# Editorial
Grigori Sidorov

**Control of Mechanical Systems with Dry Friction**
Roque Martínez and Joaquín Álvarez

**A Reorder Buffer Design for High Performance Processors**
José R. García Ordaz, Marco A. Ramírez Salinas, Luis A. Villa Vargas, Herón Molina Lozano, and Cuauhtémoc Peredo Macías

**An Operational Approach for Implementing Normative Agents in Urban Wastewater Systems**
José R. García Ordaz, Marco A. Ramírez Salinas, Luis A. Villa Vargas, Herón Molina Lozano, and Cuauhtémoc Peredo Macías

**Secure Architectures for a Three-Stage Polling Place Electronic Voting System**
Josué Figueroa González and Silvia B. González Brambila

**Incorporating Angular Ratio Images into Two-Frame Stereo Algorithms**
Pablo Arturo Martínez González and Mario Castelán

**Robust Extrinsic Camera Calibration from Trajectories in Human-Populated Environments**
Guillermo Baqueiro Victorín and Jean Bernard Hayet

**Chromatic Correction Applied to Outdoor Images**
Hayde Peregrina-Barreto, J. Gabriel Aviña-Cervantes, Iván R. Terol-Villalobos, José J. Rangel-Magdaleno, and Ana M. Herrera-Navarro

**Índice de contraste morfológico basado en el análisis de los contornos y el fondo de la imagen**
Angélica R. Jiménez Sánchez, Jorge D. Mendiola Santibañez, Gilberto Herrera Ruiz e Israel Santillan

**Diseño óptimo de transformadores de Hilbert sin multiplicadores con base en el uso de un subfiltro simple**
David E. Troncoso Romero, Miriam G. Cruz Jiménez y Gordana Jovanovic Dolecek

**Resumen de Tesis Doctoral**
Semantic Cohesion for Image Annotation and Retrieval
Hugo Jair Escalante, Luis Enrique Sucar, and Manuel Montes-y-Gómez
Editorial

This issue of the journal “Computación y sistemas” represents, as usual, great effort of the authors who are conducting excellent research, the editorial team, associate editors and reviewers who selected the best regular papers submitted to our journal. The work of all these persons makes it possible to present to the readers the high quality papers and maintain traditional high research standards of the journal.

Note that starting from this issue we adjust the numbering of issues to a calendar year, thus, the issue that corresponds to January will be the first one, etc. The journal published its first issue in July 1997, and traditionally its numbering was shifted, i.e., the first issue of the volume was published in July of each year. This situation causes difficulties in indexing/referencing and is very unnatural. Now this problem is solved.

The current issue contains nine regular papers and a report on a PhD thesis.

The paper “Control of Mechanical Systems with Dry Friction” by Roque Martínez and Joaquín Álvarez deals with the theme of control in case of dry friction and presents a technique to design a dynamic continuous controller for this situation. A method for defining of the parameters of the controller is also proposed. An example is presented that discusses the control of a 2-DOF underactuated mechanical system with dry friction in the non-actuated joint.

José R. García Ordaz, Marco A. Ramírez Salinas, Luis A. Villa Vargas, Herón Molina Lozano, and Cuauhtémoc Peredo Macías present their work “A Reorder Buffer Design for High Performance Processors”, where they design distributed reorder buffer microarchitecture by using small structures near building blocks. These blocks use the same tail and head pointer values on all structures for their synchronization. It is shown that this design allows higher performance.

In the paper “An Operational Approach for Implementing Normative Agents in Urban Wastewater Systems” by Juan Carlos Nieves, Dario García-Gasulla, Montse Aulinas, and Ulises Cortés, a model for water quality evaluation is presented. Normative agents that verify the regulations of the Catalan pollution-prevention policies are described. These agents are designed using Situation Calculus.

A special case of the electronic voting system is discussed in the paper “Secure Architectures for a Three-Stage Polling Place Electronic Voting System” by Josué Figueroa González and Silvia B. González Brambila. Secure architecture of such systems is presented that guarantee security, integrity and authenticity of the most important elements involved in an electoral process: configuration files, recorded votes and final result files. Various cryptographic protocols are discussed.

The paper “Incorporating Angular Ratio Images into Two-Frame Stereo Algorithms” by Pablo Arturo Martínez González and Mario Castelán proposes a post-processing operation on images based on slope angles related to the ratio values. Their experiments show that new angular ratio images are more robust and deliver improved disparity maps. The authors also perform an experimental evaluation of angular ratio images under the standard test bed for two-view stereo algorithms.

Guillermo Baqueiro Victoria and Jean Bernard Hayet present the paper “Robust Extrinsic Camera Calibration from Trajectories in Human-Populated Environments”, where they propose an approach to perform inter-camera and ground-camera calibration in the situation when visual monitoring is performed and the humans are the objects of this monitoring. So, the humans are tracked by the systems. Several challenging experimental setups are presented and evaluated.

An interesting issue of chromatic correction in outdoor scenes is addressed in the paper “Chromatic Correction Applied to Outdoor Images” by H. Peregrina-Barreto, J. G. Aviña-Cervantes, I. R. Terol-Villalobos, J. J. Rangel-Magdaleno, and A. M. Herrera-Navarro. Sometimes the images – especially outdoor images – are affected by a dominant color that changes its chromatic information that is called cast. So, a color correction must be applied. In
the paper, a method for correcting the color is proposed. This method consists in a complete cast processing: detection, color correction, and color improvement.

The paper “Morphological Contrast Index based on an Analysis of Contours and Image Background” by Angélica R Jiménez Sánchez, Jorge D Mendiola Santibañez, Gilberto Herrera Ruíz, and Israel Santillan proposes to use a contrast index for quantifying the perceived contrast in an image. The index is based on Weber’s law and takes into account background estimation. Experiments that evaluate its performance are presented.

In the paper “Optimal Design of Multiplierless Hilbert Transformer based on the Use of a Simple Subfilter” by David E. Troncoso Romero, Miriam G. Cruz Jiménez, and Gordana Jovanovic Dolecek, an optimal method to design a subfilter and prototype filter minimizing the number of coefficients is proposed. Two examples are presented for illustration of this approach.

In this issue, Section “Reports on PhD Thesis” contains the paper “Semantic Cohesion for Image Annotation and Retrieval” by Hugo Jair Escalante, Luis Enrique Sucar, and Manuel Montes-y-Gómez who present methods for image annotation and retrieval based on semantic cohesion among terms. They propose 1) a region labeling technique that assigns an image the label that maximizes an estimate of semantic cohesion among candidate labels associated to regions in segmented images, and 2) document representation techniques that are based on semantic cohesion among multimodal terms that compose images. Experiments that show the effectiveness of the methods are presented.

I am sure that these excellent papers will be interesting for the readers of our journal.

Grigori Sidorov
Editor-in-Chief of “Computación y sistemas”
Research professor of Centro de Investigación en computación
Instituto Politécnico Nacional,
Mexico
Control of Mechanical Systems with Dry Friction

Roque Martínez¹ and Joaquín Álvarez²

¹ Programa de Ingeniería Mecánica, Unidad Académica de Ingeniería, Universidad Autónoma de Zacatecas, Zacatecas, Zac., México
² Departamento de Electrónica y Telecomunicaciones, División de Física Aplicada, CICESE, Ensenada, B. C., México
rmartinez@uaz.edu.mx, jqalvar@cicese.mx

Abstract. A technique to design a dynamic continuous controller to regulate a class of full-actuated mechanical systems with dry friction is proposed. It is shown that the control eliminates the steady-state error and is robust with respect to parameter uncertainties. A simple method to find the parameters of the controller is also proposed. Moreover, an application of this result to control a 2-DOF underactuated mechanical system with dry friction in the non-actuated joint is described. Here, the control objective is to regulate the non-actuated variable while the position and speed of the actuated joint remain bounded. Performance issues of the developed synthesis are illustrated with numerical and experimental results.

Keywords. Stability, friction, mechanical systems, underactuated systems.

1 Introduction

Dry friction is defined as a force that resists relative motion between contacting surfaces of different bodies. The bodies "stick" when the relative velocity between the contacting surfaces is zero. If the bodies slide over each other with a non-zero velocity, we speak of a "slip". A model of dry friction must contain a description of both phases. Different models have been proposed to describe dry friction; usually they differ only in the way the stick phase is modeled [10].

A realistic approach to control mechanical systems should be able to deal with the effects of dry friction [2, 20]. Dry friction can be described by either differential inclusions or by ordinary differential equations with discontinuous right-hand side [8, 25].

Systems with discontinuous elements exhibit a wide variety of complex phenomena which must be considered in the control design process [7, 12, 17]. For instance, these complex dynamical behaviors can generally result in vibration and instability that are highly undesirable in many cases [11]. Notwithstanding the impressive development of nonsmooth and set-valued analysis, these systems have not been closely studied either computationally or analytically [24].

For full-actuated mechanical systems, an effective approach to counteract the friction phenomenon has been the use of first-order sliding mode controllers. Discontinuous friction is
regarded as a bounded disturbance of unpredictable sign and therefore counteracted by choosing adequate control amplitude. These algorithms often require a high control effort to compensate this physical phenomenon and produce control signals that commute at theoretically infinite frequency, so it is not practical in many real situations \[5\]. Some techniques have been proposed to eliminate or attenuate this effect (e.g., sliding mode algorithms of higher order). However, the anti-chattering procedures, which aim at obtaining continuous control, do not necessarily guarantee accuracy in the presence of discontinuous friction \[3, 4, 21\].

The classical approach to control underactuated mechanical systems has commonly neglected the friction effect. Thus, in recent years several works have addressed the problem of friction in this class of systems. For example, linear damping (viscous friction) in the joints is considered in \[1, 9, 26, 27\]. When dry friction is present only in the controlled joint, the problem of compensation can be solved in some cases \[15, 20, 22, 23\]. However, the problem of controlling underactuated mechanical systems with dry friction in the non-actuated joint seems to be still open (see, e.g., \[6, 13, 14, 20\]). For instance, in \[16, 17, 18\] stick-slip oscillations and sticking phenomena of a class of underactuated mechanical systems are analyzed.

In this paper, we propose a dynamic continuous controller to regulate a class of full-actuated mechanical systems with dry friction.

The proposed controller makes use of the result presented in \[13\] concerning some stability conditions of mechanical systems with discontinuous friction, but it is robust with respect to more uncertainties and it operates in such a way that the objective of control is reached faster. Moreover, we propose a simpler method to find the parameters of the controller than the method presented in \[13\]. In addition, using this result, we propose a discontinuous controller for a class of 2-DOF underactuated mechanical systems with dry friction in the non-actuated joint. We illustrate this result with numerical examples of full-actuated systems and with an application to an experimental underactuated system.

### 2 Problem Statement

Consider a $n$-DOF mechanical system represented by

$$
\begin{align*}
\dot{y}_1 &= y_2, \\
\dot{y}_2 &= f_1(y_1, y_2) - C(y_2) \text{sgn}(y_2) + u_c.
\end{align*}
$$

Hereinafter, $y_1 = (y_{11}, y_{12}, \ldots, y_{1n})^T$ and $y_2 = (y_{21}, y_{22}, \ldots, y_{2n})^T$ are the generalized position vector and the velocity vector, $f_1(y_1, y_2)$ is a smooth vector function, $u_c \in \mathbb{R}^n$ is a control input vector, and $\text{sgn}(y_2)$ is the sign vector function, defined by

$$\text{sgn}(y_2) = \begin{bmatrix} 
\text{sgn}(y_{21}) \\
\text{sgn}(y_{22}) \\
\vdots \\
\text{sgn}(y_{2n}) 
\end{bmatrix} \quad (2)$$

with

$$\text{sgn}(y_{2i}) = \begin{cases} 
1, & \text{for } y_{2i} > 0, \\
\epsilon, & \text{for } y_{2i} = 0, \\
-1, & \text{for } y_{2i} < 0,
\end{cases} \quad (3)$$

where $\epsilon \in [-1, 1]$, and $C(y_2)$ is the matrix

$$C(y_2) = \begin{bmatrix}
c_{11}(y_{21}) & 0 & \cdots & 0 \\
0 & c_{22}(y_{22}) & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & c_{nn}(y_{2n})
\end{bmatrix} \quad (4)$$

with the well-known friction model (see, e.g., \[4, 19\]) defined by

$$c_{ii}(y_{2i}) = c_{c_{ii}} + (c_{s_{ii}} - c_{c_{ii}}) \exp\left(-\frac{y_{2i}^2}{v_{s_{ii}}^2}\right), \quad i = 1, 2, \ldots, n \quad (5)$$

where $c_{c_{ii}}$ and $c_{s_{ii}}$ are the Coulomb friction level and the level of friction divided by a constant such that $0 \leq c_{c_{ii}} \leq c_{s_{ii}}$, and $v_{s_{ii}}$ is the Stribeck velocity.

The control objective is to steer to zero both $y_1(t)$ and $y_2(t)$ by means of a continuous control vector $u_c$. 

---

**Footnotes:**

1. \[5\] Roque Martínez and Joaquín Álvarez

**References:**

1. Roque Martínez and Joaquín Álvarez

**ISSN:** 1405-5546
Note that (1) has $n$ discontinuity surfaces $S_i$ characterized by $y_{2i} = 0$, $i = 1, 2, \ldots, n$. Here, the meaning of such differential equation is viewed in the Filippov sense [8].

Let us consider a control law $u_c$ with a dynamics given by

$$
\dot{u}_c = \varphi(y_1, y_2, u_c) = 
\begin{bmatrix}
\varphi_1(y_1, y_2, u_c) \\
\varphi_2(y_1, y_2, u_c) \\
\vdots \\
\varphi_n(y_1, y_2, u_c)
\end{bmatrix}
$$  \hspace{1cm} (6)

Note that $u_c$ must be a continuous vector, but not necessarily $\varphi$. The system (1) (6) is then a system with the state $(y_1, y_2, u_c)^T \in \mathbb{R}^{3n}$. The next lemma will be a key result for what follows.

Lemma 1. For a given $u_c \in \mathbb{R}^n$, define $\varphi_0 : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$, $\varphi_0(y_1, u_c) = \varphi_i(y_1, 0, u_c)$ as the restriction of $\varphi_i$ for $y_2 = 0$, $i = 1, 2, \ldots, n$. Suppose that $y_{1i} \neq 0$ implies that $\varphi_0(y_1, u_c) \neq 0$, $i = 1, 2, \ldots, n$ for any $u_c$. Therefore, if (1) (6) exhibits the sliding mode in the intersection $S_0 = S_1 \cap S_2 \cap \cdots \cap S_n \setminus \{(0, 0, u_c)\}$, then the system will leave this intersection in a finite time.

Proof. When the system (1) is in the sliding mode regime in the intersection $S_0 = S_1 \cap S_2 \cap \cdots \cap S_n \setminus \{(0, 0, u_c)\}$, it is described by

$$
y_2 = 0, \quad f_1(y_1, 0) - C(0)\text{sgn}(0) + u_c = 0. \hspace{1cm} (7)
$$

Under this condition, at least $y_{1i}$ is a constant different from zero, hence $\varphi_0(y_1, u_c) \neq 0$ while in the sliding mode. From (5), $c_{ii}(0) = c_{ii}$, therefore, there exists a finite time in which the system leaves the sliding regime, that is,

$$
\dot{u}_c + f_1(y_1, 0) = \int_{t_0}^{t} \varphi_0(y_1, 0, u_c) \, d\tau + f_1(y_1, 0) 
\leq |c_{ii}|(y_2 - u_c) 
$$  \hspace{1cm} (8)

where $t_0$ is the initial time.

Now define $x_1 = y_1$, $x_2 = y_2$, and $x_3 = f_1(y_1, y_2) - C(y_2)\text{sgn}(y_2) + u_c$, then the system (1) (6) is described, for $x_2 \neq 0$, $i = 1, 2, \ldots, n$, by

$$
x = (x_1, x_2, x_3)^T, \quad f_2 = \dot{x}_1, \quad \text{and} \quad F_4 \text{ is the matrix}
$$

$$
F_4 = 
\begin{bmatrix}
f_{411} & 0 & \cdots & 0 \\
0 & f_{422} & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & f_{4nn}
\end{bmatrix}
$$  \hspace{1cm} (10)

with

$$
f_{4ii} = \frac{2(c_{ii} - c_{ii})}{v_{x_{1i}}} |x_{2i}| \exp\left(\frac{|x_{2i}|^2}{v_{x_{1i}}^2}\right) \geq 0
$$  \hspace{1cm} (11)

Since the control objective is to steer to zero both $x_1(t)$ and $x_2(t)$ by means of a continuous control $u_c(t)$, then the control aim is to regulate at zero the state $x(t)$ of the system (9) with a hybrid “control” $\dot{u}_c$. This will be described in the next section.

3 Control Strategy

In this section, we present a control strategy that allows one to achieve the control objective. Suppose that, for the system (9), there exists a control law

$$
u_c = f_3(t, x_1, x_2), \hspace{1cm} (12)
$$

such that $\dot{u}_c = \varphi(y, u_c)$ satisfies Lemma 1, where $f_3(t, x_1, x_2)$ is a continuous vector function such that (9) with (12) can be transformed to

$$
\dot{x} = 
\begin{bmatrix}
x_2 \\
x_3 \\
x_5
\end{bmatrix},
\begin{bmatrix}
x_2 \\
x_3 \\
x_5
\end{bmatrix}
$$

$$
-\begin{bmatrix}
K_1 & -K_2 & x_5
\end{bmatrix} + F_4x_3 + \dot{f}_3 + \tau_1
$$

where $K_n, K_{n+1}, K_{n+2}$ are real-coefficient diagonal constant matrices, being $K_1$ a positive diagonal matrix,
with

\[ \frac{\partial f_5(x_1)}{\partial x_1} < \infty, \quad i = 1, 2, ..., n, \]  

(15)

\[ \tau_1 = \begin{bmatrix} \tau_{11} \\ \tau_{12} \\ \vdots \\ \tau_{1n} \end{bmatrix} \]  

(16)

and \( k_{1i} \geq 0, \quad i = 1, 2, ..., n, \) are arbitrary real constants.

The next lemma shows that when the system is in the sliding mode with \( x_1 = 0 \) then it remains there indefinitely, implying that \( \{ 0 \in \mathbb{R}^{3n} \} \) is an invariant set of the system (13).

**Lemma 2.** Suppose that the system (1)-(12) exhibits the sliding mode in the intersection of the surfaces \( S_i, \quad i = 1, 2, ..., n \) and that \( x_1 = 0 \). Then the system will not leave this intersection.

**Proof.** The system (1)-(12) in the sliding mode, in the intersection of the surfaces \( S_i, \quad i = 1, 2, ..., n \) with \( x_1 = 0 \), is given by

\[ x_2 = 0, x_3 = -C(0)\text{sgn}(0) + \int_{t_0}^{t} \tau_1 dr + f_5(0) + c_k = 0, \]  

(18)

where \( t_0 \) is the initial time and \( c_k \) is a finite constant vector. Since \( \tau_1 = 0 \) for \( x_1 = 0 \), we have that

\[ \int_{t_0}^{t} \tau_1 dr + f_5(0) + c_k \in [-c_{k1}, c_{k2}], \quad i = 1, 2, ..., n, \]  

(19)

for all time.

Therefore, if it is possible to find a function \( f_5 \) such that the system (9) with (12) can be transformed to (13) and \( K_1, K_2, K_3 \) forcing the state of (13) converge to zero, then the control objective will be attained.

Since \( \tau_1 = 0 \) for \( x_2 \neq 0, \quad i = 1, 2, ..., n \), and it is possible to decouple this system into \( n \) 1-DOF systems, the conditions for \( k_{1i}, \quad k_{2i}, \quad k_{3i}, \quad i = 1, 2, ..., n \) to accomplish the objective lim_{t \to \infty} \| x(t) \| = 0 \) were presented in [13] and can be applied to this problem. This is summarized as follows.

Consider the matrices \( P_i \) and \( Q_i \) given by

\[ P_i = \begin{bmatrix} p_{i11} & p_{i12} & p_{i13} \\ p_{i21} & p_{i22} & p_{i23} \\ p_{i31} & p_{i32} & p_{i33} \end{bmatrix}, \quad Q_i(x_1, x_2) = \begin{bmatrix} q_{i11} & q_{i12} \\ q_{i21} & q_{i22} \end{bmatrix}, \]  

(20)

where the entries of \( Q_i(x_1, x_2) \) are given by

\[ q_{i11} = -2p_{i12} + 2p_{i23} \left[ k_{2i} - \frac{\partial f_5(x_1)}{\partial x_1} \right], \]  

(21)

\[ q_{i12} = -p_{i22} + p_{i23} [k_{3i} - f_{4i}] + p_{i33} \left[ k_{3i} - \frac{\partial f_5(x_1)}{\partial x_1} \right], \]  

(22)

\[ q_{i22} = -2p_{i23} + 2p_{i33} \left[ k_{3i} - f_{4i} \right], \]  

(23)

Then the next theorem, shown in [13], can be applied.

**Theorem 3.** Suppose that matrix \( Q_i \) is positive definite for all \( (x_1, x_2)^T \in \mathbb{R}^2 \), matrix \( P_i \) is also positive definite and satisfies

\[ p_{i13} = 0, \quad \frac{p_{i11}}{p_{i23}} = \frac{p_{i12}}{p_{i33}} = k_{1i}, \]  

(24)

for \( i = 1, 2, ..., n \). Then \( x = 0 \) is a globally, asymptotically stable equilibrium point of the system (13).

In the next section, an application of this result to two systems is described.
4 Control Synthesis

In order to find the values of $k_{1,i}, k_{2,i}, k_{3,i}$, $i = 1, 2, ..., n$ easily, we propose $P_i$ and $Q_i(x_1, x_2)$ given by

$$
P_i = \frac{1}{2} \begin{bmatrix}
    \alpha_i k_{1,i} & k_{1,i} & 0 \\
    k_{1,i} & \alpha_i k_{3,i} + k_{2,i} & \alpha_i \\
    0 & \alpha_i & 1
\end{bmatrix},
$$

$$
Q_i(x_1, x_2) = \begin{bmatrix}
    q_{11} & q_{12} \\
    q_{12} & q_{22}
\end{bmatrix}
$$

(25)

where

$$
q_{11} = -k_{1,i} + \alpha_i \left( k_{2,i} - \frac{\partial f_5(x_1)}{\partial x_1} \right),
$$

(26)

$$
q_{12} = \frac{1}{2} \alpha_i f_{4,i} - \frac{1}{2} \frac{\partial f_5(x_1)}{\partial x_1},
$$

(27)

$$
q_{22} = -\alpha_i k_{3,i} - f_{4,i}.
$$

(28)

Therefore, Theorem 1 is satisfied if

$$
\alpha_i > 0
$$

(29)

$$
k_{1,i} < \alpha_i^2 k_{3,i} + \alpha_i k_{2,i} - \alpha_i^3
$$

(30)

$$
k_{2,i} > \frac{k_{1,i}}{\alpha_i} + \frac{\partial f_5(x_1)}{\partial x_1}
$$

(31)

$$
k_{3,i} > \left( \frac{1}{2} \alpha_i f_{4,i} + \frac{1}{2} \frac{\partial f_5(x_1)}{\partial x_1} \right)^2 / (\alpha_i [k_{2,i}]
$$

(32)

$$
f_{4,i} < \frac{2(c_{s,i} - c_{e,i})}{v_{s,i}} (0.429), \quad i = 1, 2, ..., n.
$$

(33)

concluding that $x = 0$ is a globally, asymptotically stable equilibrium point of system (13).

Note that in order to find the parameters $k_{1,i}$, $k_{2,i}, k_{3,i}$, $i = 1, 2, ..., n$, it is not required to know $c_{s,i}, c_{e,i}, v_{s,i}, v_{e,i}$ exactly.

In what follows, we describe the controller design procedure using two examples to regulate the position of a mechanical system. The first example is a pendulum with dry friction; the second example is an experimental torsional system with dry friction.

Example 1. A Pendulum

Let us consider the system shown in Figure 1, described by

$$
ml^2 \ddot{q} + mgl \sin(q) + ml^2 c(q) \text{sgn}(\dot{q}) = u_1,
$$

(34)

where $q \in \mathbb{R}$ is the angular position, $\dot{q} \in \mathbb{R}$ the angular velocity, $m$ the mass, $l$ is the distance, $c$ is given by (4), and $u_1 \in \mathbb{R}$.

The objective is to design a continuous control law $u_1$ so that the position $q$ converges to a given constant value $q_d$.

For this system we have

$$
f_5(x_1) = f_1(x_1, x_2) = -q \sin(x_1 + q_d)
$$

and $u_c = \frac{1}{ml} u_1$ (see Eq. (1)), where $x_1 = q - q_d$, $x_2 = \dot{q}$. If

$$
u_c = -\int_{t_0}^{t} [k_1 x_1 (\tau) - \tau_1 (\tau)] d\tau - k_2 x_1 - k_3 x_2.
$$

(35)

where $t_0$ is the initial time and $\tau_1$ is given by (16), then we arrive at the desired form (13).

Finally, if we find the constants $k_i$, $i = 1, 2, 3$, which satisfy Theorem 3, then the control objective will be attained.

Figure 2 shows numerical results, where we have set $m = 1$, $l = 1$, $g = 9.81$, $c_1 = 1$, $c_2 = 0.75$, $v_s = 0.143$. We propose $\alpha = 1$, $k_1 = 0.5$ and, from (29), (30), (31), (32), and (33), $k_2 = 11$, $k_3 = 47$, $k_4 = 20$, with $q(t_0) = 0$, $\dot{q}(t_0) = 0$ and $q_d = \pi$.

Example 2. An Experimental Torsional System

Let us consider the experimental system configuration model 205 of Educational Control Products (ECP) shown in Figure 3, described by
where \( q_i \in \mathbb{R} \) are the angular positions, \( \dot{q}_i \in \mathbb{R} \) the angular velocities, \( J \), the inertial, for \( i = 1, 2 \); \( \epsilon \) is given by (4), \( k_p, f_a \) and \( f_b \) are viscous friction coefficients, and \( u_1 \in \mathbb{R} \).

This system consists of two plates, without anchoring, coupled by a torsional spring. Control input is at plate 1, and dry friction in joint 2 is introduced through a DC motor attached to plate 2.

The objective is to design a control law \( u_1 \) so that the angle \( q_2 \) converges to a given constant value \( q_{2d} \) with \( q_1 \) and \( \dot{q}_1 \) bounded.

If we define \( x_1 = q_2 - q_{2d} \) where \( q_{2d} \) is a desired constant position, \( x_2 = \dot{q}_2 \), \( x_3 = \dot{q}_2 \).

The control input can be designed using the classic technique of sliding modes. This is convenient because if we use \( s = \dot{u}_c - \dot{u}_c^* \) as a
sliding surface then a sliding mode control designed for $u_1$ makes the control $\dot{u}_c$ converge to $\dot{u}_c^*$ in finite time. Once $\dot{u}_c = \dot{u}_c^*$, Theorem 3 ensures the convergence of $x_1 = q_2 - \dot{q}_2$ to zero.

So let us propose a control law $u_1$ be given by

$$u_1 = k_u(q_1 - q_2) + f_\alpha \dot{q}_1 + J_1 \tau_2,$$

then the actuated joint of system (36) takes the form

$$\ddot{q}_1 = \tau_2,$$

and let the sliding surface be defined by

$$s = \dot{u}_c - \dot{u}_c^* = \frac{k_{f_\alpha}}{J_2} \dot{q}_1 + k_1 x_1 + k_{c2} x_2,$$

rendering the sliding mode controller given by (41) with

$$\tau_2 = \frac{f_\alpha}{k_u} \left[ -k_5 \text{sgn}(s) - k_1 \dot{q}_1 - k_{c2} \left( \frac{k_{f_\alpha}}{J_2} \dot{q}_2 - \frac{k_5}{J_5} q_2 \right) - \frac{k_{f_\alpha}}{J_2} \dot{q}_2 + \frac{k_5}{J_5} q_1 \right],$$

with $k_5 > |k_{c2}| c_\alpha$.

