Posteriormente, se procesaron los registros obtenidos en formato binario al convertirlos a código ASCII con los cuales se formuló una base de datos que contiene: los campos que identifican el número del punto, el atributo del periodo de vibración en segundos y las coordenadas geográficas tomadas con un navegador GPS.

Representación en el SIG. El sistema de información geográfica para la identificación de sequías agrupó por medio de una geometría poligonal, las regiones definidas por las pruebas de homogeneidad. Se hizo uso de la herramienta polígonos de Thiessen, con la cual se generaron equidistancias entre los puntos que definen las estaciones climatológicas para designar el área de influencia de cada estación, como referencia para establecer los límites entre regiones y generar la capa que contiene cuatro polígonos según cada región, asignando un atributo de lluvia promedio anual que fue etiquetado sobre cada polígono para la representación de los periodos de retorno a través del uso de una leyenda que determina el valor en porcentaje del volumen de lluvia promedio según el caso.

En el despliegue de las magnitudes de los periodos de vibración se utilizó una simbología definida por un ícono en forma de rayo, en el cual, se define el tamaño de representación en función del valor de la variable asignada al campo del periodo fundamental de vibrar. Se realizó de igual manera una interpolación espacial en función de la misma variable para determinar la influencia espacial de los datos puntuales con el fin de desplegar una microzonificación representada en un rango que va de los 0.01-0.03 segundos usando una rampa del color negro al blanco respectivamente.

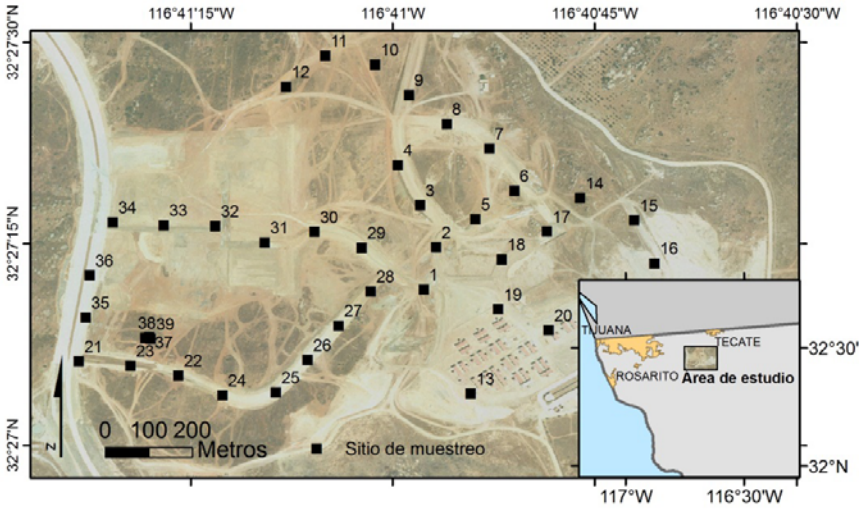


Fig. 2. Sitios de muestreo de periodo dominante del movimiento del suelo en la zona del desarrollo habitacional Valle de San Pedro, en la localidad Valle de Las Palmas Tijuana, B.C.

3 Resultados y Conclusiones

Se obtuvieron mapas de precipitación media anual (Fig. 3), donde se observa que el noroeste de Baja California cuenta con una alta variabilidad en la precipitación media anual, con un rango que va de 50 mm a 500 mm entre estaciones.

Se muestran los resultados obtenidos a través de la exhibición de los tres periodos de retorno 10, 50 y 100 años entendiendo que la posibilidad de tener un año con valores de precipitación de una tercera o una cuarta parte del promedio es de cada 10 años para la región tres y cuatro respectivamente mientras en que en la región uno en ese mismo periodo de tiempo es recurrente que se presente un año con poco menos que la mitad (44%) de lluvia que la registrada en un año promedio (Fig. 4)

Para la región dos el periodo de retorno de 100 años muestra un valor nulo, por lo que consideramos que es debido a manifestarse en la región árida y con el régimen de lluvias menor del registro por lo que puede considerarse que cada siglo existe la posibilidad de un año con precipitación mínima o sin registro de medición.

La escasez de información para la gestión, uso y manejo del agua, así como los cambios climáticos que esta sufriendo la naturaleza hacen de eventos como la sequía un problema de carácter catastrófico ya que pueden causar daños graves al suelo, cultivos, y otras actividades humanas [12]; de aquí la importancia de contar con herramientas que faciliten el monitorear su inicio, terminación y frecuencia.

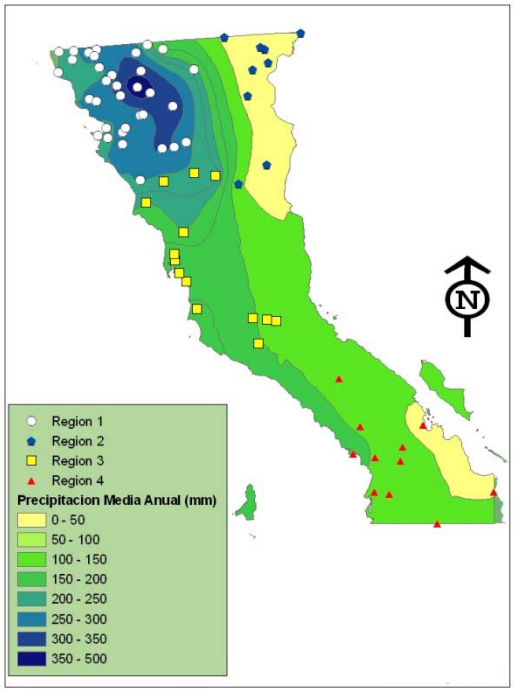


Fig. 3. Precipitación promedio anual y ubicación de estaciones climatológicas en el estado de Baja California

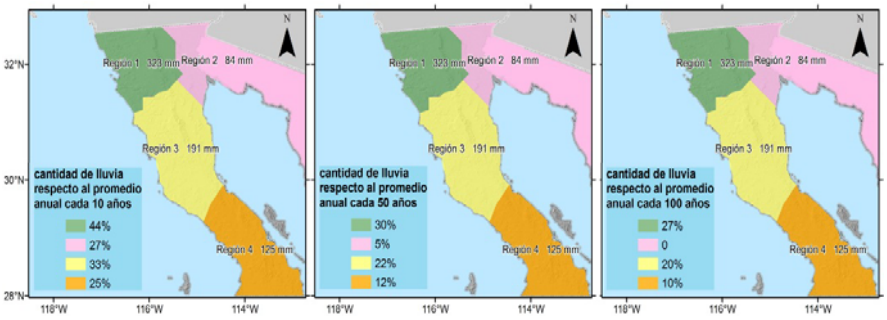


Fig. 4. Periodos de retorno a 10, 50 y 100 años en el porcentaje de precipitación respecto al valor medio anual

Se obtuvo un mapa con la distribución de los periodos dominantes del movimiento del suelo para la zona de estudio, observando la presencia de periodos relativamente cortos menores a 3.0 segundos. El período fundamental promedio resultó de 0.02 segundos sobre el suelo y en roca o donde la capa de suelo era ínfima fue de 0.01 segundos. Se muestra la distribución de los puntos registrados y el rango del valor del período fundamental de vibrar del suelo con el fin conocer donde esta vibración puede causar daño estructural donde actualmente se construyen condo-

minios de tres plantas ya que existen variaciones significativas en el valor de esta variable (Fig. 5).

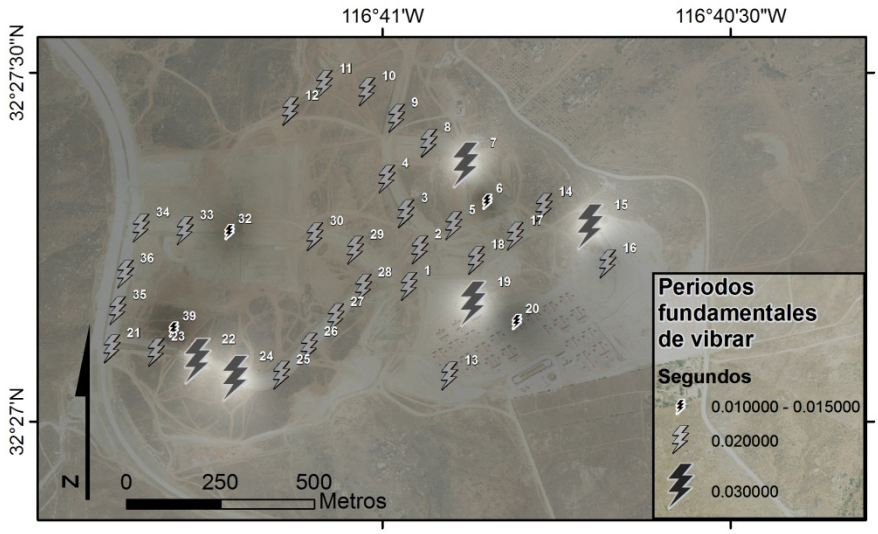


Fig. 5. Periodos dominantes del movimiento del suelo

Los mapas de microzonificación para el desarrollo habitacional completo, serán obtenidos posteriormente ya que se muestreen mayor cantidad de sitios para que pueda ser empleado directamente por los ingenieros de diseño incorporándolo en el análisis antisísmico de las diferentes edificaciones a construir en Valle de san Pedro y zonas aledañas.

Los sistemas de información geográfica proveen una manera de relacionar espacialmente una variable por lo que pueden ser usados para representar mediciones, estimaciones, cálculos, inferir valores y distribuciones de los fenómenos naturales y antropogénicos con el fin de determinar las intensidades y las probabilidades de ocurrencia.

4 Referencias

1. Keller, E. A. y Blodgett R. H. 2004.: Riesgos Naturales. Pearson Educación. Madrid. (2004)
2. Rodríguez de Acosta V.: Riesgo sísmico y Comportamiento Social Revista EIRD Informa (Estrategia Internacional para la Reducción de Desastres) http://www.eird.org/esp/revista/No6_2002/art23.htm
3. Verbist, K., Santibañes, F., Gabriels, D., & Soto, G.: UNESCO ATLAS of Arid Zones in Latin America and the Caribbean. CAZALAC, La Serena, Chile (2010)

4. Hallack-Alegria, M., y Watkins, D.: Annual and Warm Season Drought Intensity–Duration–Frequency Analysis for Sonora, Mexico. Journal of Climate **20**, 1897–1909 (2007)
5. Comisión Nacional del Agua Estadísticas del agua en México 2011, http://www.conagua.gob.mx/CONAGUA07/Publicaciones/Publicaciones/SGP-1-11-EAM2011.pdf
6. Comisión Nacional del Agua. Programa Nacional Hídrico 2007-2012, http://www.conagua.gob.mx/CONAGUA07/Contenido/Documentos/PNH_05-08.pdf
7. Guevara-Ortiz, E., Quaas Weppen, R., Fernández-Villagómez, G., Zepeda-Ramos, O., Muñoz-Hernández, E. y Torres-Palomino L.: Guía Básica para la Elaboración de Atlas Estatales y Municipales de Peligros y Riesgos. Conceptos básicos sobre peligros, riesgos y su representación geográfica, Secretaría de Gobernación, Distrito Federal. México (2006)
8. Slob, S., Hack, R., Scarpas, R., Bemmelen, B. y Duque, A.: A Methodology for Seismic Microzonation Using GIS and SHAKE - A Case Study from Armenia, Colombia in 9th Congress of the International Association for Engineering Geology and the Environment Engineering Geology for Developing Countries J. L. van Rooy and C. A. Jermy, editors. Durban, South Africa (2002)
9. Chang Kang-sung.: Introduction to Geographic Information Systems". McGraw Hill, New York (2004)
10. Nuñez-Cobo J., Verbist K., Ramírez-Hernández J. y Hallack-Alegría, M. Guía Metodológica para la Aplicación de1Análisis Regional de Frecuencia de Sequías basado en L-momentos y resultados de aplicación en América Latina**.** Programa Hidrológico Internacional UNESCO/Centro del Agua para las Zonas Áridas y semiáridas de ALC. Documento técnico No. 27 (2011)
11. Hoskins, J.R.M., y Wallis, J.R.: Regional Frequency Analysis. Cambridge. (1997)
12. Frick, D. M., Bode, D., and Salas, J. D.: Effect of drought on urban water supplies. I: Drought analysis, J. Hydrology Engineering., ASCE, 116(6), 733-753 (1990)

Design and harware implementation of Memory-Polynomial Model based on DSP board

J. R. Cárdenas Valdez[1], J. A. Galaviz Aguilar[1], A. Calvillo Téllez[1], C. Gontrand[2], and J. C. Núñez Pérez[1]

[1] Centro de Investigación y Desarrollo de Tecnología Digital, Instituto Politcnico Nacional ; Av Del Parque 1310, Mesa De Otay, Cp 22150,, Tijuana, Baja California, Mexico
[2] Université de Lyon, INSA- Lyon, INL, CNRS UMR5270, Villeurbanne, F-69621, France.
Paper received on 02/10/12, Accepted on 23/10/12.

Abstract. This paper presents a Memory-Polynomial Model as Special case of Volterra Series implemented in Hardware using a DSP board. The implementation uses Matlab/Simulink, and the DSP Development Kit Cyclone III Edition. The first stage is to develop the model in Matlab/Simulink Environment using the DSP Builder Blockset through the Signal compiler block, after that the design is downloaded to the DSP Board. The results show that this simulation technique is able to prove the effectiveness of the MPM as behavioral model for Power Amplifiers.

1 Introduction

The Power amplifier (PA) is the most important component in a Radio Frequency link, unfortunately it is inherently nonlinear generating spectral regrowth when its behavioral modeling is not properly developed, and the memory and intermodulation effects are not considered during the modeling. The Memory Polynomial Model (MPM) as special case of the Volterra Series is an adequate method to model the PA behavior. The Volterra are power series with memory [1], by its internal architecture tends to increase the required parameters to develop a behavioral modeling for Power Amplifiers, this condition leads to use a special case as MPM improving the computational processing time during the modeling, and they consider undesirable effects as memory and non-linearity before the circuit fabrication. Some works have been presented with reformulated Volterra Series as [2] reducing the processing time. The main idea of this work is not to prove the computational time cost reduction as [3] reducing the required parameter for an accurate model identification based on the MPM proving that this truncation of the Volterra Series is able to reduce the internal iterations during the modelling. This work is focused in the implementation stage in Hardware using s DSP board. A proper modeling technique for PAs is very profitable because once that the micro-device was manufactured their internal modifications are not allowed, the MPM as modulation technique for PAs assure to introduce these two main undesirable effects prior to the fabrication.

Another important factor is to prove the performance in hardware of this type of modeling, the Field Programmable Gate Array (FPGA) has flexible structure to process signals and processes related to MPM.

Mireya García-Vázquez, Grigori Sidorov (Eds.).
Advances in Computing Science and Control.
Research in Computing Science 59, 2012, pp. 277- 283

The FPGA has many advantages in digital signal processing and flexible implementation, the DSP (Digital Signal Processing) Builder Blockset in Simulink is able to convert a system architecture to VHSIC hardware description language (VHDL) code for the compilation and synthesis in Altera environment. Some works related to implementation in FPGA have been developed during recent years [4], [5], [6].

The structure of this article is organized as follows: Section 2 presents the general Volterra series and the implemented MPM into the DSP Design. The Section 3 shows the design procedure of the Memory-Polynomial Model. This section includes the implemented model into Simulink environment based in Altera DSP Builder blocks. In the Section 4 hardware implementation was done. The conclusions are drawn in Section 5.

2 Behavioral Modeling: Volterra Series

A characteristic of the Volterra Series as behavioral modeling is that if the input signal bandwidth becomes wider, and the memory effects of the power amplifier are considered, then the computational time increases in relation with the input width. The Volterra Series is a precise behavioral model to describe nonlinear HPAs [7], and can be expressed as:

$$y(n) = \sum_{k} \sum_{l_1} \cdots \sum_{l_{2k+1}} h_{2k+1}(l_1, l_2, \ldots, l_{2k+1}) \prod_{i=1}^{k+1} x(n-\tau_i) \prod_{i=k+2}^{2k+1} x^*(n-\tau_i) d\tau_{2k+1} \quad (1)$$

where

$x(n)$ is the input complex base-band signal.

$x^*(n)$ is the complex conjugate of the input complex base-band signal.

h_k are complex valued parameters.

2.1 Memory-Polinomial Model

The MPM as special case of the Volterra Series is able to consider the undesirable memory effects and non-linearities that affect the spectral regrowth during the behavioral modeling, the MPM structure is showed in equation (2):

$$y(n) = \sum_{q=0}^{Q} \sum_{k=1}^{K} a_{2k-1,q} |x(n-q)|^{2(k-1)} x(n-q) \quad (2)$$

where

$x(n)$ is the input complex base-band signal.

$y(n)$ is the output complex base-band signal.

$a_{k,q}$ are complex valued parameters.

Q is the memory depth.

K is the order of the polynomial.

Based on the equation (2) each stage of the MPM can be represented in Figure (1).

Each stage of the MPM can be subdivided depending of the sampling made for the input signal. Figure (2) shows the internal structure and the delay made for each

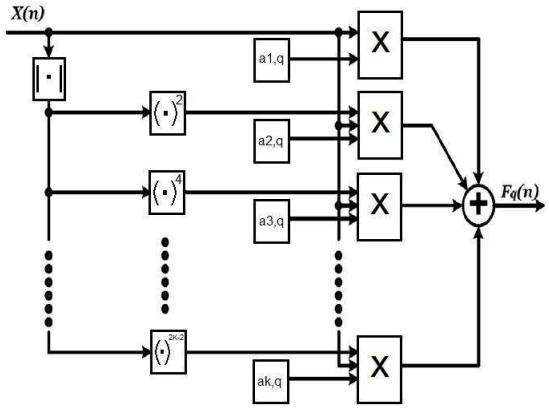


Fig. 1. MPM Internal Blocks during each stage

step. This body creates a phase offset of the signal inevitable during the first cycle and represent the general overview of this model type.

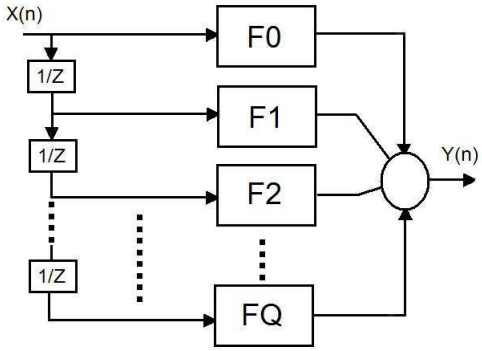


Fig. 2. MPM Subdivided for each sample of the input signal

3 Memory-Polynomial Model Implementation

The DSP Development Kit Cyclone III Edition delivers a complete digital signal processing (DSP) development environment; it includes the Cyclone III development board and Quartus II development software. The system operates with an internal clock of 50 MHz provided by a PLL (Phase-Locked Loop) circuitry. There are some works related to DSP Builder implementation [8], [9] and others applied to control systems [10].

The MPM explained in the previous section was created for a sine wave signal but can be implemented for digital or analog modulated signals as AM, FM or QAM

signals, some works have demonstrated that the MPM is able to work even for RF applications [11]. Figure (3) shows a general overview of the developed system in Simulink including the Signal Compiler icon and the Signal Tap Logic Analyzer making and interface with Quartus II Software for the synthesis, fitter and program process.

In this design, a generated sine wave is included by an internal LUT (Look up Table) and it is related with the capability of 14 bits of the DAC located in the HSMC card the signal is able to reach a value from 0 to 16384 due to the DAC sampling. This sine wave is attenuated and amplified 30 dB by the created MPM with the purpose of showing the effectiveness of the MPM using the DSP Builder Blockset, Figure (4) shows an overview of the developed MPM. Figure (5) shows the amplified signal by the MPM, as it can see there is a phase offset caused by the internal structure of the MPM, but the whole information is recovered after the second cycle.