Finally, if we find the constants $k_i, i = 1, 2, 3$, which satisfy Theorem 3, then the control objective will be attained.

Figure 4 shows the experimental results. According to the manual and the identification of system parameters, the inertia are $J_1 = 0.0193$ Nms$^2$/rad and $J_2 = 0.0187$ Nms$^2$/rad, the coefficient of elasticity of the spring is $k_{c2} = 3.2178$ Nm/rad, and the coefficients of viscous friction are $f_a = 0.1373$ Nms/rad and $f_b = 0.3$ Nms/rad. Note that $k_3 = f_b / J_2$. We propose $c_\alpha = c_\epsilon = 3.7$ rad/s$^2$, $v_\epsilon = 1$ rad/s, $\alpha = 4$, $k_1 = 4500$, and from (29), (30), (31), (32), and (33) $k_2 = 1500$, $k_5 = 5000$.  

---

Fig. 4. Position and control of the torsional system

Fig. 5. Interface of the system
\[ q_1(t_0) = 0 \text{ rad}, \; \dot{q}_1(t_0) = 0 \text{ rad/s}, \; q_2(t_0) = 0 \text{ rad}, \]
\[ \dot{q}_2(t_0) = 0 \text{ rad/s}, \text{ and } q_{3d} = 0.3 \text{ rad}. \]

Figure 5 shows the user interface of the torsional system using Matlab Simulink.

5 Conclusions

In this paper, we have proposed a continuous controller for a class of full-actuated mechanical systems with dry friction. The proposed controller makes use of the result presented in [13], but it is robust with respect to uncertainties in the parameter values of the system and it operates in the way that the objective of control is reached faster, since the system leaves the obstruction faster. Moreover, we have proposed a simpler method to find the parameters of the controller than the method presented in [13].

In addition, using this result, we have shown its application to control of an underactuated mechanical system with dry friction in the non-actuated joint. In this case, the control objective is to regulate the non-actuated joint while the position and speed of the actuated joint remain bounded. Since the term in the non-actuated joint containing dry friction must be compensated by a continuous action, a discontinuous control is designed using the classic technique of sliding modes. The proposed controller guarantees the convergence of the position error of the non-actuated joint to zero. We illustrated these results with an application to control an experimental torsional system.

References


A Reorder Buffer Design for High Performance Processors

José R. García Ordaz, Marco A. Ramírez Salinas, Luis A. Villa Vargas,
Herón Molina Lozano, and Cuauhtémoc Peredo Macías
Microtechnology and Embedded System Laboratory,
Centro de Investigación en Computación, Instituto Politécnico Nacional,
Av. Juan de Dios Bátiz, s/n, Zacatenco, 07738, México DF,
Mexico
{jgarcia, mars, lvilla, hmolina, cperedo}@cic.ipn.mx

Abstract. Modern reorder buffers (ROBs) were conceived to improve processor performance by allowing instruction execution out of the original program order and run ahead of sequential instruction code exploiting existing instruction level parallelism (ILP). The ROB is a functional structure of a processor execution engine that supports speculative execution, physical register recycling, and precise exception recovering. Traditionally, the ROB is considered as a monolithic circular buffer with incoming instructions at the tail pointer after the decoding stage and completing instructions at the head pointer after the commitment stage. The latter stage verifies instructions that have been dispatched, issued, executed, and are not completed speculatively. This paper presents a design of distributed reorder buffer microarchitecture by using small structures near building blocks which work together, using the same tail and head pointer values on all structures for synchronization. The reduction of area, and therefore, the reduction of power and delay make this design suitable for both embedded and high performance microprocessors.

Keywords. Superscalar processors, reorder-buffer, instruction window, low power consumption.

1 Introduction
Superscalar processors allow the execution of more than one instruction in a clock cycle; this goal becomes increasingly complex to achieve in hardware. The total complexity is distributed along the pipeline stages in order to make it manageable. As each stage is designed to support the parallel execution of N instructions by a processor, such a processor is referred to as an N-way processor. Modern superscalar processors implement deep pipelines by splitting the established stages (IF instruction fetch, IDi instruction decode, IR instruction rename, IDi instruction dispatch, IS issue, EX execute, WB write back, and IC instruction commitment) into sub-stages to get more clock frequency and more in-flight instructions.

Diseño de un búfer de reordenamiento para procesadores de alto desempeño

Resumen. El búfer de reordenamiento de instrucciones (ROB) fue conceptualizado para mejorar el desempeño de los procesadores al permitir ejecutar instrucciones fuera del orden original del programa y en avance al instante preciso de la ejecución secuencial, explotando el paralelismo que existe a nivel de las instrucciones ILP. El ROB es una estructura funcional de la máquina de ejecución de los procesadores para dar soporte a la ejecución especulativa, al reciclado de los registros físicos y a la recuperación precisa de excepciones. Tradicionalmente el ROB es considerado un búfer circular monolítico en donde las instrucciones entran en la dirección especificada por un apuntador de cola después de la etapa de decodificación y son terminadas en la dirección especificada por un apuntador de cabecera después de la etapa de finalización. El artículo presenta el diseño de un búfer de reordenamiento de instrucciones distribuido en pequeñas estructuras cercanas a los bloques funcionales con los cuales interactúan, usando los mismos valores de apuntadores de cola y cabecera por sincronía. La reducción de área y por consecuencia la reducción de consumo de energía y retardo hacen de este diseño apropiado para procesadores embebidos y procesadores de alto desempeño.

Palabras Clave. Procesadores súper escalares, búfer de reordenamiento, ventana de instrucciones, consumo de baja potencia.
A processor microarchitecture is divided into two sections: the front end, covering the IF, IDe, IR, and IDi stages executing in program order, and the back end, covering IS, EX, and WB executing OOO out of order; finally, IC completes the instructions in order. The OOO execution is used to exploit Instruction Level Parallelism (ILP) of in-execution code to enhance the IPC performance. To be able to perform out-of-order execution, several scheduling techniques are implemented along the processor microarchitecture. Dynamic scheduling techniques covering from IF to IC are branch prediction, register renaming, speculative execution, exception recovering, resources recycling, amount others. An important structure that makes the dynamic scheduling possible is the reorder buffer (ROB).

The ROB unit stores all instructions in execution and executed. The executed instructions wait to be committed by the processor. While instructions fly across the pipeline stages, several flags are being set in order to preserve the processor’s state because of recovering misspeculation support. Speculative execution is the execution of instructions on an optimistic code path chosen by the branch predictor unit. The instructions of the chosen path become non-speculative when the branch condition is computed and the destination address matches the speculative address offered by the branch predictor. If a mismatch takes place, an exception recovery mechanism is launched.

This paper presents a design of distributed reorder buffer microarchitecture by using small bit-vector structures near building blocks which work together, using the same tail and head pointer values of all structures for synchronization instead of a monolithic structure. The rest of the paper is organized as follows. Section 2 presents related work concerning the development of today’s processor microarchitectures. Section 3 describes the proposed design, analyzing all functions performed by the ROB unit. Section 4 analyzes simulation results, and finally, Section 5 presents the concluding remarks.

2 Related Work

Since functional units have different latencies and conditional branches may be in any position of a fetched instruction group, instruction completion may be out of order causing imprecise interrupts. Two techniques were developed to solve this problem. The first technique is to keep the state of a processor precise by allowing instructions to update the register file in program order. The second one is to tolerate the state of a processor imprecise by allowing instructions to update the register file out of order, but with a procedure for precise state recovery after an exception event.

Four methods are analyzed in [12]:

1) **Completion Order**. In this method, processor issues instructions only if all previous instructions are free of exceptions. The processor guarantees it by reserving the number of stages equal to the clock latency instructions in the result shift register. This simple approach does not make a full use of multi-latency functional units.

2) **Reorder Buffer**. This method allows out-of-order completion but stores the result of each instruction in a FIFO structure to reorder the instructions before modifying the processor’s state. Since the processor cannot issue instructions that depend on results waiting in the reorder buffer to be written to the register file, this method has a performance loss.

3) **History File**. In this method, instructions can be completed in any order and immediately updated to the register file. However, a processor needs to save the previous state of the register file in a history buffer utilized for exception recovery. The history file method uses a reorder buffer structure and a result shift register.

4) **Future File**. This method uses two structures of the register file, one called the architecture file and other called the future file. Instructions are issued and written back to the future file which provides the source for succeeding instructions. The processor updates the architectural file as in the reorder buffer method.

When an exception occurs, the architectural register file is copied to the future file in order to recover the precise processor’s state. Complexity-
performance comparison results show that the history file method should be used for high speed computations to achieve precise exceptions.

The first approaches to the ROB design were based on a monolithic multiport memory with the wakeup logic, selection logic, and the register file working together as proposed in [8]. Additionally, the future file method is implemented for precise interrupt recovering. This organization is used in the Intel P6 microarchitecture design shown in [4]. Several techniques are proposed in [6] to reduce complexity and power consumption. The first technique is to eliminate the ROB write ports by allocating small FIFO queues to store results of each functional unit. The second technique is to eliminate the ROB read ports for reading out the source operand values from FIFO queues using small sets of associative-addressed retention latches and forwarding buses to supply results to the instructions waiting in the issue queue. The second technique was motivated by the fact that only a small fraction of source operands read their values from the reorder buffer slots. The design results in low performance degradation and significant power complexity reduction.

The MIPS R10000 microarchitecture is described in [13], while [7] and [5] specify the Alpha 21264 microarchitecture. Both microarchitectures, with a few variations, represent the core of a modern superscalar processor, replacing the monolithic ROB of [8] and [3] for MIPS R1000 with a 32-entry active list (ROB), two architectural register banks of 64-entry for integer, 64-entry for floating point and 16-entry queues for integer, floating point and load-store instructions. In the case of the Alpha 21264, a monolithic-ROB was replaced with an 80-entry ROB, two architectural register banks of 80-entry for integer, 72-entries for floating point algebraic operations, and compacting queues for 20-entries for integer algebraic operations, 15-entries for floating point algebraic operations and load-store instructions. A similar ROB architecture where the register file is separated from the reorder buffer is used in the Intel Pentium 4 Burst microarchitecture [4]. Two techniques analyzed in [2] allow processors to keep thousands of in-flight instructions. In the first technique, the normal ROB structure is replaced with a mechanism to make check-pointing based on simple heuristics: 1) at the first branch after each 64 instructions, 2) after 500 instructions, and 3) after 64 stores. The second technique termed Slow Lane Instructions Queuing introduces a secondary buffer used to store instructions moved from fast instruction queues because of issue time length, freeing slots of instruction queues for more decoded instructions which will be executed quickly. These instructions are returned to the fast queue when ready to issue. With these two mechanisms, the resultant processor microarchitecture includes 128-entry pseudo-ROB, 128-entry IQ’s, and 2048-entry SLIQ, reporting a performance increase of 204% relative to a conventional processor with 128-entry ROB and 128-entry IQ’s.

It is proposed in [9] to replace the ROB with a validation buffer structure VB, two structures of register alias tables (the front-end RAT \textsubscript{fem} used in the rename stage and the retirement RAT \textsubscript{ret} for maintaining the architectural state, similar to the future file method) plus one additional table necessary to track the physical register status (RST) for recycling. This microarchitecture allows retiring instructions out-of-order of VB as soon as it is known that they are non-speculative, updating the RAT \textsubscript{ret} which contains a valid state of register mapping and is used in recovering, and updating the RAT \textsubscript{fem} table. The RST table has, in each entry, a counter for physical register successors, a valid remapping bit, and a completed bit to identify when the corresponding entry contains (0, 1, 1). These conditions ensure that a specific register can be safely recycled. Compared with in-order-commitment, the VB microarchitecture presents high IPC for FP benchmarks with 32-entry ROB size. Because of OOO retirement and early physical register recycling, VB behavior in modern superscalar processors with major size structure is more efficient.

### 3 Distributed ROB Design

The reorder buffer structures are shown in Figure 1. The ROB is composed of 1-bit vectors for dispatched, branch, branch decision, issued, and executed flags, 7-bit structures for old destination and current destination registers plus a 5-bit structure for the exception pointer, using the same tail and head pointer. All structures are of
the ROB size. The exception pointers use a 5-bit structure to index the branch ROB structure to update the branch predictor unit. All instructions are dispatched to different queues: integer, floating point, and load-store; a flag is activated (set) in the dispatched flag structure indexed by the tail pointer. At the same time, the tail pointer is stored in the in-flight tag field of IQ with the incoming instruction. When instructions are ready to be executed, they are issued to the functional units setting a flag in the issued flag structure pointed by the previously stored in-flight tag. A description of IQ’s operation can be found in [10, 11], in which the wakeup, allocation, and issue operations are presented in great detail. Each functional unit executes instructions and set a flag at execution ending in the executed flag structure entry indicated by the in-flight tag pointer. The number of 1-bit write ports in the executed flag structure is equal to the number of execution units of a processor.

3.1 Speculative Execution

The branch predictor unit is responsible for speculative execution support. In each clock cycle, the fetch unit calculates the next program counter next-PC incrementing the PC-register. Meanwhile, the branch predictor unit uses the calculated next-PC value to look for branches and their respective destination addresses in the branch history buffer BHB in order to offer a speculative program counter spec-PC for the next cycle. In the next clock cycle, instructions are fetched from a non-speculative or speculative path depending on the branch predictor decision (0-taken or 1-not taken) as it is shown in Figure 2. When branch instructions are decoded, the dispatch stage sets a flag in the branch flag structure indexed by the tail pointer.

The superscalar processor schedules branch instructions in three sub-operations: 1) calculate the branch address destination, 2) resolve the branch condition, and finally, 3) verify the decision chosen by the branch predictor.

3.1.1 Branch Address Calculation

The fetch unit uses one ACU to increment the program counter (see Figure 2) and the decode stage uses another one to compute the branch address destination (see Figure 3). Since branches are relative to a given PC, the address destination is computed using the PC and the branch instruction offset (PC+ sign extended offset). Performing the branch predictor updating at commitment requires both the PC and the offset values, and furthermore, the branch condition calculation.

The previous two values demand an area along the reorder buffer, and this space is not exploited for all instructions in the window. Our design utilizes a small structure associated to the branch predictor unit for storing these values.
The PC and the calculated destination address are stored in a structure smaller than the reorder buffer size, illustrated in Figure 3 as a Branch ROB. After the branch condition is calculated, the functional unit sets the branch decision flag (0-taken or 1-not taken) and the executed flag (see Figure 1).

Then, the branch and executed flags enable the exception pointer to select the corresponding Branch ROB entry. The PC and the branch target address are used to update the branch predictor. This action should be accomplished at the write-back stage to launch a recovery mechanism in the case of misprediction and reduce wrong path executions.

3.1.2 Resolving the Branch Condition
The speculative behavior (taken/not-taken) of a conditional branch (beq rs, rt, offset) is resolved by comparing the processor registers (rs==rt). When the condition is computed as illustrated in Figure 1, the processor writes its result in the branch decision flag structure. Then, this result is used for the branch predictor unit to update its decision machinery and to signal all structures for recovering in case of misspeculation. In both cases, the exception pointer is used.

3.1.3 Verifying Branch Predictor Decisions
Since instructions are unknown at the fetch cycle, a superscalar processor needs to resolve all branch types in the same cycle via the branch predictor unit. Subsequently, more pipeline cycles are necessary to verify if the prediction was correct. Unconditional branches and return address are resolved by the branch predictor via a branch target buffer and a return address stack. However, conditional branches need to be predicted.

The last step of turning a branch into a non-speculative instruction is to verify the decision chosen by the branch predictor. This action starts when the conditional instruction has been executed by the functional unit setting the branch decision flags and the executed flag. The branch flag is set at dispatch once a given instruction has been decoded. These two conditions (the branch flag and the executed flag) are sufficient to select the E-pointer and the branch decision flag calculated by the processor as shown in Figure 3. The exception pointer is used to index the corresponding entry of the branch reorder buffer in order to read the information in the PC and the computed branch target address. The information obtained from the exception pointer and the branch decision flags are used by the branch predictor to verify past prediction.

If the prediction was satisfactory, branch predictors set a non-speculative flag in the corresponding in-flight tag. For misspeculation, the fetch unit signals in all structures send the checkpoint for recovery.

3.1.4 Branch Predictor Unit Update
When the predictor hits or misses in the prediction of conditional branches, the processor feedbacks to the branch predictor unit with the condition and the branch destination address calculated to improve confidence for future predictions. In the case of misses, together with the update action, the exception recovery mechanism is launched to clean the reorder buffer of incorrect path instructions. In the proposed model, branch decision flags and executed flags are set by the processor on the write-back stage.

This condition is sufficient for selecting the corresponding entry of BROB to make the branch unit start updating as explained in Section 3.1.3, a fixed priority circuit can be used for the branch predictor unit update request logic.
3.2 Physical Register Recycling

Another support provided by the ROB is physical register recycling. The life time of a logical register is specified by the compiler, when the logical register is reused in the program, which means that the last value is no longer necessary in the execution code. Its associated physical register is considered old and must be recycled when the instruction is complete.

Each renamed instruction has a current destination physical register and an old destination physical register; both registers are inserted in the ROB with the instruction at dispatch. The old destination ROB section works together with the renaming unit of free register list as shown in Figure 4. At commitment, old destination register tags are recycled to the free physical register pool and are used to turn off the register ready bits which are set by wakeup events while the current destination physical register tag is used to set the register valid bit in order to update the architectural state of the processor.

3.3 Load/Store Reorder Buffer

LD/ST instructions are split into memory address calculation and the corresponding read or write action. A special address queue is used to store the immediate value, the base address register, the tag of the source or destination register, the in-flight tag, and the LS-Buffer entry assigned to memory instructions. A special LS-Buffer is used to store memory data, memory address, R/W bit, and in-flight tag as an interface to the memory port as shown in Figure 5.

Memory address computations are resolved by the ACU and the results are written to the address field of the LS-Buffer. For loads, the destination register tags appoint the register file, to write the data read from memory. For stores, the source register is read from the register file and written to a LS-Buffer entry data field. Since memory access involves multi-cycle operations, the issue flag of the reorder buffer is set when the address calculation is sent to execution meanwhile the execute flag is set when the instruction memory access is complete.

3.4 Commitment Mechanism

The superscalar processor makes a checkpoint of its state in each clock cycle through the rename units, issue queues, and register file by storing...
multiples copies in recovery structures. They map its status in shadow memories. Rename units use shadow maps, instruction queues use bit-vectors as valid entries, and the register file uses register ready bit and valid bit vectors. For example, in a 4-way processor, RATs of rename units maps four instructions in each cycle. This mapping corresponds to four instructions which will be allocated in one entry of the reorder buffer in the next cycle.

Group commitment is the mechanism implemented in our design; it assists the processor’s state management and recovery while reducing design complexities with negligible impact performance. Figure 6 shows the structures of a reorder buffer organized in groups; here the instructions are fetched in a 4-way processor to illustrate group commitment scheduling. At dispatch, four instructions are inserted in the reorder buffer and each instruction sets a bit in its corresponding bit-vector dispatch structure. Instructions could be dispatched to any queue (IQ, FPQ, or LSQ), but when issued, each queue can set the corresponding issue flag in a 1-bit vector issue structure. Executed flags are set by the functional units at the end of execution. A non-speculative flag is set by non-speculative instructions at dispatch and by the branch predictor unit when verifying the chosen decision for the branch predictor unit at write-back.

When four consecutive instructions have been dispatched, issued, executed, and are not speculative, the commitment mechanism augments the head pointer register for all structures of the reorder buffer, freeing resources such as the corresponding checkpoint copies and old destination registers. Note that the head pointer is a register of 7 bits split into a 5-bit part to index the ROB-entry and a 2-bit offset to select a precise in-flight instruction. This addressing mechanism allows fast and exact exception recovery. The 5-bit ROB entry matches the RAT copy index of the renaming unit and other checkpoint structures, while the 2-bit offset permits selecting the offending instruction exactly.

### 3.5 Exception Recovery

When misspeculation is detected by the branch predictor unit, the checkpoint index is sent to all structures including the reorder buffer unit. This index is loaded to the tail pointer register invalidating all entries between the tail pointer and the head pointer, and their corresponding checkpoint copies. Finally, the status checkpoint copies of every processor structure, indexed by

---

**Table 1. Processor configuration**

<table>
<thead>
<tr>
<th>Element</th>
<th>P1</th>
<th>P2</th>
<th>P3</th>
</tr>
</thead>
<tbody>
<tr>
<td>ROB</td>
<td>128</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B-ROB</td>
<td>08</td>
<td>16</td>
<td>32</td>
</tr>
<tr>
<td>L/S Queue</td>
<td>32</td>
<td></td>
<td></td>
</tr>
<tr>
<td>F-I-C-Width</td>
<td>4-4-4/8-8-8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Int. Functional Units</td>
<td>4-2-2/8-4-4</td>
<td>ALU-MUL-DIV</td>
<td></td>
</tr>
<tr>
<td>F.P. Functional Units</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Branch Predictor</td>
<td>gshare, 2048-Entries</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Branch Penalty</td>
<td>8-Cycles</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Memory ports</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L1 Data Cache</td>
<td>64K</td>
<td>1 Cycle</td>
<td></td>
</tr>
<tr>
<td>L1 Inst. Cache</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L2 Unified Cache</td>
<td>256K</td>
<td>10 Cycles</td>
<td></td>
</tr>
<tr>
<td>TLB</td>
<td>8-Entry, 4-Way, 8KB pages, 30 Cycles</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Memory Latency</td>
<td>100 Cycles</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
the tail pointer, are updated as the earlier processor status.

4 Evaluation

The framework for evaluation is the Simplescalar Suite [1] with modifications presented in Section 3, compiled for the PISA architecture and configured with parameters as shown in Table 1. A subset of SPEC CPU2000 benchmarks were compiled for PISA and used as input. To explore the microarchitecture behavior, a dynamic subset of instructions of each benchmark consisting of 200M committed instructions were simulated, getting statistics after 100M forward instructions.

4.1 Evaluating Commitment in Group

First, we evaluated the impact of group commitment compared with individual commitment. For comparison purposes, 08-entry, 16-entry, 32-entry, and 64-entry BROB plus 128-entry distributed ROB structures defined in Section 3 have been modeled and were compared with traditional 128-entry reorder buffer identified as 00-BROB. Figure 7 shows the average IPC performance, the results of simulating the subset SPEC CPU2000 integer and floating point benchmarks. The group commitment model allows the load-store instructions to be committed individually. The simulation results report a negligible negative impact on the processor performance. The worst cases are a 0.5% and 0.7% performance loss for 4-way and 8-way processors.

4.2 Measuring the Impact of the BROB Size

Second, we evaluated the impact of the branch ROB structure size. Our model stops the fetch
activity when BROB becomes full until it has room to allocate new branches. Figures 8 and 9 show the processor performance for the four configurations of the branch-ROB structure a) for integer and b) for floating point for 4- and 8-way processors, respectively. We can observe a little performance loss for 08-BROB and 16-BROB, but for 32-BROB model there is no performance loss.

The first consideration to select the optimal size of a Branch ROB is related with in-flight branch instruction average. Figure 10 shows the percentage of branches executed. Simulation reports (on average) 21.5 % of integer executed instructions and 12.5 % of floating point executed instructions corresponding to conditional branches.

4.3 Measuring the ROB/BROB Occupancy

The second consideration for selecting the optimal size of a branch ROB is to conserve the reorder buffer instruction occupancy similar to the traditional reorder buffer identified as 00-BROB but modeled with a BROB size equal to the ROB size in order to compare both ROB and BROB...
occupancies. Figure 11 shows the average instruction occupancy. The baseline instruction occupancy is reached in the 32-BROB model. The BROB size is answered in part by the average of executed branches and it is fully answered with similar occupancy of the baseline reorder buffer.

5 Conclusions

This paper presents a simple reorder buffer design based on distributed five 1-bit flag multiport structures (the dispatched flag, the branch flag, the issue flag, the execute flag, and the branch decision flag), two 7-bit multiport structures (the old destination register tag and the current destination register tag), and one 5-bit multiport structure (the exception pointer), which presents an easy solution for commitment and branch misprediction recovery.

The new multiport structures have 1 write port, 1 read port for the dispatched flag and the non-speculative flag, 6 write ports, 1 read port for the issue flag and 14 write ports, 1 read port for the executed flag and the branch decision flag structures for a 4-way processor. The use of the group commitment scheme assists the recovery and the processor state management while reducing design complexities.

The design proposes another hardware simplification by the use of a branch ROB, a small structure 25% of the ROB size to store the PC and the destination addresses of conditional branches. This microarchitecture allows updating the branch predictor unit as soon as the condition of the branch is resolved by the processor reducing unnecessary executions on the wrong path. The complete design does not cause a performance loss.

Acknowledgments

This work has been partially supported by grants under agreements SIP-20101320 and SIP-20101154 of the Graduate Studies and Research Department of the National Polytechnic Institute (IPN), Mexico, and by grants under agreements 124104 and 115976 of the National Council for Science and Technology (CONACyT), Mexico.

References


José R. Garcia received his B.Sc. degree in Electronic Engineering from the Autonomous University of Puebla, Mexico, in 2005. He is a M.Sc. student of the Computer Engineering Program at the Microtechnology and Embedded System Laboratory of the Center for Computing Research of the National Polytechnic Institute (CIC-IPN), Mexico. He is also an intern with GDC Intel Labs, Mexico. His research interests include high-performance computer microarchitecture and digital systems design based on HDL, FPGA and VLSI systems.

Cuauhtémoc Peredo received his B.Sc. degree in Electrical Engineering and his M.Sc. degree in Computer Engineering from the National Polytechnic Institute (IPN), Mexico. He is a member of the Center for Computing Research of IPN and a Ph.D. student of ESAII of the Polytechnic University of Catalonia, Spain. His research interests are digital systems, fuzzy logic, and digital control.

Marco A. Ramirez received his B.Sc. degree in Electronic Engineering (1995) and a M.Sc. degree in Computer Engineering (2002) from the National Polytechnic Institute (IPN), Mexico. In 2007, he received his Ph.D. degree in Computer Science from the Polytechnic University of Catalonia, Spain. Since January 1997, he has been with the Center for Computing Research of the National Polytechnic Institute (CIC-IPN), Mexico. His research interests include high-performance computer microarchitecture, digital system design based on HDL and FPGA for modeling microprocessors, and VLSI design for energy-efficient computing.

Luis A. Villa received his B.Sc. degree in Electronic Engineering (1992) and his M.Sc. degree in Computer Engineering (1994) from the National Polytechnic Institute (IPN), Mexico. In 1999, he received a Ph.D. degree in Computer Science from the Polytechnic University of Catalonia, Spain. From December 1999 to February 2001, he was with the Laboratory for Computer Science as a Postdoctoral Fellow at the Massachusetts Institute of Technology, working in the SCALE project. From October 2001 to January 2007, he was with the Mexican Petroleum Institute. Since January 2007, he has been with the Center for Computing Research at the National Polytechnic Institute (CIC-IPN), Mexico. His research interests include high-performance computer microarchitecture and VLSI design for energy-efficient computing.

Herón Molina received his B.Sc. degree in Electronic Engineering from the National Polytechnic Institute (IPN), Mexico, in 1991 and in 1995, his Ph.D. degree in Electrical Engineering from the Research Center for Advanced Studies (CINVESTAV), Mexico. From October 1998 to April 2008, he was with the Interdisciplinary School of Engineering and Advanced Technologies (UPIITA). Since May 2008, he has been with the Center for Computing Research of the National Polytechnic Institute (CIC-IPN), Mexico. His research interests include CMOS analog and digital VLSI circuit design, neuro-fuzzy systems and bioinformatics.