Figure (4) shows the Hardware compilation steps made in Quartus II Software and Figure (6) the Simulation Waveform of LUT being monitored by Signal Tap Logic Analyzer included for the DSP Board.

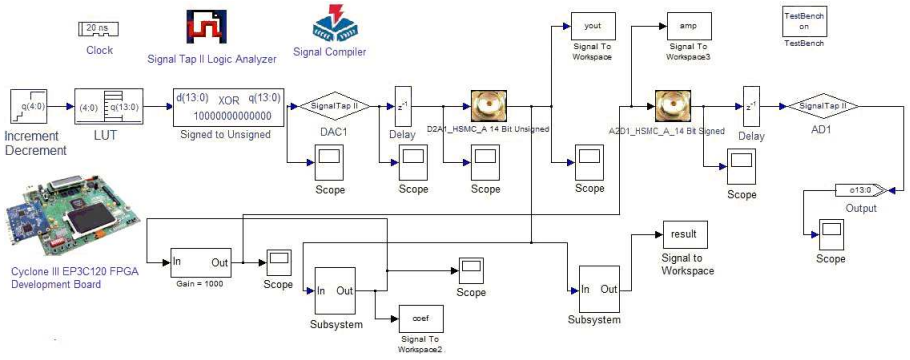


Fig. 3. Schematic Overview

The DSP Builder Signal Compiler block reads Simulink Model Files (.mdl) that are built using DSP Builder blocks and generates VHDL files and Tcl scripts for synthesis, hardware implementation, and simulation [12].

Once that the system was completed in Simulink the Signal Compiler Blockset allows to make the analyze, synthesis, fitter and programming process as a interface between Simulink and Quartus II Software, the section Export in this blockset has the property to export the valid HDL code. The VHDL or HDL code is a hardware description language used in electronic design automation to describe digital and mixed-signal systems such as field-programmable gate arrays and integrated circuits.

4 Results

The generated sine wave by an internal LUT was implemented in the developed MPM and the signal was amplified 1000 times, the generated signal by the 14 bits DAC was

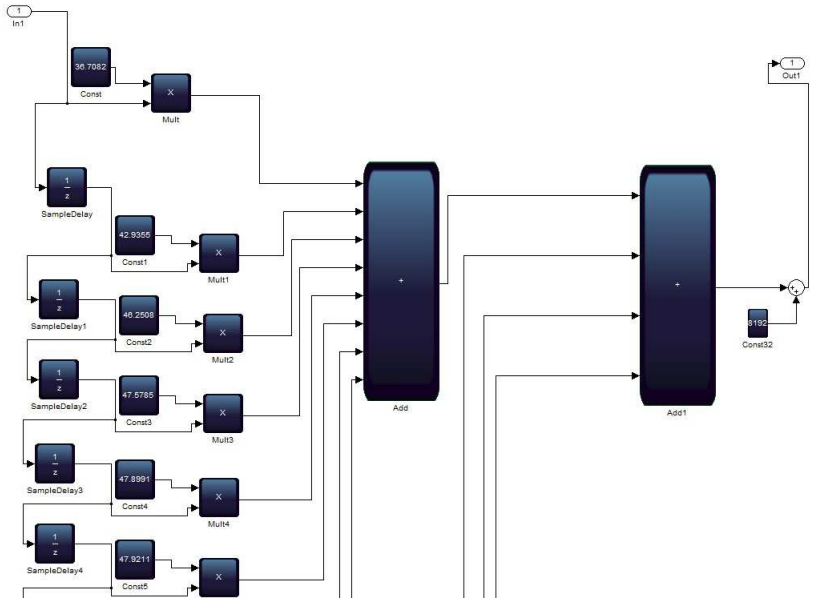


Fig. 4. Memory-Polynomial Model using the DSP Builder Blockset

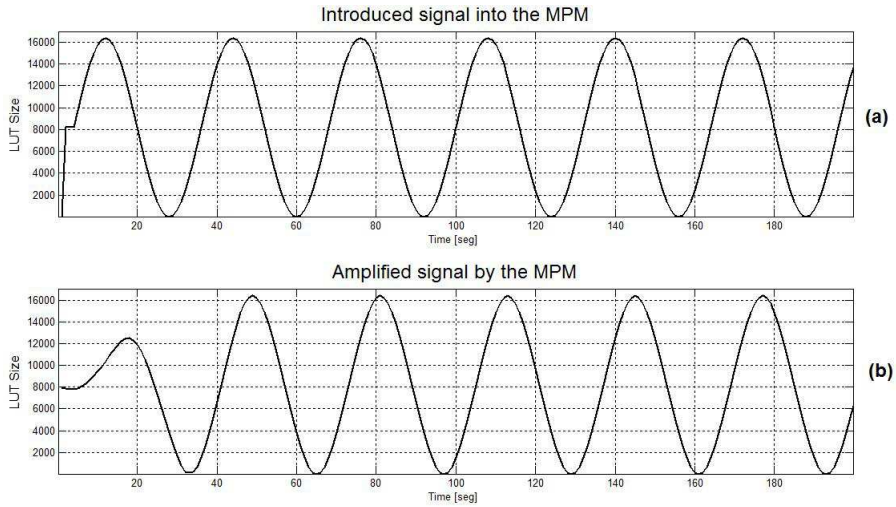


Fig. 5. (a) Generated signal by the LUT and (b) Recovered signal by the MPM

recovered by the amplification and the information passed though the communication between the DAC and ADC in the HSMC board.

The internal 50 Mhz clock was used in this work. The implemented MPM over the DSP Builder was properly developed using the DSP Builder Blockset.

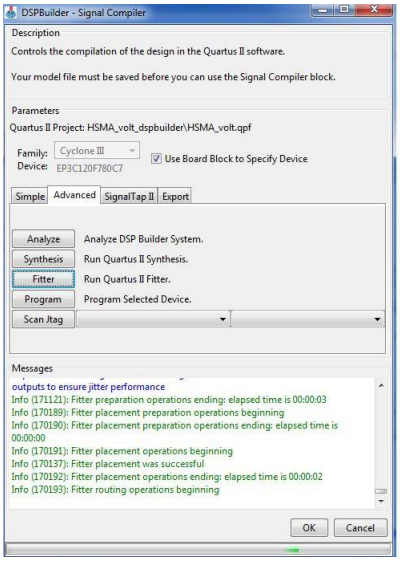


Fig. 6. Hardware compilation steps

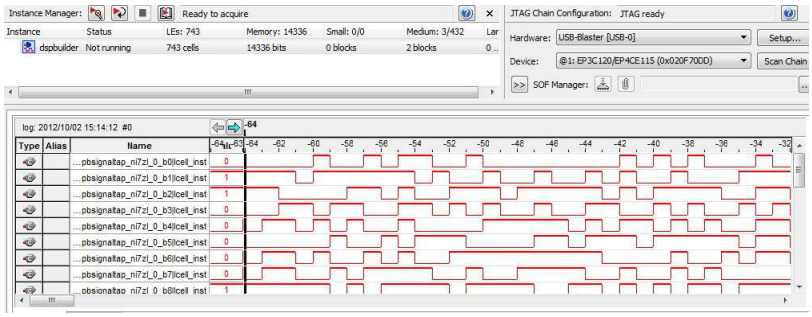


Fig. 7. Simulation waveform of LUT

5 Conclusion

In this paper, the Memory Polynomial Model as special case of the Volterra Series is implemented in the DSP Development Kit, Cyclone III Edition, the reduction of computational complexity together with the fast processing involved in the FPGA Cyclone III gives a proper behavioral modeling for HPAs and leaves open the option to introduce digitally/analogically modulated signals making wider the necessary process to modeling them.

The General Volterra Series is a precise method to create a behavioral modeling but how it was explained in this work the number of coefficients of the Volterra series increases exponentially as the memory length and the nonlinear order increase driving us to consider a Special case of the Volterra Series as the MPM, even better to create

an structure based on an FPGA as Altera Cyclone III to prove the performance of the MPM in Hardware implementation.

This stage was made using the DSP Builder technology allow us go from the system scheme to the VHDL files for synthesis and hardware implementation in the Cyclone III FPGA,. The MPM amplified the signal 30 dB using the internal clock with frequency of 50 MHz.

The computational time reduction and the eliminated internal iterations were not proved in this work, but as was referenced in the introduction section the MPM is able to achieve an accurate model of the PA reducing the processing time.

References

1. M. Schetzer: The Volterra and Wiener Theories of Nonlinear Systems. Krieger Publishing Company, Malabar, Florida, (2006).
2. A. Bennadji: Implémentation de modèles comportementaux ́damplificateurs de puissance dans des environnements de simulation système et co-simulation circuit système. Université de Limoges, École Doctorale Sciences - Technologie - Sante Faculté des Sciencies et Techniques, France, (2006).
3. K. Hyunchul, and J. Stevenson Kenney: Behavioral Modeling of Nonlinear RF Power Amplifiers Considering Memory Effects. IEEE Transactions on Microwave Theory and Techniques. vol. 51, no. 12, December 2003.
4. L. Guan and A. Zhu: Low-Cost FPGA Implementation of Volterra Series-Based Digital Predistorter for RF Power Amplifiers. IEEE Transactions on Microwave Theory and Techniques. vol. 58, no. 4, April (2010).
5. N. Laskarian and C. Dick: FPGA Implementation of digital predistortion linearizers for wideband Power Amplifiers. Proceeding of the SDR 04 Technical Conference and Product Exposition, (2004).
6. Q. Luo, M. Pirola, V. Camarchia, R. Quaglia, R. Tinivella, S. Shen and G. Ghione: FPGA implementation of adaptive baseband predistortion for FET-based wireless power amplifiers. (2209). The 1st International Conference on Information Science and Engineering.
7. X. Wu, J. Shi and Huihuang Chen : On the Numerical Stability of RF Power Amplifiers Digital Predistortion. 15th Asia-Pacific Conference on Communications. (2009). pp. 430–433.
8. H. Alasady and M. Ibnkahla: Design and hardware implementation of Look-Up Table predistortion on ALTERA stratix DSP board Canadian Conference on Electrical and Computer Engineering. (2008). pp. 1535-1538.
9. G. Xiong, X. Zhou and Peirong Ji: Implementation of the Quadrature Waveform Generator Based on DSP Builder International Symposium on Intelligent Information Technology Application Workshops. (2008). pp. 773 – 776.
10. Dai Bin and Z. Run-lin : Implementation of sliding mode variable structure control for Induction heating power supply based on DSP builder International Conference on Electric Information and Control Engineering. (2011). pp. 5933 – 5936.
11. J. R. Cárdenas-Valdez, M.Palafox, C. Gontrand and J.C. Núñez: Performance Evaluation of a Memory-Polynomial Model for Microwave Power Amplifiers. Programación Matemática y Software. Mag. vol. 4, no. 1, art. 2. June. (2012) pp. 13–23.
12. DSP Builder, Altera Inc. August 26, 2012. Available at¿ http://www.altera.com/products/software/products/ dsp/dsp-builder.html. Last accesed on October 26, 2012.

GPU-based parallel solution for a phase field model

Juan J. Tapia[1], Rigoberto Alvarado[1], and Fernando A. Villalbazo[1]

[1]Instituto Politécnico Nacional, CITEDI Research Center
Av. del Parque no. 1310, Mesa de Otay. Tijuana, México 22510
{jjtapia,ralvarado,fvillalbazo}@citedi.mx
http://www.citedi.mx
Paper received on 22/09/12, Accepted on 17/10/12.

Abstract. In this work, the implementation of an algorithm for the explicit solution of nonlinear equations based on the finite difference method in a Graphic Processing Unit is presented. The work focus on the solution for the Allen-Cahn equation that describes a phase-field model. Our algorithm implementation represents an easy way to solve in a GPU a nonlinear problem for interface phenomena, taking advantage of the characteristics of the problem.

Keywords: GPGPU, CUDA, Phase-field, FTCS

1 Introduction

Graphic Processing Units (GPU) were originally developed as an acceleration unit for graphic and video operations. However, they have been used in the last years for General Purpose Computing, called GPGPU. The key success of the use of GPUs in general computing applications lies on their relation cost/performance when compared with other processor types, i.e. multicore architectures. The GPUs constitute a relatively cheap tool that offers a peak number of FLOPS that is almost 10 times higher than the one reached by multicore processors [10]. This greater computational capacity is full exploited in many investigations and scientific fields.

The nonlinear problems are of especial interest in the scientific field because most physical systems and phenomena, e.g. chemical reactions, ecology, biomechanics and population growth, are nonlinear in nature. A nonlinear problem is described by a nonlinear equation which can be solved using an implicit method or an explicit one. The implicit method for a nonlinear equation involves the iterative solution of a nonlinear system; the explicit method involves the iterative solution of an equation. Generally, the implicit methods offer better numerical stability than the explicit methods and allow the use of bigger time steps in the numerical solutions. For these reasons, an ever increasing proportion of modern scientific research in the nonlinear field is devoted to the design of implicit solutions for nonlinear equations. Among this research, the develop of parallel algorithms that solve large nonlinear systems is a field of ongoing investigations, in particular, GPU-based algorithms that allow a faster and reliable solution for large nonlinear systems.

In [4], Dechevsky *et al.* developed a GPU-based nonlinear system parallel algorithm based on wavelet analysis. Garcia [7] implemented a GPU-based algorithm based on the

Mireya García-Vázquez, Grigori Sidorov (Eds.).
Advances in Computing Science and Control.
Research in Computing Science 59, 2012, pp. 284- 294

biconjugate gradient method. Wei, Feng and Lin [13] made an algorithm for nonlinear systems based on the Newton-Raphson method implemented on a GPU. In [6], Galiano *et al.* derived a GPU-based solution that uses a nonlinear conjugate gradient method.

In this work, we focus on an explicit parallel solution for the Allen-Cahn (AC) equation via a GPU-based algorithm. The AC equation is a nonlinear second order differential equation that describes a phase-field model. Many scientific problems involve multiphase and multicomponent flows, e.g. the impact of a droplet on a solid surface, bubbly and slug flows in a microtube, petroleum engineering, combustion and reaction flows, realistic simulations for computer graphics among other applications. The phase-field models are used to model and simulate multiphase and multicomponent flows.

Based on the characteristics of the AC equation and the Courant-Friedrichs-Lewy stability criterion, an explicit solution is an option to solve it. Furthermore, an explicit solution makes unnecessary the storage of large matrices involved with an implicit solution. Due the non data-dependencies and the number of arithmetic operations involved in an explicit method, a parallel solution is a good approach to solve the problem. These characteristics makes a GPU-based implementation an option to solve the AC equation.

In section 2 an introduction to the theory of the phase field models is presented and the Allen-Cahn equation is described. The CUDA programming model and the GPU general architecture are discussed in section 3. Section 4 presents the finite difference method and a detailed description of our algorithm. The results are presented in section 5 and finally, these results are discussed in section 6, where some ideas for future work are highlighted.

2 Phase field model: the Allen-Cahn equation

Phase field models are used to describe the behavior of multiphase and multicomponent flows. In a phase field model, the interfaces are replaced by extremely thin transition regions. The main idea is to introduce an order parameter ϕ that varies sharply but continuously across the thin interfacial region and has an almost uniform value on the bulk phases [2]. This concept is illustrated in Fig. 1, in which ϕ changes its value from phase 1 to phase 2 in a continuous manner across the interface region. In fact, the interface region is kind of a linear interpolation of ϕ, which is used to characterize the phases and could be density, concentration or mass fraction among others parameters.

Assume a binary fluid made of A and B particles in which diffusion is the transport mechanism. Denote by $\phi = -1$ a phase completely made of A particles and by $\phi = 1$ a phase completely made of B particles. A free energy can be defined for times when the system is not in equilibrium. The system evolution is driven by the minimization of this free energy which is given by the functional [9]

$$E[\phi] = \int \left[F(\phi) + \frac{1}{2}\varepsilon |\nabla\phi|^2 \right] d\Omega \tag{1}$$

where Ω is the volume of the system under consideration. The term $F(\phi)$ is the bulk energy potential defined as a classical double-well potential function with two minima for $\phi = -1$ and $\phi = 1$ corresponding to the two phases of the system (the function shape is shown in Fig. 1). The term $|\nabla\phi|^2$ is the gradient energy, called capillary term,

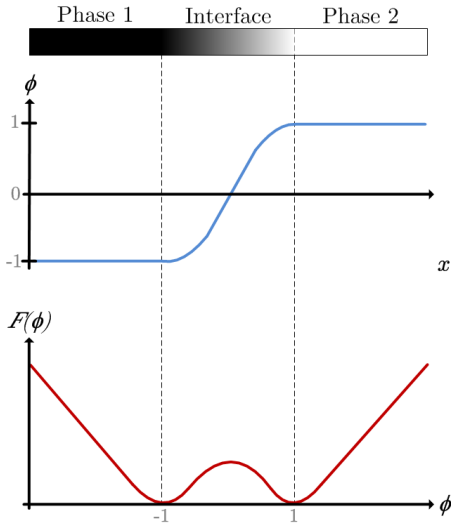


Fig. 1. Concept of phase field model.

which acts as a penalty for sharply varying concentration of ϕ; ε is its coefficient. The bulk energy, also called Helmholtz free energy, describes the entropy of the system. The gradient energy term describes the energy of the interactions between particles of type A and B [2, 9].

Variation of E with respect to ϕ is quantifying how the energy changes when particles change position, i.e. the chemical potential of the system [2]. This chemical potential is found by applying variational calculus to (1)

$$\frac{\delta E}{\delta \phi} = \mu = F'(\phi) - \varepsilon \nabla^2 \phi. \tag{2}$$

The evolutionary equation for the system is obtained by making the chemical potential (2) a time-dependant system. The result is the Allen-Cahn equation [1]

$$\frac{\partial \phi}{\partial t} = \varepsilon \nabla^2 \phi - F'(\phi) \tag{3}$$

where $F(\phi) = \frac{1}{4}\left(1-\phi^2\right)^2$. Replacing $F'(\phi)$ in eq. (3) we get

$$\frac{\partial \phi}{\partial t} = \varepsilon \nabla^2 \phi + \phi - \phi^3. \tag{4}$$

To completely specify the model of equation (4), it is assumed that the boundary and initial conditions are known [8].

The Allen-Cahn equation (4) is used to model the separation process of a binary fluid, in which the fluid is decomposed into a fine-grained mixture of particles. This process is known as spinodal decomposition [2].

3 Graphic Processing Unit

Recently, semiconductor industry has settled on two main design lines for microprocessors: the multicore processors and the manycore processors. The Graphic Processing Units (GPUs) are an example of the former.

The ratio between GPUs and multicore processors for peak floating-point operations is about 10 to 1. The reason for this gap in their performance lies in their architecture [10].

3.1 CUDA programming model

CUDA is an architecture of parallel computing for general purpose that offers high level access to the GPU. CUDA allows the use of a GPU to solve computationally intensive problems in a more efficient way than with a CPU [12].

The CUDA programming model allows the transparent scalability of applications through GPUs with different number of cores [12]. The base to get this transparent scalability are three key abstractions: a hierarchy of threads groups, shared memory and barrier synchronization. These abstractions allow the programmer to partition the problem into thread blocks that can be solved independently. The thread blocks can be considered as sub-problems and are executed in the available GPU cores, in any order and in parallel or sequential manner. This independence of the thread blocks execution is the characteristic that allows the scalability of the CUDA applications [12, 5].

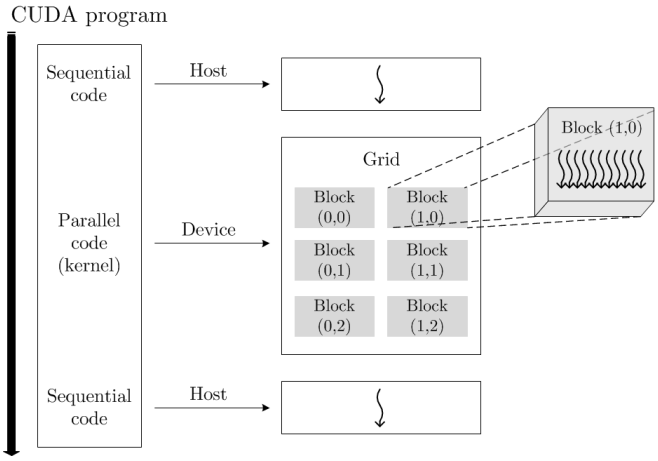


Fig. 2. CUDA heterogenous programming model.

The CUDA programming model let the use of a GPU as a co-processor of the CPU. In this context, GPU is called device and CPU is called host, as it is shown in Fig. 2. A CUDA program is composed of sequential code sections for the host and parallel code

sections for the device. In the parallel code sections, thousands of threads are executed concurrently in order to reduce the computation time [10].

In a CUDA program the main thread is executed in the host, as it is shown in Fig. 2. When a kernel is invoked, the execution is moved from the host to the device where a massive number of threads are executed concurrently to perform the parallel operations. A kernel is a subroutine that is executed K times by K threads within the device [10]. The threads generated by a kernel are grouped in thread blocks, and the total of thread blocks within a kernel is called a grid. The kernel calls are asynchronous, which implies that after the invocation of a kernel, the host can execute the rest of the sequential code or just wait to the termination of the kernel execution [10].

3.2 Optimization strategies

The optimization of a CUDA program is based on three main aspects: maximize parallel execution; optimize memory usage to achieve maximum memory bandwidth; and optimize instruction usage to achieve maximum instruction throughput [12, 3].

The most important rule to optimize the memory space usage of the GPU is to minimize the data transfer operations between the CPU and the GPU, because these memory operations have a lower memory bandwidth than the internal transfers within the GPU. It is also important to minimize the kernel access to the global memory and maximize the use of the shared memory [3].

4 GPU implementation of the explicit nonlinear solver

For the explicit solution of eq. (4) we use the Forward-time Central-Space scheme (FTCS), which is a finite difference method used for numerically solve time-dependant partial differential equations (PDE).

4.1 Allen-Cahn equation in 1-D

The scheme for the 1-D FTCS method is show in Fig. 3, where the domain is $[-1, 1]$, N is the number of points of the spatial grid and M is the number of iterations. The black dots represent the initial value for ϕ^0. The red dots are the boundary conditions, which are constant throughout all the iterations. The unknowns that will be calculated are represented as white dots.

In the FTCS scheme, the eq. (4) is discretized in space using the centered finite difference equation for the second order derivative

$$\frac{\partial^2 \phi_i}{\partial x^2} = \frac{\phi_{i-1} - 2\phi_i + \phi_{i+1}}{\Delta x^2} + O(\Delta x^2), \tag{5}$$

and is discretized in time with the forward Euler method

$$\frac{\partial \phi_i}{\partial t} = \frac{\phi_i^{t+1} - \phi_i^t}{\Delta t} + O(\Delta t). \tag{6}$$

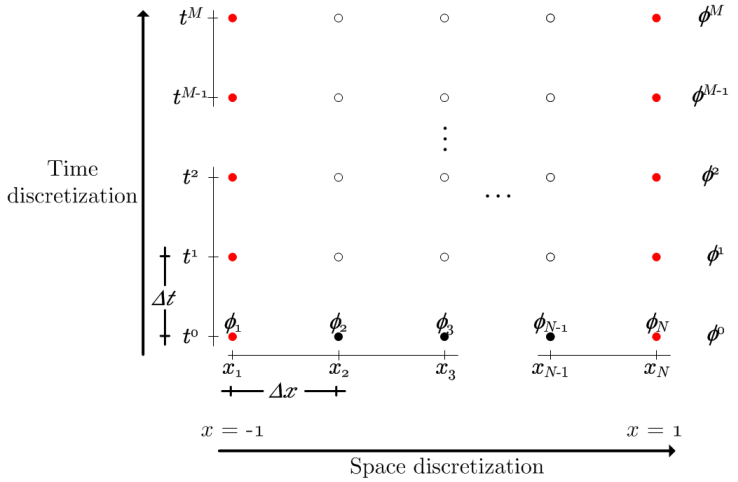


Fig. 3. FTCS scheme.

Therefore, the FTCS scheme is first-order of convergence in time and second order error in space.

Replacing eq. (5) and eq. (6) in eq. (4) we get the explicit discrete version of the AC equation

$$\frac{\phi_i^{t+1} - \phi_i^t}{\Delta t} = \varepsilon \left[\frac{\phi_{i-1}^t - 2\phi_i^t + \phi_{i+1}^t}{\Delta x^2} \right] + \phi_i^t - \left(\phi_i^t\right)^3. \tag{7}$$

Clearing the term ϕ_i^{t+1} we get

$$\phi_i^{t+1} = r\phi_{i-1}^t + r\phi_{i+1}^t + (-2r + 1 + \Delta t)\,\phi_i^t - \Delta t \left(\phi_i^t\right)^3, \tag{8}$$

where

$$r = \frac{\varepsilon \Delta t}{\Delta x^2}. \tag{9}$$

The eq. (8) is the iterative explicit equation to solve the AC equation in 1-D. This equation is numerically stable as long as it meets the Courant-Friedrichs-Lewy (CFL) stability criterion for the time step [11]. The CFL is defined as

$$\frac{\varepsilon \Delta t}{\Delta x^2} \leq 1. \tag{10}$$

From eq. (10), we can see the reason that justifies our explicit solution for the AC equation. Since ε weights the gradient energy, i.e. defines the interface thickness (interface area width of Fig. 1), it always has relatively small values, so we can use relatively large time steps (Δt values) that allow get faster to the solution without violating the CFL stability criterion. This advantage plus the non-necessity of storage for large matrices involved with an implicit solution make the explicit solution a good approach to solve the AC equation.

For the 1-D case, we use the Dirichlet boundary conditions

$$\phi(-1,t) = -1, \phi(1,t) = 1. \tag{11}$$

The structure of the GPU implementation of eq. (8) is shown in the algorithm 1.1.

Algorithm 1.1 CUDA FTCS pseudocode for the 1-D Allen-Cahn equation

main() {
allocate memory
copy CPU to GPU
$u_{new}[0] = u_{old}[0] = -1$ // Boundary conditions
$u_{new}[N-1] = u_{old}[N-1] = 1$
kernelFTCS1D$<<< 1, N-2 >>> (u_{old}, u_{new})$
copy GPU to CPU
free memory
}
__global__ **kernelFTCS1D** (u_{old}, u_{new}) {
$dx = 2/(N-1)$
$r_1 = \epsilon dt/dx^2$
$r_2 = 1 - 2r_1$
$id = threadIdx.x + 1$
for $t < T$ **do**
$u_{new}[id] = r_1 u_{old}[id-1] + r_2 u_{old}[id] + r_1 u_{old}[id+1] + dt(u_{old}[id] - (u_{old}[id])^3)$
$__syncthreads()$
$u_{old}[id] = u_{new}[id]$
endfor
}

As explained in section 3, the first step to use a GPU is copy the necessary data from the CPU memory to the GPU memory. After that, the kernel that implements the eq. (8) is implemented by N threads simultaneously. This means that we only need a loop for the time iterations, because all the threads represents the space discretization. This is the main advantage of the GPU implementation and the root for the better performance.

4.2 Allen-Cahn equation in 2-D

The scheme for the 2-D FTCS method is similar to the 1-D case, except that this time we need to discretize in space through a bidimensional mesh; the time discretization is done in the same way as in the 1-D case. We used an $N \times N$ square shaped domain to solve the eq. (4) for our 2-D experiments, where the interval for both the x and y axis is $[-1, 1]$, N is the number of points and M is the number of iterations. The initial value for all the points of the mesh except the ones at the boundaries are given by the initial condition; the values of the points at the boundaries are given by the boundary conditions through all the iterations.

The 2-D AC equation is discretized in time using the eq. (6). The space discretization is done with eq. (5), but for the 2-D case, we need to calculate $\Delta_2\phi$ with respect to

the x and y axes. This means that we need to add another subindex, unlike eq. (7), so we can denote the respective derivatives of a point.

Replacing eq. (6) and eq. (5) in eq. (4) we get the explicit discrete version for the 2-D AC equation

$$\frac{\phi_{i,j}^{t+1}-\phi_{i,j}^{t}}{\Delta t}=\varepsilon\left(\left[\frac{\phi_{i-1,j}^{t}-2\phi_{i,j}^{t}+\phi_{i+1,j}^{t}}{\Delta x^{2}}\right]+\left[\frac{\phi_{i,j-1}^{t}-2\phi_{i,j}^{t}+\phi_{i,j+1}^{t}}{\Delta y^{2}}\right]\right) +\phi_{i,j}^{t}-\left(\phi_{i,j}^{t}\right)^{3}. \tag{12}$$

Clearing the term ϕ_{i}^{t+1} and assuming $\Delta y^{2}=\Delta x^{2}$ for the mesh we get

$$\begin{aligned}\phi_{i,j}^{t+1} &= r\phi_{i-1,j}^{t}+r\phi_{i+1,j}^{t}+r\phi_{i,j-1}^{t}+r\phi_{i,j+1}^{t} \\ &\quad +\left(-4r+1+\Delta t\right)\phi_{i,j}^{t}+\Delta t\left(\phi_{i,j}^{t}\right)^{3},\end{aligned} \tag{13}$$

where r is defined in eq. (9), so the same CFL criterion can be applied in this case, meaning that the explicit solution is a good approach to solve the 2-D AC equation.

For the 2-D case, we use Dirichlet boundary conditions at the bottom and top of the square domain, i.e.

$$\phi(x,-1,t)=-1,\phi(x,1,t)=1. \tag{14}$$

For the left and right sides of the domain, we use the homogeneous Neumann boundary condition, using centered finite difference with ghost points.

$$\phi_{1,i}^{t+1}=r\phi_{1,i}^{t}+2r\phi_{2,i}^{t}-\phi_{1,i}^{t}, \tag{15}$$

$$\phi_{N,i}^{t+1}=r\phi_{N,i}^{t}-2r\phi_{N,i}^{t}-\phi_{N-1,i}^{t}. \tag{16}$$

The physical interpretation of these boundary conditions and domain implies a container in which the bottom and top sides correspond to different phases or components of a fluid, and the right and left sides represent the walls of the container.

The structure of the GPU implementation of eq. (4.2) is shown in the algorithm 1.2.

5 Results

In this section we present the numerical results for the algorithms showed in the section 4.

The Fig. 4 shows the process of phase separation in 1-D, i.e. the numerical solution for eq. (8). The Fig. 4(a) shows the initial condition, which is made of random values -1 and 1. Intermediate steps are shown in Fig. 4(a) and (b). The final phase separation is shown in Fig. 4(d), where the final shape resembles the interface area of Fig. 1.

Moving one step forward, the Fig. 5 shows the process of phase separation in 2-D, i.e. the numerical solution for eq. (4.2). The Fig. 5(a) shows the initial condition, which is made of random values -1 and 1 and represents a non-homogeneous fluid made of two components, one represented by the value -1 and the other represented by 1. Intermediate steps are shown in Fig. 5(a) and (b), where we can see the first separation steps. The final phase separation is shown in Fig. 5(d), where the two fluids are completely separated and the interface area is represented by the middle line.

Algorithm 1.2 CUDA FTCS pseudocode for the 2-D Allen-Cahn equation

main() {
allocate memory
copy CPU to GPU
kernelFTCS2D$<<<1,(N,N)>>>(u_{old},u_{new})$
copy GPU to CPU
free memory
}
__global__ **kernelFTCS2D** (u_{old},u_{new}) {
$dx = 2/(N-1)$
$r_1 = \epsilon dt/dx^2$
$r_2 = 1 - 2r_1$
$i = threadIdx.x$
$j = threadIdx.y$
$id = j * blocksize.y + i$
for $t = 1 : T$ **do**
if $(id < 0$ AND $id < N)$ **then**
$u_{new}[id] = u_{old}[id] + 2r(u_{old}[id+1] - u_{old}[id])$ // Boundary condition
endif
if $(id > N(N-1)$ AND $id < N^2)$ **then**
$u_{new}[id] = u_{old}[id] + 2r(-u_{old}[id] + u_{old}[id-1])$ // Boundary condition
endif
$u_{new}[id] = r_1 u_{old}[id-1] + r_2 u_{old}[id] + r_1 u_{old}[id+1] + dt(u_{old}[id] - (u_{old}[id])^3)$
$__syncthreads()$
$u_{old}[id] = u_{new}[id]$
endfor
}

6 Conclusions and future work

6.1 Conclusions

In this work, we present an explicit parallel method to solve the AC equation in a GPU. Our approach represents an easy way to solve in a GPU a nonlinear partial differential equation that models a nonlinear phenomena, in this case, a phase field model.

Our proposed algorithm is based on the assumption that the CFL criterion is always met, due the nature of our problem. As the value of ε is always relatively small, because it represents the interface thickness, is possible to use an explicit approach and relatively big step values without violating the stability criterion.

Our approach avoid the storage of large matrices, due its explicit nature, and this characteristic allows us to solve large systems in the relatively small GPU memory.

6.2 Future work

A detailed comparison and numerical analysis between our explicit solution and an implicit one is part of our ongoing investigations. We think that this way we can establish the relation of performance and numerical accuracy between the two methods.

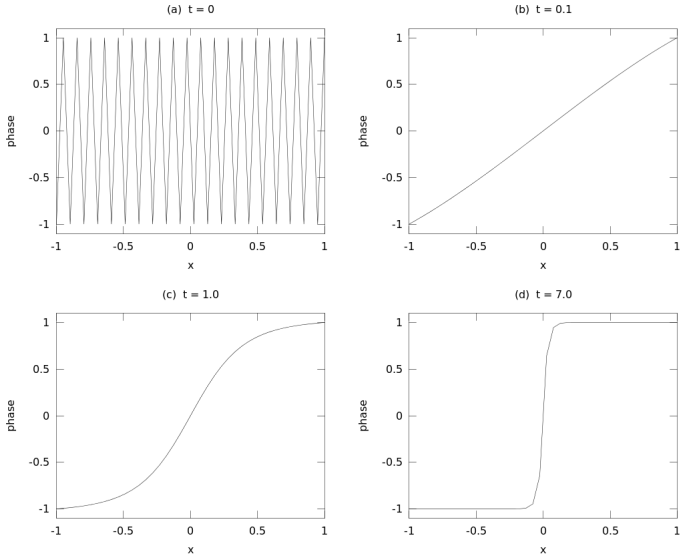


Fig. 4. FTCS scheme for 1-D Allen-Cahn.

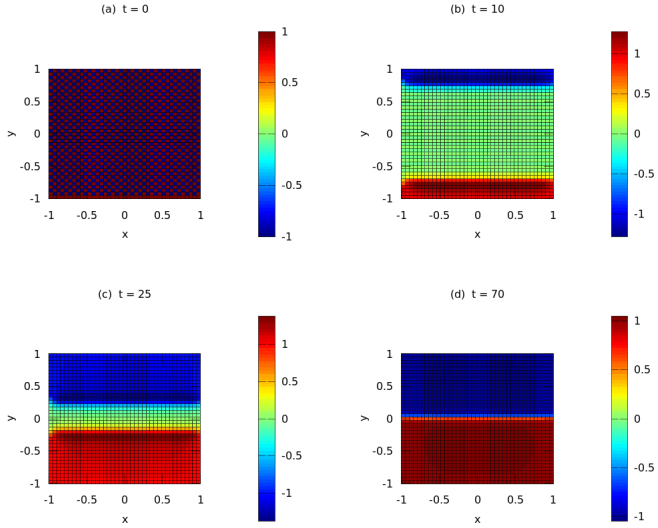


Fig. 5. FTCS scheme for 2-D Allen-Cahn.

As we stated in section 4, ϵ controls the interface thickness of the phase field. However, when we use very small ϵ values, the regular grid for the numerical solution is not longer numerically accurate. At this point, is necessary the use of an adaptive mesh refinement method in order to get a good numerical solution over the interface area. This is a future step of our investigation. In particular, we will seek for Conjugate Gradient methods as the adaptive mesh function.

A comparison between the execution time of an algorithm implemented in a CPU and in a GPU is not the best way to evaluate a CUDA application, although it is the most common way. It is not the best choice, because is not objective, fair nor qualitative. Many factors are involved such as the capacity of the processors, available resources, among others. That is why we think that an execution time comparison as well as an objective measurement of the GPU and CPU resources utilization is the best way to measure a GPU application. For this reason, the objective measurement of the kernels and C code that are used in this work is the next step of the project. To accomplish it, we need to establish adequate metrics and find tools for debugging and profiling the CUDA programs and the C programs, e.g. Compute Visual Profiler and CUDA-gdb.

Acknowledgments. This work has been partially supported by COFAA-IPN and grant IPN-SIP-20120606.

References

1. Allen S. M., Cahn J. W., A microscopic theory for antiphase boundary motion and its application to antiphase domain coarsening, Acta Metallurgica 27 (1979), pp. 1085–1095
2. Badalassi V. E., Ceniceros H. D., Banerjee S., Computation of multiphase systems with phase field models, J. Comput. Phys. 190 (2003), pp. 371–397
3. CUDA C Best Practices Guide (2011)
4. Dechevsky L., Bang B., Gundersen J., Laks A., Kristoffersen A.R., Solving nonlinear systems of equations on graphic processing units, Lect. Notes Comput. Sci. 5910 (2010), pp. 719-729
5. Farber R., CUDA Application Design and Development. Elsevier (2011)
6. Galiano V., Migallon H., Migallon V., Penadés J., GPU-based parallel algorithms for sparse nonlinear systems, J. Parallel Distrib. Comput. 72 (2012), pp. 1098-1105
7. Garcia, N., Parallel power flow solutions using a biconjugate gradient algorithm and a Newton method: A GPU-based approach, In 2010 IEEE Power and Energy Society General Meeting, pp. 1–4 (2010)
8. Huang P., Adbuwali A., A numerical method for solving Alen-Cahn equation, J. Appl. of Math. and Informatics 29 (2011), pp. 1477–1487
9. Kim J., Phase-Field Models for Multi-Component Fluid Flows, Commun. Comput. Phys. 12 (2012), pp. 613–661
10. Kirk D. B., Hwu W. W., In Praise of Programming Massively Parallel Processors: A Hands-on Approach. Elsevier (2010)
11. Lui S. H., Numerical Analysis of Partial Differential Equations. John Wiley and Sons, 2011.
12. NVIDIA CUDA C Programming Guide 5.0 (2012)
13. Wei F., Feng J., Lin H., GPU-based parallel solver via Kantorovich theorem for the nonlinear Bernstein polynomial system, Comput. Math. Appl. 62 (6) (2011), pp. 2506-2517

Sistema de Monitoreo y Control de Dispositivos en Hogares Usando Plataformas Arduino

Dueñas, D, Hernández, V, Iriarte, J, Cervantes, A, Vejar, H, López, J.*

Universidad Tecnológica de Tijuana.
Carretera Libre Tijuana-Tecate Km. 10
C.P. 22253 Tijuana, Baja California.

*david.duenas@uttijuana.edu.mx, viusanyi@hotmail.com, ces_jc@hotmail.com, cervantes.arturo@hotmail.com, humberto.vejar@uttijuana.edu.mx, jenrique.lopez@uttijuana.edu.mx

Paper received on 22/09/12, Accepted on 16/10/12.

Abstract. Se presentan los resultados obtenidos del desarrollo de un prototipo electrónico que monitorea y controla la temperatura de un entorno de hogar, los datos son procesados y transmitidos por medio de una plataforma Arduino, la cual tiene incorporado un módulo EthernetShield que permite transmitir los datos y ser almacenados en una base de datos SQL 2008. Los valores de la temperatura pueden ser visualizados en tiempo real a través de una interfaz gráfica desarrollada en Visual Basic.NET y fijar el valor de la temperatura deseada mediante un control ON/OFF, además permite activar y desactivar equipos de casa tales como luces, ventiladores, refrigeradores, etc. Todo esto a través de un solo acceso a Internet. El prototipo cuenta con dos Arduinos modelo UNO, el primero de estos se encarga de monitorear las condiciones de temperatura y controlar la temperatura en base a un setpoint que puede ser fijado desde Internet. El segundo Arduino se encarga de mantener un control de los demás equipos que se quieren activar remotamente. La interfaz desarrollada permite al usuario ingresar a una dirección de internet pública y visualizar que dispositivos se encuentran activados o desactivados.