*Article received on 01/02/2010; accepted on 15/04/2011.*
An Operational Approach for Implementing Normative Agents in Urban Wastewater Systems

Juan Carlos Nieves¹, Dario Garcia-Gasulla¹, Montse Aulinas², and Ulises Cortés¹

¹ Knowledge Engineering and Machine Learning group (KEMLg), Universitat Politècnica de Catalunya (UPC), Barcelona, Spain
² Laboratory of Chemical and Environmental Engineering, University of Girona, Girona, Spain
{jcnieves,dariog,ia}@lsi.upc.edu, aulinas@lequia.udg.cat

Abstract. Water quality management policies on a river basin scale are of special importance in order to prevent and/or reduce environmental pollution caused by human sources. Industrial effluents are a priority issue particularly in Urban Wastewater Systems (UWS) that receive mixed household and industrial wastewaters, apart from rainfall water. In this paper, we present an analysis and implementation of normative agents that capture concrete regulations of the Catalan pollution-prevention policies. The implementation of the normative agents is based on Situation Calculus.

Keywords. Rational agents, environmental decision support systems, practical normative reasoning, situation calculus.

Un enfoque operacional para implementar agentes normativos en sistemas urbanos de aguas residuales

Resumen. Las políticas de gestión de la calidad del agua a nivel de cuenca hidrográfica son especialmente importantes para la prevención y/o reducción de la polución originada por el hombre en el medio ambiente. Los efluentes industriales son un elemento prioritario particularmente en los Sistemas Urbanos de Aguas Residuales (SUAR) que reciben mezcladas las aguas residuales provenientes de viviendas particulares y de industrias, así como el agua de lluvia. En este artículo, presentamos un análisis y una implementación de agentes normativos que capturan las regulaciones específicas de las políticas Catalanas de prevención de la polución. La implementación de los agentes normativos está basada en el Cálculo de Situaciones.

Palabras clave. Agentes racionales, sistemas de ayuda a la toma de decisiones, razonamiento práctico normativos, cálculo de situaciones.

1 Introduction

Environmental decision-making is a complex, multidisciplinary, and crucial task. Water managers have to deal with complex problems due to characteristics of the processes that occur within environmental systems and possible consequences for the environment. In addition to this, water managers have to deal with normative regulations that have to be considered in any decision.

At the European level, [9] was developed to apply an integrated environmental approach to the regulation of certain industrial activities. This means that, at least, emissions to air, water (including discharges to sewer), and land must be considered together. It also means that regulators must set permit conditions so as to achieve a high level of protection for the environment as a whole. Several national and regional efforts are being made in order to improve water quality management as well as to comply with the European regulations. Specifically, we analyze the Catalan experience as a realistic example of adapting the European guidelines to manage water taking into account the local/regional reality.

In order to analyze the context of pollution-prevention policies in Catalonia, we consider a concrete regulation [5]. It is a regional regulation developed to follow the Catalan sanitation...
program. One of the aims of this updated program is to directly link the urban wastewater treatment program with the industrial wastewater treatment program. It pays special attention to the industrial component of urban Wastewater Treatment Plants (WWTPs) in order to facilitate the connection to the public system of those industries and/or industrial parks that accomplish the requirements.

This regulation is not easy to interpret since each wastewater management case to be solved is different and has its particular complex peculiarities (e.g., flow, loads, frequency, location, polluting potential, involved agents, etc.). Water managers have to ensure that their decisions, given a particular problematic case, are taken on time and comply with specific regulations to prevent pollution and to ensure high water quality standards. For this reason, tools to simplify and shorten the experts’ decision-making task, specifically those aimed at understanding and applying regulations, are required.

Following the perspective that a software agent is an active entity whose behavior is described by such mental notions as knowledge, goals, abilities, commitments, etc. [25], we have been exploring the definition of intelligent agents provided with the normative knowledge in order to manage concrete normative regulations which are in the context of UWS [2,20].

A central issue to successfully implement a normative agent is the selection of formalisms for performing practical normative reasoning. By practical normative reasoning, we mean a computable normative inference able to infer a statement α from a normative knowledge base Σ at a low computational cost. Although several powerful formalisms exist, finding the right one is a non-trivial challenge, as it must provide a level of expressiveness that serves the practical problems at hand in a tractable way. In the literature, one can find several approaches for performing normative reasoning such as deontic logic, temporal logic, dynamic logic, etc. [16]; however, these approaches have a high computational cost. An important feature of formal methods is that they do help in the long run to develop a clearer understanding of problems and solutions; hence, the definition of computable tractable approaches based on formal methods is relevant for performing practical normative reasoning.

Since norms in real world are usually defined at an abstract level [23], the modeling process of real norms is not straightforward. Therefore, a relevant issue for performing normative reasoning is to find an operational representation of norms. In [1], a declarative representation of norms was introduced. The main ingredient of this representation is the consideration of the norm conditions, which define a life cycle of each norm, to infer when a norm is violated. According to [1], the detection of norm violations depends on two properties of inspection to be done:

1. Observability: the conditions or actions can be checked by internal agents, given the time and resources needed;
2. Computability: the conditions of actions can be checked in a feasible and low cost manner.

Using these two properties, we can analyze their impact on the implementation of norm enforcement. The hypothesis is that by observing the items which affect the lifecycle of a norm, one can infer the state of violation of the norm. In this paper, we explore the life cycle of a norm in terms of state machines such that each state represents a state of the world. An important issue in our approach is a representation of the world in terms of states/situations. We follow the approach introduced in [14,15] for observing the world and then inferring the state of a norm.

In this paper, we extend the work of the earlier paper [18] in order to present an implementation of normative agents based on Situation Calculus for performing practical normative reasoning in the domain of WWTP. The normative knowledge structure follows the approach introduced in [1]. Unlike the approach presented in [20], which extends the action language A to capture norms, in this paper we explore a norm’s lifecycle by considering states of the world (situations/sets of fluents). This means that our main concern will be to monitor the states of norms being either active or inactive (monitoring if a norm applies to a
certain agent), and violated or respected (regulating if a certain agent respects a norm’s content). We will present an analysis of the existing specific Catalan regulation of providing normative knowledge to normative agents. Also, we will describe an implementation of these normative agents using Situation Calculus.

The rest of the paper is organized as follows. In Section 2, a realistic hypothetical scenario is described in order to illustrate the role of some regulations for managing industrial discharges. In Section 4, a brief introduction to Situation Calculus is presented. This section also describes how to introduce normative knowledge in a Situation Calculus specification. In Section 5, we explain how to implement the approach of Section 4. In Section 6, an operational execution example of the prototype is presented. Section 7 gives a summary of related work, and in the last section, we outline our conclusions.

2 Realistic Scenario

In this section, a realistic hypothetical scenario is described in order to illustrate the role of some regulations for managing industrial discharges.

At the municipality of Ecopolis, a new industry called MILK XXI expects to be set up. As a result of its production processes, the main characteristics of its wastewater will be as follows:

- Flow: 60 l/s (5184 m³/day),
- Suspended Solids (SS): 130 mg/l,
- Biochemical Oxygen Demand (BOD₅): 450 mg/l,
- Chemical Oxygen Demand (COD): 800 mg/l,
- Oils and greases: 275 mg/l.

The Milk factory plans to work 16 h/day (two shifts), 225 days per year. It plans to get connected to the municipal sewer system which collects wastewater from a population of 12 000 inhabitants and transports it to the municipal WWTP. WWTP complies with regulations strictly. The owner of the industry submits a request to obtain authorization to discharge into the municipal sewer system, which is compulsory by law (Decree 130/2003 establishes the public sewer systems regulations). Moreover, the Milk industry plans to apply BAT² in order to reduce water consumption, so this fact is also declared in the request for a final authorization decision.

The industry intends to reduce 30% on water consumption, and consequently, the increment of pollutant concentrations is projected to be as follows:

- Flow (reduction): 42 l/s (3628.8 m³/day),
- SS: 200 mg/l,
- BOD₅: 600 mg/l,
- COD: 1000 mg/l,
- Oils and greases: 357.5 mg/l.

Several rules are launched to manage this case, which are included in the regulation analyzed in this work. In Tables 1 and 2, we list the agents and the norms involved directly in the case.

² BAT: Best Available Techniques [4].
We omit a detailed description of each agent; however, a version of these can be found in [2]. Note that the behavior of each agent is fixed by a set of norms that the agent has to comply with. In Section 3, we describe how these norms are expressed in mental notions of an agent.

3 Normative Specification Based on Situation Calculus

This section describes our approach to introducing normative knowledge in Situation Calculus. We begin with a brief introduction to Situation Calculus.

3.1 Situation Calculus

Situation Calculus [15] is a first order language for axiomatizing dynamic worlds. Nowadays, it has been considerably extended beyond the classical language to include concurrency, continuous time, normative knowledge, etc. [6, 14, 21, 22]; however, in all cases its basic ingredients are actions, situations and fluents.

- Actions: Actions are first-order terms consisting of an action function symbol and its arguments. In the scenario described in Section 2, a possible action of the industry is make_spill(spill_init,milkXXI).

- Situations: A situation is a first-order term denoting a sequence of actions. These sequences are represented using a function symbol do: do(a,s) denotes the sequence resulting from adding the action a to the sequence s. The constant s0 denotes the initial situation, namely an empty action sequence.

- Fluents: Relations whose values vary from state to state are called fluent; they are denoted by functions and predicates symbols. For relating the values of a fluent in a given situation, the binary relation hold(f,s) denotes the value of the fluent f in a situation s.

As any approach for temporal reasoning, Situation Calculus must deal with the Frame Problem to make its implementation consistent [15, 22]. Being aware of that, we present a specification that fully asserts the effects of all actions on every norm.

3.2 Norm Specification

In this section, we describe how to express norms in a specification of situation calculus, that is, a modeling process of norms at a high level. Since environmental domains are dynamic, that is, truth-values change with time, the described approach must deal with the specification of norms in dynamic domains. For this purpose, we follow the approach by states for specifying the world: this means that each state of the world will be reached by a finite sequence of actions in terms of Situation Calculus.

Before working on how to specify norms, we will analyze them, following [1, 19, 20] and keeping Situation Calculus particularities in mind. To fully specify a norm, several aspects must be identified:

Type of norm: norms that oblige to do something, norms that allow/permit something or norms that forbid something.

Conditions and content: the norm conditions and the norm content must be separated in order to study the characteristics of situations, in which the norm is active and in which the norm is violated.

States: a set of variables the norm refers to. For each possible value of those variables, the norm has one and only one activation and violation state, e.g., if a norm is applied to an agent then it should have one variable such as IdAgent, and if it is applied to an agent’s spill then it should have two variables, IdAgent and IdSpill.

Actions: a complete list of domain actions which may influence the activation state and the violation state, separately, for each norm.

Preconditions: preconditions for each action must be defined, that is, in which situations it can be executed and the requirements regarding its parameters.

Having all that information in mind, we can start to specify our norms. For the normative domain, we follow an adapted version of Reiter’s solution to the frame problem presented in [21]. We propose to split the specification into two
parts corresponding to the main properties that all norms have:

- Situations in which the norm is active;
- Situations in which the norm is violated.

To specify the situations in which a norm is active or violated, we will declare a value of fluents that will define unequivocally a set of situations which represent those set of states. The first part of the specification is meant to contain all possible states in which the norm must be taken into consideration (it is active). The second one comprises all the states in which the norm's content is violated. In what follows, we present the first part of the specification.

In our proposal, a norm $N$, after doing an action $A$ in a situation $S$, is active if and only if it fits one of three cases:

i. **$N$ was not active before doing $A$.** There is a set of conditions under which $A$ changes the activation state of $N$ from inactive to active. The conditions needed for $A$ to activate $N$ are fulfilled in $S$.

   *The Activation Condition*: Given a certain norm in a situation where the norm is inactive, the range of $A$ is a set of actions that may modify the values of the fluents on which the activation state of the norm depends, in a way that the resultant situation (defined by the resultant value of the fluents) could belong to the situations in which the norm is active.

ii. **$N$ was active before doing $A$.** There is a set of conditions under which $A$ changes the activation state of $N$ from active to inactive. The conditions needed for $A$ to deactivate $N$ are not fulfilled in $S$.

   *The Inertial Condition*: Given a certain norm in a situation where the norm is active, the range of $A$ is the actions that may modify the values of the fluents on which the activation state of the norm depends, in a way that the resultant situation could belong to the situations in which the norm is inactive.

iii. **$N$ was active.** There is no set of conditions that can make $A$ change the activation state of $N$ from active to inactivate.

The **Non-Termination Condition**: Given a certain norm and a situation where the norm is active, the domain of $A$ is the actions that may modify the values of the fluents in a way that the resultant situation could change the state of the norm from active to inactive or from inactive to active.

If we analyze these three rules, we can assure that every state in which the norm is active fits into one and only one of these three rules. By checking a certain situation with a proposed action, we can assert the activation state of any norm after that action has been performed in the situation.

The second part of the specification contains the situations in which a norm is violated. In this case, we have decided to make a simpler specification. In the specification of the activation condition, we had the temporal progression integrated into it by the use of a variable that represents the action just performed (variable $A$). This variable allows us to represent temporariness by joining the current state with the past state. In the case of the violation state, we propose a static specification. In it we will omit the variable $A$ and have the specification of the violation state solely based on the situation's fluents. It is possible to do so without losing expressivity since the temporal progression in our domain is represented as well in the fluents definition (whose specification looks very much like the Activation Condition specification), which contains the variable $A$ as well. Otherwise, we would lose the concept of time. By deleting this action variable, the specification becomes much simpler as only the fluents that define the states where the norm is violated have to be stated.

In our proposal, a norm $N$ is violated in a situation $S$ if and only if it fits one of these two cases:

i. $N$ is active in $S$, $N$ obliges to the value of one or more fluent, and $S$ does not fulfil all of those obliged fluents. This rule is intended to cover violations performed upon norms that oblige to something.

ii. $N$ is active in $S$, $N$ forbids the value of one or more fluent, and $S$ fulfils one of those forbidden fluents. This rule is intended to
cover violations performed upon norms that forbid something.

With those two rules, we cover all possible violations which can come upon a norm, as norms that allow something cannot be violated. Since a norm cannot be forbidding and obliging at the same time, those two cases are mutually exclusive. Once having the two parts of the specification of each norm, we can implement them to see how they work in a real life domain.

3.3 Permission Norms

Before seeing the resultant implementation, there is a particular case that must be studied, as it implies specification and implementation particularities: it is the case of permission norms. Following the Deontic Logic definition of permission, permission norms cannot be violated. That is because formally the permission deontic operator specifies something that can be done, but does not implicitly specify that the opposite cannot be done. On the other hand, considering that our objective is to faithfully reproduce an existing legal framework that regulates a real world domain, our world is closed such that everything is allowed unless otherwise regulated. These two facts together make implicit knowledge of permission norms an issue, which must be dealt with specifically.

As an example of the implicit forbidding content of a permission norm, we can see that the norm It is allowed to spill black waters into the river if one has the required authorization implicitly includes the forbidding norm It is forbidden to spill black waters into the river without authorization. When that forbidding part of a norm does not appear on its own among the norms and only appears implicitly, it must be made explicit in order to capture the complete meaning of a normative system.

Once we have analyzed all implicit norms, which must be made explicit (ideally with the help of a legal expert), it is necessary to coordinate them with the rest of the norms, which cover the same situation. The permission norms which are applicable to the same situation as forbidding norms (the norm may originally contain the implicit forbidding norm or not) will work as exceptions to the generic forbidding norm. It is important to make sure that the activation states of all those norms are dependent on each other. That is, one or more permission norms will be active for a given agent when the forbidding norm is not active, and no permission norm will be active for a given agent when the forbidding norm is active. One of the norms (a permission norm or a forbidding norm) must always be active, and two (one permission norm and one forbidding norm) must never be active at the same time in all possible situations of the norm to avoid self-contradiction.

To achieve such merging of norms, it is necessary to analyze and integrate together all norms regulating the same situation. One of the advantages of the specification and implementation approach presented here is that it defines generic norms first and allows later addition of exceptions (usually represented as permission norms) to the existing norm. Therefore, it is very easy to integrate new exceptions to generic norms, thus extending and enhancing the normative knowledge base.

4 Normative Implementation Based on Situation Calculus

In the previous section, we proposed a specification approach for norms working on dynamic domains using Situation Calculus and specifying a norm’s lifecycle as active, inactive, violated, and respected. Now we will show how this specification can represent real laws in standard Prolog. First, we consider and justify the code, which represents the states of a norm, and then we present an example of its application in our prototype.

We will discuss in detail only the code of one norm here, but the interested reader can find the whole Prolog domain and normative knowledge base at http://www.lsi.upc.edu/~jcnieves/software/NormativeKnowledge-PAAMS-2010.pl.

In this example, we are going to consider Decree 130/2003 of the Catalan Water Agency. Remember, the motivation of this decree was explained in Section 2. Norms regarding bureaucracy issues will be avoided deliberately, as we consider them to be less relevant to the
In order to make the formalizing process easier to understand, we will see a commented implementation of only one of those norms; specifically, we will see the regulation included in article 7.1. Following the analysis explained in Section 4, we know that:

Norm 7.1 is an obliging norm.
- The norm is active in the following situations:
  - A non-domestic agent is considered pollutant.
  - An agent spills > 6000 m³/year.
- The activation state of the norm can be changed by the following actions:
  - Make a spill: the execution of a spill is started by the agent thus increasing the agent’s m³/year production.
  - Add substance: a certain amount of a substance is added to a spill of the agent thus increasing the agent’s m³/year production.
  - Set agent type: the type (domestic, industrial...) of the agent is changed.
  - Set agent activity: the activity of the agent is changed; agents performing polluting activities are considered pollutant.
  - Set/Unset an activity as pollutant: the pollutant consideration of an activity is changed; agents performing polluting activities are considered pollutant.
  - Cancel a spill: the execution of a spill by the agent is terminated thus decreasing the agent’s m³/year production.
  - Delete substance: a certain amount of substance from a spill of the agent is deleted thus decreasing the agent’s m³/year production.
- Preconditions of each action will be defined by a predicate termed \( \text{poss}(A,S) \) following the Situation Calculus syntax, where \( A \) is an action and \( S \) is a situation.

### 4.1 A Norm Implementation Example

To implement the previously analyzed norm, we start with a representation of the activation state which may be active or inactive depending on the previous state of the world and what action occurred in it. In the code given further in this section, we specify three ways for a norm to become active in a given situation after an action occurred, as an adapted version of Reiter’s simple solution [21]. The code may look too extensive at first glance, but it is quite simple. After a thorough look, it can be observed that since we need to assert all the cases in which the

---

The mentioned Annex I and II in the laws refer to Annex I and II of the Catalan Decree 130/2003 that regulates the public drainage system. Annex I includes a list of limited substances and specifies these limitations; Annex II includes a list of forbidden substances.
norm is active so that the unasserted cases can be automatically classified as inactive, understanding the meaning and purpose of each line of the code is easy. The requirement of defining unequivocally all the cases in which the norm is active generates a simple and explicit representation of each of the possible states in which the norm is active.

The resultant Prolog implementation of norm 7.1 after applying the specification schema given before for the activation state of a norm (split into the three cases; Activation Condition, Inertial Condition and Non-termination Condition as explained in Section 4) can be seen in Tables 3 and 4.

### Table 3. Prolog code for the activation state of norm 7.1

<table>
<thead>
<tr>
<th>Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>holds(norm(71,IdAgent),do(A,S)):-</td>
<td>Norm 7.1 is active (holds) for an agent ( \text{IdAgent} ) after doing action ( A ) in situation ( S ) (do(A,S)) if and only if one of the following situations (Activation, Inertial or Non-termination conditions) takes place.</td>
</tr>
</tbody>
</table>

#### ACTIVATION CONDITION: Actions which could activate the norm and did so when the norm was inactive.

| A=make_spill(IdSpill,IdAgent), holds(spill_total_size(IdSpill,SizeS),S), holds(agent_spills_total_size(IdAgent,SizeA),S), Total is SizeA+SizeS,Total>6000, \( \vdash \text{holds(norm(71,IdAgent),S), poss(A,S)}; \) | The action is **make a new spill**, and with the size of this spill, the agent's total spills are bigger than 6000 \( \text{m}^3/\text{year} \). |
| A=add_substance(IdSpill,Sub,Qua), holds(agent_spill(IdAgent,IdSpill),S), holds(agent_spills_total_size(IdAgent,Size),S), \( 6000<\text{Qua}+\text{Size} \), \( \vdash \text{holds(norm(71,IdAgent),S), poss(A,S)}; \) | The actions is **add a substance** to a spill of the agent, and with the added substance, the agent's total spills are bigger than 6000 \( \text{m}^3/\text{year} \). |
| A=set_agent_type(IdAgent,non_domestic), holds(pollutant_agent(IdAgent),S), \( \vdash \text{holds(norm(71,IdAgent),S), poss(A,S)}; \) | The action is **set the agent's type** as non-domestic, and such agent is pollutant. |
| A=set_agent_activity(IdAgent,Activity), holds(pollutant_activity(Activity),S), holds(agent_type(IdAgent,non_domestic),S), \( \vdash \text{holds(norm(71,IdAgent),S), poss(A,S)}; \) | The action is **set the agent's activity** to the one considered as pollutant and the agent is non-domestic. |
| A=set_pollutant_activity(Activity), holds(agent_activity(IdAgent,Activity),S), holds(agent_type(IdAgent,non_domestic),S), \( \vdash \text{holds(norm(71,IdAgent),S), poss(A,S)}; \) | The action is **set an activity to pollutant**, this activity is of the agent, and the agent is non-domestic. |

#### INERTIAL CONDITION: Actions which could deactivate the norm but did not do it when the norm was active.

<table>
<thead>
<tr>
<th>Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>holds(norm(71,IdAgent),S), A = set_agent_type(IdAgent,non_domestic), poss(A,S);</td>
<td>The norm was active for the agent, and the action is <strong>set its type</strong> as non-domestic.</td>
</tr>
<tr>
<td>holds(norm(71,IdAgent),S), A=set_agent_type(IdAgent,AgentType), AgentType\neq \text{non_domestic}, holds(agent_spills_total_size(IdAgent,SizeA),S)</td>
<td>The norm was active for the agent, the action is <strong>set its type</strong> not as non-domestic, and the</td>
</tr>
</tbody>
</table>

---

4 Only the most significant inertial conditions are seen here. The rest can be found in: [http://www.lsi.upc.edu/~jcnieves/software/NormativeKnowledge-PAAMS-2010.pl](http://www.lsi.upc.edu/~jcnieves/software/NormativeKnowledge-PAAMS-2010.pl)
This implementation, as justified in Section 4, represents fully and in a computable way the activation states of a norm because it includes all possible situations in which the norm is active. Table 4 presents an implementation of the violation state of the norm as explained in Section 4.

5 An Operational Example

After having seen how the code of a norm is implemented within a dynamic domain, we can consider how agents working on that domain can interact with it. To achieve this, we developed a
prototype representing our domain and its norms. Using it, we can define the events that happen and observe how the state of world and what norms change as a result. The internal operation of the prototype makes use of functions introduced in Section, i.e., \(do(A,S)\) and \(holds(X,S)\). With them, the prototype is able to represent the world, its norms, interactions available in it, and the effects caused by the latter. Nevertheless, in order to interact with the prototype, we do not need any technical knowledge because most part of it is handled internally. To run it, we just need to define an initial situation, and based on it, a list of one or more actions to be performed sequentially on the initial situation. The prototype will then show us how the world and its norms change, affected by the actions of the user.

Next, we give an execution example based on a situation similar to the realistic scenario introduced in Section 2. In this example, we have an industry called MILK XXI, which intends to connect to the local WWTP. The new industry predicts the main characteristics of its wastewater to be as follows:

- Flow: 60 l/s (5184 m\(^3\)/day),
- SS: 130 mg/l,
- BOD\(_5\): 450 mg/l,
- COD: 800 mg/l,
- Oils and greases: 275 mg/l.

An industry agent named milkXXI defines the initial situation. This agent is of type non\_domestic and has no activity assigned. The activity dairy\_farming is considered pollutant in the Catalan Classification of Economic Activities, and there is an agent representing the local water agency called water\_agency. There is one spill, termed spill\_init still not associated with milkXXI, containing the substances described previously. Figure 1 presents this situation.

In that initial situation, no norm is violated and only one is active:

- 8.1 It is forbidden to spill forbidden substances.

In the active fluents of the initial state, it is declared that the spill spill\_init does not respect the limitation for the substance oils\_and\_greases (which is a forbidden substance). In the rest of the fluent, certain things can also be certified, which are true in the initial situation, i.e., that the spill, its total size (5184 m\(^3\)/year), and its substances are registered in the census.

In this situation, we will set the agent’s activity as dairy\_farming which corresponds to the action:

\[
set\_agent\_activity(milkXXI,dairy\_farming).
\]

As a result, we will reach the situation described in Figure 1b. In this situation, norm 7.1 is activated because the first condition is satisfied.

- 7.1 For the following agents, it is obligatory to obtain an authorization and to respect the restrictions of Annex I and II:
  - Non-domestic users whose activity is included in C, D and E sections of the Catalan Classification of Economic Activities (Decree 97/1995) and who are considered potential pollutant agents.
  - Those who generate spills > 6000 m\(^3\)/year.

If we look at the code of norm 7.1, in the part of the norm activation commented previously, the Activation Condition part, we can see one rule to justify this:

\[
A= set\_agent\_activity(IdAgent,Activity),
holds(pollutant\_activity(Activity),S),
\text{and} +holds(norm(71,IdAgent),S),
\text{and} poss(A,S);
\]

As with the action set\_agent\_activity, we set the agent’s activity to the one considered pollutant, and from now on, the agent is considered pollutant. Since the agent was non\_domestic, which is the other requirement of the first condition, the norm activates.

The next step is to associate the spill to the agent. Our intent is to simulate the scenario in which the company milkXXI produces the spill defined as spill\_init, in our prototype and based on the previous situation we execute the following action:

\[
make\_spill(spill\_init,milkXXI).
\]

\[\text{Norm}\ 8.1\ is\ always\ active\ for\ each\ agent.\]
By doing this, we reach the situation of Figure 2a, where there are five norms active, two of which are violated. The only active norms are the following:

- **7.31.** Active for agent *milkXXI* and *spill_init*: only if the pertinent agent considers it best to spill to the environment, then it is possible to not spill to the sewage system.

- **8.1.21.** Active for agent *milkXXI* and *spill_init*: it is allowed to dilute in order to approach best levels; if there is an emergency or an imminent risk, it is possible to dilute with a previous warning given to the competent agent.

- **10.21.** Active for agent *milkXXI* and *spill_init*: if one has obtained the authorization, this agent may spill black waters to the public sewer system according to the established regulations.

The violated norms are the following:

- **8.1 Violated for agent milkXXI** and substance *oils_and_greases*: forbidden substances must not be spilled.

---

6 Since this is a permission norm and the pertinent agent does not consider it best to spill to the environment, it is forbidden to not spill "spill_init" to the sewage system.

7 Since this is a permission norm, the activated unit here is the forbidding part of the norm. For this agent, it is forbidden to dilute "spill_init" because the situation is normal.

8 Since this is an permission norm and "milkXXI" does not have authorization, it cannot spill black waters.
7.1 Violated for agent milkXXI: for the following agents, it is obligatory to obtain an authorization and to respect the restrictions of Annex I and II:
- Non-domestic users whose activity is included in C, D and E sections of the Catalan Classification of Economic Activities (Decree 97/1995) are considered potential pollutant agents.
- Those who generate spills > 6000 m³/year.

We will focus on norm 7.1 because we have already analyzed its code. Norm 7.1, after being set active in the previous situation, is active again in the situation obtained after the execution of the action make_spill. We can find the reason for that in the activation state code, specifically, in the Non-termination condition:

\[
\begin{align*}
&\text{holds}(\text{norm}(71, \text{IdAgent}), S), \\
&\quad \triangledown A = \text{set_agent_type}(\text{IdAgent}, \text{Type}), \\
&\quad \triangledown A = \text{set_agent_activity}(\text{IdAgent}, \text{Activity}), \\
&\quad \triangledown A = \text{unset_pollutant_activity}(\text{Activity2}), \\
&\quad \triangledown A = \text{del_total_substance}(\text{IdSpill}, \text{Sub}), \\
&\quad \triangledown A = \text{del_substance}(\text{IdSpill}, \text{Sub}, \text{Qu}), \\
&\quad \triangledown A = \text{delete_agent}(\text{IdAgent}), \\
&\quad \triangledown A = \text{cancel_spill}(\text{IdSpill}, \text{IdAgent}), \text{poss}(A, S).
\end{align*}
\]

Since the norm was previously active and the executed action could not deactivate the norm (which is why make_spill does not appear in the Non-termination condition code), the norm will be active after the action is performed. Regarding the violation state, we can see that one of the conditions is fulfilled, specifically, the following:

The norm is active for the agent \text{IdAgent} and it produces a spill which violates a substance limitation.

\[
\begin{align*}
&\text{holds}(\text{norm}(71, \text{IdAgent}), S), \\
&\text{holds}(\text{agent_spill}(\text{IdAgent}, \text{IdSpill}), S), \\
&\text{holds}(\text{spill_violates_limitation}(\text{IdSpill}, \text{Substance}), S).
\end{align*}
\]

This is why the norm is active and violated in that state.