Keywords: Arduino, Sensor, Red, Dirección IP, Aplicación.

1 Introducción

En la actualidad, existen tecnologías diversas que permiten el monitoreo y actuación en aplicaciones relacionadas a la domótica [12], control de procesos, redes inalámbricas de sensores [1], agricultura de precisión entre otras. El avance y la disponibilidad de la tecnología han hecho posible la adquisición y experimentación en áreas educativas e industriales, hoy en día es común contar con tarjetas de adquisición de datos, radios transreceptores, controladores entre otros. En el presente trabajo se usa la plataforma Arduino, la cual está basada en diseños de hardware y software

Mireya García-Vázquez, Grigori Sidorov (Eds.).
Advances in Computing Science and Control.
Research in Computing Science 59, 2012, pp. 295-301.

libre, se tienen una gran variedad de modelos que pueden adaptarse a prácticamente la mayoría de las aplicaciones. Para este caso, se usa el modelo Arduino Uno junto con un modulo EthernetShield [6], el cual le brinda conectividad con redes de área local o amplias, dicho dispositivo se muestra en la Figura 1.

Fig. 1. Arduino con módulo EthernetShield

Arduino está diseñado para ser usado por personas que cuentan con poca o nada de experiencia en aplicaciones relacionadas a los micro controladores [5], sin embargo también los utilizan personas con experiencia que deseen usar equipos que evitan el uso de licencias particulares que en muchas ocasiones los hacen difícil de adquirir, es por éstas razones que las plataformas Arduino son muy convenientes en actividades educativas y de investigación.

La aplicación de este proyecto es la de contar con una interfaz, desde donde sea posible monitorear variables de interés presentes en una casa [11], así como contar con la posibilidad de activar aparatos de corriente alterna tales como luces, aires acondicionados, televisores, etc. En este caso se usó un sensor de temperatura LM35, pero se planea incluir una diversidad de sensores tales como de consumo eléctrico, consumo de agua, presencia de gas, etc. El desarrollo de este tipo de trabajos forma parte de los contenidos estudiados en la carrera de tecnologías de la información y comunicación de la Universidad Tecnológica de Tijuana, de los alumnos de noveno cuatrimestre. Los resultados obtenidos se lograron durante el cuatrimestre Mayo-Agosto del 2012.

2 Desarrollo

En esta sección se explican los procedimientos que se siguieron para el desarrollo del proyecto.

2.1 Etapa de sensado

El sensor usado en esta aplicación es el LM35, el cual es un sensor de circuito integrado que produce una salida de voltaje aproximadamente lineal al cambio de la temperatura. Es un sensor sencillo de usar, debido a que no es necesaria una calibración externa cuando el dispositivo se encuentra entre -55 a 150°C, con variaciones de ±3/4 °C [7]. La Figura 2 muestra los pines de conexión del sensor.

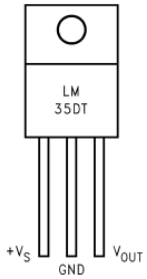


Fig. 2. Sensor de temperatura LM35.

La salida del sensor está conectada al Arduino en el pin A_o, como se mencionó anteriormente, el valor del voltaje producido por el sensor, depende directamente de la temperatura y viene definida en la ec. 1. Como se ve en la ecuación, a una temperatura de 25°C, el sensor produce un voltaje de aproximadamente 250mV. La característica de linealidad se exhibe para temperaturas entre 2°-150°C.

$$V_{sensor} = 10T \tag{1}$$

donde:
V_{sensor} = Voltaje producido por el sensor [mV]
T=temperatura [°C]

Los Arduinos tienen convertidores analógicos/digital que cuentan con una precisión de 10 bits, por lo que para aplicaciones en las que la variables a estudiar no varían de manera abrupta, son adecuados.

En esta aplicación se muestrea el valor de temperatura, cada 5 segundos, los datos se envían del Arduino a una base de datos montada en un Servidor SQL 2008.

2.2 Etapa de Interacción Web

El micro controlador va unido a un módulo que le permite conectarse a una red LAN o a Internet por medio del puerto Ethernet, de esta manera éste puede ser manipulado por medio de Internet, utilizando una interfaz gráfica montada en Visual Basic.NET, además en el módulo se puede insertar una tarjeta micro SD en la cual se aloja la página web desde la cual se controla y vigila la temperatura en el termóstato y se tiene la opción de activar o desactivar salidas digitales para controlar aparatos domésticos, por lo que el micro controlador hace también la función de servidor Web. La Figura 3 ilustra el esquema de interacción entre los dispositivos cliente y el módulo Arduino.

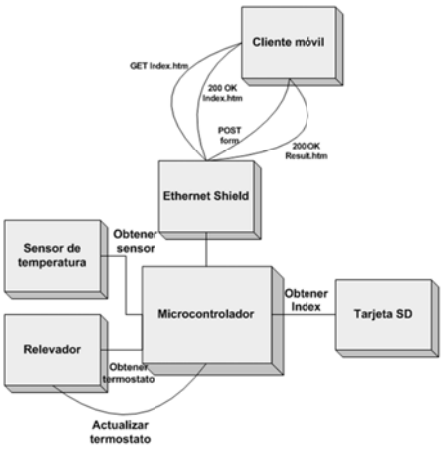


Fig. 3. Proceso de interacción en el proyecto.

2.3 Etapa de configuración de Arduinos

Para este proyecto se utilizó la versión 0022 de Arduino Alpha, plataforma donde se codificó la funcionalidad del sensor de temperatura, la tarjeta SD donde se aloja el servidor web, los parámetros de red para el Arduino Ethernet Shield y las salidas de los pines digitales del Arduino Uno. Es muy importante detectar el puerto COM por donde se ha establecido la comunicación. La tabla 1 muestra un resumen de los pines configurados en los Arduinos.

Table 1. Configuraciones de los pines en Arduino.

Salida Digital	Función
2	Encender Luz Sala
3	Encender Luz Comedor
4	Encender Luz Cocina
5	Encender Luz Recamara
6	Apagar todas las luces
7	Encender todas las luces

El Arduino tiene la dirección IP 192.168.1.177/24, la cuál es asignada dependiendo de la red a la que esté conectado el equipo, también se tiene que especificar la dirección MAC cuando se ejecuta el Sketch, la cual para éste caso es DEAD.BEEF.FEED.

2.4 Etapa de desarrollo de la aplicación Web

La idea principal de este proyecto es ofrecer al usuario un panel de control, donde pueda ver el estatus de la temperatura actual de un cuarto, así como establecer la temperatura deseada desde Internet, además que se permita controlar el encendido y/o apagado de las lámparas de la sala, comedor, recámara, y cocina; al mismo tiempo

conocer la tendencia del clima, a través de un gráfico en tiempo real, simplemente ingresando a la dirección IP de la página Web.

La aplicación desde donde se visualiza el panel de monitoreo/activación fue desarrollada en Visual Basic.NET 2008, se configuraron los puertos COM22 para la comunicación con el Arduino EthernetShield, encargado de la lectura de la temperatura y el COM29 para la comunicación con el Arduino Uno encargado de la manipulación de los pines digitales para el control del apagado/encendido de las luces. La Figura 4 ilustra las opciones de configuración para los puertos.

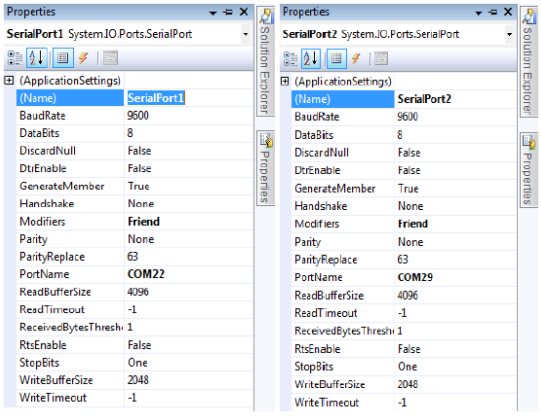
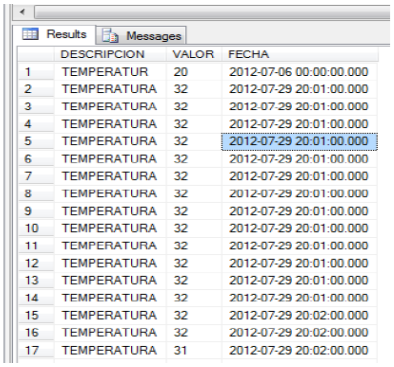


Fig. 4. Configuración de puertos COM.

La Figura 5 muestra la página principal del sitio desarrollado, desde donde se puede consultar el estado de temperatura en el momento, fijar el valor de temperatura deseado, además se cuenta con una interfaz desde donde se puede seleccionar la luz que se desee activar. La interfaz es muy sencilla de usar, pero para fines prácticos cumple con los objetivos planeados.

Fig. 5. Página principal del proyecto.

La base de datos contiene los campos de interés, entre los que se tienen: El valor de temperatura y la hora en que se recibieron dichos datos. Estos mismos son enviados a una gráfica que muestra las variaciones de temperatura a través del tiempo, además con los datos se pueden generar informes de variables de interés durante el día.



Results | Messages

	DESCRIPCION	VALOR	FECHA
1	TEMPERATUR	20	2012-07-06 00:00:00.000
2	TEMPERATURA	32	2012-07-29 20:01:00.000
3	TEMPERATURA	32	2012-07-29 20:01:00.000
4	TEMPERATURA	32	2012-07-29 20:01:00.000
5	TEMPERATURA	32	2012-07-29 20:01:00.000
6	TEMPERATURA	32	2012-07-29 20:01:00.000
7	TEMPERATURA	32	2012-07-29 20:01:00.000
8	TEMPERATURA	32	2012-07-29 20:01:00.000
9	TEMPERATURA	32	2012-07-29 20:01:00.000
10	TEMPERATURA	32	2012-07-29 20:01:00.000
11	TEMPERATURA	32	2012-07-29 20:01:00.000
12	TEMPERATURA	32	2012-07-29 20:01:00.000
13	TEMPERATURA	32	2012-07-29 20:01:00.000
14	TEMPERATURA	32	2012-07-29 20:01:00.000
15	TEMPERATURA	32	2012-07-29 20:02:00.000
16	TEMPERATURA	32	2012-07-29 20:02:00.000
17	TEMPERATURA	31	2012-07-29 20:02:00.000

Fig. 6. Datos almacenados.

La base de datos fue desarrollada totalmente en SQL Server 2008 de forma simple, ya que solo interesaba saber el valor de temperatura y la hora en que fue tomada la muestra, esto se puede apreciar en la Figura 6, aunque se tiene planeado incluir en un trabajo futuro una variedad de sensores que permitan ampliar el panorama de aplicación de este proyecto.

3 Resultados

En base a la implementación del proyecto, se tiene una página Web, desde donde es posible consultar la temperatura actual de un sitio de interés, como un cuarto de una casa, el site de servidores de una empresa, o cualquier lugar desde donde se desee monitorear la temperatura. También es posible fijar una temperatura deseada desde la aplicación y llegar al valor deseado a través de un control ON/OFF de un aire acondicionado u otro aparato usado para disminuir la temperatura. Además es posible activar o desactivar salidas digitales del Arduino desde Web. La etapa de potencia para permitir a los equipos de corriente alterna se implementó mediante actuadores tipo PowerSwitchTail [9], los cuales aíslan la electrónica de baja potencia del Arduino para permitir conmutar aparatos a corrientes de 10A/120AC.

Cabe señalar que las pruebas realizadas se hicieron en un servidor local para la página Web por motivos de no contar con una dirección IP pública y el tiempo del desarrollo del proyecto de sólo 4 meses.

Entre los resultados más notables se tiene la interfaz gráfica desde donde se puede fijar una temperatura deseada, monitorear temperatura actual, y activar/desactivar salidas digitales.

4 Conclusiones

La implementación del proyecto deja un gran aprendizaje de cómo la integración dispositivos y herramientas tecnológicas puede ayudar a propuestas que se apliquen en necesidades de la región. El prototipo fue probado de manera local durante la expo

proyectos de la Universidad Tecnológica de Tijuana en Agosto del presente año. Las expectativas son incluir una variedad de sensores y nuevas tecnologías que permitan escalar a necesidades variantes de clientes que deseen contar con un mecanismo para activar/desactivar aparatos a distancia. Entre los sensores que se planean incluir son de consumo eléctrico y de agua, por ser éstos comunes en la mayoría de las viviendas, se está enfocando esfuerzos para brindar un servicio que pueda ser adquirido por personas de clase media interesadas en contar con este tipo de posibilidades. Todavía queda mucho trabajo por realizar, pero con una visión fija se pueden lograr los objetivos.

Entre los problemas que se presentaron se tiene el hecho de que los microcontroladores Arduino de la serie UNO no posean la capacidad de efectuar tareas múltiples al mismo tiempo, lo que llevó al uso de dos dispositivos. Se exploraran nuevas plataformas para realizar el sensado y control de salidas al mismo tiempo. Aunque existen dispositivos comerciales que permiten realizar las tareas que se implementaron en este proyecto [10], la aportación principal es que se están usando tecnologías de hardware libre que pueden ser adaptadas a una gran variedad de aplicaciones, el costo estimado de este proyecto fue de alrededor de 180 dólares, que al compararlo con sistemas que proporcionan solo algunas funciones es económico. Ya que con estas mismas plataformas se pueden activar hasta 14 entradas/salidas digitales, lo que lo hace escalable en diversos entornos.

5 Referencias

1. K. Sohraby, D. Minoli, T. Znati. Wireless Sensor Networks: Technology, Protocols and Applications. Ed. Wiley. USA, 2007.
2. T. Igoe, Making Things Talk, Ed. O´Reilly, USA, 2010.
3. R. Faludi. Building Wireless Sensor Networks. Ed. O'Reilly. California, 2010.
4. M. Margolis. Arduino Cookbook. Ed. O'Reilly. USA, 2011.
5. http://www.arduino.cc
6. http://arduino.cc/en/Main/Hardware
7. https://www.national.com/ds/LM/LM35.pdf
8. R. Coughlin and F. Driscoll. Circuitos integrados lineales y amplificadores. Ed. Prentice-Hall. México, 1995.
9. http://www.powerswitchtail.com/Pages/default.aspx
10. http://www.smarthome.com/2441TH/INSTEON-Thermostat/p.aspx
11. J. Paz, J. Dávila, R. Pérez, Casa Inteligente y segura, Ed. Dirección General de Difusión Cultural y Divulgación Científica, México, 2011.
12. R. Piyare, M. Tazil, Bluetooth Based Home Automation System Using Cell Phone, IEEE 15th International Symposium on Consumer Electronics, 978-1-61284-842-6, 2011.

Diseño en VHDL de un transceptor de la interfaz de línea digital E1 y su implementación en un FPGA

Yudith Florencia Gonzalez Padilla[1,2], Topacio Osuna Altamirano[2], Josué López Leyva[2].

[1]Instituto Superior Politécnico José Antonio Echeverría, Cujae, Calle 114 No. 11901. e/ Ciclovía y Rotonda. Marianao 15. La Habana, Cuba. CP 19390.
[2]Centro de Investigación Científica y de Educación Superior de Ensenada; Carretera Ensenada-Tijuana No. 3918, Zona Playitas, C.P. 22860, Ensenada, B. C. México.
pupi_bonita@yahoo.es,tosuna@cicese.edu.mx, jalopez@cicese.edu.mx
Paper received on 01/10/12, Accepted on 25/10/12.

Abstract. Se presenta el diseño de un módulo de hardware embebido con las funcionalidades de una interfaz digital E1. El sistema diseñado realiza las tareas de un transmisor-receptor, encargándose de acciones como la codificación y decodificación del flujo digital, el análisis del sincronismo de la trama y la generación de relojes. También incorpora una interfaz con un microprocesador para la configuración de los parámetros iniciales del funcionamiento o selección de los modos de operación. Cada uno de los bloques presentes en el sistema se ha diseñado a través del lenguaje de descripción de hardware VHDL empleando el software Xilinx ISE Design Suite 10.1; e implementado en la tarjeta de desarrollo SPARTAN3 Starter Kit. La principal contribución del trabajo es un módulo de hardware reutilizable que aumenta el rendimiento, reduce la latencia de tratamiento de datos en comparación con soluciones de circuitos integrados existentes y tiene integrado un mecanismo de análisis de fallos en la trasmisión.

Keywords: FPGA, Dispositivo programable, flujo E1, trama, multitrama.

1 Introducción

A inicios de la década de los noventa, las telecomunicaciones y las redes de datos incorporaron el uso de los FPGAs a sus aplicaciones, dado el tamaño y sofisticación de los mismos. Sin embargo, no es hasta principios del 2000 cuando se llegó a tener FPGAs de alto rendimiento con sistemas de millones de compuertas y bloques embebidos de microprocesadores, procesadores digitales de señales, e interfaces de entrada/salida de muy alta velocidad pudiendo sobrepasar los 5 Gbps. Lo anterior justifica que estos dispositivos sean aptos para casi cualquier aplicación, como por ejemplo sistemas de comunicaciones, procesamiento de imágenes y video, redes neuronales artificiales y procesamiento digital de señales, por mencionar algunos.

Mireya García-Vázquez, Grigori Sidorov (Eds.).
Advances in Computing Science and Control.
Research in Computing Science 59, 2012, pp. 302-313.

Los módulos de propiedad intelectual, también conocidos como IP CORE, son descripciones software de componentes hardware, que implementados en un dispositivo programable como un FPGA, pueden sustituir la circuitería instalada, con ventajas tales como: la reducción en los costos y la capacidad de reutilización.

En las comunicaciones digitales el estándar más empleado es la interfaz de línea digital de 2.048 Mbps, denominada E1, dentro de la jerarquía de dicha red. Esta tecnología se utiliza principalmente en pizarras telefónicas (PBX), centrales locales o en la trasmisión de datos a través de un enlace dedicado que puede ser desde 64 kbps hasta 2Mbps, y constituye el nivel primario para jerarquías de mayor encapsulado. Se describe en las recomendaciones UIT-T G.732, UIT-T G.703, UIT-T G.704, UIT-T G.706.En este trabajo se presenta el diseño e implementación de un módulo de propiedad intelectual capaz de sustituir el hardware de una interfaz de línea digital E1.

2 Especificaciones de las recomendaciones G.732, G.703, G. 704 y G.706 de la UIT

La trama de trasmisión E1 se conforma a través del multiplexado en tiempo de 32 canales de 8 bits cada uno. Cada trama dura 125μs y según los 256 bits que contiene se alcanza una velocidad de 2048kbps. De los 32 intervalos de tiempo, 30 son canales telefónicos de 64kbps y 2 de igual velocidad destinados a la señalización y sincronismo de trama. Según la información que viaja en el intervalo cero, se pueden distinguir dos tipos de tramas: las pares y las impares. Todas las tramas pares contienen la bandera de alineación, dada por la secuencia “X0011011” que permite la recuperación del sincronismo de 8 kHz en el receptor. Las tramas impares no llevan esta información en dicho canal, en cambio contienen bits de supervisión y alarma.[1]

Para aumentar las facilidades de gestión en la trama E1 de la RDSI (Red Digital de Servicios Integrados) de banda estrecha y para proporcionar a los usuarios canales transparentes con el fin de conservar la integridad de los bits, existe el concepto de multitrama. La multitrama consiste en una agrupación de 16 tramas E1, las cuales a su vez se divide en dos submultitramas (SMT) de 8 tramas, para computar los coeficientes de CRC-4 denominadas SMT I y SMT II. Dichos coeficientes se ubican en el primer bit de cada trama impar de la submultitrama que se envía detrás. Además, en la SMTII las últimas tramas en la misma posición proporcionan información sobre la existencia de errores en la transmisión de los canales de usuario.[2-3]

Un código de línea es usado en un sistema de comunicaciones como soporte para la transmisión. El estándar G703 específica muchas opciones para la transmisión física, especificando de forma casi exclusiva el formato del código de línea HDB3. En HDB3 un ‘1’ se representa con polaridad alternada mientras que un ‘0’ toma el valor cero. Este tipo de señal no tiene componente continua ni de bajas frecuencias, pero presenta el inconveniente que cuando aparece una larga cadena de ceros se puede perder el sincronismo al no poder distinguir un bit de los adyacentes. Para evitar esta situación, el código establece que en las cadenas de cuatro ceros o más, se reemplace el cuarto ‘0’ por un bit denominado “bit de violación” el

cual tiene el valor de un '1' lógico e igual polaridad que el ultimo '1' y si hay cadenas continuas de cuatro ceros o cantidad de violaciones pares además del "bit de violación", se le agregan primero un bit de balance que es un pulso con distinto signo que el anterior. [4]

3 Descripción del diseño del transceptor

El transceptor de flujo E1 quedó concebido para transportar en un sentido un flujo de datos desde su entrada serie de 2048 kbps (STBUSTX) hacia su salida a la línea E1, incorporándole en los canales correspondientes las banderas que requiere, así como la codificación necesaria para la transmisión.

En el otro sentido, el transceptor conduce la trama E1 adquirida hacia su salida serie de 2048 kbps (STBUSRX), a la cual llegará luego de ser decodificada y evaluada detalladamente. Ambos sentidos cuentan con el suministro de las fuentes de reloj necesarias para ejecutar sus funciones y además, se debe proporcionar una interfaz a la que se entregue toda la información requerida por el microprocesador.

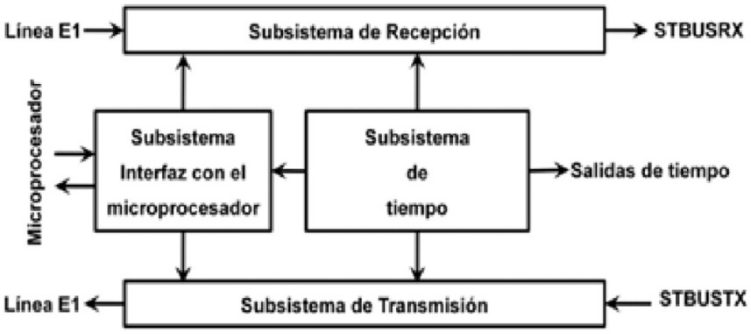


Fig. 1. Estructura del sistema transceptor de flujo E1.

En la figura1 se observa el esquema general del sistema y la relación existente entre los cuatro subsistemas que lo conforman. El subsistema de transmisión, transporta los datos entregados de forma serie por la entrada STBUSTX hacia la línea con la codificación HDB3 y las banderas que necesita la trama para ser transmitida. El subsistema de recepción, recibe los datos de la línea E1 codificados en HDB3 y los entrega por la salida STBUSRX como una cadena de bits decodificados en forma serie. El subsistema de reloj, obtiene las bases de tiempo necesarias en todos los bloques y las ofrece como salidas del transceptor, con el fin de lograr la sincronización con otros dispositivos. El subsistema de interfaz con el microprocesador, brinda la posibilidad de controlar y de verificar el funcionamiento del transceptor.

3.1 Subsistema Trasmisor

En el trasmisor se selecciona la señal de sincronismo (FT sincronismo externo o FO sincronismo interno) y se pone en '1' la señal que habilita el proceso de tras-

misión (STBUSTH). Así, se van tomando los bytes del STBUSTX y se les incorpora la bandera e información de señalización, los bytes conformados van saliendo por la señal DATOBUS2. A continuacion se llenan los canales de la trama que sale por la señal E1, con los datos de la señal DATOBUS2 o del microprocesador (DATOSUP), según el modo de trabajo seleccionado. Por la señal E1 van saliendo las tramas a las cuales se les calcula los coeficientes CRC que serán colocados en la próxima submultitrama. Además, como salida tiene también a la señal DATOTX que entrega el canal que se le realiza la prueba en caso de que el sistema esté configurado en modo de prueba lazo interno selectivo. En la figura 2 se puede observar las señales descritas y los bloques que desempeñan estas funciones. Por último se le realiza la codificación HDB3 a los datos que salen por la señal E1Tx colocando los bits de violaciones y balance correspondientes. Finalmente la señal bipolar es puesta en la salida a través de dos señales SalirR1 y SalirR2.

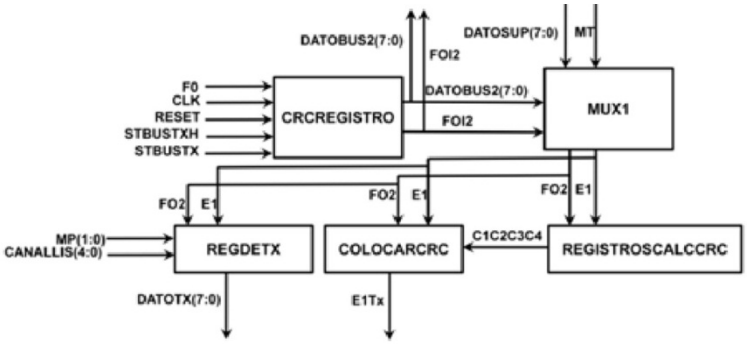


Fig. 2. Estructura del bloque SISTEMACRC del Subsistema Trasmisor.

3.2 Subsistema Receptor

El flujo E1 que circula por la línea pasa primero por un acoplador que le entrega al sistema dos señales EC1H y EC2H que contienen los pulsos positivos y negativos respectivamente de la codificación HDB3. Según el modo que este configurado el microprocesador (normal o lazo de prueba) se toman las señales EC1H y EC2H o las provenientes del trasmisor SalirR1 y SalirR2. Asi, de la entrada seleccionada se crea una señal (Sd) con los bits pertenecientes a la decodificación. Despues se realiza la búsqueda de la bandera de alineación de la trama bit a bit, mientras esto no suceda, el bloque de recepción no tomará ninguno de los datos que se reciban como parte del flujo E1. En caso de pérdidas de alineación, se activa una señal LOS a la salida y se inicializa todo los procesos del sistema. A continuación se realiza la comprobación de CRC y en caso de haber error se activa una señal ERROR a la salida, además los bits de indicación de error de la multitrama son colocados a la salida con la señal CE. Finalmente, la salida STBUSRX estará en tercer estado hasta que el microprocesador la active para poner los canales provenientes de la línea o información del micro (DATOTX). En la figura 3 se muestran los bloques y señales involucradas en este subsistema.

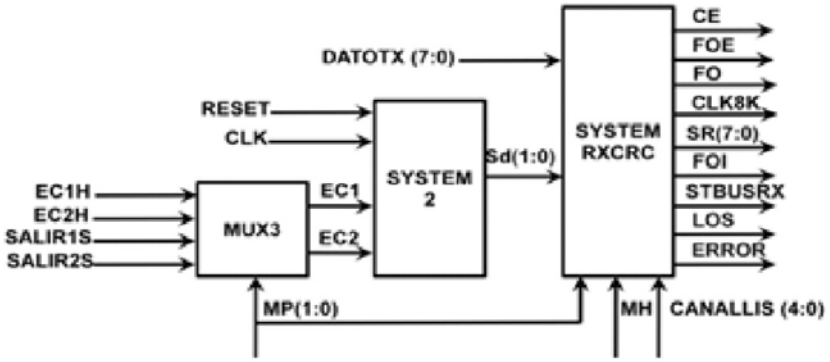


Fig. 3. Diagrama en detalle del Subsistema de Recepción.

3.3 Subsistema Tiempo

Este subsistema ofrece una salida de reloj a una frecuencia de 4096 kHz de la que se deriva la fuente de reloj de 2048 kHz; además debe ser capaz de sincronizarse con la fuente de reloj que se seleccione. Se utilizó el bloque CDR, que constituye una aplicación que brinda el fabricante Xilinx para la recuperación del reloj de los datos recibidos y la generación de un reloj libre de valor especificado cuando no se introduzcan datos a la entrada.

3.4 Interfaz con Microprocesador

La interfaz con el microprocesador constituye un grupo de bloques formado por memorias RAM, donde se almacenan los datos entregados por el microprocesador para la configuración del transceptor y los datos entregados por este (incluye los datos recibidos de la línea E1 y los datos a transmitir por ella), permitiendo el acceso a toda esta información por parte del microprocesador.

El microprocesador configura cada intervalo de tiempo del flujo de salida E1TX en dos modos de trabajo: modo transparente y modo mensaje. Esto se hace a través de dos memorias, una de 32x8 y otra de 32x9. La primera se llena con los datos a transmitir, que se entregan por STBUSTX, y la segunda con los datos de configuración que brinda el microprocesador. Cada vez que se vaya a transmitir un canal, se analiza el bit 9 de la memoria de configuración (‘0’ modo transparente, ‘1’ modo mensaje) y en función de esto se transmite el dato entregado por STBUSTX o el dato existente en los 8 bits restantes de la memoria configurada por el microprocesador.

El microprocesador indica también modos de prueba a través de una memoria de 1x7 donde los 2 bits más significativos indican el modo de trabajo a utilizar (‘00’ modo normal, ‘01’ lazo interno completo, ‘10’ lazo interno selectivo). En el lazo interno completo, la salida del bloque transmisor se conecta a la entrada del bloque receptor. En el lazo interno selectivo, se recibe de la línea una trama E1 a la que es incorporado un canal de la transmisión. Este canal es indicado con los 5 bits menos significativos de la memoria.

4 Comprobación del funcionamiento del transceptor

Se comprobaron los subsistema de manera individual, a través de simulaciones y mediciones capturadas luego de la descarga del software confeccionado en el FPGA Spartan3. Para la realización de las pruebas se utilizan algunas herramientas que permiten la verificación y validación del diseño.

4.1 Comprobación del trasmisor

Para el análisis del subsistema trasmisor se confecciono la tabla 1 que predice el resultado de la simulación. En ella aparece la entrada de datos STBUSTX y el resultado esperado.

Table 1. Entradas y salidas del boque de trasmisión.

STBUSTX	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	1	0	0	1	1	1	0	0	0	0
Salir2	0	0	0	1	0	0	1	0	1	0	1	0	0	1	0	0	0	0	1	0	0	0	1	0	1	0	0	1
Salir1	0	0	0	0	1	0	0	1	0	1	0	0	0	0	1	0	0	1	0	0	0	1	0	1	0	0	0	0

Las señales RESET y STBUSTXH se ponen en ‘1’, lo que equivale a la habilitación del equipo y de la transmisión respectivamente. Estas condiciones establecen que las salidas SALIR1 y SALIR2 deben ofrecer la información entregada por STBUSTX en una trama E1 codificada en HDB3. En la figura 4 se muestra el resultado de la simulación que corrobora el funcionamiento del subsistema.

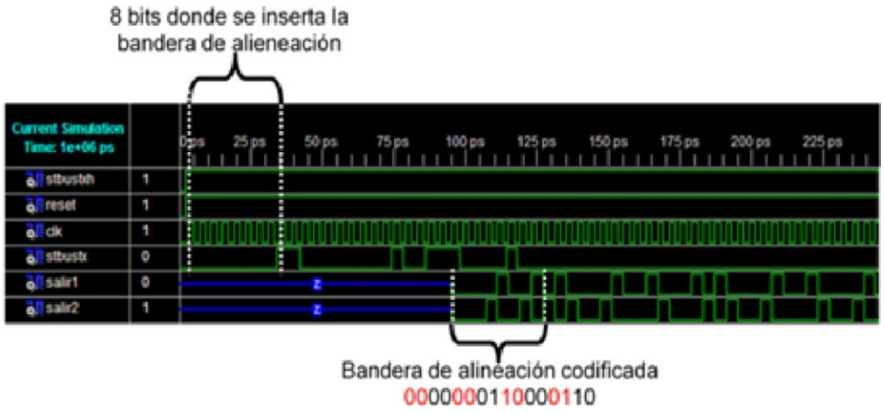


Fig. 4. Simulación del Subsistema Trasmisor.

Para la verificación del comportamiento de este sistema en el FPGA, (ver figura 5) se emplearon algunos interruptores de la tarjeta de desarrollo para indicar la inicialización de la transmisión con la activación de la señales: RESET y STBUSTXH. También se cuenta con el interruptor SFO que indica cual pulso de sincronismo será utilizado para marcar el comienzo de cada trama (FT o FO).

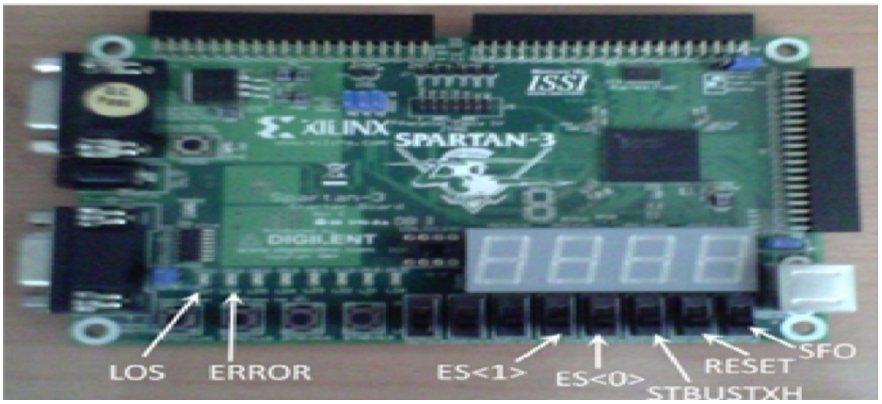


Fig. 5. Interruptores del sistema transceptor de flujo E1.

Una vez puestos a ‘1’los interruptores RESET y STBUSTXH, se inicia la captura de los canales de información proporcionados por el STBUSTX, sincronizados con el pulso de sincronismo seleccionado por SFO (FT para esta prueba).A continuación se trasfirieron a los canales por la entrada STBUSTX ‘0’ y ‘1’alternamente y se empleó el analizador de protocolos para verificar el funcionamiento a nivel de multitramas como lo muestra la figura 6.

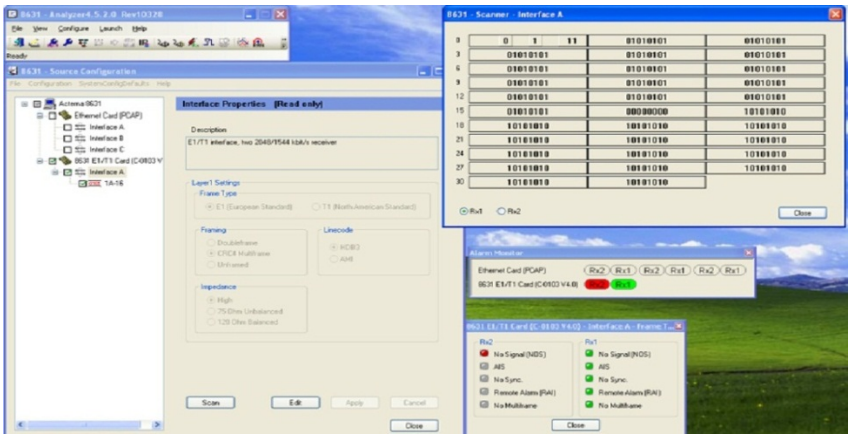


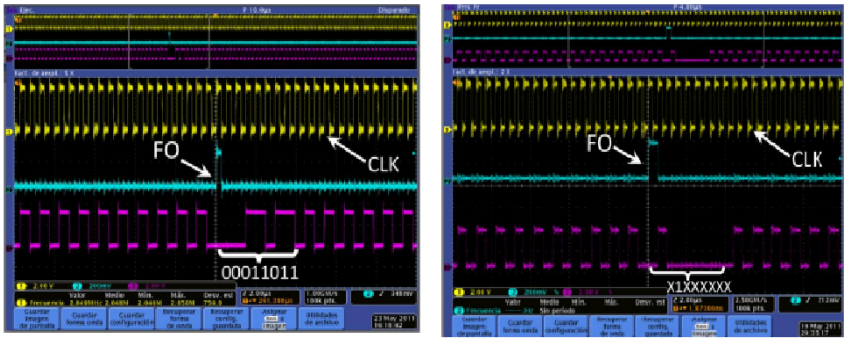
Fig. 6. Analizador de protocolo.

El analizador de protocolos permite configurar los parámetros que se desean verificar. En la ventana de configuración fuente (Source configuración) se marca el análisis de la interfaz E1 de 2048 kbps, el comportamiento de la multitrama CRC4 y la codificación HDB3. El monitor de alarmas (Alarm Monitor), no posee ninguna alarma activa en rojo por el canal de recepción al que está conectado el flujo transmitido; por lo que se puede asegurar que el sistema transmisor realiza correctamente la alineación tanto de trama como de multitrama, la comprobación de los errores y la codificación HDB3.Se puede corroborar que a la salida en los canales la información son los ‘1’y ‘0’ en la ventana que muestra la exploración de

las tramas (Scanner). En el canal cero se muestra la alineación de multitrama CRC y en el canal 16 el mensaje de señalización que en este caso está en cero.

4.2 Comprobación del Receptor

Para la comprobación del subsistema receptor, son de importancia las señales LOS y ERROR porque en dependencia de estas el receptor comienza todo su proceso. Para la comprobación del receptor en el FPGA, se activa el interruptor destinado al RESETpara el comienzo de la recepcion. Inicialmente el sistema tiene 2 LEDs encendidos (ver figura 5), destinados a las senales LOS y ERROR respectivamente, que constituyen las alarmas del sistema. Cuando la señal obtenida de la decodificación ingresa en el detector de sincronismo, se busca la bandera de alineación de trama. Una vez encontrada y comparados los coeficientes CRC4 recibidos con los calculados en tramas anteriores a la que está siendo detectada, estos leds serán apagados. Para la detección de la alineación en las tramas pares debe encontrarse la bandera: X0011011 y en las tramas impares la secuencia X1XXXXXX como se muestra en la figura 7.



Trama Par **Trama Impar**

Fig. 7. Canal cero de las tramas pares e impares del bloque receptor**.**

4.3 Comprobacion de la Interfaz con el Microprocesador

La interfaz con el microprocesador esta estrechamente relacionada con los subsistema de trasmisión y recepción; es por ello que para su validación las pruebas están destinadas unas a las memorias que interactúan con el subsistema trasmisor y las otras a las que interactúan con el subsistema receptor.

Con el sistema de transmisión interactúan las memorias: MEMOR1 y MEMORM. La memoria MEMOR,1 se utiliza para almacenar las tramas que se forman con los canales proporcionados por el STBUSTX que el procesador desea leer. La memoria MEMORM, se emplea para escribir en el canal (o los canales) que defina el procesador el mensaje (o los mensajes) que quiera insertar en el flujo a transmitir.

A continuación se presenta la comprobación de la memoria MEMORM. El BUS_DIR_M activa la memoria de mensajes del microprocesador con la secuen-

cia “01000001” y además selecciona el canal 1 para la inserción del mensaje. El contenido del mensaje se extrae de BUS_DATOS_E_M, que para esta prueba es ‘00001111’. Como se muestra en la figura 8 la señal STBUSTX presenta otra información en ese canal, mientras que E1TX la sustituye por el mensaje dado por el microprocesador.