As it can be noted in the parameters of the violated norms, both refer to agent milkXXI. If we analyze both norms, we will realize that they regulate the same issue, and therefore, both violations can be solved by a single action. Such action deletes the amount of substance causing the violation. The action to achieve that would be: \text{del_total_substance}(\text{spill_init}, \text{oils_and_greases}).

After executing the action \text{del_total_substance}, we reach the situation of
Figure 2b. In this situation, even though all five norms are still active for milkXXI, none of them is violated. This can be proved by the fact that the fluent representing the violation spill_violates_limitation(IdSpill,Substance) is not true in that situation, thus, the deactivation of the norm is justified. Since the situation and the action performed satisfy the Non-termination condition again, the norm is still active.

This example demonstrates how to represent the world and its norms with fluents, how actions performed on the world change those fluents, and how a norm’s state is altered as a result of these changes. The presented prototype was developed with the objective of making it intuitive for the user. UWS managers, potential users of the tool, would require a very short introduction to the software because they are experts on the specific subject it works on. The prototype would help in the decision making process by considering, thanks to its computational power, all involved variables and all existing normative details. In particular, if the chosen actions to be performed on a UWS are tested with the prototype before their execution, potential unexpected violations of norms can be detected, undesired side effects can be avoided, and future situations can be analyzed.

6 Related Work

In the literature, different approaches for performing normative formalization can be found [7, 8, 13, 17, 24]. Papers more related to our work are those which use a state machine to represent the world and its norms, as we do. Such approaches come from the use of logic formalisms like Situation Calculus and Event Calculus, because their actions and fluents support that kind of representation. Event Calculus is formalism similar to Situation Calculus. The former uses actions (or events) happening on the domain as its main representing element, while the latter involves mostly with the world states (or situations). In [3], a normative formalization based on Event Calculus with the main objective to detect conflicts between policies is proposed. It is an interesting approach, and it deals with contradictory norms – one of the problems we came across when formalizing the laws regulating WWTPs. Also based on Event Calculus, Fornara and Colombetti [10] develop a more agent-orientated approach to deal with normative frameworks. They consider methods of communication between agents and work with such organizational elements as agent’s institutions. They define a norm’s state similar to our norm’s life cycle, although in our case, we focus on the world situation to define it, while Fornara & Colombetti use mostly roles and their available actions for the same purpose (which is explained by differences between Situation Calculus and Event Calculus).

Even though the system presented here could be used for both normative monitoring and reasoning, we consider normative reasoning as its main use. In particular, we consider of a special interest the practical reasoning process for analyzing the scope of a set of norms with respect to a sequence of actions. In fact, this is one of the main objectives of the prototype described in Section 5. Concerning that, practical normative reasoning can be applied to many scenarios, from a human society regulating the behavior of groups of people, like the WWTPs presented in this article, to a digital interaction unit controlling, for example, the interaction of Web Services. Kagal et al. [12] use their own specification language to define and manage the policies and constraints regulating the interaction of Web Services.

7 Conclusions and Future Work

We discussed an integrated approach towards building real normative systems to be deployed in scenarios were decision-making is constrained by the norms in use. The main contribution of this paper is that our research is a proof of concept for the advantages of using a normative system to make decisions in complex real life environments. In this sense, the most relevant principle is Accountability: agent-based services should know what they are doing and why, and they should be able to explain their actions or recommendations. Also, the separation of the logic layer from the user-interface and the dialogue layer is important. Since norms in real world are usually defined at
an abstract level [23], modeling real norms is not a straightforward process. Some authors have already pointed out that an instantiation of norms in a context domain helps to represent norms in a normative knowledge base [23].

In order to capture the scope of a norm in a dynamic domain like UWS, we have shown that one can fix the observable items that affect the lifecycle of a norm (see Section 4). In particular, the representation of these items in terms of fluents/predicates can help to infer the state of a norm. Since the state of a norm will be affected by changes in observable items, one can analyze the lifecycle of a norm in parallel to the changes of the observable items (see Section 4). We considered the use of Situation Calculus for implementing our approach. Note that the context domain can be clearly delimited by a set of fluents (a situation). This fact has been one of the main reasons for us to use Situation Calculus. As a running example, we analyzed the Catalan Decree 130/2003. It is a realistic example for managing UWS (see Sections 2 and 5).

In order to incorporate normative knowledge in a Situation Calculus specification, we proposed to split the specification of norms into two parts: 1) situations in which a norm is active and 2) situations in which a norm is violated.

The first part of the specification is meant to include all possible states in which the norm must be taken into consideration (the norm is active). The second one comprises all the states in which the norm’s content is violated. Since the norms are represented in terms of the fluents in a given domain, the proposed specification represents a natural extension of a Situation Calculus specification. Although the integrated framework has not been completely realized, we expect our work to lead to a methodology of systematic development of normative systems for decision-making in complex real life environments.

Here are some open issues we will pursue in future:

1. **Lifecycle of actions**: at the moment, we have assumed actions as atomic events. This assumption has its limitations in capturing temporal aspects as deadlines. Preliminary results with respect to this issue are presented in [11].

2. **Conflicts between norms**: to consider this issue, we will explore a partial order of norms.

**Acknowledgements**

Numerous discussions with J. Vázquez-Salceda helped us to clarify our ideas. This work has been partially supported by the FP7 European project ALIVE IST-215890. The views expressed in this paper are not necessarily those of the ALIVE consortium.

**References**


Juan Carlos Nieves is a postdoctoral researcher in the Knowledge Engineering and Machine Learning Group at the Department of Llenguatges i Sistemes Informàtics (LSI) of the Technical University of Catalonia (UPC). He explores reasoning approaches for defining reasoning skills of individual agents in such domains as Wastewater Treatment Plants. His general research domain is knowledge representation and reasoning, mainly using the argumentation theory and answer set programming, a declarative logic programming approach.

Dario Garcia-Gasulla is a Ph. D. student at the Technical University of Catalonia (UPC) and a junior researcher in the Knowledge Engineering and Machine Learning Group of the Department of Llenguatges i Sistemes Informàtics (LSI). He is interested in dynamic domain representation, normative reasoning and knowledge discovery in semantic networks.

Montse Aulinas has been a partial time associate teacher of the University of Girona (UdG) since 2008. Since 2004, she has been working on several areas of environmental engineering: solid waste management, water and wastewater management. Her interests include intelligent environmental decision support systems and use of knowledge representation techniques for environmental data and knowledge. She has a Ph. D. in Environmental Sciences, particularly, in multi-agent systems and argumentation processes to improve water management in river basins.

Ulises Cortés has been a professor of the Technical University of Catalonia (UPC) since 2007. Since 1982, he has been working on several areas of Artificial Intelligence: knowledge acquisition and concept formation in knowledge-based systems, machine learning, and autonomous intelligent agents. He was awarded with the CLUSTER chair at the École Polytechnique Fédérale de Lausanne (EPFL) for 1998-1999. He has been a guest editor of special issues in several international journals: AI Communications, Journal of Computer-aided and Civil Engineering, Computación y Sistemas, Environmental Modeling & Software, Applied Intelligence, Neural Networks, and Complex Problem-Solving Technologies, International Journal of Approximate Reasoning. From July 2002 to 2008, he was a member of the European Coordinating Committee for Artificial Intelligence (ECCAI).

Article received on 27/03/2010; accepted on 14/12/2010.
Secure Architectures for a Three-Stage Polling Place
Electronic Voting System

Josué Figueroa González and Silvia B. González Brambila
Departamento de Sistemas, Universidad Autónoma Metropolitana,
Azcapotzalco, México DF, Mexico
josue.figueroa@gmail.com, sgb@correo.azc.uam.mx

Abstract. Security on electronic voting systems is fundamental; it must assure the integrity of all the elements involved or generated during a voting process. This paper presents a design of secure architectures for providing security, integrity and authenticity of the most important elements involved in an electoral process: configuration files, recorded votes and final result files. Also, different cryptographic protocols for assuring security properties of configuration and final result files are presented as a part of one of the layers of the architectures. We consider a polling place electronic voting system composed by three stages and the use of three different systems during the whole process. Our analysis of architectures and protocols shows that the designed elements assure the secure properties which an electronic voting system must fulfill.

Keywords. Cryptographic protocol, electronic voting, integrity, secure architecture, security.

Arquitecturas de seguridad para un sistema de voto electrónico presencial de tres etapas

Resumen. La seguridad en los sistemas de voto electrónico es fundamental, esta debe asegurar la integridad de todos los elementos involucrados o generados durante el proceso de votación. Este trabajo muestra la creación de arquitecturas de seguridad para satisfacer la seguridad, integridad y autenticidad de los elementos más importantes involucrados en un proceso electoral: archivos de configuración, votos almacenados y archivo de resultados finales. Adicionalmente, como parte de una de las capas de las arquitecturas, se desarrollaron diferentes protocolos criptográficos que aseguran las propiedades de seguridad de los archivos de configuración y de resultados finales. Se considera un sistema de voto electrónico presencial formado por tres etapas y el uso de tres diferentes equipos durante todo el proceso. El análisis de las arquitecturas y de los protocolos muestra que los elementos diseñados aseguran las propiedades de seguridad que un sistema de voto electrónico debe satisfacer.

Palabras clave. Protocolo criptográfico, voto electrónico, integridad, arquitectura de seguridad, seguridad.

1 Introduction

This section presents a general description of an electronic voting system and a voting process for which the secure architectures presented in this paper were designed.

Electoral processes have been carried out using traditional methods like ballots or telephone calls, and many of the results are obtained through manual counting. As time passed and technology rose, first electronic voting systems appeared but they were considered only as electronic vote counters [9]. Nowadays, the use of these systems is quite common in many countries.

Electronic voting systems are divided into two groups: remote voting systems with voting performed in a location other than that of the voting center, and polling place voting systems with voting performed at the location point of voting equipment known as the electronic ballot box.

To complete the whole process, three systems have been created, each one fulfilling a special function:

- Configuration file generator that creates files containing the configuration of the elections to be carried out;
Ballot box system that collects votes:

− Total result generator that collects all the results generated for each one of the ballot boxes and generates the final results for different elections.

The ballot box system issued in this paper is composed by three stages: the pre-voting stage, related to the installation of configuration files transported through a non-secure communication channel; the voting stage responsible for collecting votes, and finally, the post-voting stage that generates the final results for a particular equipment, which will be transported by a non-secure communication channel to a system that collects the final results of election.

Two kinds of users interact with the system: the functionary, a user responsible for the system (turns it on, configures it, and disables it) during the electoral process, and the voter, a principal user who casts his/her vote with the system.

1.1 Voting Process

A voting process begins with the creation of configuration files which contain information about elections. These files are produced by an agency responsible for conducting elections. These files are created before elections (the time depends on the regulations of the responsible agency).

The basic data of these files are: the name of elections to be performed and the options to choose from, which may include the names of political parties, candidates, questions of a poll, or others. Additionally, they can include information of the location of the ballot box, i.e., electoral district, state, etc. Validation of this information is made by the responsible agency and the persons involved in the election (candidates, representatives, volunteers, etc.). Digital signatures, public and private keys for the security of configuration files are generated in the same location where the configuration files were created.

Once configuration and security files have been generated, they are installed in the electronic ballot on the day of elections, during the electoral journey.

When the process of collecting votes is completed, a counting process begins, which can be of two types: total counting and partial counting. Partial counting gives the number of votes recorded in each one of the electronic voting machines; this is done at the place where the equipment is located. The results are displayed on a screen and also printed out so that the concerned audience can view them.

Total counting is performed in a place other than the location point of the ballot box, and with another system. The results of each partial counting are gathered and added to obtain the final results for each election. The results of each ballot box can be sent to this system through Internet or delivered using a storage device.

2 Related Work

This section discusses work related to the security of electronic voting systems.

When an electronic voting system is constructed, cryptography is not a problem; there exist many cryptographic techniques which have been efficiently tested. The problem appears when a system is developed under a non-secure platform or architecture; such problem is known as “the secure platform problem” [8]. For an electronic voting system, the data to be protected are votes considered as the fundamental element of the system, and also the so-called critical data which include configuration files and final result files [5]. According to [10], the menaces to consider at the moment of developing an electronic voting system are the following:

− External attacks. Until this moment, external attackers have not had enough time to access the systems for altering them; this is explained mainly by a lack of external ways to access the systems;
− Malicious voters. A voter might try to obtain an improper access to the system and vote more than once or affect the system performance.

There are a lot of papers about security on electronic voting systems including such topics as security protocol development [6], secure architectures [2, 10, 11] and the right way for performing an electoral process [4, 12]. Also,
there are papers that analyze existing voting systems [5].

A protocol similar to the one presented in this paper can be found in [6]; however, the latter work is focused more on remote voting. It manages the security of votes during a voting process efficiently, but it does not make any reference to the security of critical data (after or before a voting process), which are an important part of the system.

The same is true of the approaches presented in [2, 10, 11]. They are focused on remote voting and management of security for votes during their transmission to another system. The referenced papers do not consider the security of the pre- and post-voting stages.

Polling-place voting is explored in [1]; here, the security of the votes is linked to the activation method. Also, the security of data after the election ends is assured by a cryptographic protocol; however, this approach does not consider the security of configuration files.

3 Development

In this paper, the following attacks to be solved are considered:
- Modification of configuration files, which can alter the way votes are recorded;
- Relationship between the vote and the voter, so that someone could know if a particular voter votes for a particular candidate or option;
- Possibility for a voter to cast more than one vote;
- Modification of final result files, which alters the results of election.
- We do not consider such attacks as:
  - Supplanting a voter (this attack can be realized on a system that activates the ballot box) or cheating a functionary;
  - Changing a vote. This attack does not affect votes inside the system, which are secure because an attacker has no way to access them. However, when the results are outside the system, they can be threatened.

Table 1 presents the nomenclature used in the protocols. Note that when a -$f$ is used for any key in the protocols means that it is used for decrypting information.

After reviewing the architecture in [7], it was found that its layers and operation were appropriate for an electronic voting system. The cryptography layer is composed of cryptographic primitives and protocols that assure secure properties of an electronic voting system [2].

<table>
<thead>
<tr>
<th>Equipment</th>
<th>Key</th>
<th>Nomenclature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Generation of configuration files</td>
<td>Encrypted public key</td>
<td>$^e_{GM}$</td>
</tr>
<tr>
<td></td>
<td>Private key</td>
<td>$d_{GM}$</td>
</tr>
<tr>
<td></td>
<td>Symmetric key</td>
<td>$k_{GM}$</td>
</tr>
<tr>
<td></td>
<td>Special key</td>
<td>$k_{ESP}$</td>
</tr>
<tr>
<td></td>
<td>Public key</td>
<td>$e_U$</td>
</tr>
<tr>
<td></td>
<td>Private key</td>
<td>$d_U$</td>
</tr>
<tr>
<td></td>
<td>Symmetric key</td>
<td>$k_U$</td>
</tr>
<tr>
<td></td>
<td>Private key for signing</td>
<td>$d_{UF}$</td>
</tr>
<tr>
<td></td>
<td>Public key for verifying</td>
<td>$e_{UF}$</td>
</tr>
<tr>
<td>Ballot box</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Obtaining total results</td>
<td>Public key</td>
<td>$e_R$</td>
</tr>
<tr>
<td></td>
<td>Private key</td>
<td>$d_R$</td>
</tr>
<tr>
<td></td>
<td>Symmetric key</td>
<td>$k_R$</td>
</tr>
<tr>
<td></td>
<td>Special key</td>
<td>$k_{ESP}$</td>
</tr>
<tr>
<td></td>
<td>Captured result key</td>
<td>$r_T$</td>
</tr>
</tbody>
</table>

3.1 Initial Considerations

For developing secure architectures and protocols, an electronic voting system of three stages mentioned in Introduction was considered. For a correct implementation of protocols, it is necessary to install some keys on different equipment before the electoral process begins. This initial distribution is shown in Table 2. Symmetric and asymmetric keys not shown in Table 2 are generated when necessary. Such generation does not impact the system.
performance. Public and private keys take longer time to be generated, but there are only few of them. There are much more symmetric keys — about 500-700 times — that are generated and stored really fast. Tests on a PC take less than 2 seconds, and in an embedded system, less than 5 seconds. The fact that symmetric keys are generated when they are needed increases the security of stored votes.

<table>
<thead>
<tr>
<th>Table 2. Initial location for different keys involved in the protocol</th>
</tr>
</thead>
<tbody>
<tr>
<td>Generation of configuration files</td>
</tr>
<tr>
<td>(e_U)</td>
</tr>
<tr>
<td>(d_{GM})</td>
</tr>
</tbody>
</table>

There is no limit for the amount of equipments to be installed; also, there is no restriction which ties a physical device with the configuration information that can be installed on it.

3.2 Pre-Voting Stage Security

At this stage, security must assure the integrity and authenticity of received configuration files, that is, these files must not have been modified or substituted. Security of generated records during system configuration has to be guaranteed also. Secure architecture for this stage is composed of two layers: the control access layer and the cryptography layer.

The control access layer allows or denies access to the configuration interface and sends the generated elements to the cryptography layer. Components of the control access layer are the following:

- Subject: a functionary;
- Secure object: configuration interface;
- Authorization: allows or denies access to the configuration interface;
- Restrictions: the turning-on date and hour must be posterior (within certain limits) to the ones registered in the system for beginning its operation.

Using a cryptographic protocol, the cryptography layer validates the integrity and authenticity of configuration files.

It can detect if these files have been modified or if they proceed from a different source than the authorized one. Also, using symmetric ciphering, it assures the security of generated elements.

Cryptographic Protocol

The cryptographic protocol is divided into two main steps: generation and verification.

Generation Protocol

This step is performed by the configuration file generator, and the protocol must guarantee that any alteration of files will be detected. Also, if an attacker creates a new set of files, these must be detected as non-valid.

Generating a set of digital signatures \( (s_1, s_2, s_3, \ldots, s_n) \) for each one of the configuration files (1), a special key \( (k_{ESP}) \) is created by taking parts of these signatures (2). Using the special key, the public key of the configuration file generator \( (e_{GM}) \) is encrypted (3). The configuration files are ciphered \( (a) \) using the symmetric key of the configuration file generator \( (k_{GM}) \) (4); this key is protected by its ciphering with the ballot box public key \( (e_U) \) (5). Once finished, the files which will be sent are: the encrypted data \( (c) \), the digital signatures \( (s) \) and the encrypted symmetric key \( (p) \). Here are the steps of the protocol:

1: \( s = d_{GM} (a, H(a)) \).
2: \( k_{ESP} = s_1 + s_2 + s_3 + \ldots + s_n \).
3: \( ^*e_{GM} = k_{ESP} (e_{GM}) \).
4: \( c = k_{GM} (a) \).
5: \( p = e_U (k_{GM}) \).

Verification Protocol

When the set of files \( \{c, s, p\} \) is received, the encrypted symmetric key \( (p) \) is decrypted with the ballot box private key \( (d_U) \) so that the symmetric key \( (k_{GM}) \) is obtained (1); the latter deciphers the encrypted data \( (c) \) so that the configuration files \( (a) \) are obtained (2). Using the set of digital signatures \( (s_1, s_2, s_3, \ldots, s_n) \), the special key \( (k_{ESP}) \) is created (3). This key deciphers the public key...
of the configuration file generator \((e_{GM})\) (4). If the digital signatures have not been altered, the special key \((k_{ESP})\) will correctly decipher the public key \((e_{GM})\), and the integrity of data is thus assured (5). Here are the steps of the protocol for verifying the integrity and authenticity of data:

1: \(k_{GM} = d_U(p)\).
2: \(a = k_{GM}^{-1}(c)\).
3: \(k_{ESP} = s_1 + s_2 + s_3 + \ldots + s_n\).
4: \(e_{GM} = k_{ESP}^{-1}(\sqrt[3]{e_{GM}})\).
5: \(e_{GM}(H(a), s)\).

### 3.3 Voting Stage Security

A secure architecture for this stage must guarantee the integrity and confidentiality of stored votes. Such architecture is composed of three layers: the authentication layer, the control access layer, and the cryptography layer.

The authentication layer validates the identity of a user determining if the user is allowed to participate according to the following restrictions: the user must be registered in the list containing voters allowed to vote. Also, the user must not have participated previously. These conditions are verified on another system not considered in this paper.

The control access layer is based on information provided by the authentication layer; it determines the type of the user who will interact with the system and at this stage is expected to be a voter. Once the voter has completed his/her participation, the system is disabled so that this voter cannot vote again. The elements of the control access layer are the following:

- **Subject**: a voter;
- **Secure object**: a voting interface and a file with registered votes;
- **Authorization**: allows the access to the voting interface according to some restrictions;
- **Restrictions**: a voter must be validated for participation and can participate in voting only once; the system can record votes only until a specified hour.

There are two possibilities to prevent double participation of a voter.

- The first possibility involves the activation step, when the voter must be validated before voting by a functionary in a separate system which possesses the information of the voters.
- The second possibility involves the final step, when the voter finished his participation.

The fact of the voter’s participation is registered by the system, and if he wants to participate again, the system will reveal this intent, and the functionary will not enable the ballot box. Thus, when the voter is permitted to vote, this enables the system to register the vote. After that, the system is disabled and does not allow the voter to access the interface.

The cryptography layer uses symmetric ciphering algorithms to assure the integrity of registered votes. A different key, of an appropriate length [3], is used for ciphering each vote. The votes are stored using random storage in order to avoid the relationship ‘vote – voter’.

The keys used for ciphering votes are generated when the system is configured during the pre-voting stage. When a vote is registered, the symmetric key is chosen randomly, and the vote is encrypted before being stored.

### 3.4 Post-Voting Stage Security

At this stage, a secure architecture assures that once the system has been turned off, it cannot be turned on for introducing more votes. It also manages the security, authenticity and integrity of the files that will be sent to another system and of the file of generated records. Such architecture is formed by three layers: the authentication layer, the control access layer, and the cryptography layer.

The authentication layer validates the identity of a user determining if the user is a functionary, and allows such user to access the administration interface in order to finalize the electoral process.

The control access layer is based on the information sent by the authentication layer and determines if a particular user is a functionary. When the electoral process ends, this layer disables the system so that it cannot be used again. The elements of this layer are the following:

- **Subject**: a functionary;
- **Secure object**: a management interface and a file with registered votes;
Authorization: allows the access to the administration interface according to some restrictions;
Restrictions: only a functionary can access the management interface, and optionally, it can be accessed only after a specified hour.

The cryptography layer deciphers stored votes for final counting and obtaining the results of each election. This layer applies a cryptographic protocol for assuring the integrity and authenticity of the result file which will be sent to the final result generating system.

**Cryptographic Protocol**

At this stage, the encrypted items are those generated during the whole process, i.e., votes, final result, and records which indicate that the electoral process has ended.

The cryptographic protocol is divided into the same steps as the pre-voting stage protocol.

**Generating Protocol**

This protocol must assure that received data are the same that the ballot box generated. A set of asymmetric keys is created in the ballot box \((d_{Uf}, e_{Uf})\) (1) for generating a digital signature \((s)\) of the results file \((r)\) (2). This file and the private key \((d_{Uf})\) are encrypted using the symmetric key \((k_{Uf})\) (3) obtaining \((c)\) and \((^d_{Uf})\) (4). After that, the symmetric key \((k_{Uf})\) is encrypted using the public key \((e_{R})\) producing \((p)\) (5). A special key \((K_{ESP})\) is formed with the digital signature (6) which is used for ciphering the public key \((e_{Uf})\) (7). After these steps, the set of files to be sent are: \((c, ^d_{Uf}, ^d_{Uf}, p)\). The steps of the protocol are the following:

1: \(d_{Uf}, e_{Uf}\)
2: \(s = d_{Uf} (r, H (r))\).
3: \(c = k_{Uf} (r)\).
4: \(^d_{Uf} = k_{Uf} (d_{Uf})\).
5: \(p = e_{R} (k_{Uf})\).
6: \(K_{ESP} = s(r)\).
7: \(^e_{Uf} = K_{ESP}(e_{Uf})\).

**Verifying Protocol**

Once the set of files \((c, ^e_{Uf}, ^d_{Uf}, p)\) is received, the symmetric key \((k_{Uf})\) is decrypted using the final result generating system private key \((d_{R})\) (1). Then with \((k_{Uf})\), the encrypted results \((c)\) and result digital signature are decrypted (2). Also using this key, the private key \((d_{Uf})\) is decrypted (3). After that, results are captured from the record \((r_T)\) and their digital signature \((s)\) is obtained using \((d_{Uf})\) (4). This key is used for creating \((K_{ESP})\) (5) which deciphers the public key \((e_{Uf})\) (6) which in its turn verifies the integrity of the results (7). The steps for assuring authenticity and integrity are the following:

1: \(k_{Uf} = d_{R} (p)\).
2: \(r = k_{Uf} (c)\).
3: \(d_{Uf} = k_{Uf} (^d_{Uf})\).
4: \(s = d_{Uf} (r_T)\).
5: \(K_{ESP} = s\).
6: \(e_{Uf} = K_{ESP}(e_{Uf})\).
7: \(^e_{Uf} = H(r), s\).

**4 Security Tests**

In order to test the performance of security architectures and cryptographic protocols, an electronic ballot box was developed, in which the architecture’s elements and protocols were implemented.

**4.1 Pre-Voting Stage Security**

The control access layer did not allow the system to be used before the indicated hour and showed a message indicating that the system was turned on at a wrong moment and turned it off automatically.

For testing efficiency of the cryptographic protocol, a new set of configuration data, public and private keys were created. All possible cases of combining this data were tested, even assuming that an attacker obtained the original set of data. The goal was that the system recognized the altered data as valid. The results for each case are presented below, cases 1-6.
As it can be seen, the only case in which the system recognized data was when the entire original set of data files was used. Any other combination produced an error which was detected by the system.

4.2 Voting Stage Security

The authentication layer never allowed a voter to participate more than once, and the control access layer disabled the system each time a voter finished his/her participation to prevent double participation.

The cryptography layer encrypted each vote with its own key thus raising the security levels of the system, and random storage did not allow the relationship between a voter and his/her vote.