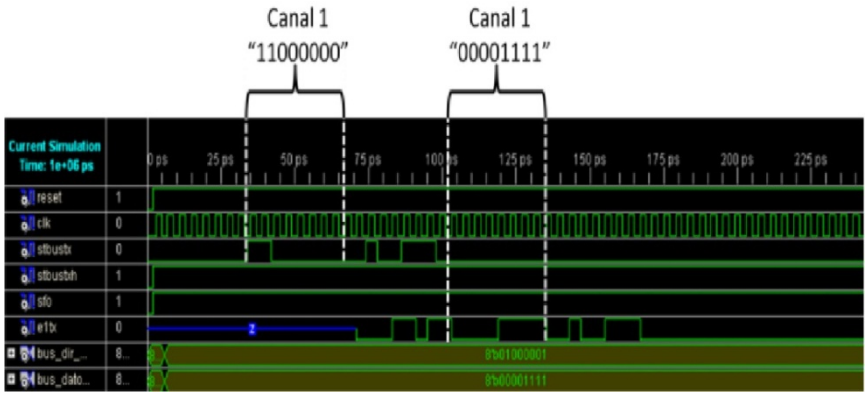


Fig. 8. Simulación de la interfaz con el microprocesador MEMORM.

Con el sistema receptor interactúan las memorias: MEMOR2 y MEMORM2. La memoria MEMOR2 almacena las tramas que se reciben para que el procesador pueda leerlas. La memoria MEMORM2, almacena la configuración de los modos de habilitación para cada canal del bloque recibido. En la interfaz existe una memoria de propósitos generales que incluyen tanto al sistema transmisor como al receptor. Esta memoria (MEMORM3), se encarga de almacenar los modos de prueba configurados por el microprocesador.

Para comprobar el modo de prueba lazo interno completo se escribe en el BUS_DIR_M la secuencia “100XXXXX”, la cual permite seleccionar con los 3 primeros bits, la memoria que almacena los modos de prueba. Luego se le entrega por el BUS_DATO_E_M ‘X01XXXXX’ para indicar con los bits 5 y 6 el modo de prueba a utilizar. Con el fin de restablecer el funcionamiento normal, se vuelve a direccionar la memoria, pero esta vez el procesador debe escribir por el BUS_DATO_E_M la secuencia “X00XXXXX”.

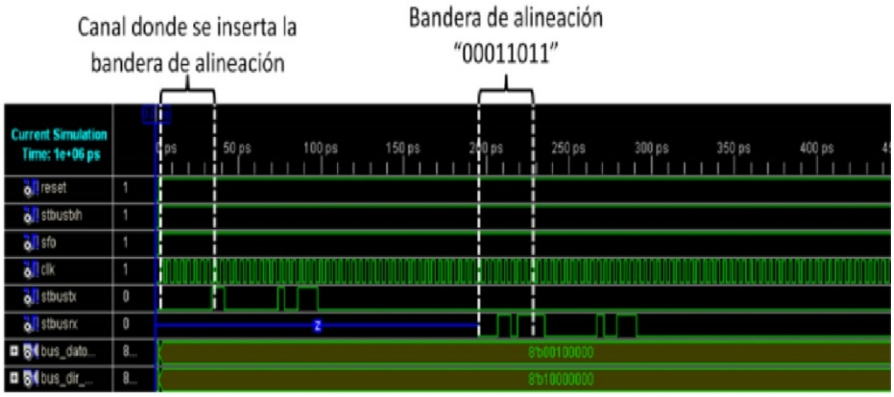


Fig. 9. Simulación de la interfaz con el microprocesador MEMORM3.

Como se observa en la figura 9 todos los canales de información que ofrece el STBUSTX inmediatamente después del canal que reserva para la inserción de la bandera de alineación, son los mismos que se reciben por el STBUSRX. Por tanto es posible notar que el sistema se encuentra funcionando en un lazo que conecta su salida transmisora a la entrada de recepción.

Para validar el grupo de resultados que ofrece la interfaz, se encapsuló todo el sistema trasceptor en un periférico con un procesador Microblaze. Además se empleó un programa desarrollado en el ambiente SDK con comunicación serie (RS-232) hacia una PC con el fin de efectuar la lectura y escritura de las memorias que conforman la interfaz. En la figura 10 se puede observar el ambiente SDK.

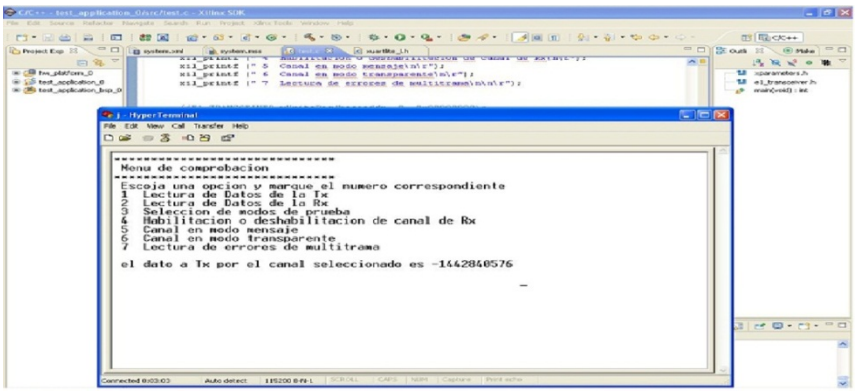


Fig. 10. Implementación de la interfaz con el microprocesador MEMOR1.

La primera prueba realizada consistió en la lectura de los datos a transmitir. Para ello se escogió la primera opción del menú y luego se seleccionó el canal 18 para leer. En la figura 10 se observa el valor que devuelve el programa, el cual en hexadecimal es FAA000000H, de este valor para separar el dato del canal, desechamos el primer simbolo hexadecimal y nos quedamos con los otros dos que

le sigue (AA). El valor entregado por lo tanto es AAH el cual coincide con el valor entregado por el STBUSTX.

A continuación en la figura 11 se muestra la capacidad de habilitar y desahabilitar canales; se deshabilito el canal 13 por lo que en todos va información menos en el 13.

Fig. 11. Implementación de la interfaz con el microprocesador MEMORM2.

Otra prueba que se realizo fue declarar un canal de los datos a trasmitir en modo mensaje.

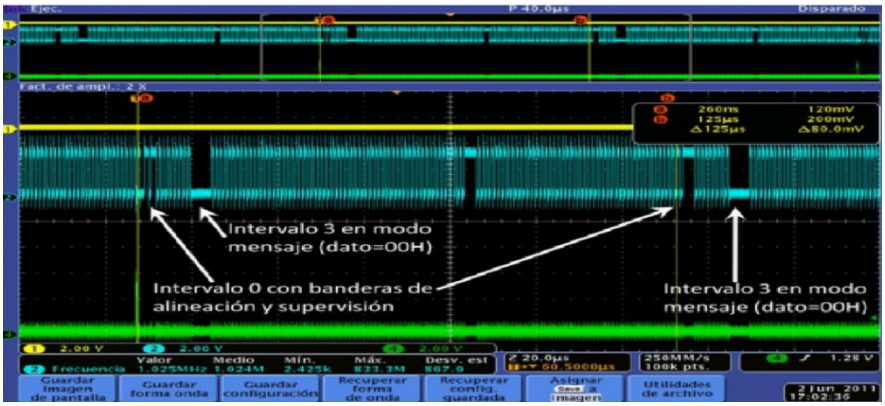


Fig. 12. Implementación de la interfaz con el microprocesador MEMORM.

En la figura 12 se muestra el siguiente caso. El STBUSTX, el cual contiene en todos sus canales un valor predefinido (55H ó AAH), se le declara en modo mensaje el canal 3 para generar el valor 00H. El canal declarado en modo mensaje se mantendrá con esta configuración hasta que se le indique nuevamente modo transparente.

5 Conclusiones

Mediante la utilización de las herramientas de diseño de Xilinx ISE Design Suite 10.1 y la tarjeta Spartan3 Starter Kit, quedó confeccionado un módulo de propiedad intelectual capaz de transmitir y recibir tramas E1, así como realizar todo el procesamiento necesario para el funcionamiento interno con mejor rendimiento y menor latencia que las soluciones comerciales, con su adecuada sincronización. La interfaz con el microprocesador que tiene incluido es una de las contribuciones de este trabajo, porque brinda un mecanismo de detección de problemas en la trasmisión, a través de las pruebas que permite hacer tanto a nivel de trama como a nivel de canal. El análisis de los datos registrados demostró, que tanto en la línea de transmisión como en la de recepción, se llevaron a cabo satisfactoriamente los procesos de alineación de trama y multitrama, la codificación y decodificación HDB3 y la detección de los errores de la multitrama mediante el cálculo de los coeficientes CRC4. Además se logró validar el diseño de la interfaz con el microprocesador integrando el sistema con un Microblaze. La verificación de este módulo expuesto en este trabajo permite la creación de un prototipo en desarrollo que integra al transceptor con un microblaze y un controlador hdlc (control de enlace de datos de alto nivel). Ademas, también se puede usar como base para la confección de un módulo de propiedad intelectual T1 (su estándar homologo norte americano) con un esfuerzo mínimo de diseño.

6 Referencias

1. UIT G.732 : Características del equipo múltiplex MIC primario que funciona a 2048 kbit/s.:In: articulo No. E 11161.
2. UIT G.706 : Procedimientos de alineación de trama y de verificación por redundancia cíclica (VRC) relativos a las estructuras de trama básica definidas en la Recomendación G.704.:In:articulo No. E 2101.
3. UIT G.704 **:** Estructuras de trama síncrona utilizadas en los niveles jerárquicos 1544, 6312, 2048, 8448 y 44736 kbit/s.In:articulo No E 15440.
4. UIT G.703 : Características físicas y eléctricas de los interfaces digitales jerárquicos.:In: articulo No. E 31727.

Corrosion in control systems decrease the lifetime of the electronic devices of the industrial plants of Mexicali, BC, Mexico

Gustavo López Badilla, Benjamín Valdez Salas, Michael Schorr Wiener

[1]Investigador-Académico, Universidad Politécnica de Baja California, Mexicali, B.C., México.
[2] Investigador-Académico, Instituto de Ingeniería, Universidad Autónoma de Baja California, Mexicali, B.C., México.
Paper received on 04/10/12, Accepted on 24/10/12.

Abstract. The principal factor to obtain the economical value of the products manufactured in the electronics industry is due to the production yielding. In this city are around an 80% of companies which fabricate electronic devices and systems or have electronic systems and machines to the manufacturing process. Mexicali is located in the border with the California State of the United States of America (USA). A study was conducted in indoors of three industrial plants to determine the grade level of deterioration of the electronic control systems (ECS) used in the electronics industry of this city. The results showed that to major air pollution detected by specialized methods, the lifetime of the ECS decrease by the generation of corrosion in their electrical connectors and connections at 75% in winter and 50% in summer for this electrochemical phenomenon.

Keywords: Corrosion, electronic devices, control systems, industrial plants

1. Introduction

The permanence of an industrial plant in the world market depends of its planning of the manufacturing, including the methods of the production, specialized people and the ECS used in the industrial equipments and machines [1]. The ECS are very important in the manufacturing process with control operations to assembly parts of a product, detect defects in articles fabricated, count the partial and total products manufactured, make risky operations with toxic chemical, control humidity and temperature that originate the corrosion process, generation of clean environments to avoid the deterioration of the articles fabricated and the ECS, and repair defective articles fabricated, principally. All factors of the functions of ECS

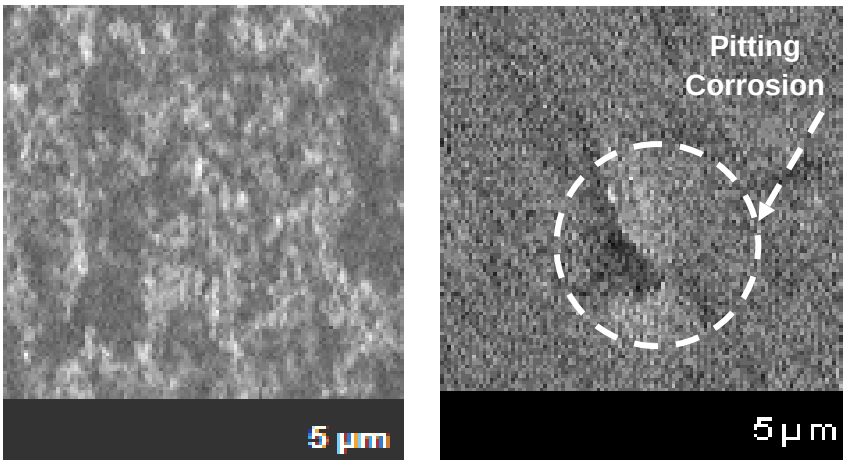
Mireya García-Vázquez, Grigori Sidorov (Eds.).
Advances in Computing Science and Control.
Research in Computing Science 59, 2012, pp. 314-325.

are necessary verify in some periods of the day to are sure that the electronic equipments and machine with ECS are operating correctly. The uncontrolled operations mentioned above, can decrease the production yielding and the low the value of the article fabricated, derived by the capacity of the ECS in the industrial process. The damage of the ECS can decrement their functions and reduce their electrical properties, decreasing their lifetime and causing economical losses. This is affected in the major times, by the uncontrolled microclimate of indoor in the industrial plants, causing uncontrolled chemical processes that affect the operation of the ECS [2]. The presence of air pollutants which penetrate by inlets, air conditioning systems and holes, principally; are deposited in the electrical connections and connectors. The corrosion that occurs in industrial plants generates great economic losses, which concern to the owners of companies and managers and specialized people of the electronics industry. Various research institutions works a lot about this phenomena in different methods to know the origin of the different types of corrosion that can occur. The two more One of the principal types of corrosion which occurs in the companies located in Mexicali, are the uniform and pitting corrosion, where the first type of corrosion is detected very easy because appears in the majorly of the metallic surface. The pitting corrosion is more difficult to detect, because in this type are formed corrosion products over the pitting, and its can run at low or fast velocity to into the metal and not outside this, causing the deterioration of the metals of the electrical connectors and connections and therefore the electrical failures [3].

1.1 The global electronics industry

The world market competition, has led to develop technological advances, particularly in the area of the electronics, which are increasingly being manufactured, smaller components, robust and with a greater number of operations. The tendency to miniaturization of technology is a leader in the development of equipment electronics such as cell phones, electronic organizers personal and microcomputers, primarily. Materials used in the electronic components are alloys based on aluminum (integrated circuits with at micro and nano size of wire), copper contacts nickel electro-plated with gold to improve the resistance to corrosion, but increase the cost of this [4]. The characteristics of modern electronic equipment with high voltage contemplate three principal factors as speed of operations, very small currents and the miniaturization and extremely sensitive to corrosive agents. This can lead to some electrical failures in the microelectronic components, generating technical problems in the ECS, caused by the atmospheric pollutants and the variations of humidity and temperature. Its attack the principal metallic materials mentioned above, used in the electronic devices. The microchips are protected by specialized coatings to prevent that not suffer any type of damage immediately by exposure to aggressive environments and climate changes decreasing its functionality and their lifetime. The corrosion damage tolerance of electronic components is very small order of magnitude (10-12 grams, pico-grams). According to research studies, the order of the width of the films of electronic boards is 50 microns. In hybrid circuits (HC), the space should

be 2.5 microns. This type of microelectronic devices corroded is showed in the figure 1.



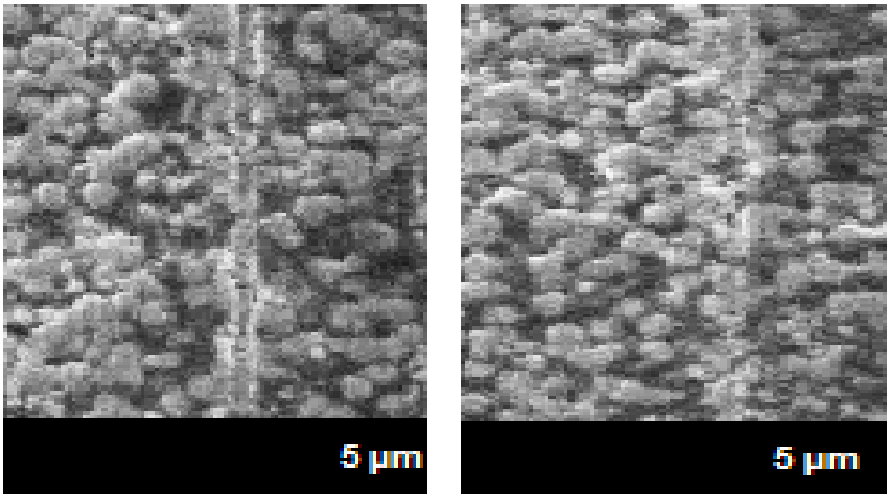
(a) (b)

Fig 1. Microphotographs of metallic surface deteriorate of ECS caused by (a) uniform corrosion and (b) uniform and pitting corrosion in a industrial plant in Mexicali (2011).

As show in the figure 1, the metallic surfaces of the microelectronic devices were damaged by the corrosion process, indicating in each microphotography the presence of uniform corrosion, but in the section b, appeared pitting corrosion with a little pit where not was observed to the naked eye, and only was detected with the microanalysis [5].

1.2 Corrosion in the electronics industry

The electronic boards suffer a variety of problems in their electrical conduction surfaces by the presence of the air pollutants as sulfurs principally and the levels of relative humidity (RH) and temperature. When its parameters are combined, the atmospheric agents react very easy with the metallic surfaces, decreasing the resistance to the corrosion of the connectors and electrical conductive paths. This causes the formation of metallic filaments that grow between routes where not debit growth, causing electrical conductivity between terminals metal (pins), or metallic unions [6]. The conditions required for this type generating a combination of ionic contamination, humidity (> 70%) and temperature (> 30 ° C) and the voltage application. The ECS used in very dry conditions almost not suffer from the corrosion, but at high RH and low in some months of winter and high RH and temperature in some periods of summer temperatures this process, originates the condensation forming visible or invisible thin films of water that occurs on the surface of the metallic connectors and connections, beginning the corrosion. This originates the absorption very fast of the air pollution by the micro and macro electronic components, decreasing the metallic surfaces by the deterioration and for consequence the strength of the material forming metallic dendrites (Figure 2).



(a) (b)

Fig 2. Microphotography of metallic dendrites formed in pathways of electronic devices in (a) summer and (b) winter.

As show in figure 2, in the both microphotographs (a and b), the pathways where flow the electrical current was corroded, indicating that corrosion products cover not totally the metallic surface and causes a pitting corrosion under the corrosion products.

1.3 Atmospheric corrosion

The climatic factors generate the dispersion of air pollutants such as fine particles and gases: hydrogen sulfide, sulfur dioxide, carbon monoxide and nitrogen oxides. This air pollutants, in sometimes exceed the standards of air quality in this region. These contaminants are detected by specialized teams of Environmental Monitoring Stations (EMS) installed in strategic places of the Mexicali city. These chemical agents are emitted by external sources such as traffic, industrial plants, geothermal fields, soil erosion and microorganisms, which penetrate to the indoor of the electronics industry. At levels above 80% of RH and 25°C, that is common in some periods of the year in Mexicali, were obtained the time of wetness (TOW), indicating the periods in which the metal surfaces of electronic devices kept moist for at least one day, causing the electrochemical corrosion process. The pollution levels are evaluated according to standard regulations that indicate the concentration levels regulated by environmental institutions in each country. In Mexico, the Secretaría del Medio Ambiente y Recursos Naturales (SEMARNAT), in mutual agreement with the Environmental Protection Agency (EPA) of USA [3, 7], are responsible for the regulation of the pollution emissions of the outdoor and indoor of the industrial plants. With international agreements, on the border of Mexico and the United States, SEMARNAT and the EPA, have installed Environmental Monitoring Stations (EMA) and Meteorological Monitoring

(MMS), to obtain information about the outdoor environment that have effect in the levels of pollution and the variations of the RH and temperature showing diverse values in different seasonal periods. The chemical agents which generate the air pollution penetrate to the indoor of industrial plants, where are monitored principally in this city the hydrogen sulfide (H2S), sulfur dioxide(SO2), carbon monoxide (CO), nitrogen oxides (NOX) and the ozone (O3) and the particulate matter (PM10 and PM2.5), reported by the National Ambient Air Quality Standards (NAAQS, 2006). Additionally are monitored the volatile organic compounds (VOC) generated in indoors of the companies, like ammonia that is inorganic agent and have an effect in the deterioration of metallic connectors and connections of the ECS, and even at low concentrations, have a significant effect on the corrosion phenomenon [2].