4.3 Post-Voting State Security

The authentication layer validated a functionary correctly in all cases, and the control access layer never allowed a voter to access the management interface. The elements of the cryptographic layer maintained the security of generated elements. The protocol for assuring integrity and authenticity

<table>
<thead>
<tr>
<th>Case 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original data: ---</td>
</tr>
<tr>
<td>Modified data: dGm, eU, s, a</td>
</tr>
<tr>
<td>Verifying</td>
</tr>
<tr>
<td>kGm ≠ du(p)</td>
</tr>
<tr>
<td>ERROR: The attackers’ public key doesn’t match with the original private key.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Case 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original data: eU</td>
</tr>
<tr>
<td>Modified data: dGm, s, a</td>
</tr>
<tr>
<td>Verifying</td>
</tr>
<tr>
<td>kGm = du(p)</td>
</tr>
<tr>
<td>a = kGm⁻¹(c)</td>
</tr>
<tr>
<td>eGm = KESP⁻¹(eGm)</td>
</tr>
<tr>
<td>eGm(a, s)</td>
</tr>
<tr>
<td>The public key matches with the private key.</td>
</tr>
<tr>
<td>Data are correctly decrypted.</td>
</tr>
<tr>
<td>ERROR: The public key (eGm) is not deciphered correctly because the signature creates a different special key than the expected one.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Case 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original data: eU, s</td>
</tr>
<tr>
<td>Modified data: dGm, a</td>
</tr>
<tr>
<td>Verifying</td>
</tr>
<tr>
<td>kGm = du(p)</td>
</tr>
<tr>
<td>a = kGm⁻¹(c)</td>
</tr>
<tr>
<td>eGm = KESP⁻¹(eGm)</td>
</tr>
<tr>
<td>eGm(a, s)</td>
</tr>
<tr>
<td>The public key matches with the private key.</td>
</tr>
<tr>
<td>Data are correctly decrypted.</td>
</tr>
<tr>
<td>The public key is correctly decrypted.</td>
</tr>
<tr>
<td>ERROR: Modified data do not match with the original signature.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Case 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original data: eU, dGm</td>
</tr>
<tr>
<td>Modified data: s,a</td>
</tr>
<tr>
<td>Verifying</td>
</tr>
<tr>
<td>kGm = du(p)</td>
</tr>
<tr>
<td>a = kGm⁻¹(c)</td>
</tr>
<tr>
<td>eGm ≠ KESP⁻¹(eGm)</td>
</tr>
<tr>
<td>The public key matches with the private key.</td>
</tr>
<tr>
<td>Data are correctly decrypted.</td>
</tr>
<tr>
<td>ERROR: The public key (eGm) is not correctly decrypted because the modified signature does not create the right special key.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Case 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original data: eU, dGm, s</td>
</tr>
<tr>
<td>Modified data: a</td>
</tr>
<tr>
<td>Verifying</td>
</tr>
<tr>
<td>kGm = du(p)</td>
</tr>
<tr>
<td>a = kGm⁻¹(c)</td>
</tr>
<tr>
<td>eGm = KESP⁻¹(eGm)</td>
</tr>
<tr>
<td>eGm(a, s)</td>
</tr>
<tr>
<td>The public key matches with the private key.</td>
</tr>
<tr>
<td>Data are correctly decrypted.</td>
</tr>
<tr>
<td>The public key is correctly decrypted.</td>
</tr>
<tr>
<td>ERROR: Verification fails because data do not match with the original signature.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Case 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original data: eU, dGm, s, a</td>
</tr>
<tr>
<td>Modified data: ---</td>
</tr>
<tr>
<td>Verifying</td>
</tr>
<tr>
<td>kGm = du(p)</td>
</tr>
<tr>
<td>a = kGm⁻¹(c)</td>
</tr>
<tr>
<td>eGm = KESP⁻¹(eGm)</td>
</tr>
<tr>
<td>eGm(a, s)</td>
</tr>
<tr>
<td>The public key matches with the private key.</td>
</tr>
<tr>
<td>Data are correctly decrypted.</td>
</tr>
<tr>
<td>The public key is correctly decrypted.</td>
</tr>
<tr>
<td>SUCCESS: Data match with the original signature.</td>
</tr>
</tbody>
</table>
of the result file was tested with the same kind of tests used at the pre-voting stage.

The protocol was subjected to the following tests with the results presented in cases 7-11.

<table>
<thead>
<tr>
<th>Case 7</th>
<th>Original data: ---</th>
<th>Modified data: s, r, d_U, e_U, e_R</th>
</tr>
</thead>
<tbody>
<tr>
<td>Verifying</td>
<td>k_U ≠ d_R(p)</td>
<td>ERROR: The public key does not match with the private key.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Case 8</th>
<th>Original data: e_U, e_R, d_U, e_R</th>
<th>Modified data: s, r, d_U</th>
</tr>
</thead>
<tbody>
<tr>
<td>Verifying</td>
<td>k_U = d_R(p)</td>
<td>The public key matches with the private key.</td>
</tr>
<tr>
<td></td>
<td>r = k_U^-1(c)</td>
<td>Results are correctly decrypted.</td>
</tr>
<tr>
<td></td>
<td>d_U = k_U^-1(*d_U)</td>
<td>The private key is correctly decrypted.</td>
</tr>
<tr>
<td></td>
<td>s = d_U(r_T)</td>
<td>Results from the final record are captured and signed.</td>
</tr>
<tr>
<td></td>
<td>k_ESP = s</td>
<td>The special key is created.</td>
</tr>
<tr>
<td></td>
<td>e_U ≠ k_ESP(*e_U)</td>
<td>ERROR: The special key is not the expected one because data have been modified and the public key cannot be decrypted.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Case 9</th>
<th>Original data: e_U, e_R, d_U, s</th>
<th>Modified data: r</th>
</tr>
</thead>
<tbody>
<tr>
<td>Verifying</td>
<td>k_U = d_R(p)</td>
<td>The public key matches with the private key.</td>
</tr>
<tr>
<td></td>
<td>r = k_U^-1(c)</td>
<td>Results are correctly decrypted.</td>
</tr>
<tr>
<td></td>
<td>d_U = k_U^-1(*d_U)</td>
<td>The private key is correctly decrypted.</td>
</tr>
<tr>
<td></td>
<td>s = d_U(r_T)</td>
<td>Results from the final record are captured and signed.</td>
</tr>
<tr>
<td></td>
<td>k_ESP = s</td>
<td>The special key is created.</td>
</tr>
<tr>
<td></td>
<td>e_U = k_ESP(*e_U)</td>
<td>The special key is the expected one and deciphers the public key.</td>
</tr>
<tr>
<td></td>
<td>e_U(r,s)</td>
<td>SUCCESS: Verification is valid because all data are original and have not been modified.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Case 10</th>
<th>Original data: e_U, e_R, d_U, s, r</th>
</tr>
</thead>
<tbody>
<tr>
<td>Verifying</td>
<td>k_U = d_R(p)</td>
</tr>
<tr>
<td></td>
<td>r = k_U^-1(c)</td>
</tr>
<tr>
<td></td>
<td>d_U = k_U^-1(*d_U)</td>
</tr>
<tr>
<td></td>
<td>s = d_U(r_T)</td>
</tr>
<tr>
<td></td>
<td>k_ESP = s</td>
</tr>
<tr>
<td></td>
<td>e_U ≠ k_ESP(*e_U)</td>
</tr>
</tbody>
</table>

The obtained results show that only when the original set of signatures, keys, and results are used, the system recognizes the data as valid.

### 5 Conclusions and Future Work

The electronic voting process involves more than collecting and counting of votes but also management of files involved during the whole process which is important too. Security of all elements used and generated during the configuring stage, vote gathering and final counting is fundamental for these types of systems. The secure architectures designed for each stage of the voting process assure the most important secure properties to be fulfilled by any electronic voting system. They guarantee that a
voter cannot vote more than once and that a vote cannot be related with the voter who issued it. The tests of the protocols were focused more on the steps that conform these protocols than on the security of the algorithms that are used at each step. It is important to take into account that polling place systems are considered secure because it is difficult for an attacker to obtain control of them; however, they are vulnerable at the moment of sending the configuring information or when results are sent to another system. As it is shown in the section devoted to the tests, the designed protocols are able to detect any modification of this critical data, even when an attacker gains access to them or to different keys used during the whole process.

The main difference between our research and other works related to electronic voting security, besides the fact that the latter are more focused on remote systems, is that most of these papers deal only with vote security, but few of them consider the so-called critical files. For an attacker, it can be difficult to access votes during the system operation; however, accessing the files while these are outside the system may be easier. An attack on configuration or result files can alter results without even altering votes. This paper presents a method to prevent such attacks, especially considering the fact that configuration or result files can be transported through an insecure communication channel. The security management of all elements generated throughout the process of voting is the main contribution of our work.

Future work within this approach may include development of secure architectures for the stage of configuration file generation and the stage of obtaining total results, and studying the way these are related to the architectures presented in this paper.

References

Silvia Beatriz González Brambila
holds a Computer Science degree from the Metropolitan Autonomous University (UAM), campus Iztapalapa, Mexico, a M.S. degree in Computer Science from UAM, campus Azcapotzalco, Mexico, and a Ph.D. in Computer Science from the Technological Research Institute of Monterrey, Morelos, Mexico. She works as a professor and researcher at the Systems Department of UAM, campus Azcapotzalco. She was the coordinator of the Computer Science Master Degree Program, the Computer Engineering Program, a teaching coordinator at the Basic Sciences and Engineering Department of UAM, campus Azcapotzalco, and EGEL of Informatics and Computer. Her research interests include electronic voting, IT governance and visual learning for robots.

Article received on 17/11/2010; accepted on 29/09/2011.
Incorporating Angular Ratio Images into Two-Frame Stereo Algorithms

Pablo Arturo Martínez González and Mario Castelán

Robótica y Manufactura Avanzada, CINVESTAV – Unidad Saltillo
Carretera Saltillo-Monterrey Km. 13.5, C.P. 25900, Ramos Arizpe, Coah.
Mexico

{pablo.martinez, mario.castelan}@cinvestav.edu.mx

Abstract. Light Transport Constancy (LTC) asserts that the reflectance ratio obtained from two different illumination variations remains constant for any given view of the observed scene. LTC was proposed in [21] as a rank constraint for solving the correspondence problem in multiple view stereo. In two-frame stereo, the simplest setting for LTC requires only two illumination variations and a single light source. Under this scenario, the rank constraint can be formulated through ratio images, and standard stereo algorithms can be applied in order to obtain a disparity map. Unfortunately, a ratio image may be subject to saturated pixel values, and this may diminish the quality of disparity maps. To solve this problem, as a first contribution in this work, we propose a post-processing operation based on slope angles related to the ratio values. Experiments show that new angular ratio images are more robust and deliver improved disparity maps. A second contribution of this paper consists in performing an experimental evaluation of angular ratio images under the standard test bed for two-view stereo algorithms, i.e., using different aggregation and optimization approaches. The results of our research are consistent with previously reported conclusions for two-view stereo surveys. It means that LTC may benefit from a vast variety of existent methods to solve the two-view stereo problem.

Keywords. Light Transport Constancy, two-frame stereo, ratio images.

1 Introduction

Acquiring a three-dimensional surface of objects is an important problem in computer vision because a 3D surface simplifies modeling of the object’s appearance. A 3D surface can be obtained using contact devices such as laser scanners. Other possibilities imply information provided by one or more cameras. This

Incorporación de las imágenes de relación angular en algoritmos de estéreo binocular

Resumen. La Constancia de Transportación de la Luz (LTC) establece que la relación de reflectancia obtenida de dos diferentes variaciones en iluminación permanece constante para cualquier vista dada de la escena observada. En [21] LTC fue propuesta como una restricción de rango para resolver el problema de la correspondencia en estéreo de múltiples vistas. En estéreo binocular, el escenario más simple para LTC requiere solamente dos variaciones en iluminación y una sola fuente de luz. Bajo este escenario, la restricción de rango puede ser formulada a través de las imágenes de relación y los algoritmos estéreo estándar son aplicados con el objeto de obtener un mapa de disparidad. Desafortunadamente, una imagen de relación puede ser sujeta a valores de pixeles saturados, los cuales pueden disminuir la calidad de los mapas de disparidad. Para superar este problema, como una primera contribución en este artículo presentamos una operación de post-procesado basada en los ángulos de pendiente relacionados a los valores de relación. Los experimentos muestran que las nuevas imágenes de relación son más robustas y ofrecen mejores mapas de disparidad. Como una segunda contribución, realizamos evaluación experimental de las imágenes de relación angular bajo una cama de pruebas estándar para algoritmos de estéreo binocular, i.e., usando diferentes enfoques de agregación y optimización. Los resultados de esta investigación son consistentes con conclusiones previamente reportadas en estudios sobre estéreo. Esto significa que LTC puede beneficiarse de una vasta variedad de métodos existentes para el problema de estéreo binocular.

Palabras clave. Constancia de Transportación de la Luz, estéreo binocular, imágenes de relación.
methodology is known as image-based 3D shape recovery. Although the image-based approach is appealing, the nature of the image acquisition process makes input images prone to errors. With respect to this, lighting manipulation represents a way to pose constraints on image-based shape recovery techniques. For instance, the light intensity can be regulated in order to obtain a 3D shape. This is the core idea of Light Fall-off Stereo (LFS) [12], where a number of images are gathered from a stationary camera as the illumination source moves away from the scene. Based on the inverse square law for light intensity, ratio images are directly related to the scene depth from the perspective of the light source. Controlling the geometric position of the light source is another option to attack the problem. The Photometric Stereo Method (PSM) [6, 22] is a classical technique in this respect. Here, a single camera captures images while the light source moves around the object in a fixed pose.

When more than one camera is required, binocular stereo (two-frame stereo) is obtained using the image-based 3D shape recovery method with the simplest setting. Here, only two cameras are needed to capture a still scene, and the correspondence problem is solved between the two views in order to obtain depth information. Unfortunately, when Lambertian reflectance and color/brightness constancy are not observed, calculation of correspondences becomes a difficult task. In binocular stereo, the manipulation of lighting has also been proposed. For example, structured light patterns may be projected over the surface of an object [17]. This is normally done using a projector, but colored laser rays can also be projected if more accurate results are needed. Another approach based on lighting variations is Helmholtz stereopsis. This method allows matching arbitrary Bidirectional Reflectance Distribution Functions (BRDF) and uses reflectance function reciprocity as an invariant [13, 24]. By collocating point light sources with each camera, it is possible to record reciprocal pairs using two different lighting conditions. Due to the reciprocity, the reflected light to the cameras will be equal. This method, however, requires the light sources to be collocated with respect to the optical center of each camera.

Other approaches combine photometric and geometric cues. For instance, in multiple-view photometric stereo, a number of images of an object are obtained from multiple viewpoints under varying lighting directions. Here, the silhouette of the object is used to recover camera motion. The correspondence problem, however, is not solved by means of illumination variations [10, 18].

Recently, Light Transport Constancy (LTC) [21] has been proposed as a correspondence clue in multiple-view stereo. LTC is used to formulate a rank constraint matching cost when the scene is observed in several lighting variations (changes in light intensity). LTC asserts that the reflectance ratio obtained from two different illumination variations remains constant for any given view of the observed scene. LTC does not require the position of light sources to be precisely calibrated or even known. In two-frame stereo, the simplest setting requires only two illumination variations. Under this scenario, the rank constraint can be formulated through ratio images, and standard stereo algorithms can be applied in order to obtain a disparity map. Unfortunately, a ratio image may be subject to saturated pixel values, noise, and occlusions, which may diminish the quality of disparity maps.

On the other hand, unlike classical grayscale (or color) image pairs which usually assume brightness/color constancy, ratio images rely on LTC and have demonstrated to provide improved disparity maps. For these reasons, Wang et al. pointed out a potential use of ratio images in future as the most feasible and robust way to deal with the two-frame stereo problem. Using ratio images for such a task may borrow ideas from the extensive literature related to two-frame stereo algorithms [4, 8, 11, 16]. For example, from the taxonomy of Scharstein and Szelisky [16], different matching costs, aggregation support, and optimization approaches can be applied to make calculation of dense disparity maps from ratio images more robust.
1.1 Aim and Contribution

The aim of this paper is to provide a new insight into the use and performance of the LTC constraint in binocular stereo approaches. The topic has been practically unexplored due to the original multi-view formulation of the LTC constraint, which differs from the binocular approach in the methodologies used to address the correspondence problem. With respect to our objective, the contribution of this paper is twofold. First, we introduce a post-processing operation based on the slope angles related to ratio values, which we call the angular ratio image. This operation attempts to overcome the unavoidable problem of either highly saturated or too dark ratio values in traditional ratio images. Experiments show that new angular ratio images are more robust and deliver improved disparity maps in comparison with their traditional counterparts.

Second, we perform an experimental evaluation of angular ratio images under the standard test bed for two-view stereo algorithms, i.e., under different aggregation and optimization approaches. To the best of our knowledge, this is the first work to report a detailed experimentation related to LTC under the well-known two-view stereo test bed. The results of this research are consistent with previously reported conclusions in stereo surveys, and this fact suggests that angular ratio images conserve some properties of intensity images and therefore are eligible to be put into any binocular stereo frameworks. In other words, LTC, in the form of pairs of angular ratio images, may benefit from a vast variety of existent methods for the two-view stereo problem.

The paper is organized as follows. In Section 2, LTC and use of angular measures for improving ratio images are explained; Section 3 presents experimental evaluation of angular ratio images based on the standard stereo taxonomy; finally, Section 4 gives conclusions and outlines further research directions.

### 2 Light Transport Constancy and Angular Ratio Images

For an easier comprehension of notation used in this article, Table 1 presents expressions referring to cameras, pixels, light source intensity variations, and ratio images. Light Transport Constancy states that the percentage of light reflected by a surface patch (the BRDF) remains constant for any given viewing direction of a static scene. Following the explanation in [21], let us denote a particular point in the scene as \( x_i \). This point will reflect light to cameras \( C_1 \) and \( C_2 \) according to \( I_{C_j}(x_i) = L(x_i)R(x_i, L, C_j) \), where \( I_{C_j}(x_i) \) is the reflected intensity in the direction of \( C_j \) from the point \( x_i \), \( L(x_i) \) is the incident light intensity at point \( x_i \), and \( R(x_i, L, C_j) \) is the reflectance function or BRDF at point \( x_i \), indexed by vectors.
The Lambertian assumption states that the reflected light is equal in the directions of a and b, i.e., the BRDF is shared and \( R(x_i, L, C_j) = R(x_i, L, C_j) \). Thus, we have \( I_{c_j}(x_i) = I_{c_j}(x_i) \). However, this relation will not hold in general for arbitrary (non-Lambertian) BRDFs. Light transport constancy assumes that the surface reflectance function, \( R(x_i, L, C_j) \), remains constant under variable illumination. If we vary lighting conditions so that the incident illumination varies by a factor of \( k(x_i) \), then the observed reflected light, \( I_{c_j}(x_i) \), will also vary by a factor of \( k(x_i) \) because

\[
I_{c_j}(x_i) = k(x_i)L(x_i)R(x_i, L, C_j) .
\]

(1)

Wang et al. [21] have shown how LTC can be used in multiple-view stereo to impose a rank constraint on the matrix

\[
I_{cv} = \begin{pmatrix}
I_{c_1 v_1} & I_{c_2 v_1} & \cdots & I_{c_m v_1} \\
I_{c_1 v_2} & I_{c_2 v_2} & \cdots & I_{c_m v_2} \\
\vdots & \vdots & \ddots & \vdots \\
I_{c_1 v_m} & I_{c_2 v_m} & \cdots & I_{c_m v_m}
\end{pmatrix},
\]

(2)

where \( I_{c_j v_k} \) is the observed grayscale value by the \( j \)th camera under the \( k \)th lighting variation. Note that, for the sake of simplicity, we omit the expression \( (x_i) \). However, each of the remaining equations in the paper is related to a single pixel at the position \( x_i \). The matrix with the minimum rank is therefore sought, i.e., if LTC is observed through different camera viewing positions and lighting variations, then the dimension of the column space of the matrix \( I_{cv} \) should be minimal.

The rank constraint holds only when the number of light sources is less than both the number of lighting variations and the number of cameras. Then the rank of \( I_{cv} \) is equal at most to the number of light sources. Since \( I_{cv} \) will be corrupted with noise, it is impossible to calculate the rank exactly. The Singular Value Decomposition of \( I_{cv} \) may be used for rank approximation. A matrix with most of their energy in the first few principal components is preferred, and moments can be used to approximate the notion of the minimum rank, as

\[
M = \sum_i i \cdot \sigma_i^2 / \sum_i \sigma_i^2 ,
\]

(3)

where \( \sigma_i \) are singular values of \( I_{cv} \). For multiple-view and multiple-lighting stereo, the minimum score is used as the matching cost.
Incorporating Angular Ratio Images into Two-Frame Stereo Algorithms

Let us now consider the simplest setting for LTC-based stereo, which is the case of interest in this paper: a single light source and two cameras. For each pixel in the left and right images, the intensities observed with the first lighting variation can be explained in terms of the intensities observed with the second lighting variation, as 

\[ I_{c_1v_2} = I_{c_1v_1} k_1 \quad \text{and} \quad I_{c_2v_2} = I_{c_2v_1} k_2. \]

Therefore, the relation between lighting variations is given by the ratio:

\[ \frac{I_{c_1v_2}}{I_{c_1v_1}} = \frac{I_{c_2v_2}}{I_{c_2v_1}} = k_1. \]

\[ \frac{I_{c_1v_2}}{I_{c_1v_1}} = \frac{I_{c_2v_2}}{I_{c_2v_1}} = k_2. \]

The matrix of intensities, \( I_{CV} \) can be now defined as

\[ I_{CV} = \begin{pmatrix} I_{c_1v_1} & I_{c_2v_1} \\ I_{c_1v_2} & I_{c_2v_2} \end{pmatrix}. \]

Note that LTC holds only if the second singular value of \( I_{CV} \) is zero. This means that the minimum rank of \( I_{CV} \) is one (the number of light sources) if and only if \( k_1 = k_2 \). Therefore, minimizing the second singular value is equivalent to minimizing...
Eq. 3. The so-called ratio image is a function \( R(v_1, v_2, x_i) = k_1 \), where \( v_1 \) and \( v_2 \) are two different lighting variations.

The ratio image is defined only for the two frame/two lighting variations scenario, and at most two ratio images can be recorded for a given stereo image pair as in Eq. 4 and Eq. 5. Note how the minimization of Eq. 3 can also be carried out using a simple absolute difference matching cost over the ratio image pair. In this sense, a wide variety of two frame stereo algorithms provide extensive ways to calculate dense disparity maps through ratio images.

In practice, unfortunately, the intensities of pixels do not necessarily observe \( k_1 = k_2 \). This is due to several reasons, among which insufficient lighting and a poor camera response are most common. Moreover, in some regions of an image, i.e., where specularities and edges occur, the ratio is likely to be either a value close to zero or an overly saturated value, that is, much greater than 100%, and as a consequence, correspondence cannot be solved. The unwanted effect of these pixel values can be reduced if the ratios are redefined as:

\[
k_1' = |\tan^{-1}(k_1)|, \quad k_2' = |\tan^{-1}(k_2)|,
\]

where \(| \cdot |\) is the absolute value. The angular data constrain the ratio values from the interval \([0, \infty)\) to the interval \([0^\circ, 90^\circ] \). We can now define a new ratio \( R'(v_1, v_2, x_i) = k_1' \), which will be referred to as the angular ratio image.

### 2.1 Experiments on Angular Ratio Images

For image acquisition, a Bumblebee stereo camera, 9cm baseline was used. The size of grayscale images was 640 × 480 pixels. A 20W Halogen-bulb desk lamp was mounted on a rotating ruler in order to capture illumination variations around a range of 180°, with 20° increment as shown in Figure 1 (right). The acquisition setting is shown in Figure 1 (left).

Before ratio values are converted into image values (grayscale), a normalization operation has to be performed. Let \( R(v_1, v_2, x_i) \) be the ratio values obtained from a pair of left images with lighting variations \((v_1, v_2)\) (the same observations hold for its corresponding right pair), i.e., using

---

**Fig. 4.** Disparity results on brightness constancy and angular ratio images. The first row presents disparity maps, the second row shows the lines of interest with constant y-axis along the disparity map.
Eqs. 4 and 5. Ratio images are stored after the following filter has been applied:

\[ R(v_1, v_2, x_i) = \begin{cases} \tau & \text{if } k_i \geq \tau \\ k_0 & \text{otherwise} \end{cases} \]

(8)

where \( \tau \) is a cut ratio value. Once high values are filtered, the final image is generated from the normalized values:

\[ R(v_1, v_2, x_i) = \frac{R(v_1, v_2, x_i)}{\tau}. \]

(9)

Note that, once Eq. 8 is applied, \( \tau \) becomes the maximum value of \( R(v_1, v_2, x_i) \). As far as angular ratio images are concerned, i.e., \( R'(v_1, v_2, x_i) \), Eq. 8 is not required, since Eq. 9 can be directly applied with \( \tau = \pi/2 \) (radians). Once a ratio (or angular ratio) image is generated, dense disparity maps are calculated using the standard platform developed in [21]. The sum of absolute differences (SAD) and the mean filter \( 9 \times 9 \) window were respectively used as matching cost and aggregation support parameters. The disparity maps were finally calculated under the winner-takes-all (WTA) criteria.

The experimental analysis commences with Figure 2, where the difference between ratio and angular ratio images can be appreciated. From top to bottom, the rows of Figure 2 show the lighting variation pairs \((-60^\circ, 60^\circ)\) and \((-40^\circ, 40^\circ)\). The left-camera image for lighting variations 1 and 2, the ratio image \( (\tau = 1) \), and the angular ratio image are shown row-wise. Recall that ratio and angular ratio images are calculated for a single view (i.e., left or right camera images) and two variations. In Figure 2, it is noticeable that the angular ratio images reveal a more robust adjustment of values than the ratio images, where the cut value has set the ratios to saturated values. This is the main problem in ratio image generation, i.e., choosing an optimum cut.
value. This effect can be visualized in Figure 3, where different values of \( \tau \) are applied to different lighting variations. Again, a generalized optimum value of \( \tau \) is not clear, since the figure shows that \( \tau = 5 \) favors the ratio obtained from the pair \((-60^\circ, 60^\circ)\), but over-darkens the ratio obtained from the pair \((-40^\circ, 40^\circ)\).

As far as the disparity results are concerned, these are demonstrated in Figure 4. Here, the first row includes disparity maps, while the second row shows the lines of interest with constant \( y \)-axis along the disparity map. For visualization purposes, a mask has been applied for isolating the objects of interest from the background. The first column demonstrates the results for brightness constancy, i.e., neither ratio images nor angular ratio images are used here. Instead, a usual grayscale image left/right image pair is used as stereo input. The rest of the columns present results from the angular ratio images with lighting variation pairs of \((-60^\circ, 60^\circ)\), \((-40^\circ, 40^\circ)\), and \((-20^\circ, 20^\circ)\). As expected, there is an improvement in disparity calculation for the angular ratio images over the brightness constancy, that is, the disparities are located over more continuous regions. Interestingly, there is little difference between the results related to angular ratio images, which suggests that the angles may represent a robust way to obtain similar disparity maps through different lighting variations. Additional scenarios are presented in Figure 5, where again, the disparity results for angular ratio images are better than for those relying on the brightness constancy assumption.

We explore the use of two alternative adjustment functions for ratio values, namely, the standard histogram equalization for ratio images and a linear mapping which fits the elements of vector \( x \) into the upper and lower boundaries \( b_u \) and \( b_l \), respectively. This linear mapping is given by the formula:

\[
f(x, u_l, b_l) = \frac{b_l - b_u}{\min(x) - \max(x)} x_l + \left( b_u - \left( \frac{\min(x) - \max(x)}{\max(x)} \right) \cdot b_u \right)
\]

where \( x_l \) is the \( i_{th} \) element of \( x \). Figure 6 shows the behavior of different adjustment methods as a function of grayscale values, i.e., grayscale values ranging from 0 to 1 (\( x \)-axis) are divided by a fixed grayscale value in order to obtain a ratio (\( y \)-axis) . From left to right, Figure 6 presents the results for the fixed grayscale values of 0.1, 0.5, and 1, respectively. The purpose of this figure is to provide a perspective of possible ratios which can be obtained from a fixed value. The linear mapping, equalization, arctangent and
ratio functions are shown as dark dotted, dark dashed, dark solid and gray solid plots, respectively. The arctangent function was bounded from $(0, \pi)$ to $(0, 1)$ for visualization purposes.

There are several features to note in Figure 6. First, the ratio function assigns values between 0 and 1 only if the fixed value is 1. This means that many pixels are prone to become saturated or too dark. Second, the remaining functions are able to bind the ratio values between 0 and 1. Third, the arctangent function appears as an interpolated plot between both equalization and linear mapping, allowing a more continuous mapping between different ratio values, thus avoiding the probabilities of assigning the extreme ratio values. Only when the fixed value tends to one (right-most diagram), the ratio function (which is identical to linear mapping) and the equalization function seem suitable for value assignment.

The last step of the analysis in this section is presented in Figure 7. Here, the top and bottom rows display the results using histogram equalization and linear mapping, respectively. In this figure, a standard histogram equalization operation was performed on a raw ratio image with the variation $(-60^\circ, 60^\circ)$.

Despite the equalized ratio image looks similar to its angular ratio image counterpart (top right corner of Figure 2), the disparity results are rather different, favoring the use of angular ratio images again. Similarly, the results of linear mapping present a deep discontinuity at the end of the depth line plot. Note that the linear mapped image is much darker than both the histogram equalized image and the angular ratio image.

3 Angular Ratio Images and the Standard Stereo Taxonomy

In [16], Scharstein and Szeliski proposed a taxonomy developed to study and classify a wide variety of two-view stereo methods. Their work has become a benchmark for testing and reviewing novel stereo approaches. Recently, some papers also used this taxonomy to evaluate the robustness of stereo algorithms on changes in illumination and color [9, 11]. Roughly, the taxonomy establishes the principle that any stereo algorithm can be divided into the following main steps:

- **Matching cost.** A cost of correspondence is calculated in order to determine pixel
disparity. Sum of Absolute Differences (SAD) and Sum of Square Differences (SSD) are typical examples of matching cost functions.

- **Aggregation.** The initial costs of correspondence are spatially aggregated over support regions. Square windows of fixed and varying shape and size are typical examples used as aggregating regions.

- **Optimization.** A disparity for each pixel is chosen as the result of minimizing a local or global objective function. Graph cut and dynamic programming are amongst the most popular optimization approaches.

- **Refinement.** The generated disparity maps are post-processed in order to remove errors, i.e., filling regions where disparity could not be determined.

The above steps can be combined into a specific sequence. Different stereo algorithms consist of different sequences of steps. For example, in local algorithms (window-based), the calculation of disparity at a given point depends only on the intensity values within a finite window (i.e., the aggregation step). These algorithms usually make implicit the softness assumptions due to aggregation. Typically, only matching cost with aggregation is used in these approaches [2, 23]. Global algorithms, on the other hand, make explicit the softness assumptions and solve the optimization problem. These algorithms do not usually include the aggregation step. Instead, they assign the disparity that minimizes the objective function. In some cases, this idea is realized with the help of a function that combines data of the first step with regularization terms [3, 14, 15].
The purpose of this section is to explore the influence of different aggregation and optimization approaches on disparity maps delivered from pairs of angular ratio images. The outcome of this experiment is relevant, since it may show if angular ratio images can be incorporated into standard two-view stereo frameworks. Following the experimental protocol in [16], we compare the results of applying the following aggregation methods to ratio images: shiftable window [1, 20], iterated binomial [5], regular diffusion [15], and membrane diffusion [15]. Likewise, we studied the performance of the following optimization methods: dynamic programming [1], scanline optimization [16], graph cut [3], simulated annealing [7].

Figure 8 illustrates different scenarios used in the experiments. Stereo image pairs were gathered under the same acquisition setting as explained in Section 2.1. From left to right, the scenarios are named Fps, Bot, Gift, and Edy. Fps and Bot were taken on a plain black background. A 20W halogen desk lamp was used as a light source.

The lamp was located approximately 50cm away from the closest object to the camera.

These scenarios were used in Section 2.1. For scenarios Gift and Edy, a textured background was used. A 150W halogen reflector functioned as a light source. The reflector was located approximately 80cm away from the closest object to the camera. In Figure 8, only the left image with central illumination configuration (i.e., 0° according to the diagram in Figure 1 from each stereo pair is presented.

By using two different light source intensities, we aim to provide comparisons among different illumination scenarios. It is important to mention that this evaluation does not include a quantitative analysis, since no range scanner or structured-light equipment was available for the experiments.

Fig. 10. Four different aggregation approaches in scenario Gift. The figure presents disparity maps calculated under angular ratio images using different aggregation methods: square shiftable window, iterated binomial, regular diffusion, and membrane diffusion

Fig. 11. Close-up images for regions of high discontinuity. From left to right, close-up images of a specific high discontinuity region of the disparity maps from Figure 10 are shown
Nonetheless, we believe that qualitative results provide a fair idea of the performance of different methods. Also, no information about a public database suitable for LTC tests was available.

### 3.1 Evaluating Aggregation Methods on Angular Ratio Images

In this section, we present experiments aimed to compare the performance of different aggregation methods applied to angular ratio images. For all experiments, a matching cost of SAD and a WTA optimization were used. Following the suggested experiments in [16], we selected the following aggregation parameters: square shiftable window size $9 \times 9$, iterated binomial with 6 iterations, regular diffusion with 30 iterations, membrane diffusion with 150 iterations, and $\beta = 0.2$. In order to show a classical aggregation example, i.e., $9 \times 9$ square window aggregation, we start with Figure 9. It presents, column-wise, a left image with a $40^\circ$ illumination variation, a left image with a $-40^\circ$ illumination variation, an angular ratio image, and the obtained disparity map from the angular ratio image in scenario Gift. The results of applying the aggregation steps mentioned previously are shown as disparity maps in Figure 10. Here, square shiftable window, iterated binomial, regular diffusion and membrane diffusion are presented from left to right. At a first glance, Figures 9 and 10 reveal the fact that there is no significant difference between aggregation methods and the classical one. However, it can be noticed that the most successful aggregation method over regions of high discontinuity (located along the boundaries of the different boxes and the ball) is the square shiftable window. This observation is clearly consistent with the conclusions reported in [16], where it was shown that the best approach to deal with discontinuity errors was the square shiftable window. In order to emphasize the former observation, a close-up of the bottom-right corner area of different disparity maps of Figure 10 is shown in Figure 11. Here, it is noticeable that the shiftable window aggregation outperforms the rest of the
aggregation methods, i.e., the error surrounding the contour of the ball does not affect the disparities obtained from the square shiftable window approach. Besides, object-to-background transition is sharper for the square shiftable window result, which is more noticeable around the boundaries of the boxes, that is, the shiftable window presents a less diffuse transition around such areas. Note that the rest of the aggregation approaches present a rather softened object-to-background transition.

As far as the remaining scenarios are concerned, only comparisons between the standard $9 \times 9$ square window and the square shiftable window are presented in Figure 12, since no major changes were observed using iterated binomial, regular diffusion, and membrane diffusion. The results in scenarios Fps, Edy, and Bot are shown row-wise. Note that in scenario Fps, the square shiftable window helps to improve the disparity map by diminishing the error around the shaded area of the vases. A similar phenomenon can be noticed in Edy and Bot scenarios in the shaded area of the mannequin’s head and the pots, respectively.

Finally, in order to demonstrate that LTC provides a substantial help to solve the correspondence problem in stereoscopy, we included Figure 13. The figure presents the results of different aggregation methods using only the brightness constancy constraint, i.e., no angular ratio images were used as inputs for these experiments. Instead, intensity images taken at $0^\circ$ light source direction were used. From left to right, Figure 13 presents disparity maps recovered using square shiftable window, iterated binomial, regular diffusion and membrane diffusion. For all aggregation cases, only recovered disparity maps of less quality than those obtained with angular ratio images can be observed. Basically, this is due to the presence of irregularities in the recovered disparity maps. Such irregularities are not noticeable when angular ratio images are used (Figure 10). Recall that the purpose of Figure 13 is not to compare different aggregation methods, but to compare the results obtained from using LTC (in the form of angular ratio images) against a usage of brightness constancy (in the form of traditional grayscale images).

### 3.2 Evaluating Optimization Methods on Angular Ratio Images

In this section, we present experiments aimed to compare the performance of different optimization methods when applied to angular ratio images. Following the suggested experiments in [16], we used the following optimization parameters: dynamic programming with a softness weight...
\( \lambda = 20 \) and occlusion cost \( \lambda = 20 \), scanline optimization with \( \lambda = 20 \), graph cut with \( \lambda = 20 \), simulated annealing with \( \lambda = 20 \) and 500 iterations.

Figure 14 shows four optimization approaches in different scenarios. From top to bottom, the figure presents disparity maps obtained from angular ratio images in scenarios Edy, Fps, Gift, and Bot. From left to right, the following optimization approaches are shown: dynamic programming, scan line optimization, graph cut, and simulated annealing. Observing the figure, interesting facts can be noticed. First, graph cut seems to group different regions of the disparity maps into clusters of disparity. These results resemble a clustered version of the standard approach (Figure 12); among the four methods, graph cut delivers the sharpest object segmentation. Unfortunately, this method is prone to errors due to self-shadowing in objects, as it can be seen mainly in scenarios Edy and Fps.

Dynamic programming, on the other hand, seems to cope well with self-shadowing, but the boundaries between different objects and the background are not defined as well as in the graph cut case. Scan-line optimization seems to be an intermediate option, between dynamic programming and graph cut.

Finally, the worst performance is delivered by simulated annealing, especially regarding the smoothness of the recovered depth maps. These results are consistent with the observations reported in [16], [3], and [19], where similar conclusions were derived for brightness constancy in grayscale images. Optimization methods were also applied to intensity images, i.e., to test the performance of brightness constancy against LTC and angular ratio images. The results of different optimization methods are presented in Figure 15; in these experiments, only the brightness constancy constraint was applied, i.e., no angular ratio images were used as inputs.
Incorporating Angular Ratio Images into Two-Frame Stereo Algorithms

4 Conclusions

We have focused on a potential issue in LTC-based two-frame stereo algorithms: generating ratio images suitable for standard stereo algorithms. To this end, we have shown that the arctangent function can transform the original ratio image values into a stabilized range of values, outperforming alternative adjustment functions such as histogram equalization or linear mapping. We have also fulfilled an experimental evaluation of performance of the standard benchmarking method in two-frame stereo algorithms over several angular ratio image scenarios. The main results of this evaluation confirm that aggregation and optimization techniques in two-view stereo algorithms applied to angular ratio images present a behavior comparable to the behavior of aggregation and optimization approaches applied to grayscale images. The former fact suggests that the standard taxonomy for binocular stereo may be applied to angular ratio images, keeping specific properties of different aggregation and optimization approaches as when applied on grayscale imagery. As LTC is an emerging tool whose potential use in the stereo vision area is mainly related to the two-frame case, this study represents the first attempt to validate the penetration of ratio images into the extensive two-frame stereo vision literature. Further promising directions of research include a rigorous study and, possibly, an appropriate automatization of light source intensity and position for two-frame stereo vision using LTC. This is motivated by the idea that an automatic control of the intensity of a light source based on information provided by cameras, i.e., the number of pixels out of normal ratio values, can deliver more appropriate disparity maps in accordance with the specific illuminations needs of the observed scene.

Acknowledgements

This work has been supported by Project Conacyt Ciencia Básica 61593.

References


Pablo Arturo Martínez González obtained BSc from Technological Institute of Fresnillo Zacatecas (ITSF) in 2007 and MSc in Robotics and Advanced Manufacturing from the Research Center for Advanced Studies of National Polytechnic Institute (CINVESTAV-Saltillo) under the supervision of Dr. Castelán in 2010. Currently, he occupies a teaching position at Polytechnic University of San Luis Potosí (UPSLP). At present, his research interests are related to the area of computer vision for robotics applications.
Mario Castelán obtained a BSc from the University of Veracruz (UV) in 1999 and a MSc in Artificial Intelligence from the University of Veracruz and the National Laboratory of Advanced Informatics (LANIA) in 2002. He obtained a Ph. D. in Computer Science from the University of York, U.K., in 2006. Currently, he is a full-time researcher at the Robotics and Advanced Manufacturing Research Group of CINVESTAV-Saltillo. Dr. Castelán has published over thirty papers in refereed international journals and international conferences in the area of computer vision, 3D shape analysis, statistical learning, and vector space analysis for robotics applications.


Article received on 01/02/2010; accepted on 10/01/2011.
Robust Extrinsic Camera Calibration from Trajectories in Human-Populated Environments

Guillermo Baqueiro Victorín¹ and Jean Bernard Hayet²

¹ Digipro, Distrito Federal, Mexico
² Centro de Investigación en Matemáticas (CIMAT A.C.), Guanajuato, Mexico
{baqueiro, jbhayet}@cimat.mx

Abstract. This paper proposes a novel robust approach to perform inter-camera and ground-camera calibration in the context of visual monitoring of human-populated areas. By supposing that the monitored agents evolve on a single plane and that the cameras intrinsic parameters are known, we use the image trajectories of moving objects as tracked by standard trackers in a RANSAC paradigm to estimate the extrinsic parameters of the different cameras. We illustrate the performance of our algorithm on several challenging experimental setups and compare it to existing approaches.

Keywords. Calibration, computer vision, tracking, video-surveillance and multiple camera systems.

1 Introduction

In spite of its spectacular development, video surveillance is yet largely dependent on human agents in charge of monitoring up to dozens of TV screens, which may be a source of negative detections. Recent years have seen the emergence of automatic, computer-aided video surveillance systems in the computer vision community. Typically these systems use state-of-the-art tracking algorithms in each camera of the network and fusion techniques to recover the 3D trajectories of the moving objects in the scene [3]. Then, this information feeds pre-defined or unusual event detection and may trigger alarms for the agents. An important element for a widespread use of such systems is an automatic calibration algorithm that would not require the costly intervention of an expert and that would allow the data collected throughout all the video streams to be fused properly.

This article presents an algorithm that estimates the extrinsic parameters of a set of different cameras involved in a surveillance network, i.e. the 3D transformations between pairs of cameras and between each camera and the reference plane. The assumptions we make are that (1) the targets are moving on a planar scene, which is a common setup in surveillance systems, (2) we have an estimation of the intrinsic parameters of each camera, and (3) the cameras are static. An important characteristic of such camera networks is that the viewpoints may be dramatically different from one camera to another, e.g., in the frames from the two sequences of cameras depicted in Fig. 1. In particular, this
prevents us from using traditional feature-based matching techniques based on local descriptors around interest points [7] for estimating the underlying geometric transforms.

Instead, in the vein of the seminal work of [6], we rely on the output of motion detection and motion tracking to guess correspondences at the level of motion blobs or motion tracks and to infer the corresponding geometry. In other words, we use the dynamic part of the scene for registering views instead of the static part.

The organization of this paper is as follows: in Section 2, we highlight noticeable related work in the literature; in Section 3, we describe our algorithm for robust inter-camera homography estimation; in Section 4, we see how to recover all the extrinsic parameters from homographies and show how to calibrate the whole camera network; in Section 5, we comment results obtained on different setups and compare our algorithm to other existing works in the literature; finally, Section 6 draws conclusions and introduces future work.

2 Related Work and Contributions

The seminal work of Lee et al. [6] uses the centroids of blobs extracted with standard background subtraction techniques to perform homography fitting with a least median square (LMS) approach, that is further refined in a second step. Its main drawback is that the number of putative correspondences grows very fast with the number of targets simultaneously detected at a given time-stamp, so that the number of inliers for the LMS optimization drops dramatically in proportion, making the algorithm unsuitable for regularly crowded scenes. Obviously, the dimension of the search space is reduced drastically when instead of motion detection blobs one forms the correspondences from tracking sequences [1, 9]. In [9], the authors present a RANSAC-like approach that performs, as we do here, non-uniform sampling in the set of putative sequences. It sequentially tests homographies from two pairs of sequences (two pairs in each video) and keeps the best homography. However, the likelihood functions that ponder each sample are not clearly defined.

The work in [1] is more general in a sense, as it is extended to fundamental matrix estimation. It is also based on RANSAC, but does not make particular distinction between samples to guide the consensus to the most promising pairs of sequences. In another paradigm, the work of [8] uses perspective invariants, namely the cross ratio of five points, to match trajectories between video sequences. The algorithm also allows calibrating the time offset between video streams. However, in most situations, it is quite difficult to isolate non-degenerate trajectories - i.e., sufficiently far from straight lines - to compute stable cross-ratios, so that the possible applications of this work are limited. Among the most recent works in the area, the one of [4] is interesting as it also takes radial distortion into account. However, the correspondences are determined on the base of control points manually selected on trajectories, which may make it more adapted for expert users. A common inconvenience of these previous approaches is that they use tracking trajectories directly as they come from the tracking algorithm, which causes problems of robustness in the case that the
tracking fails - and that the system is not aware of it. In many situations, e.g., because of occlusions in crowded scenes, tracking algorithms may be unable to distinguish one object from another and may assign a wrong identity to some tracked object. This may be catastrophic for the estimation of scene geometry.

Our approach takes some of the ideas developed in [1, 9] to cut down the algorithmic complexity of the correspondence problem and brings several contributions including (1) robustness with respect to possible failures of the tracking algorithms, (2) more reliable guidance of the optimization process to the correct geometry, and (3) an optimization process to find the calibration of a set of N cameras. As far as notations are concerned, we will use bold capital letters for matrices, regular letters for scalars and vectors. The indices will generally refer to the camera(s) to which the variable is related.

3 Inter-Video Homography Estimation

3.1 Problem Formulation

The problem setup and notations are detailed in Fig. 2: several cameras $C_i$, $1 \leq i \leq N$, with different degrees of overlap, monitor a scene where people or other mobile objects move. We suppose that this scene is laid on a reference plane $\Pi$, which induces a homography between any pair of cameras $(i,j)$ monitoring the scene, i.e., if $p_i=(u_i,v_i)^T$ is an image point in camera $C_i$, the projection of a point $P$ of the plane $\Pi$, and $p_j=(u_j,v_j)^T$ is the projection of this same point on camera $C_j$, then we have the classical relationship in homogeneous coordinates [4],

$$p_j = \begin{pmatrix} u_j \\ v_j \\ 1 \end{pmatrix} \sim H_{ij} p_i = \begin{pmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{pmatrix} \begin{pmatrix} u_i \\ v_i \\ 1 \end{pmatrix} \quad (1)$$

where $\sim$ means that a relation of equality holds for any multiplication factor $\lambda>0$, so that $H_{ij}$ has in fact only 8 degrees of freedom. The problem consists in estimating (1) these transforms and (2) the homographies $H_i$ that map points $P$ of the reference plane to their projection $p_i$,

$$p_i = \begin{pmatrix} u_i \\ v_i \\ 1 \end{pmatrix} \sim H_i P = \begin{pmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{pmatrix} \begin{pmatrix} X \\ Y \\ 1 \end{pmatrix} \quad (2)$$

where $(X,Y,0)^T$ are the coordinates of point $P$ in a frame $(X,Y,Z)$ (depicted in Fig. 2) such that $Z=0$ is the equation of $\Pi$. Traditional methods estimate homographies $H_{ij}$ by searching for point correspondences $(p_i,p_j)$ and by using them to solve the linear system directly induced by all the instances of Eq. 1. As these correspondences are difficult to find with static scene points and appearance whenever the viewpoint changes strongly, we rely on tracks from video trackers.

Our algorithm can be summarized as follows: (1) collect trajectories in each stream $V_i$ with a tracking algorithm, (2) pre-process trajectories to eliminate ambiguities at occlusion points (we will refer to the trajectory parts built in this way as trajets), (3) apply the RANSAC-like robust optimization process with a likelihood-guided sampling process between all pairs of cameras, and (4) refine the whole camera set calibration by non-linear optimization techniques.

3.2 Collect and Pre-Process Trajectories

In the first step of our algorithm, we collect tracking trajectories from all available video streams. We conceived our algorithm to be robust w.r.t. the properties of the 2D tracker, so that
which tracker to use is not very relevant here. Practically, we used some of the 2D tracker algorithms implemented in the OpenCV library. The result, for a video stream \( i \) of camera \( C_i \), is an initial set of trajectories \( L_i = \{l_i^{(m)}, m \geq 0\} \), encoding the position of one target centroid along the time. The centroids are chosen here instead of feet positions because most of the authors seem to agree that they are not as sensitive to noise as the feet position [1, 6, 10]. However, a consequence is that the computed homography will correspond to a plane passing through target centroids (i.e., not \( \Pi \)). We denote it as \( \mathbf{H}_{ij} \) and will see that the feet-to-feet homography \( \mathbf{H} \) can be estimated from \( \mathbf{H}_{ij} \).

In the second step, we form what we call trajlets, i.e., pieces of trajectories smaller than the initially collected ones in \( L_i \). The idea is to get a second set of trajectories that are not too short, in order for the optimization to remain tractable computationally, but at the same time cut in such a way that they would not be susceptible to be contaminated by errors from the tracking algorithm, e.g., occlusion errors. The latter case is quite common for most tracking algorithms: two tracks that intersect at some point may exchange their respective target identity. In that case, the result is that both tracks are unusable for establishing correspondences. To avoid this, for all the pairs of collected trajectories \( \{l_i^{(m)}, l_j^{(n)}\} \) for which some of the points \( p_{i,s}, p_{j,t} \) are close in the image (at some timestamp \( t \)), we simply cut off the ambiguous parts within a given radius \( \delta \).

For each cut on an initial pair \( \{l_i^{(m)}, l_j^{(n)}\} \), this process creates four sub-trajectories (trajlets) \( \{l_i^{(m)}, l_j^{(n)}\} \), such that,

\[
\begin{align*}
  l_i^{(m)} & = l_i^{(m)} - \bigcup \{p_{i,t-\delta}, \ldots, p_{i,t+\delta}\} \bigcup l_i^{(m)+}, \\
  l_j^{(n)} & = l_j^{(n)} - \bigcup \{p_{j,t-\delta}, \ldots, p_{j,t+\delta}\} \bigcup l_j^{(n)+}.
\end{align*}
\]

Then, we smooth these trajectories by using local filtering based on Bezier curves so as to reduce the impact of noise in the objects' position (this smoothing is evaluated in Section 5). The result of this processing is, again, for each video stream \( i \), an - a priori larger - set \( L_i' = \{l_i^{(m)}, m \geq 0\} \). Some of these trajlets are drawn in the upper part of Fig. 1.

### 3.3 Robust Homography Estimation

The estimation of \( \mathbf{H}_{ij} \) is done in a RANSAC-like scheme described in this section. A priori, a homography candidate for explaining the two images of the same scene can be derived from just one correspondence between a trajectory in \( i \) and a trajectory in \( j \), since it is entirely defined by 4 point correspondences [1, 8]. However, most of the trajlets appearing in usual video-surveillance contexts are close to degenerate, i.e. linear. This is why we generate here the candidate homographies from two trajlets correspondences instead of one. This, in turn, has an inconvenience, since, if we have an order of magnitude of \( \tau \) trajlets appearing at intersecting windows of time, then the probability for a sampled pair to match is roughly \( 1/\tau \). Then, the probability for two consecutively sampled trajlets to match is \( 1/\tau^2 \). Hence, the number of sampling iterations needed in RANSAC to ensure (in stochastic expectation) that at least one correct pair is sampled is quadratic in \( \tau \), which can be problematic with crowded scenes.

A solution is to avoid a uniform sampling process by assigning likelihood values to all possible pairs of trajectories and by sampling the trajectories according to these values of likelihoods. We define them through

\[
p(\mu; \nu) \propto \frac{1}{\max(N_j(l_i^{(m)}), N_i(l_j^{(n)}))} \Delta(l_i^{(m)}, l_j^{(n)}),
\]

where \( N_j(l_i^{(m)}) \) stands for the number of trajlets in \( L_j' \) that have a time overlap with trajectory \( l_i^{(m)} \) (the value being defined as if there were no time overlap) and \( \Delta(l_i^{(m)}, l_j^{(n)}) \) measures the time overlap between trajectories \( l_i^{(m)} \) and \( l_j^{(n)} \). These two terms (1) penalize the sampling of trajectories that could result ambiguous to match (large \( N_i \) or \( N_j \)) and (2) favor the trajectories with large overlap which improves homography estimation by using larger sequences. Another important point in the sampling process is a geometrical consistency check made on pairs of trajectories, according to which the polygon formed by the extremities of two trajectories should just keep or inverse the order of its vertices in other views.

\[\text{ISSN 1405-5546}\]
Both criteria make the required number of samples much lower than the aforementioned quadratic term above. In practice, we need a few dozen iterations to get a pair of correctly matched trajectories. The remaining part of the process is based on the classical RANSAC scheme and described in Algorithm 1:

**Algorithm 1.** Computation of homography $\hat{H}_{ij}$ between cameras $C_i$ and $C_j$.

\[ \hat{S} \leftarrow 0 \]

**repeat**

1. Sample a pair of *trajectories* $(l^{(m)}, l^{(n)})$ according to the likelihoods $p(l^{(m)}, l^{(n)})$.
2. Compute the candidate homography $\hat{H}_{ij}^{mn}$ from the correspondences between all points of $l^{(m)}$ and $l^{(n)}$ by using the classical DLT [5], and compute its inverse $\hat{H}_{ij}^{-1}$.
3. For all *trajectories* pairs $(l^{(r)}, l^{(s)})$, compute the residual symmetric error $\varepsilon(r,s)$:

\[
\varepsilon(r,s) = \frac{1}{2\|l^{(r)}\|\|l^{(s)}\|} \sum_{p_r \in l^{(r)}, p_s \in l^{(s)}} \left[ d^2 (\hat{H}_{ij}^{mn} p_r, p_s) + d^2 (\hat{H}_{ij}^{-1} p_r, \hat{H}_{ij}^{mn} p_s) \right].
\]

4. In the residual matrix $\varepsilon^2(r,s)$, identify elements that are (1) below a given threshold and (2) minima on the line $r$ and column $s$.
5. Sum in $S$ the lengths of the trajectories corresponding to the identified elements.
6. if $S > \hat{S}$:
   \[ \hat{S} \leftarrow S; \]
   \[ \hat{H}_{ij} \leftarrow \hat{H}_{ij}^{mn}. \]

**end**

until a given proportion of trajectories from video streams $V_i$ and $V_j$ have been explained by $\hat{H}_{ij}$ or a given number of iterations have been done.

if $\hat{H}_{ij}$ explains enough trajectories in $V_i$ and $V_j$
1. Consider $\hat{H}_{ij}$ as recovered;
2. Refine $\hat{H}_{ij}$ by a few Levenberg-Marquardt iterations on the residual symmetric error minimization.

else
   Consider the two views as unregistered.

**end**

### 4 Extrinsic Parameters Estimation

In this section, we describe how to recover a geometry of the scene (i.e., the relative position of two cameras), from an inter-image homography. Then, we propose an optimization scheme to calibrate a set of N cameras.

**Homography decomposition.** Once the homographies $\hat{H}_{ij}$ have been recovered as we saw in the previous section, we estimate the extrinsic parameters, i.e., the parameters $R_{ij}, t_{ij}$ of the rigid 3D transform between the two cameras acquiring video streams $i$ and $j$. For this purpose, we use the following decomposition of matrix $\hat{H}_{ij}$ into intrinsic and extrinsic parameters [4],

\[
\hat{H}_{ij} \sim K_i R_{ij} [t_{ij} \bar{n}^{T}_{ij} - d_{ij} I_{3 \times 3}] K_i^{-1},
\]

where the matrices $K_i$ are the intrinsic parameters of cameras $i$, supposed known here, and where $d_{ij}, n_{ij}$ give the equation of the plane (here, the centroids plane) in camera $i$ frame, i.e., its equation is $n_{ij}^T Q = d_{ij}$, where $Q$ are the coordinates of 3D points in the camera $i$ frame. The indices $ij$ may seem superficial in $n_{ij}$ and $d_{ij}$ as the plane equation is expressed just in the camera $i$ frame. However, we will use it to distinguish these estimates from other estimates of the same quantities. For example, vectors $n_k$ resulting from the decomposition of the computed homographies $\hat{H}_{ik}$ are also estimates of the normal to $\Pi^k$ expressed in the frame of $C_i$.

Note that Eq. 3 is given *only up to a scale factor* that we will determine in a second time. We use Triggs’ algorithm [10] to determine the decomposition values of $\hat{H}_{ij}$. Note that this algorithm gives two possible pairs for $R_{ij}, t_{ij}$, but one of them can be easily discarded. Here, we just select the one that corresponds to the most horizontal configuration of the camera.