1.4 Corrosion types in the electronics industry

According to the humidity originated in indoors of industrial plants, material of electronic devices generates different corrosion processes, such as those explained below [1]:

A). Corrosion voltaic. This type of corrosion is generated in a small space between electronic circuit components, when a voltage is applied to a device generating the voltage gradients in order of mega-volt circulating component surfaces, resulting accelerated electrochemical corrosion reactions and ion migration. The integrated circuits are very susceptible to corrosion. The combination of electrical fields, with atmospheric humidity and air pollution are the main promoters of this type of corrosion in metals used in the microelectronic devices. The corrosion can occur only with the presence of low humidity in sometimes. The pH generated by a cathodic reaction, by reduction of water. High pH levels result in a solution in the passive film the surface of the metal oxide and aluminum substrates and increases driving resistance, generating an open circuit.

B). Electrolytic metal migration. In presence of moisture and an electric field, silver ions migrate to a charged surface cathodically, forming dendrites, which grow as bridges between contacts, possibly causing short circuits to reduce the production yielding of the ECS and decreasing their life time, and in sometimes causing fires. Always a small volume of he dissolved metal forms the dendrites in form large. Other materials which are susceptible to metal migration ions are the gold, tin, lead, palladium and copper. Dendrites can be silver, copper, tin, lead or combination of these metals and cause electrical failures in industrial electronic equipment for short circuits. The dendrite growth can be generates very fast the electrical failures of ECS, and always is known to can be have effect at least after 30 minutes or sometimes up to several days, weeks or months until originates great damage in the electrical connections and connectors. The growth rate of dendrite depends on the applied voltage, the amount of contamination atmospheric combined with the moisture, which affect the metal surface.

C). Formation of pores and cracks in electrical contacts and connections metal. To prevent the connectors and electrical contacts stained by the process of corrosion metals such as gold, the process is performed on the surface of silver contacts and connectors. However, defects in protective coatings of metals can expose and corrode the substrate material and create imperfections like pores and cracks. If the substrate is copper or silver, and it is exposed to environments with sulfates or chlorides, the products corrosion, cracks will generate pores, and if gold added to silver forms a high resistance in the conductive layer, whereby open circuits are generated.

D). Fretting corrosion on separate jacks with fine finishes. It results in the formation of tin oxides boards electronic or electrical contacts. The problem starts very often when the tin is used or replaced by gold metal being a more economical. The possible solution is to replace the hand or use a little more expensive metals.

E). Galvanic corrosion. It occurs when two different metals, such as aluminum and gold come together, as it is in the encapsulation of integrated circuits. The polymers used are packaging as a pair porous and plastic tapes that seal metal joints are manufactured as electronic devices ceramic or metallic. Sometimes, humid environments generate galvanic corrosion conditions.

F). Corrosion in industrial processes. Integrated circuits are environments exposed to numerous aggressive attack ionic or wet and aluminum, which are the main compounds. The ion etching requires a combination of gas and if this is formed as chlorides aluminum covering the metallic structures of the ECS, which are acidic agent and generates acid moisture. The ionic contamination occurs in soldering processes and handling of materials used to fabricate microelectronic devices with very thin films, originated by dust and changes climate.

G). Microcorrosion in the manufacture of integrated circuits. The metallization of aluminum and copper alloys can form inter-metallic compounds like Al_2Cu with a large grain boundaries. This is the beginning of the dissolution of metals which form a micro-pitting during the etching.

H). Corrosion by chlorinated solvents or halogenated. Liquid solvents steam or used in the manufacture of integrated circuits, mainly corroding aluminum components. Contaminated water of solvents increases the time of the presence of corrosion and also increases the corrosion rate, which is the speed at a metal which disintegrates. The stabilized solvent dissolution with alcohol or aromatic solvents is the main cause of halogenated solvents breaking and forming chlorine ions, corroding the aluminum and copper alloys, with major effect in the aluminum.

I). Corrosion in welding. Corrosion resistance joints of tin and lead in aqueous and gaseous environments is a function of the alloy. This significantly improves when increases more than 2 times the rate of the alloy. Lead forms unstable oxide, which reacts readily with chlorides, borates and sulfates.

1.5 Mexicali is an arid region

The topography of the Mexicali city and the levels of air pollutants mentioned above, penetrate to indoor of the electronics industry where there are about 156 industrial plants according to the AMAQ [8]. These contaminants in certain seasonal periods generate aggressive climates mainly in production areas and warehouses where are used and store the microelectronic devices and the ECS. The climate is an important factor, and in the operational functions of the ECS of this city. The RH of the city is around from 50% to 90% and the temperature ranges are near of 0 ° C and the higher value are around the 45 ° C. These climate changes vary the outdoor and indoor of the industrial plants, contributing to the deterioration of materials. Dust and chemicals in indoor environments, are added to the microelectronic components of ECS such as computers, measuring instruments and industrial machinery as showed in the figure 3. A computer operating properly when it is free of contaminants and the microclimates are controlled [9, 10].

Fig 3. Electric fan controlled by ECS of a industrial machine contaminated by dust and chemical agents

1.6 Mathematical simulation

MATLAB is a software called Matrix Laboratory, being a mathematical software that offers integrated development environment (IDE) with a proprietary programming language. Its basic features are: array manipulation, data representation and functions, implementation of algorithms, creation of user interfaces (GUI) and communication with programs in other languages and other hardware devices [11]. The MATLAB package has two additional tools that expand their services: Simulink and Guide Commands. In this study, this software was used to made correlations about the climatic factors and the air pollutants related to the corrosion rate (CR), which represents the deterioration levels of the materials used in the electrical connectors and connections of the ECS.

2. Methodology

Humidity and temperature are the two most important climatic parameters related to the origination of the corrosion process. For this study was used information of the air pollutants mentioned above, and also the temperature and RH of 2010 to 2011 obtained of the specialized equipments. Based on the statistical information of the climate factors was made an analysis of the generation of the different types of corrosion in the ECS, which use microelectronic devices. According to the emission supplies, the levels of air pollution are concentrated in some areas of Mexicali, principally where are located the electronics industry. In Mexicali the temperature is very hot in summer until major of 45 °C in some periods of this season and very cold nights until near of 0 °C in the winter. The maximum and minimum recorded monthly periods of relative humidity and temperature of indoor show ranges varied be under 10% and above 90% for RH, and temperatures of 0 ° C in the winter to nearly 50 ° C in the months of June to August. The TOW monthly values correspond to low and medium levels of corrosivity, with ISO 9223 and ISO 11844-1 [12, 13, 14]. The study was made in three industrial plants of this city principally of the electronics area from 2010 to 2011.

3. Results

The levels of RH and temperature promote the process and influence the dispersion of air pollutant from natural and anthropogenic sources, causing the corrosion process. In the summer the atmospheric pollution, is based on the temperatures in the range from 30 °C to 42 °C and the RH levels ranging from 55% to 85% with a higher intensity is observed of 90%, representing low dispersion of air pollutants remaining near of the industrial plants and penetrating to its companies. This originates in some periods of the year and some areas, a high concentration of pollutants, increasing the CR and the deterioration of the metallic surfaces of ECS. The improvement of this job is to save costs in the electronics industry which give the opportunity to make the study, for not generate economic losses by the attention of warranty in the periods where the products manufactured in this industrial plant and others with the same production process, and debit be operating in good conditions. Before this study around of 15% of these products were returned to this industry to the warranty.

3.1 Correlation analysis of CR in ECS and climatic and environmental factors

The CR was influenced by the exposition to different levels of the air pollutants mentioned above, damaging the microelectronic devices. In this study a correlation analysis was made to know the grade of deterioration of the micro components of the ECS. At different levels of RH, temperature and concentration levels of air pollution, was obtained diverse values of correlation, indicating the major indices in winter where occur the condensation phenomenon and the fast and easy

adherence of the air pollutants, as show the table 1. The decreasing of the ECS in sometimes for the corrosion process, generates low productive yielding and every time the microelectronic devices suffer damage decreasing their lifetime. The correlation was divided in four ranges.

Table 1. Correlation analysis of CR in different seasonal of the year

	Ranges		Ranges			
	T, 0 °C– 20 °C	RH, 0% - 40%	T 21 °C– 35°C	RH, 41% - 75%	36 °C– 45°C	76% - 90%
Spring	0.76	0.78	0.77	0.79	0.79	0.78
Summer	0.82	0.83	0.84	0.86	0.84	0.85
Autumn	0.75	0.74	0.77	0.75	0.76	0.77
Winter	0.89	0.87	0.90	0.90	0.88	0.89

Table 1 shows the levels in each seasonal period where are correlated the climatic factors with the CR. In the spring season, the major CR was 0.79 at the ranges from 36 °C to 45 °C and 0.78 from 76% to 90%. In this season, the presence of corrosion was low, without high effects in the deterioration of metallic surfaces and the lifetime of ECS. The lower value of CR was 0.76 at leveles from 0 °C to 0.78 from 0% to 40%. In summer the higher value of CR was 0.84 at levels from from 36 °C to 45 °C, where was presented pitting corrosion at value higher than 75% and 35 °C, and the CR of 0.85 was the highest value at ranges from 76% to 90%. In autumn was indicated the lowest value of the analysis with 0.75 from 0 C to 20 C and the higher level was a intensity 0.77 of CR at ranges of 21 °C to 35 °C. In this season the corrosion appear at low intensity as same as the spring season. In winter the CR was represented the highest value with 0.90 at ranges from 21 °C to 35 °C and 0.90 at levels from 41% to 75%. In this season was presented uniform corrosion at values higher of 70% and 30 °C. This represented that in winter the metals suffer of more damage than in others periods of the year.

3.2 MatLab simulation

A factor showed in this study was that at low and medium concentrations of air pollutants, the CR was very fast, but at high concentration of air pollution, the CR was low, as mention the standard concentration level to the sulfurs is 75 ppb, according to the EPA. Furthermore, in winter, levels near of 10°C and ranges of RH from 35% to 70%, there was a lower incidence of condensation of water on the metal surface and the CR was high. In the winter time, it has higher air pollution in the temperature range of 2 °C to 13 °C and RH levels from 34% to 70%. At values of temperature from 15 °C to 20 °C and RH of 45%, 75%, the CR was higher. At temperatures of 25 °C to 30 °C, with RH levels from 30% to 75%, there is a low dispersion of air pollutants, showing a CR high and more impact in the deterioration of the metallic surfaces of ECS. The analysis in this research in the summer for the dispersion of air pollutants, indicating that focus on larger scale, at temperatures below 20 °C, showing a CR low at 35% to 55%, being a high dispersion of the air pollutants in indoor of the industrial plants located in

Mexicali. The corrosion was stabilized at 40 °C and was a CR very low as show figures 4 and 5.

In figure 4, the maximum CR of the correlation was 391 mg.m^2/year at 28 °C and 78% temperature and RH levels showed fast deteriorate of the metallic surfaces of the electrical connectors and connections of the ECS. This causes lack of electrical current and not function the ECS, originating damage in some microelectronic components of the ECS and decreasing their lifetime. The minimum CR was 9 mg.m^2/year at 11 °C and 54% temperature and RH levels. In figure 5, the maximum CR was 209 mg.m^2/year at 42 °C and 78% of temperature and RH levels and the minimum value of CR was 6 mg.m^2/year at 19 °C and 58% of temperature and RH levels.

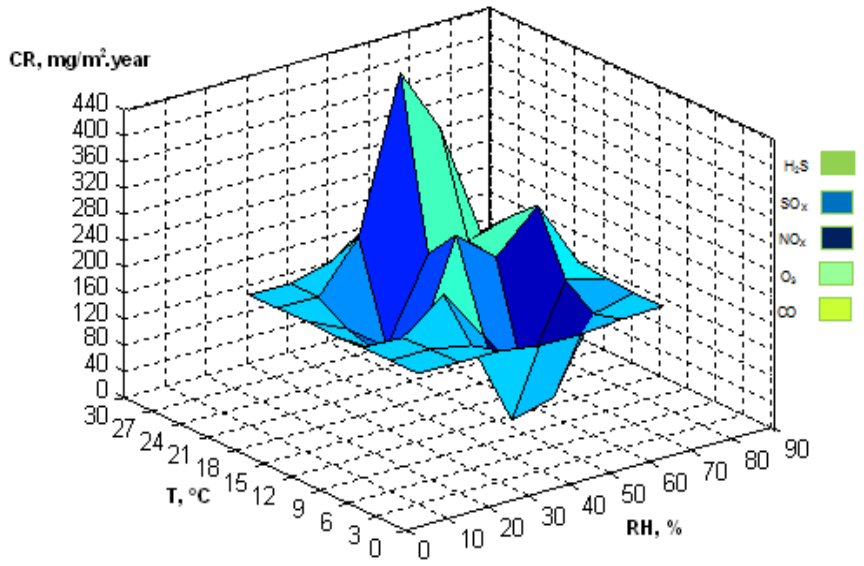


Fig 4. Correlation of CR and climatic factors with air pollutants monitored in Mexicali in winter (2010-2011).

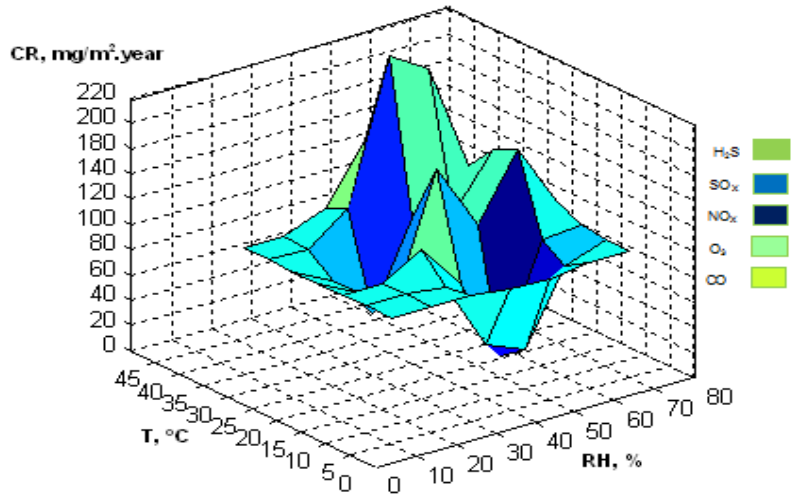


Fig 5. Correlation of CR and climatic factors with air pollutants monitored in Mexicali in summer (2010-2011).

4. Conclusions

With increasing miniaturization of systems electronics and the explosive increase in its availability, it is estimated that the corrosion and deterioration of metal materials electronic will increase causing incalculable consequences. The corrosion phenomena have an effect in the operation of ECS with microelectronic used in the electronics industry. The presence of acidic substances in indoor of industrial plants to damage the electrical connectors and connections of ECS, originates aggressive environments, which generates very fast the deterioration and decrease the lifetime of the ECS of industrial equipments and machines. This affects the functionability every time and originates electrical unoperation that causes the defect before they present their standard lifetime, and begins to fail electrically. This concern to the specialized personas and managers because generates unnecessary costs. The types of corrosion detected in the connections of electronic devices as microchips were the uniform and pitting corrosion, which are generated by the different values of humidity and temperature in the diverse periods of the year. The climatic factor with major effect was the humidity, which was changing in according to the seasonal period. The uniform corrosion was formed in the winter season and the pitting corrosion appeared more frequently in the summer period. This was evaluated to determine the periods of the year, which are a corrosion rate (CR) high in some periods of the year, decreasing the manufacturing process until 70 %.

5. References

1. López Badilla Gustavo, "Caracterización de la corrosión en materiales metálicos de la industria electrónica en Mexicali"", Mexicali, B.C., Mexico, 2008, 115 pag (Tesis Doctoral).
2. B.G. Lopez, S.B. Valdez, K.R. Zlatev, P.J. Flores, B.M. Carrillo and W.M. Schorr; "Corrosion of metals at indoor conditions in the electronics manufacturing industry", Anti-Corrosion Methods and Materials, 54/6, 2007, 354–359, ISSN 0003-5599.
3. Annual Book of ASTM Standards (G4, G15, G31), Section Three; Metals Tests Methods and Analytical Procedures, Vol. 03.01; 2001.
4. Ashrae, Handbook, Heating, Ventilating, and Air-Conditioning, Applications; American Society of Heating, Refrigerating and Air-Conditioning Engineers, Inc.; 1999.
5. Nishikata A. and Ichihara Y, The effect of Time of Wetness (TOW) in metallic components; **Corrosion Science**; No. 37, 1995, paginas 134-138.
6. Velasco, L., La contaminación atmosférica de las grandes urbes, Ciencia y Desarrollo, marzo-abril, 1996.
7. NAAQS- National Ambient Air quality Standards; U.S. Environmental Protection Agency (EPA); consultado en: http://www.epa.gov/air/criteria.html; [consultado en agosto de 2012].
8. Asociación de Maquiladoras de Mexicali-AMAQ, Departamento de Estadística, Reporte Anual de la Industria en Mexicali, Gobierno Municipal; 2008.
9. Camuffo, D. and Bernardi, A., Controlling the Microclimate and the Particulate Matter inside the Historic Anatomic Theatre, Padova. Museum Management and Curatorship, 1997, paginas 285-298.
10. Morcillo, M., Evaluación de la polución ambiental en interiores de empresas, Madrid, España; Revista Iberoamericana de Corrosión; No. 17, Vol 42, 1996, paginas 112-115 .
11. B. H. Duncan, Walsh An Engineer's Guide to MATLAB, 2e: with Applications Electrical Systems, Prentice Hall, 2005.
12. ISO 9223:1992, Corrosion of metals and alloys, Corrosivity of Atmospheres, `Classification.
13. ISO 11844 PART 1. Corrosion of metals and alloys- Classification of low corrosivity f indoor atmospheres- Determination and estimation of indoor corrosivity.
14. ISO 11844 PART 2. Corrosion of metals and alloys- Classification of low corrosivity of indoor atmospheres- Determination and estimation attack in indoor atmospheres.

Electronic system to save water due to their decrease in Mexicali by the coating of the all American canal in USA

Gustavo López Badilla[1], Elizabeth Romero Samaniego[2], Sandra Luz Toledo Perea[2], Miriam Maleni García Castrellon[3], Luis Alberto Gameros Rios[3]

[1] Investigador-Académico del Instituto Tecnológico de Mexicali (ITM), Mexicali.
[2] Investigador-Académico del Instituto Tecnológico de Ensenada (ITE), Ensenada.
[3] Alumnos de la Carrera de Ingeniería Industrial del Instituto Tecnológico de Mexicali (ITM), Mexicali.

Paper received on 06/10/12, Accepted on 25/10/12.

Abstract. Efficient use of water is an essential part in the development of the entire area of each country. We consider the implementation of automated systems to maintain adequate fluid flow to vital activities performed in each area of a city or the countryside. The main automated control systems are developed with basic electronic devices coupled application geese sometimes complex components or systems; the cost should not be too high and be developed in an easy, to be operated quickly and easily. In cities or regions of the world where you have water shortages, government authorities are working flat out to supply water to reduced or restricted. In this research project is evaluated considering the issue of the possible restructuring of the All American canal coming from the northern United States (Wyoming), through Mexico and southern reaches of the continent. The U.S. government intends to take the channel, which involves placing a foundation plate, whereby certain areas of the border region of Baja California and California, they have problems of lack of precipitation into the ground and thus not to generate wells in Mexicali valley regions, where water is drawn for supply to farms. Efficient use of water is an essential part in the development of the entire area of each country. We consider the implementation of automated control systems (ACS) to maintain adequate fluid flow to vital activities performed in each area of a city or the countryside. The main automated control systems are developed with basic electronic devices coupled application geese sometimes complex components or systems; the cost should not be too high and be developed in an easy, to be sold quickly and easily. In cities or regions of the world where you have water shortages, government authorities are working flat out to supply water to reduced or restricted. In this research project is evaluated considering the issue of the possible restructuring of the All American canal coming from the northern United States (Wyoming), through Mexico and southern reaches of the continent. The U.S.