**Determining a frame on the reference plane.** Once the normal $n_{ij}$ to the plane $\Pi^i$ (and $\Pi$, which is parallel) has been computed, a base $(e_1', e_2')$ of vectors generating the centroid plane can be chosen, for example $e_1' = (e_1 \wedge n_{ij}) / \|e_1 \wedge n_{ij}\|$ and $e_2' = (e_1' \wedge n_{ij})$, where the $e_k$ form the canonical base $(e_1 = (1,0,0)^T)$. This allows defining a frame associated to $\Pi^i$ as described in Fig. 3, centered on $D_i$, the orthogonal projection of the center of projection of $C_i$ onto $\Pi^i$. As a consequence, the vector of coordinates of any point $Q$ given in the camera $i$ frame can be written as

\[
\begin{pmatrix}
Q_x \\
Q_y \\
Q_z
\end{pmatrix} =
\begin{pmatrix}
D_i \\
0 \\
0
\end{pmatrix} +
\begin{pmatrix}
Q_{x'} \\
Q_{y'} \\
Q_{z'}
\end{pmatrix},
\]

where $Q_{x'}, Q_{y'}, Q_{z'}$ are expressed in the camera $i$ frame on the reference plane.
\[ Q = d_{ij}n_{ij} + Q_{n} = d_{ij}n_{ij} + \alpha e'_1 + \beta e'_2, \]

where \((\alpha, \beta)\) are the coordinates of \(Q\) in the defined frame.

**Image plane to ground plane homography.**

From the computed image-to-image homography \(H_{ij}\) and its decomposition, one recovers (up to a scale factor for \(d_{ij}\)) an homography to the centroids plane by deriving from the projection equation on \(C_i\) of point \(Q\) onto point \(q = (u_q, v_q)^T\)

\[(u_q, v_q, 1)^T \sim K_i (d_{ij}n_{ij} + \alpha e'_1 + \beta e'_2), \]

from which one derives in terms of the spatial coordinates on the (real) centroid plane, \((\alpha, \beta),\)

\[(u_q, v_q, 1)^T \sim K_i (d_{ij}n_{ij} + \alpha e'_1 + \beta e'_2), \]

i.e. \(H_{ij} = K_i (e'_1 \mid e'_2 \mid d_{ij}n_{ij})\) acts as a homography from the centroid plane \(\Pi^c\) (coordinates \((\alpha, \beta)\)) to the image plane in camera \(C_i\). However, an ambiguity remains in this definition as \(d_{ij}\) is computed only up to a scale.

**Scale recovery.**

As \(t_{ij}\) and \(d_{ij}\) are computed from Eq. 3 only up to a scale, we use some knowledge about the scene to compute the scale factor. One option is to assume a constant, fixed velocity for the object in the scene that has the median velocity. Another one is to assume a known half-height between people’s centroids and feet. In both cases, the scale recovery is straightforward.

We will explain it hereafter for the second case, and it can be proven in a similar way for the first one. Let us suppose that we have the knowledge of the half-height of a person, as the quantity \(L\) (e.g., 80 cm).

We will denote the different intrinsic parameters of camera \(C_i\) as \(\alpha_{u,i}, \alpha_{v,i}, u_{0,i}\), and \(v_{0,i}\).

\[ K_i = \begin{pmatrix} \alpha_{u,i} & 0 & u_{0,i} \\ 0 & \alpha_{v,i} & v_{0,i} \\ 0 & 0 & 1 \end{pmatrix} \]

If \(F\) and \(Q\) are the 3D points corresponding respectively to the feet and centroid of a tracked target (see Fig. 3), then we may write

\[ F = \begin{pmatrix} \alpha \\ \beta \\ L \end{pmatrix} \quad \text{and} \quad Q = \begin{pmatrix} \alpha \\ \beta \\ 0 \end{pmatrix} \]

The projections of these two points will be denoted by \(f\) and \(q\), and the half-height as measured in the image of camera \(C_i\) is \(l = ||q-f||\). By using the projection equations, \(Q\) is projected onto \(q\) through

\[ q \sim K_i (\alpha e'_1 + \beta e'_2 + d_{ij}n_{ij}) \sim \begin{pmatrix} \alpha_{u,i} \alpha e'_1 + \beta \alpha e'_2 + d_{ij}n_{ij} + u_{0,i} \\ \alpha_{v,i} \alpha e'_1 + \beta \alpha e'_2 + d_{ij}n_{ij} + v_{0,i} \\ 1 \end{pmatrix} \]

and, similarly for \(F,

\[ f \sim K_i (\alpha e'_1 + \beta e'_2 + (L + d_{ij})n_{ij}) \sim \begin{pmatrix} \alpha_{u,i} \alpha e'_1 + \beta \alpha e'_2 + (L + d_{ij})n_{ij} + u_{0,i} \\ \alpha_{v,i} \alpha e'_1 + \beta \alpha e'_2 + (L + d_{ij})n_{ij} + v_{0,i} \\ 1 \end{pmatrix} \]

We can express the observed distance in the image by taking the norm of \(q-f\), \(l\), which leads, by denoting \(q=(u_q, v_q)^T\), to

\[ l = \sqrt{[(u_q - u_{0,i})n_{ij}^2 + \alpha_{u,i} n_{ij}^2]^2 + [(v_q - v_{0,i})n_{ij}^2 + \alpha_{v,i} n_{ij}^2]^2} \]

\[ l = \sqrt{\frac{\alpha e'_1 + \beta e'_2 + (L + d_{ij})n_{ij}}{1}} \]

Then, by using Eq. 6, we get two equations in terms of \(x_q=(u_q-\mu_{0,i})/\alpha_{u,i}\) and \(y_q=(v_q-\nu_{0,i})/\alpha_{v,i}\),

\[ \begin{cases} x_q (\alpha e'_1 + \beta e'_2 + d_{ij}n_{ij}) = \alpha e'_1 + \beta e'_2 + d_{ij}n_{ij} \\ y_q (\alpha e'_1 + \beta e'_2 + d_{ij}n_{ij}) = \alpha e'_1 + \beta e'_2 + d_{ij}n_{ij} \end{cases} \]

from which we express \(\alpha = K_{\alpha} d_{ij}\) and \(\beta = K_{\beta} d_{ij}\), where \(K_{\alpha}\) and \(K_{\beta}\) are deduced from the known quantities \(x_q\), \(y_q\), \(e'_1\), \(e'_2\), and \(n_{ij}\). Now, by plugging...
these expressions for \( \alpha \) and \( \beta \) into Eq. 7, one can deduce an expression for \( d_{ij} \).

\[
L = \left[ -n_{ij}^T + \sqrt{\left( (u_i - u_0)x_j + s_0 \sigma_x n_{ij} \right)^2 + \left( (v_i - v_0)y_j + s_0 \sigma_y n_{ij} \right)^2} \right] - \frac{1}{L}
\]

\[
d_{ij} = \frac{K_0 e_i^T + K_0 e_j^T + n_{ij}^T}{L}
\]

(8)

Each blob \((q,f)\) giving an estimate of \( d_{ij} \) with the previous development and adult people forming the vast majority of the tracked blobs but not the entirety, we take the median among the different computed estimates for all the estimated inliers of \( H_{ij} \). In our experiments, this option assuming the people’s heights constant (as opposed to assuming the velocities constant) gave better and much more stable results\(^1\).

Once the scale is recovered, one finally gets either a ground plane-to-image \( H \) (as in Eq. 2) or an image-to-image \( H_{ij} \) induced by \( \Pi \) (as in Eq. 1). Most results below illustrate the second form. Note that for any homography we estimate, we will get a different definition of a frame relative to \( \Pi \). Also note that once the centroid-to-reference plane \( H_i \) is estimated, we get immediately the feet-to-plane \( H_i \)

\[
H_i = K_i[e_1^T, e_2^T, (d_{ij} + L)n_{ij}^T].
\]

**Calibration of a network of cameras.** As described in the previous paragraphs, each pair of cameras \((C_i, C_j)\) gives an estimate of the relative scene geometry for these cameras. Now, to get an estimate of the geometry of the set of \( N \) cameras, we proceed as follows. For each possible pair of videos, a calibration process following the previous method is done. The result is a set of external calibration data: \( \{(n_{ij}, d_{ij}, R_{ij}, t_{ij})\}_{ij} \) for all pairs of cameras \((i,j)\) that we could calibrate. For all these pairs, we also get a set of inlier points pairs \((p_i^{[k]}, p_j^{[k]})\) from matched trajectories, i.e., \( I_{ij} = \{(p_i^{[k]}, p_j^{[k]})\}_k \), as mentioned before, we may get several estimates of the same plane parameters, from different pairs of video streams. For example, the estimation of \( H_{12} \) and \( H_{13} \) both lead to estimates of \( \Pi \) in the frame of camera 1.

Also, among all the estimates, some constraints should apply. For example, if the equation of \( \Pi \) is estimated in both frames \( i \) and \( j \) respectively by homographies \( H_{ij} \) and \( H_{jk} \), then we should be able to get the plane parameters expressed in \( C_i \) from the ones expressed in \( C_j \) and the 3D transformation \((R_{ij}, t_{ij})\)

\[
\left\{ \begin{array}{l}
n_i = R_{ij}n_j, \\
d_i = d_j + n_j^T R_{ij}t_{ij}
\end{array} \right.
\]

(9)

To use efficiently all this redundancy, we propose an optimization scheme that cycles over all video streams \( i \), for which it alternates the following steps:

- **Step ‘P’** refines the plane parameters \((n_i, d_i)\) expressed in camera \( i \), with transformations \((R_{ij}, t_{ij})\) fixed, over all cameras \( j \) that have been registered with \( i \) (\( I_{ij} \neq \emptyset \));
- **Step ‘R’**, allows to enforce Eq. 9,

\[
\begin{align*}
n_i &= \gamma_i n_i + \sum_j \gamma_j R_{ij}n_j, \\
d_i &= \gamma_i d_i + \sum_j \gamma_j (d_j + n_j^T R_{ij}t_{ij})
\end{align*}
\]

(10)

where the sums are taken on all cameras \( j \) such that an homography has been computed between \( i \) and \( j \), and the \( \gamma_j \) are weights summing to one that are proportional to the total number of inlier points contributing to the estimation of the plane parameters through video \( j \) (i.e., a rough level of confidence over the estimation of plane parameters);

- **Step ‘T’**, for each camera \( j \) registered with camera \( i \) in the first steps, refine the transformations \((R_{ij}, t_{ij})\) with the plane parameters \((n_i, d_i)\) remaining fixed.

The two refinements (steps ‘P’ and ‘T’) are made respectively over the union of all sets of inliers \( I_{ij} \) and over individual sets \( I_{ij} \). The step ‘P’ tries to minimize a function \( F_i \)

\[
F_i(n_i, d_i) = \frac{1}{2} \sum_{i,j} \sum_{(p_i^{[k]}, p_j^{[k]}) \in I_{ij}} \| h(K_j R_{ij}[f_3 x_3 - \frac{f_3 (n^T)}{d - L}] K_i^{-1} p_i^{[k]} - p_j^{[k]} \|^2
\]

under the constraint \( ||n||=1 \), where \( h(a,b,c) = (a/c, b/c) \). The step ‘T’ tries to minimize, for each registered pair \((i,j)\), a function \( F_{ij} \):

\[
\]
where $\gamma$ is a triple of angles, and $t$ is the translation vector. The outputs of the steps 'T' are then used again in the following 'P' steps (as new values for $R_{ij}$ and $t_{ij}$) in the next iterations of this cycle. For each of these steps, we used a classical Levenberg-Marquardt (LM) approach.

Algorithm 2. Scene calibration.

1. $I_{ij} \leftarrow$ Apply algorithm 1 between all pairs $(i,j)$.
2. Decompose all $H_{ij}$'s into $R_{ij}, t_{ij}, n_{ij}, d_{ij}$ using [10].
3. Determine the scale of $t_{ij}, d_{ij}$ by using Eq. 8 on each blob in $I_{ij}$ and taking the median.
4. Initialize $n_i, d_i$ by averaging all $n_{ij}, d_{ij}$.

repeat
  for camera $C_i$ do
    Refinement LM steps on $n_i, d_i$ on the objective function given by Eq. 11.
    Relaxation scheme on $n_i$ (Eq. 10).
    Refinement LM steps on $R_{ij}, t_{ij}$ on the objective function given by Eq. 12.
  end

until convergence

Choose reference camera $i$ with the highest $\sum |h_{ij}|$, compute a reference frame on $\Pi$ and a homography $H_i$ from this frame to $C_i$, (Eq. 4).

For all cameras $j$ connected to $i$ by a path of homographies, compute the reference plane-to-image $H_{ij}$.

The scene calibration is summed up in algorithm 2, which makes use of the previously described non-linear optimization steps. Once these steps are done, we simply select the most promising camera $i$ in terms of inliers contributing to the estimations of $H_i$ and for all cameras $j$ that can be connected to $i$ by a path of homographies (e.g., $H_{ij}, H_{ik}, H_{kj}$), we deduce $H_j$ from Eq. 4 and the estimated relative transforms.

5 Results

We tested and compared our algorithm on the PETS [11] benchmark data. This database provides several datasets, with increasing levels of difficulty for the tracking algorithms, i.e., different levels of people density. Each dataset gives eight video streams which makes it suitable for our evaluation needs. For our calibration purposes, we used the medium density crowd dataset (labelled as S0).

As mentioned before, this work has been implemented entirely in C++ with the OpenCV library. It uses a classical tracking algorithm from OpenCV (color-based particle filter). Moreover, we use a somewhat more efficient variant of the RANSAC algorithm, LO-RANSAC [2] that locally optimizes the estimation of the model after each improvement of the current winner model.

We first give some qualitative results of camera-to-camera homography registrations $H_{ij}$ in Fig. 9. We depict the registrations obtained with three of the computed homographies $H_{ij}$ (left) and their inverse $H_{ij}^{-1}$ (right) by warping the image $j$ onto the image plane $i$, all of these without applying yet the optimization from Section 4. Ideally, if the homographies were correct, all the elements of the scene belonging to the plane $\Pi$ should coincide. One can note that most of the recognizable roads, lines, spots are correctly warped in the other view. Fig. 1 depicts the inliers $\textit{trajectories}$ (i.e., the matched data in the RANSAC process of Alg. 1) that allowed the estimation of the third homography in Fig. 9 (i.e., between views 2 and 7). In addition to these three examples of Fig. 9, the algorithm calibrates 15 of the possible 28 camera pairs. When pairs are not calibrated, it is mainly because of a too little overlap between the cameras or due to some large off-the-plane obstacles (e.g., trees) that make the track correspondences difficult to find.

Another qualitative result is depicted in Fig. 4: the output of Algorithm 2 for cameras 1, 2, 3, 4, 5, 6, and 7, i.e. computed homographies $H_1, H_2, H_3$...

The images at some timestamp $t$ from all these video streams $i$ are projected with the homographies $H_i$ onto the bird view over the real scene.
The field of view of each camera is materialized through yellow angular sectors (or red angular sectors if the registration is not done directly by one homography to the reference cameras, but by several ones). To compare it with the ground truth data given with the dataset, we manually superimposed an aerial view of the zone where all camera positions are indicated by blue numbers. Ideally, the vertices of the field of views should indicate the positions of the cameras. As it can be observed, the position errors are in the order of dozens of centimeters.

Quantitatively speaking, in Fig. 5 we compare our approach with different algorithms in the literature, as far as computational efficiency is concerned. To evaluate this efficiency, we plot the average numbers of iterations the RANSAC loop has to perform (y-axis) before reaching a given level of precision (x-axis), i.e., satisfying decreasing levels of quality. These numbers are averaged over 11 runs and presented in log scale for better visualization. In black, we plot the number of iterations necessary for the blob-based algorithm of [6], which is far above all the others, as it has been pointed out already, because of its intrinsic higher complexity. In red, we plot the standard RANSAC algorithm such as [1]. In green, we plot the result of [9] and in blue the one of our algorithm: the last three have a rather similar behavior, but ours gives systematically lower time requirements to reach a given precision bound because of (1) the non-uniform sampling scheme and (2) the pre-processing of trajectories into trajlets.

Conversely, in order to evaluate our algorithm in terms of precision and robustness, we compared the image-to-image homographies estimated by Alg. 1 to the ones computed through (a) a naive implementation of RANSAC (similar to [1]) and (b) a better implementation with non-uniform sampling (similar to [9]). These three schemes are evaluated for fixed numbers of iterations (x-axis) in Fig. 6, for the pairs 1-2 (left) and 2-7 (right). The results we obtained for the other pairs are similar. The comparison is made on per-pixel average re-projection errors, measured by the Euclidean distance between the projections by \( H_{ij} \) of points in images \( i \) onto images \( j \) and the projections of these same points with ground-truth homographies \( G_{ij} \). We took the mean error over 15 different runs. One can notice that the mean error for the first two schemes are quite high and unstable because of the presence of outlier trajectories in the RANSAC scheme that may be inserted in the estimation, whereas ours converges much more stably to its best value.

In Fig. 7, we depict two outputs of the non-linear optimization scheme of Section 4. In the left part, we plot the error residuals of functions \( F_{ij} \) (y-axis) that, as expected, tends to drop first and then remains stable with the number of steps (x-axis). In the right part, we plot the norm \( n_1 R_{12} n_2 \).
which, as we explained it previously, is extracted from two separate estimations, but theoretically should vanish. It can be observed that because of our relaxation step in Alg. 2, this geometrical consistency term $n_1 - R_{12} n_2$ tends to be minimized with the number of steps.

Finally, in Fig. 8 we present the distributions of the same reprojection error of Fig. 6, but for two versions of our algorithm, i.e., with (right) and without (left) smoothing of the trajectories. We which, as we explained it previously, is extracted show the results for pair (1,6) on the left, while on the right for pair (2,7). What can be shown is that the median (bold horizontal line) error is not necessarily better with smoothing, but there are much less outlier situations (unfilled dots), that is why we have chosen to smooth trajectories.

Fig. 6. Comparison of the reprojection errors (i.e., registration quality) obtained as a function of the number of iterations in the RANSAC scheme. The plots are mean values over 15 registration intents and are relative to two pairs of videos (1-2 on the left, 2-7 on the right). In red, results with the standard RANSAC ([1]), in green, with the algorithm of [9], and in blue, with our approach

Fig. 7. Two outputs of the non-linear optimization scheme applied to scene calibration, based on camera pairs 1-2, 1-6, 1-7, and 2-7, that makes the set of estimates of pairwise geometries evolve towards consistent values. Left, evolution of the value of the objective functions $F_{ij}$ on a per pixel basis. Right, evolution of the norm of the geometric coherence term $n_1 - R_{12} n_2$. 
6 Conclusions

We presented an algorithm for the calibration of a set of video surveillance cameras. It has several advantages over comparable algorithms in the literature as follows:

(1) by pre-processing trajectories into trajectories and by assigning likelihood values to pairs of them so that unambiguous matches are favored, it keeps the computational complexity reasonable;

(2) as we do not rely on entire trajectories, but instead on smaller parts less susceptible to be erroneous, it is much more robust to the occlusion problems accompanying standard 2D tracking algorithms;

(3) its non-linear optimization step incorporates all geometrical consistency constraints between the cameras, so that the final estimates between the pairs are not contradicting each other.

We presented rectification results in challenging situations where the viewpoint changes makes it nearly impossible to register the views by traditional point correspondences techniques, and we showed that our own approach outperforms existing ones when efficiency is concerned, i.e., the number of iterations needed for the calibration process to reach a given level of precision is always lower with our algorithm. Moreover, to our knowledge, there is no other work in the literature that goes further the pairwise camera calibration to perform the calibration of a whole network of cameras while taking the consistency of the computed estimates into account.

As ongoing and future work, we aim to model and compute probabilistic uncertainties on the estimated homographies based on the uncertainties on measured trajectories. We plan to use them

(1) in optimization, to favour less uncertain homographies and
(2) in a 3D people tracking scheme over the camera network.

References


Fig. 9. Video registrations on the PETS2009 [11] sequences. Each row depicts one pair \((i,j)\) and the corresponding homography \(H_{ij}\) (left) and its inverse (right). The processed pairs are, from up to down, \((1,2)\), \((1,3)\), and \((2,7)\).


Guillermo Baqueiro graduated in Mathematics from the Autonomous University of Yucatan in 2007. He got his Master degree in Computer Science at CIMAT in 2010 and joined Digipro SA de CV in the same year.

Jean-Bernard Hayet graduated from ENSTA (Paris), University Paris VI and got his Ph.D. from U. of Toulouse (2003). He has been post-doc fellow at U. of Liege. Since 2007, he has worked at CIMAT in Guanajuato, Mexico. He teaches computer science and does research in robotics and computer vision.

*Article received on 11/03/2010; accepted on 04/02/2011.*
Chromatic Correction Applied to Outdoor Images

Hayde Peregrina-Barreto¹, J. Gabriel Aviña-Cervantes¹, Iván R. Terol-Villalobos², José J. Rangel-Magdaleno¹, and Ana M. Herrera-Navarro³

¹ Universidad de Guanajuato, Guanajuato, Mexico
² Centro de Investigación y Desarrollo Tecnológico en Electroquímica, Querétaro, Mexico
³ Universidad Autónoma de Querétaro, Querétaro, Mexico

hperegrina@ieee.org, avina@ugto.mx, famter@ciateq.net.mx, jjrangel@ieee.org, anaherreranavarro@gmail.com

Abstract. The color of an image may be affected by many factors such as illumination, complex and multi-spectral reflections, and even the acquisition device. Especially in outdoor scenes, these conditions cannot be controlled. In order to use the information of an image, the latter must present the information as closer as possible to the original scene. Sometimes images are affected by a dominant color (cast) that changes its chromatic information. In order to avoid this effect, a color correction must be done. In this work, a novel method for correcting the color of outdoor images is proposed. This method consists in a complete improvement process of three steps: cast detection, color correction, and color improvement.

Keywords. Cast detection, color correction, chromatic adaptation, natural outdoor images, color enhancement.

1 Introduction

Information analysis through images is widely used nowadays. Sometimes images are acquired under light controlled conditions, but it is not always possible. When images are taken in outdoor environments, acquisition conditions such as illumination, light scattering, surrounding reflections, etc. cannot be controlled. All these factors, including characteristics of capturing devices, may affect the quality of images [2, 19, 29]. Taking these factors into account, one of the most affected properties of an image is color, which is a useful characteristic as it provides additional information related to an object and permits to distinguish among objects with the same physical characteristics (size, shape, etc.). In image processing, color is used as a discrimination parameter on some low level tasks such as segmentation [3, 12], object tracking and robotics navigation [33], etc. Some processes are designed to associate a particular color with some kind of information (e.g., a green region associated with vegetation), However, if an image is taken under an illuminant which generates a reddish appearance, it may provoke a wrong
association or conduct to an inaccurate decision (e.g., move on or stop in robot navigation).

The human visual system has an ability to maintain the chromatic features of an object relatively constant despite illumination changes. It means that even when there is a slight difference in the observed color in the scene, the chromatic concept is maintained [25]. This visual mechanism is called chromatic adaptation and permits to recognize the same object or scene when lighting condition changes. Some acquisition devices emulate this mechanism through sensor calibration, yet it is necessary to adapt the sensor to each lighting change. Another solution consists in correcting an image after its acquisition, i.e., by image processing (color balance). Color correction improves not only the appearance of the image but it may also be used as a pre-processing step in order to get a balanced color image for further processing. Color balance is the global adjustment of the intensities of the primary colors (red, green, and blue). An important goal of this adjustment is to correctly render specific colors, particularly neutral colors (gray to white nuances).

This paper is focused on improvement of images which have been affected by an illuminant color. The proposed correction involves three steps: detection of the dominant color, chromatic correction, and color enhancement. These processes take into account the information provided by the image itself; thus, the correction is made to an accurate extent according to the original image characteristics. Section 2 describes related work available in the literature. The basic concepts and comparative methods used in this work are discussed in Section 3. Section 4 presents the proposed methodology and describes each step individually. Some experimental results are given in Section 5. Finally, Section 6 presents our conclusions.

2 State of the Art

Modern chromatic adaptation methods are usually based, conceptually and mathematically, on the Von Kries assumption. It asserts that color constancy can be achieved if the three cone signals are regulated through their respective gain coefficients [16]. Then any change in adaptation conditions is translated into a simple recalculation of the sensibility spectral curves of the fundamental mechanism expressed as

$$\begin{pmatrix} L' \\ M' \\ S' \end{pmatrix} = \begin{pmatrix} k_L & 0 & 0 \\ 0 & k_M & 0 \\ 0 & 0 & k_S \end{pmatrix} \begin{pmatrix} L \\ M \\ S \end{pmatrix}$$ (1)

where $L$, $M$, and $S$ represent the initial cone response; $k_L$, $k_M$, and $k_S$ are the coefficients used to recalculate the initial responses; $L'$, $M'$, and $S'$ are the resulting responses after adaptation [16]. With this hypothesis, it is possible to obtain a controlled gain since each coefficient has an individual gain for its respective cone; each cone response is considered independent from the others. In order to make a chromatic correction, it is necessary to discount the effects of the illuminant [26, 38]. Then, the coefficients are calculated to be the cone inverse responses and they are estimated as the ratio between the maximum response under the reference illuminant and the response to the current one. The reference illuminant is usually the D65 CIE (Comission Internationale de l’Eclairage) standard illuminant due to its similarity to the noon daylight conditions in open-air. For practical purposes, the R, G, and B values are used hereinafter instead of the $L'$, $M'$, and $S'$ values.

Several strategies have been proposed in the literature for color correction. One of the most known algorithms is the gray-world algorithm which assumes that the spatial average of the surface reflectance is achromatic [19, 22]. It means that, when there is a dominant color in the light and this impinges on an achromatic surface, the surface uniformly changes the intensity of its component [8]. A luminance value, relatively uniform and closer to the daylight, is obtained when the intensities of each chromatic component are averaged [28]. Thus, this algorithm recovers an estimate of the original spectral distribution of the illuminant and the surface reflectance [17]. In order to make a color correction, the gray-world algorithm works together with the Von Kries method. First, the average of each RGB channel
is obtained \((\mu_R, \mu_G, \mu_B)\) in order to determine the intensity with which these were affected and the minimum of them \((\mu_{\text{min}})\) which represents the less dominant channel and is the reference to attenuate the other channels. The coefficients are calculated with these parameters \(\mu_R, \mu_G, \mu_B, \text{ and } \mu_{\text{min}}\) as shown in (2). Color correction is reached by using these coefficients in (1).

Retinex was the first attempt to develop a computational model in order to emulate the color constancy process of human vision [30]. This theory is considered as an improvement to the Von Kries hypothesis. Retinex improves the visual representation of images when light conditions are not good [34, 36] and is based on the biological mechanism of the human eye for chromatic adaptation. This method combines color constancy with contrast and local luminance of each pixel in order to approach the real appearance of a scene; it also gives a better appreciation of dark regions [7]. The algorithm calculates the luminance of a point \(x_p (L_{x_p})\) influenced by \(N\) points \((x_i)\) chosen randomly (4). There are many Retinex versions like the Brightness-based Retinex or the Change-based Retinex [10, 11]; in this work the Retinex version of Frankle-McCann [20] is used.

\[
\begin{align*}
    k_R &= \frac{\mu_{\text{min}}}{\mu_R} \\
    k_G &= \frac{\mu_{\text{min}}}{\mu_G} \\
    k_B &= \frac{\mu_{\text{min}}}{\mu_B} \\
    k_R &= \frac{1}{k_N} \\
    k_G &= \frac{1}{k_{iW}} \\
    k_B &= \frac{1}{k_{iW}}
\end{align*}
\]
Finlayson et al. [19] proposed a use of histogram equalization for color constancy. This method consists in histogram transformation which distributes the values all the range long in a more uniform way and is applied to each color channel. Other proposals associate one or more correction methods to an image according to its characteristics in order to achieve the best result [4, 13, 23, 27].

There is a big difference between a cast and a wide homogenous area. A cast affects the image in a global uniform manner and it is more evident in lighter regions. For example, Fig. 1(b) shows an image where one can observe that there is a bluish layer over the entire image, which is affecting the other colors. That layer is the cast and it is better reflected on the lighter regions such as the principal building. The observer understands that the building is white; however, it has a bluish appearance that must be corrected.