Mireya García-Vázquez, Grigori Sidorov (Eds.).
Advances in Computing Science and Control.
Research in Computing Science 59, 2012, pp. 326-335.

government intends to take the channel, which involves placing a foundation plate, whereby certain areas of the border region of Baja California and California, they have problems of lack of precipitation into the ground and thus not to generate wells in Mexicali valley regions, where water is drawn for supply to farms.

Keywords. Electronic systems, save water, All American canal, control systems

Introduction

The efficient use of water is important in the development of any country. To maintain an ecological behavior to water saving in any region is necessary install ACS. The northwest of Mexico, where they are located cities in Sonora as San Luis Rio Colorado (SLRC) and Mexicali in Baja California, are desert areas that require water for domestic, industrial and agricultural activities. For this reason, we propose the use of ACS to save water with electronic devices at low costs. This study was performed from 2007 to 2011 to support to people in the cities mentioned above. People of SLRC and Mexicali, is concerned because water levels will be decrease in the next five years by the covering of the All American Canal, which is a source to this region. When people use polluted water as water recycled of agricultural operations to domestic activities generate stomach and respiratory infections. The reduction of water in underground aquifers (UA) in this region, damage the soil and generate environmental deteriorate and droughts that affect the agricultural and economic operations [1].

1. The use of water

Experts in the field of water protection, think that water shortages in some countries have a negative effect in the health of people. The World Health Organization (WHO) indicates that more than half the world population is victim of water scarcity, which has contributed to climate change in some regions and the generation and propagation of actual and new diseases [2]. In Mexico some regions has by water shortages and decrease the productivity in agricultural, commercial and industrial operations. The northwest of Mexico is prone to suffer the negative effects of water scarcity [3]. Specialists in this area, consider that the developing of environmental policies and engineering support are good methods to contribute to save water. With water scarcity, decrease the productivity of food vegetables, legumes, fruits, wheat, corn and food for animals that produce food the people, increasing the costs of its products, generating economic problems [4]. According to a report of the Comisión Nacional del Agua in Mexico (CNA), in this region, water decreased, and declined in UA and wells in the last 20 years and about of 12,000 hectares was damaged and is infertile zones to agricultural activities and to be used for people. Water of this All American Canal comes from the northern of United States, in Wyoming, and is a source to SLRC and Mexicali

cities and valleys. This has been reducing the commercial and industrial operations in the last five years.

1.1 Water for life. Water is the key to the survival of life, and is essential for the viability and development of any civilization, in order to respond to the requirements and basic needs of communities. Some difficulties have been analyzed and proposed solutions for the provision of water to small and large populations to agricultural and industrial activities [5]. In the last 30 years water has been scarcited and is not enough to supply to people. A whole range of actions has been important factors to avoid the damage in sources of water used in human activities [6].

1.2 Ecological use of water. Is very important approach the water resource adequality, involving to public and private institutions and the people. Is very significant that government agencies, universities and public and private institutions know about it problem, and make new activities to present and future periods to ensure the sources of water [7]. Is necessary protect our environment to ensure to the present and future generations the sufficient food and water, that are the most important tasks to any society [8].

1.3 Domestic consumption. Many houses in developed countries consume already of 1500 liters of water at week [9]. At the same time more than 1,700 million people not have access to this water. The WHO considers an ideal consumption of 100 liters per day [2]. Two thirds of Mexico have water shortages, while in other zones exist excessive use of this resource [4]. This unequal distribution, generate serious environmental problems, originating climate change and economic and social lack of balance**.**

1.4 Irrigation programs. Whit low productivity of water, the negative effects are the salinization, desertification and erosion [6]. This result in infertile soils that immediately leads to food shortages that are a serious problematic situation in the world with a population, that growth very fast around 90 million people per year. According to the Food Agricultural Organization (FAO) since 1950, the water consumption increase at three times in the world. While the consumption was increased in 50% in the last 50 years, the sources have less levels of water. In the agricultural activities water is used at 70%, in industrial plants at 20% and only 10% to people. Today, a quarter of the world's countries have insufficient water in both cases: quantity and quality, by the inappropriate use, increasing the risks of health in the population, principally of stomach and respiratory infections [9]. Some studies estimate that 80% of all illnesses and 33% of deaths in developing countries are related to the inadequate quality water. According to the United Nations (UN) mentioned in a study called "four of five endemic diseases in developing countries are caused by dirty water or lack of health facilities", health of people is down every day. The WHO reported that water scarcity is responsible for three quarters of the 49 million of deaths that occur on the world each year, and 2500 world half-million people suffer from diseases associated with water

pollution and lack of hygiene, indicating a strong correlation between the failure and quality of the resource and the occurrence of diseases [3].

1.5 Water scarcity and global economy. Simultaneously with the environmental degradation, the economic conditions of the population in most developing countries have stagnated or decayed with the reduction of water [8]. While in the 60's, water was enough to cities and towns, at the beginning of 21th century, populations were immigrated to new regions by water scarcity. This produces problems in the economy and social factors. At this time, developed countries, with 20% of world population, poor countries have the 80% of population that around 60% have water shortages. Currently, more than one billion people live below the poverty line, with a dollar at day [10]. The reconfiguration process involves a strong economy and continued polarization in the distribution of water to have good life and avoid economic problems. Will it be possible to have life of high quality and can control the natural resources of this world?

2. Methodology

The method used in this study was a examination of use of water in the cities mentioned above and propose the use of ACS to save water, to be used in agricultural, industrial, commercial and domestic activities. The analysis has four steps:

a). Sonora-Baja California analysis in the use of water. According to the CNA, principally in San Luis, Sonora and Mexicali, Baja California, some areas of this region have an inappropriate use of water and estimates that can extend the problematic situation of lack of water in the next five years. For reason this public institution consider that is very important this study.

b). Covering of All American Canal and the negative effects in SLRC and Mexicali. It is of great concern, knowing that the precipitation of water into the subsoil and therefore in some parts of the valley of SLRC and Mexicali will reduce the water levels and is necessary to care water and decrease the food production.

c) Microscopic soil evaluation. Evaluation of different areas of this region, where was applied and nor used the ACS, with Scanning Electron Microscopy (SEM) in various areas of the cities evaluated.

d). Manufacture the ACS. In some areas of SLRC and Mexicali, is estimated that about in five years, will be a reduction of up to 80% of the water sources. For this reason is important use this electronic equipment to save water, at low cost with an efficiency of 100%, tested for five years in the cities.

2.1 Proposed water conservation. In this study, there is a proposal of the use of ACS, which operates with solar energy, at low cost (around $10 dlls) and maintains the high levels of the sources of water. The first step involved an

analysis of polluted water by solid residues in the cities evaluated with values of 43% in Mexicali and 39% in SLRC. The second step was developed to manufacture an ACS and the third step to probe the operation of the ACS.

2.2. Analysis in domestic activities. The ACS was probed in domestic activities with an efficiency of 100% in the five years of the study. There were some tests to develop with the 100% of efficiency of ACS (Figure 1), such as:

- Evaluation of water consumption in agricultural, commercial, domestic and industrial activities.
- Use a methodology for developing the ACS to promote the environmental awareness.
- Analysis of electronic devices with an efficient design at low cost.
- Proposal to be applied in domestic and educational activities to experimental testings.

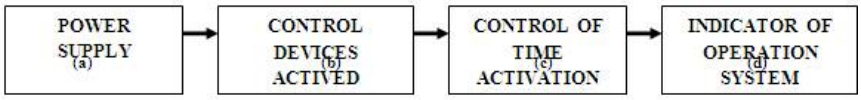


Fig 1. Steps of operation of the ACS to save water, probed in Mexicali

Fig 2. ACS equipment used in Mexicali

The ACS for water saving consists of four parts that generate the process of the operation: (a) with a power supply that generates the necessary energy to activate (b) the electronic devices (Integrated Circuits, IC) and activated, according to a timer device with the (c) electronic components to control the operation, and generated by electrical actuators. Automatic IC timer programmed is good to use to this activity, but is very expensive and not easy to program, only by specialists. Finally we have the signal indicators on and off, controlled by a device that identifies the periods in which automated equipment is activated and not activated. With this automatic system ensures a savings of approximately 80% of water in

domestic, commercial, industrial, agricultural activities in the cities mentioned above from 2007 to 2011. The power supply is of 12 volts of direct current to provide the electrical energy to turn on the electronic system. The control devices to active and turn off is with electronic components used in the electronic sensors and the process to maintain operating the electronic system depends of a value of electronic resistor, being represented by a luminescent indicator to indicate when is turn-on and turnoff the ACS.

3. Results

The ACS contributed to saving water for five years in the Sonora-Baja California region. This prevents water shortages, as they are proven to perform an ecological awareness in the society of this zone of Mexico. Table 1 shows percentages of efficiency of use or not of ACS from 2000 to 2009 in Mexicali and SLRC, indicating the necessity of use the ACS that probed a saving process of water at 55% exactly. The water saving percentage was 45% in San Luis Rio Colorado with a population of 350,000 and Mexicali with 1.000, 000 was 55% [11]. These populations required the use of ACS, which decreased by use of water in order of average of 50%. Water supply varies depending on the activities in each city in this arid region, where the consumption was evaluated in domestic activities, increase the water saving. Table 2 represents the use of water in industrial activities, where the consumption was high. Table 3 indicates the use of water in agricultural operations, that is the second activity that consume a lot quantities of water, less than industrial and more than commercial and domestic areas. The cities of SLRC and Mexicali were the cities where people have inadequate use of water. Water consumption is a bit higher than diagnosed by WHO for use right, which is 1000 m^3 per week on average for each person. The values were obtained by a process of study and statistical estimation in the five cities evaluated, and 50 colonies in each city, with data from 2007 to 2011.

3.1 Numerical analysis

This analysis was made to know the relation between the environmental concretization and the use of water (liters / week by person) in each city. In the both cities evaluated the water consumption, can be observed the increase of water principally in the summer season, being Mexicali the city with more use of hydraulic resources every day, more than in SLRC. The analysis was made in house of citizens of both cities.

Figure 3 shows at 73% of relative humidity (RH) and 42 °C of temperature with 3489 liters every three days by house water consumption, being the higher water consumption in SLRC in summer (in July and August) of 2011 used to watering plants, wash cars, use in a domestic activities (in the kitchen, wash clothes and in the bath).The less level was 3510 liters every three days at RH of 40% and temperature of 10 °C. The color marked, represents the levels of levels of awareness of water conservation measured by the consumption of water and the

grade of levels of schooling to stand that at high grade of school, better awareness on water conservation, but it's not was really, principally in Mexicali people consume more than in SLRC. The red color indicates low awareness on water conservation and green-blue the good awareness. Figure 4 indicates that at 68% of RH and 26 °C, with a water consumption of 956 liters every three days by house in winter (December and January). The less level was 425 liters every three days, at 38% of RH and 7 °C of temperature. As represents the color, the level of awareness not was very several.

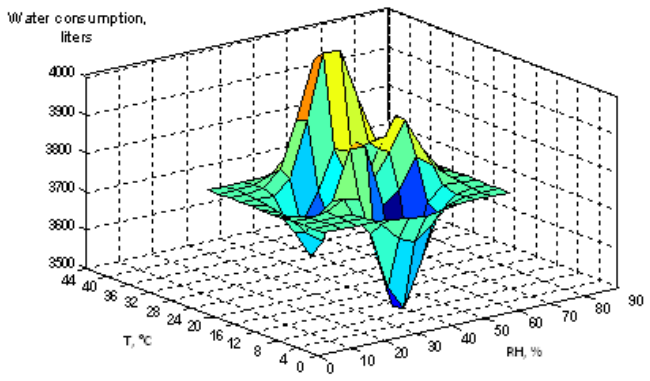


Fig 3. Correlation of use of water by houses in SLRC in summer in 2011.

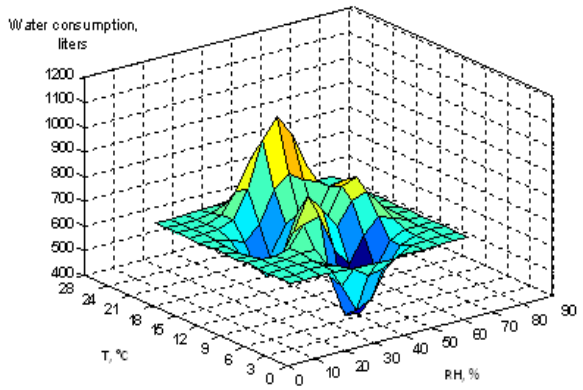


Fig 4. Correlation of use of water by houses in SLRC in winter in 2011.

Figures 5 and 6 represents the water consumption evaluated in Mexicali indicated that in this city the awareness is less that in SLRC. Some people was studying for several year this topic, but not are persons who make any control, principally by ACS, as show this research. Figure 5, shows value of water consumption of 3879 liters every three days at 70% of RH and 42 °C in summer. The less intensity was 3356 liters every three days at 74% of RH and 26 °C of

temperature. In this figure the level of awareness was very low indicated by red color.

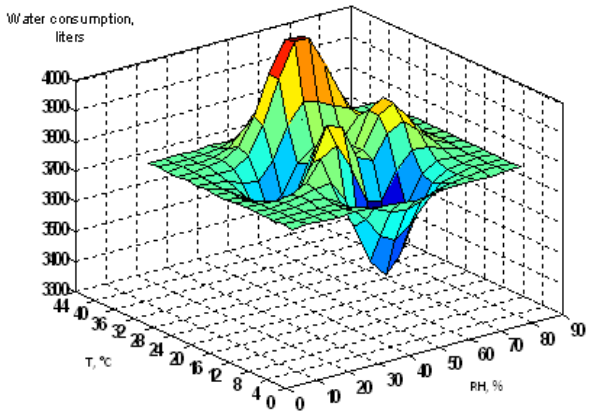


Fig 5. Correlation of use of water by houses in Mexicali in summer in 2011.

Figure 6 shows the same correlation of the parameters mentioned above indicating the high level of water consumption of 1306 liters every three days by house and the less intensity was 714 liters every three days by house at 67% of RH and 20 C of temperature. In each house evaluated live around four persons. This figure represents the color red as the lowest level of awareness of save water in domestic activities.

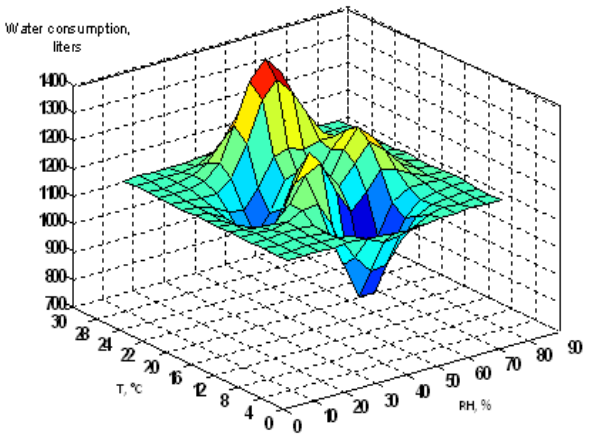


Fig 6. Correlation of use of water by houses in Mexicali in winter in 2011.

3.2 SEM evaluation. The advantage to use the ACS for saving water allowed having more fertile soils in agricultural, green areas in industrial, commercial and domestic zones and reduces at low percentage the high temperatures in summer. We can see that when not use the ACS, can damage the soil cracking (Figure7),

and change with the use of ACS observed in the figure 8, that remains wet the soil and this require less water to irrigate.

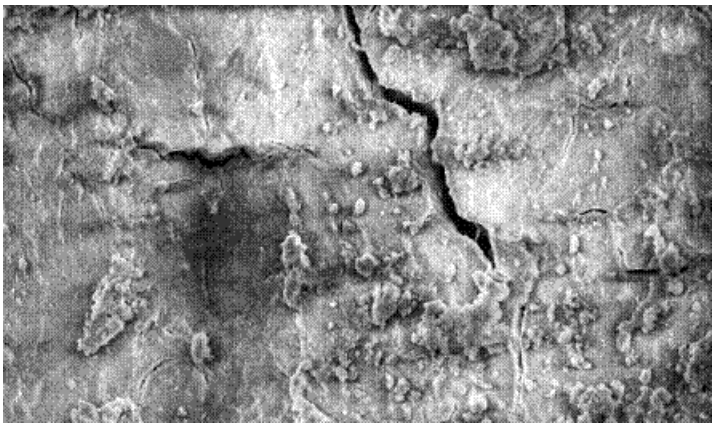


Fig 7. Microphotography (100μm) of soil before use the ACS

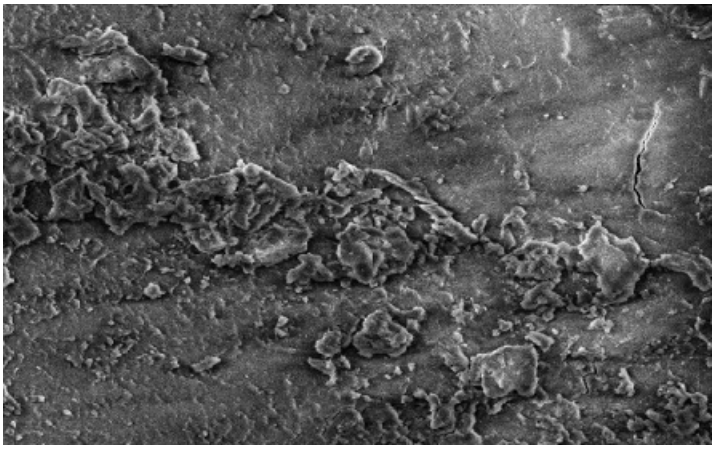


Fig 8. Microphotography (100μm) of soil after six months using the ACS.

4 Conclusions

The water saving systems are very important in the care of water. It must be realized in the use of water because it is a nonrenewable resource, and in certain regions of the world have the problem of water scarcity. It begins to have serious social problems, and could be a possible cause of a war between countries, by the vital fluid. The water in Mexicali is consumed with a great lack of awareness, in some activities such as irrigation in fertile zones and green areas. There were two important reasons to develop this study: In this zone of Mexico is necessary have good awareness to save water, otherwise water will be decrease their level and will be more expensive and all products and operations to need water. We propose some improvements to save water:

- Create an environmental assessment in the use of water, in cities evaluated and principally in SLRC and Mexicali, where will be problems in the next years for the covering of the All American Canal.
- The design and the cost of ACS is very accessible.

5 References

1. Arévalo Germán; Conciencia en el uso del recurso hidráulico del siglo XXI; Editorial Trillas, 2000.
2. World Heatlh Organization, WHO; Use of water in the World, 2004.
3. Romero A. & González R.. Efectos del revestimiento del Canal All American en los valles de San Luis y Mexicali, XX, 2005.
4. Sánchez P., Páez A. & Flores R., Evaluación zonas de cultivo sustentables y el uso adecuado del agua, Editorial Panamericana, 2006.
5. Rueda S. & Domínguez B., Niveles de concientización del uso adecuado del agua, Editorial Trillas, 2002.
6. Beltrán H. & Leyva C., Deterioro del suelo por el mal uso del agua, Editorial Oceánica, 2001.
7. Herrera Martín & Sánchez Raúl; Uso de la electrónica con sistemas solares para el cuidado del agua y medio ambiente; Editorial Panamericana; 1998.
8. Soriano Gonzalo, Torres Armando & Zamudio Joaquín; Aspectos para el cuidado del uso del agua y la economía de cada región mundial; Editorial McGraw-Hill;1996.
9. Zarate Oscar; Uso básico de dispositivos electrónicos en la industria; Editorial Oceánica; 2000.
10. La economía y el desarrollo sustentable, Editorial Oceánica.