The difference between these two conditions - cast and self-cast - must be taken into account in image analysis in order to arrive to the conviction that color correction is necessary. In this work, cast detection based on the image information analysis is proposed.

3.1 Step 1: Cast Detection by Mean RGB Distances

The first step in color correction is to detect whether color correction can be made and whether a cast is present. As it was previously mentioned, the light regions of an image reflect the cast existence better and also provide information to determine if correction is possible. In order to make a correction, an image must have certain quantity of pixels with enough luminance. In a study concerning cast detection, Gasparini and Schettini [21] suggested to evaluate the luminance values (L*) in the range 30>L*<95; if at least 20% of pixels are within this interval then color correction is possible. For example, Fig. 1(c) shows a blue dominant color, evident on a lighter region; nevertheless, the majority of its pixels have a low luminance level that does not give enough information about the illuminant for an optimal correction. With the aim to analyze the luminance, the L*a*b* color space was used. To transform RGB values to L*a*b* values, a middle step is necessary. First, RGB values must be translated to XYZ tristimulus values by using the transformation matrix in (5). Later, L*a*b* values are obtained as in (6), where $X_n$=95.047, $Y_n$=100.0, and $Z_n$=108.883 according to the illuminant CIE D65, L* is the luminance channel, and a*, b* are the chromatic channels [16, 24].

$$L_{sp} = \frac{1}{N} \sum_{i=1}^{N} \left( \log(I(x_p)) - \log(I(x_i)) \right)$$  \hspace{1cm} (4)

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} 0.4124 & 0.3576 & 0.1805 \\ 0.2126 & 0.7152 & 0.0722 \\ 0.0193 & 0.1192 & 0.9505 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix}$$  \hspace{1cm} (5)

$$L^* = 116f(Y/Y_n - 16)$$
$$a^* = 500[f(X/X_n) - f(Y/Y_n)]$$
$$b^* = 200[f(Y/Y_n) - f(Z/Z_n)]$$  \hspace{1cm} (6)

If the Gasparini criterion is fulfilled, then the image is analyzed by the proposed cast detector with the aim to determine if color correction is necessary. Since the luminance values change from one image to another, the proposed cast detector is based dynamically on the individual information of an image.

This detector consists in evaluating a cast factor (factor) which indicates whether a cast is present. The factor is calculated by using the RGB values from the suspicious cast color and consists basically in determining if this cast color is homogeneously distributed on the image. Thus, the cast color is surely determined if this color is uniformly distributed on the image which is mathematically computed by the expected values of this color on the image. The expected values ($\mu X$ and $\mu Y$) are interpreted as the centroid coordinates of the cast color. If these coordinates are close to the center of the image, the cast color is detected (the suspicious color was present entirely in the image); on the contrary, if the coordinates are in another position, it means that...
the object contains enough representative color (self-cast) to alter the basic cast detector. The factor parameter is computed by the following equation:

\[
\text{factor} = \sqrt{\frac{(H_X - \frac{W}{2})^2 + (H_Y - \frac{H}{2})^2}{(H_X - \frac{W}{2})^2 + (H_Y - \frac{H}{2})^2}}
\] (7)

where \(X, Y\) are the image axis, and \(W, H\) are the image width and height, respectively. According to several experimental results derived from outdoor images, if the factor \(\geq 0.07\), the image has a cast color, and a self-cast color otherwise. Besides the analysis of luminance, the color distribution of an image (channels \(a^*b^*\)) gives information about the cast if they are represented on a bidimensional histogram, hereinafter referred to as AB histogram. Consider the case of Fig. 1(a) and its AB histogram in Fig. 2(a). It can be observed that there is a wide and distributed peak and that the values are far from the neutral axis \((0, 0)\) which indicates that the colors are intense.

Texture and close-up images have these characteristics. When there is a cast, the histogram presents two characteristics: narrow peaks and a far distribution of the neutral axis; it means that the farther the distribution is, the more intense the cast is. Figure 1(b) and its AB histogram (Fig. 2(b)) show an example of the cast condition. Thus, by analyzing the AB histogram of an image, it is possible to detect a cast and its presence can be confirmed by calculating the factor value.

3.2 Step 2: Chromatic Correction by Neutral Values

The lightest values of an image not only give information about the cast but sometimes are used for color correction (as in the white patch algorithm). However, lighter regions do not always represent information of the original scene because some processes can change their values. It is common that acquisition devices make an automatic correction of the image when it is taken; this correction is called white balance.
and it forces the darkest points to become black and the lightest points to become white. If a process affects the information of the lightest regions, color correction based only on them may result inaccurate.

A chromatic correction method using the highest and lowest values of luminance is proposed in this work. This method is based on the assumption that if the highest and lowest values of luminance are considered in order to find the accurate value (neutral value) for color correction, which works in a similar way as the white of reference for the white patch hypothesis and the RGB average values for the gray-world hypothesis, then it is possible to reach balanced color in an image. This process works over the luminance of an image by combining a percentage of the lightest and darkest values with the aim to find the most accurate percentages.

We used a database of 40 outdoor images of texture, close-up, and with a cast for testing different percentages.

\[
\text{neutral} = [R_n, G_n, B_n] = \frac{1}{N + M} \left( \sum_{i=1}^{N} L_i + \sum_{j=1}^{M} L_j \right) \tag{8}
\]

\[
k_R = \frac{\max_{\text{neutral}}}{R_n}, \quad k_G = \frac{\max_{\text{neutral}}}{G_n}, \quad k_B = \frac{\max_{\text{neutral}}}{B_n} \tag{9}
\]

More satisfactory results were obtained by taking those pixels within the lighter 20% \(L^+\) and those within the darkest 12% \(L^-\) of the luminance range of the image. For example, if the luminance range of an image is [0-255], then \(L^+\) contains all indexes of the pixels that 204\(\leq L_{i}\)\(\leq 255\) and \(L^-\) contains all the indexes of the pixels that 0\(\leq L_{i}\)\(\leq 31\). Finally, to obtain the neutral value, we calculate the average of RGB values of the indexes contained in both sets \(L^+\) and \(L^-\) as shown in (8), where \(N\) and \(M\) are the number of indexes in \(L^+\) and \(L^-\), respectively, and the neutral contains the RGB values used for color correction. Based on the RGB values of the neutral, the coefficients of correction are calculated as it is shown in (9), where \(\max_{\text{neutral}}\) is the maximum value of \(R_n, G_n\) and \(B_n\).

### 3.3 Step 3: Color Enhancement by Luminance

Sometimes after a chromatic correction is made, the image presents grayish colors induced by this procedure. For example, in color segmentation processes, it is more difficult to distinguish one region from another if there are grayish tones, and since this application is focused on image improvement, color enhancement is an important stage.

Nevertheless, color enhancement is not an easy task because false colors may be created thus changing the natural appearance of an image. Color is a delicate feature and a uniform distribution is not always enough for its enhancement. Color enhancement by luminance, which takes into account the individual luminance of each pixel, is proposed as a complement step for chromatic correction. This method consists in increasing color saturation without changing the hue. In a first iteration, color quotients are defined, dividing each RGB channel by its respective luminance, see equation (10).

\[
\alpha(i) = \frac{1}{k} \frac{C(i)^n}{L(i)} \tag{10}
\]

where \(C \in \{R, G, B\}\), \(L\) is the luminance, and \(k\) is a scale factor defined by \(k=\max\{C(i)^n/L(i)\}\) to avoid data overflow. These quotients modulate the original RGB values and \(n\) can be adjusted to obtain different levels of color saturation, in our experiments \(n=2\) is used. In a second iteration, more saturated and enhanced colors \(C_{enh}(i)\) are obtained by (11). This set of equations works in a similar way as gamma color correction.

\[
C_{enh}(i) = \alpha(i)C(i) \tag{11}
\]
4 Experimental Results

Figure 3(a) shows a wide homogenous area (flower), and although it has an intense color because its AB histogram (Fig. 3(b)) is far from the center, the analysis determines a factor=0.01, which indicates that the image is not affected by any cast but it could be a texture or a close up image, therefore, no chromatic correction method is applied. The Gasparini analysis is used as a comparison method and in this case it also determines the presence of a self-cast.

Figure 4(a) demonstrates a global blue cast, its AB histogram shows the values located far from the neutral axis (Fig. 4(b)), and it has a \textit{factor}=0.08.

\textbf{Fig. 3}. (a) A self-cast image and (b) its AB histogram

\textbf{Fig. 4}. (a) Image affected by a blue cast, (b) its AB histogram, and (c) its RGB histogram
Fig. 5. Chromatic correction of Fig. 3(a) by using (a) histogram equalization, (c) gray world, (e) white patch, (g) Retinex and (i) neutral values (b, d, f, h, j) with their respective RGB histogram.

Fig. 6. (a) Second correction by neutral values and (b) its color enhancement by luminance.
Statistics show that the distance to the neutral axis is big enough and the cast is intense; both detectors agree to apply a correction. Bluish appearance is explained by the fact that the B channel is affecting the lightest pixels and its peak must be attenuated in order to balance the color (Fig. 4(c)).

Figure 5(a) shows a correction by histogram equalization where the blue cast was eliminated; however, the result presents new colors, which could be observed in its RGB histogram (Fig. 5(b)), and this is not accurate because it implies a change to the original chromatic information. The GW correction has attenuated the blue cast (Fig. 5(c)) and its result looks more natural. This correction often presents grayish colors and low saturation which is also observed in (Fig. 5(d)). When an image has a good level of brightness, the WP method may not present strong changes because the correction is made based on the lightest region information. In this case, WP increases the cast intensity instead of attenuating it (Fig. 5(e)(f)). Retinex has a good correction by attenuating the cast and improving the brightness (Fig. 5(g)). Note that, unlike the previous method, Retinex increases the R and G channels on lightest region instead of attenuating the B channel (Fig. 5(h)), and although the

Fig. 7. (a) Original image, (b) its chromatic correction by neutral values with color enhancement by luminance, and (c, d) their respective RGB histograms
improvement is notorious, some regions have been vanished. This method also causes a halo effect evident on the tower and the white building. The neutral values method provides an adequate result since the blue peak on the lighter region was attenuated (Fig. 5(i)). The R and G channels were increased but their distribution was maintained.

Sometimes, even when the cast is attenuated, the resulting image may still have a weak cast. For this reason, the resulting image is analyzed again and if necessary, a new correction is applied. Again, the coefficients are computed for a new correction according to several tests; significant changes are reached only after the third correction. Figure 4(a) was computed two times by the neutral values method because the image presents a deep and very intense cast. The first correction (Fig. 5(i)) still conserves a little blue cast which is removed in the second correction (Fig. 6(a)). After the chromatic correction, color enhancement by luminance is applied owing to the grayish appearance resulting from the cast attenuation. Observe that the color is better balanced and has a good saturation level, and although there is a wide blue region, it belongs to the natural appearance of the sky and the river but not to the cast (Fig. 6(b)). Another result is demonstrated in Fig. 7 which also combines chromatic correction and color improvement with the methods proposed in this work.

5 Conclusions

The elimination of dominant colors in images not only permits visual improvement, but also provides better results for further processing. Outdoor images are more complicated because illumination conditions cannot be controlled. In this work, a chromatic correction method posterior to image capture was proposed. Before processing the image, an analysis with the objective to detect a cast is made in order to determine if chromatic correction is necessary and possible. Additionally, the proposed cast detector was compared with the detector suggested by Gasparini and both match in the casting detection 94% of the times. Once the cast is detected, a chromatic correction is made based on the image information and its neutral value which permits to have a result more similar to the individual condition of each image. The proposed method is based on simple operations which permit a short processing time. Results were compared, in a qualitative way, with some of the most popular methods for chromatic correction making a best comprehension of its operation, results, and efficiency possible. With the aim of complementing the improvement process, a color enhancement method using luminance as a modulator element was also proposed. Color enhancement is important because processes, as well as color correction, may generate low saturation on the resulting image. Finally, it was demonstrated that all three steps together - cast detection, color correction, and color enhancement - can improve the quality of an image in a complete way.

References


enhancement. *Journal of Electronic Imaging*, 13(1), 100–110.


---

**Hayde Peregrina-Barreto** received her B.Sc. from the Technological Institute of Cuautla, Mexico, her M.Sc. in Engineering from the University of Guanajuato (FIMEE), Mexico, and her Ph.D. in Engineering from the Autonomous University of Queretaro, Mexico. Her current research interests in image processing include mathematical morphology, color appearance models, segmentation, and human visual perception.

**J. Gabriel Aviña-Cervantes** received his B.Sc. and M.Sc. degrees from the University of Guanajuato, and his Ph.D. in Informatics and Telecommunications from the Institut National Polytechnique de Toulouse, France. His research interests include artificial vision for outdoor robotics, pattern recognition, object classification, and image processing. He is currently a researcher at the Engineering Division of the Campus Irapuato-Salamanca.

**Iván R. Terol-Villalobos** received his B.Sc. from the National Polytechnic Institute (IPN), Mexico, his M.Sc. in Electrical Engineering from the Research Center for Advanced Studies of IPN, Mexico, and a DEA in Computer Science from the University of Paris VI, France. He is currently a researcher at CIDETEQ, Mexico. His research interests include morphological image processing, morphological probabilistic models, and computer vision.

**José J. Rangel-Magdaleno** received his B.Sc. and M.Sc. in Engineering from the University of Guanajuato (FIMEE), Mexico, and his Ph.D. in Engineering from the Autonomous University of Queretaro, Mexico. His current research interests include signal and image processing with FPGA, vibration monitoring, and vibration analysis.

**Ana M. Herrera-Navarro** received her B.Sc. from the Technological Institute of Queretaro, Mexico, her M.Sc. in Engineering from the Autonomous University of Queretaro, Mexico, and she is currently a Ph.D. student at the Autonomous University of Queretaro. Her research interests include mathematical morphology and material characterization.

---

*Article received on 10/03/2010; accepted on 17/12/2010.*
Índice de contraste morfológico basado en el análisis de los contornos y el fondo de la imagen

Angélica R. Jiménez Sánchez, Jorge D. Mendiola Santibañez, Gilberto Herrera Ruíz e Israel Santillan

1 División de Investigación y Posgrado de la Facultad de Ingeniería, Universidad Autónoma de Querétaro, Centro Universitario S/N, C.P. 76010, Querétaro, México

2 Universidad Politécnica de Querétaro, Carretera Estatal 420 S/N, El Rosario, El Marqués, C.P. 76240, Querétaro, México

{rosariojs, mendijor, gherrera}@uaq.mx, santilis@gmail.com

Resumen. En este artículo se presenta un índice para cuantificar el contraste que se percibe en una imagen. El índice está basado en la ley de Weber y toma en consideración una estimación del fondo de la imagen mediante la erosión de la apertura por reconstrucción. El desempeño de la propuesta se ilustra con un conjunto de imágenes procesadas por un mapeo de contraste y se compara con dos medidas de contraste dadas en la literatura.

Palabras clave. Contraste morfológico, apertura por reconstrucción, índice de contraste, ley de Weber.

Morphological Contrast Index based on an Analysis of Contours and Image Background

Abstract. In this work, a contrast index for quantifying the perceived contrast in an image is proposed. The index is based on Weber’s law and takes into account background estimation through the erosion of opening by reconstruction. The performance of our proposal is illustrated with a set of images processed by contrast mapping and compared with two contrast measures given in the literature.

Keywords. Morphological contrast, opening by reconstruction, contrast index, Weber’s law.

1 Introducción

El contraste se define como la diferencia de intensidad luminosa entre un punto de una imagen y sus alrededores. Bajo esta definición el contraste que se percibe en una imagen con tonalidades similares es menor que el que se percibe en una imagen con tonalidades diferentes. En muchos campos es importante contar con un índice del contraste, por ejemplo para medir la calidad de una imagen o bien para indicar cuando es necesario corregirlo. La medida del contraste involucra directamente el concepto de luminancia como lo indican las leyes de Weber y Michelson [5]. Estas leyes físicas son utilizadas para modelar el contraste. Por definición, la luminancia de una superficie es la intensidad luminosa emitida por unidad de superficie en una dirección dada y se mide en candelas por metro cuadrado (cd/m²). Sin embargo, en el procesamiento de imágenes y en particular en imágenes en niveles de gris, la luminancia se asocia con el nivel de gris que corresponde a cada pixel, por lo que cambios importantes de la luminancia se encontrarán alrededor de los contornos de la imagen, mientras que en los espacios de color existen ecuaciones específicas para el cálculo de la luminancia que en este trabajo no revisamos.
Por otro lado, medir la mejora de una imagen después de que esta ha sido procesada por algún operador no es una tarea fácil, ya que esto es algo subjetivo y depende de la aplicación. En la práctica, existen varias definiciones que permiten medir el realce o mejora en las imágenes, como ejemplo ver: [1, 2, 3, 7, 11, 12, 14, 15].

En [7] se menciona que no existe una medida universal que pueda validar tanto la parte objetiva y subjetiva del mejoramiento de la imagen, de aquí parte el surgimiento de diversas propuestas. Por ejemplo, en [3] la medida se obtiene a través del promedio de los niveles de gris en dos ventanas rectangulares centradas en cierto píxel. En [2] se realiza la medida a partir del análisis de los contornos de la imagen. También existe una medida estadística propuesta en [12] en la cual se obtiene un parámetro a partir del histograma de la imagen. Otro trabajo interesante se presenta en [1] en el cual se analizan las intensidades máximas y mínimas dentro de una ventana que se desplaza a través de la imagen. Por otra parte, el contraste ha sido tratado de manera sistemática dentro del campo de la Morfología Matemática (MM), ver por ejemplo, mapeos de contraste [9, 19, 26, 27], filtros morfológicos por pendiente [24, 26], el centro morfológico [18, 23], y el top-hat. Sin embargo, existen pocos trabajos dentro de la MM en donde se cuantifica el mejoramiento de una imagen procesada. En [11] se introduce un índice de medida de contraste que involucra a un operador que trabaja de manera similar al laplaciano, no obstante esta propuesta tiene la desventaja de no considerar el fondo de la imagen dentro de la medida. Partiendo de este punto y de que en [6] se propuso un operador que permite obtener el fondo de la imagen a diferentes escalas (multiback-ground en inglés); en la presente investigación se introduce un índice de contraste que indica cual de las imágenes procesadas por cierto operador presenta el mejor contraste visual. Dicho operador basa su funcionamiento en la ley de Weber y considera el fondo de la imagen. El índice propuesto utiliza solamente operadores morfológicos, además de ser simple de implementar y de aplicar.

Este artículo está organizado de la siguiente manera: en la sección 2 se presenta una breve descripción acerca de algunas transformaciones morfológicas, la ley de Weber y un mapeo de contraste. En la sección 3 se muestra la aproximación al fondo de la imagen por medio de la erosión de la apertura por reconstrucción. En la sección 4 se introduce el índice para cuantificar el contraste. En la sección 5 se presenta un ejemplo donde se calcula el índice propuesto para un conjunto de imágenes y se realiza la comparación con otras medidas reportadas en la literatura. La sección 6 corresponde a las conclusiones.

2 Transformaciones morfológicas y ley de Weber

La MM es una metodología de procesamiento de imágenes que surge a finales de los años 60 en Francia [17]. La descripción básica de la MM se apoya en la teoría de conjuntos, es decir, toda operación morfológica es el resultado de una o más operaciones de conjuntos (unión, intersección, complementación, etc.). Las transformaciones en MM utilizan un conjunto geométrico conocido como elemento de estructura, el cual posee forma, tamaño definido a priori y un origen. El elemento de estructura se traslada sobre la imagen bajo estudio con la finalidad de determinar el conjunto de puntos, respecto al origen del elemento de estructura, que intervendrán en cada operación que se efectúe con las componentes de la imagen. La forma que adquiere el elemento de estructura puede ser variada, por ejemplo, en el caso bidimensional, puede ser un disco, un cuadrado, una línea recta, entre otros. En el caso tridimensional se definen diferentes clases de poliedros, por ejemplo, el cubo, el prisma hexagonal, etc. La talla del elemento de estructura se denota como \( \mu_0 \lambda \), los cuales son un factor de tamaño que determina la dimensión de la estructura geométrica de \( B \). En este artículo se usa un elemento de estructura cuadrado \( B \) de \( 3 \times 3 \) píxeles, el cual contiene su origen en el centro, es decir, un elemento de estructura simétrico. También se define \( \tilde{B} = \{ -x : x \in B \} \) como el conjunto transpose de \( B \) respecto a su origen. A continuación se presentan algunas definiciones de ciertas transformaciones morfológicas extendidas a niveles de gris.
2.1 Operadores morfológicos básicos: erosión y dilatación

La erosión morfológica está dada por el conjunto de los orígenes de $B$ siempre que $B$ está incluido en el conjunto $X$. Cuando esto no ocurre el resultado de la erosión es el conjunto vacío. La definición anterior puede extenderse directamente al caso de imágenes en escala de grises. La erosión de una imagen $f$ por un elemento de estructura $B$ se denota por $\varepsilon_{\mu B}(f(x))$ y se define como:

$$\varepsilon_{\mu B}(f(x)) = \{ f(x) : x \in \mu B_x \}$$ (1)

Donde, $\wedge$ es el operador ínfimo, que en el caso práctico corresponde al elemento mínimo del conjunto analizado. La dilatación $\delta_{\mu B}(f(x))$ es la operación dual de la erosión $\varepsilon_{\mu B}(f(x))$, es decir:

$$\delta_{\mu B}(f(x)) = (\varepsilon_{\mu B} (f^c)(x))^C$$ (2)

Donde $c$ es el operador complemento. Las transformaciones morfológicas que cumplen con las propiedades de ser crecientes e idempotentes también se conocen como filtros morfológicos [4, 17, 23]. Los filtros morfológicos básicos son la apertura $\gamma_{\mu B}(f(x))$ y el cierre $\varphi_{\mu B}(f(x))$ usando un elemento de estructura dado. Formalmente, la apertura $\gamma_{\mu B}(f(x))$ y el cierre $\varphi_{\mu B}(f(x))$ morfológicos se expresan como sigue:

$$\gamma_{\mu B}(f(x)) = \delta_{\mu B} (\varepsilon_{\mu B}(f))(x)$$ (3)

$$\varphi_{\mu B}(f(x)) = \varepsilon_{\mu B} (\delta_{\mu B}(f))(x)$$ (4)

2.2 Apertura y cierre por reconstrucción

Para algunas aplicaciones de análisis de imágenes es conveniente restringir el campo de acción de una transformación a ciertas regiones de interés. Esta idea dio origen a una nueva clase de transformaciones conocidas como geodésicas.

A diferencia de las transformaciones morfológicas que actúan sobre toda la imagen, las geodésicas sólo actúan sobre alguna parte de la imagen, es decir, sobre un subconjunto denominado máscara geodésica. Las transformaciones por reconstrucción se obtienen a partir de las transformaciones geodésicas, y son filtros que permiten modificar los mínimos y máximos de la imagen sin cambiar considerablemente la estructura de las demás componentes [21, 22, 28].

La dilatación geodésica $\delta^1_{\mu f}(g)$ y la erosión geodésica $\varepsilon^1_{\mu f}(g)$ de tamaño uno están definidas como $\delta^1_{\mu f}(g) = f(x) \wedge \delta(g)$ con $g(x) \leq f(x)$ y $\varepsilon^1_{\mu f}(g) = f(x) \vee \varepsilon(g)$ con $g(x) \geq f(x)$, respectivamente. Cuando la función $g(x) = \varepsilon(f)$ o $g(x) = \delta(f)$, se obtiene la apertura $\gamma_{\mu B}(f(x))$ (o cierre $\varphi_{\mu B}(f(x))$) por reconstrucción, es decir:

$$\gamma_{\mu B}(f(x)) = \lim_{n \to \infty} \delta^n_{\mu f}(\varepsilon_{\mu B}(f(x)))$$

$$\varphi_{\mu B}(f(x)) = \lim_{n \to \infty} \varepsilon^n_{\mu f}(\delta_{\mu B}(f(x)))$$ (5)

2.3 El gradiente

El gradiente es una medida de cambio en una función, y una imagen puede considerarse como un arreglo de muestras de alguna función continua dada en términos de los niveles de intensidad de la imagen Por analogía, los cambios significativos en los valores de gris en una imagen pueden ser detectados mediante una aproximación discreta del gradiente. A continuación se presenta la definición del gradiente morfológico [16]:

$$\text{gradm}_{\mu B} f(x) = \delta_{\mu B}(f(x)) - \varepsilon_{\mu B}(f(x))$$ (6)
2.4 Mapeo de contraste

Un mapeo de contraste selecciona para cada punto de la imagen un valor de nivel de gris entre los diferentes patrones (primitivas) utilizados de acuerdo a un criterio de proximidad. A continuación se presenta el criterio de proximidad $\rho(x)$ ($\rho(x) \in [0,1]$) junto con un mapeo de contraste de tres estados [9, 19]:

$$\rho(x) = \frac{\varphi_{\mu_1}(f)(x) - f(x)}{\varphi_{\mu_2}(f)(x) - \gamma_{\nu_2}(f)(x)}$$  \hspace{1cm} (7)

el mapeo está definido como:

$$w_{\mu_1,\mu_2,\alpha,\beta}(f)(x) = \begin{cases} 
\varphi_{\mu_1}(f)(x) & 0 \leq \rho(x) < \alpha \\
 f(x) & \alpha \leq \rho(x) < \beta \\
\gamma_{\mu_2}(f)(x) & \beta \leq \rho(x) \leq 1
\end{cases} $$  \hspace{1cm} (8)

Cuando el nivel de gris del cierre es igual al de la imagen original el valor del criterio es igual a 0, mientras que cuando el valor del nivel de gris de la apertura es igual al de la imagen original, el criterio tiene valor 1. En la ecuación 8 note que $\mu_1$ y $\mu_2$ representan el elemento de estructura para la apertura y el cierre respectivamente, mientras que $\alpha$ y $\beta$ definen el intervalo $[\alpha, \beta]$ para el mapeo de contraste. El problema a resolver en algunos trabajos recientes consiste en encontrar los valores adecuados de $\alpha$, $\beta$ y de $\mu_1, \mu_2$ tales que la imagen de salida presente un mejor contraste [9].

2.5 Ley de Weber

La ley de Weber es una ley bien conocida relacionada directamente con el estudio de la percepción [5]. Ernst Heinrich Weber formuló esta ley que establece lo siguiente: “El incremento en la intensidad del estímulo necesario para provocar un cambio en la sensación es proporcional a la intensidad del estímulo inicial”.

Matemáticamente la ley de Weber se representa como sigue: $dp = k \frac{dS}{S}$

donde $dp$ corresponde al cambio percibido en el estímulo $S$, $dS$ corresponde al cambio de magnitud del estímulo y $k$ es una constante.

Integrando la ecuación anterior se obtiene que: $p = k \log S + Cte$

El parámetro $Cte$ se obtiene cuando $p=0$ y $S=S_0$ en la ecuación anterior, con $S_0$ el nivel de estímulo por debajo del cual no se percibe sensación, es decir $Cte = -k \log S_0$:

por lo tanto,

$$p = k \log S - k \log S_0 k = \log (\frac{S}{S_0})$$  \hspace{1cm} (9)

La ley de Weber se utiliza en estudios psicovisuales para medir el contraste percibido $C$. El contraste $C$ de un objeto con luminancia $L_{\text{max}}$ y luminancia de sus alrededores $L_{\text{min}}$ es definida como sigue (Peli, 1990): $\Delta C = \frac{L_{\text{max}} - L_{\text{min}}}{L_{\text{min}}}$

Si $L = L_{\text{min}}$ y $\Delta L = L_{\text{max}} - L_{\text{min}}$, la ecuación anterior puede rescribirse como:

$$\Delta C = \frac{\Delta L}{L}$$  \hspace{1cm} (10)

Aplicando la ley de Weber, el contraste percibido $C$ puede expresarse como sigue [5]:

$$C = k \log L + b , \quad L > 0$$  \hspace{1cm} (11)

Donde $k$ y $b$ son constantes, siendo $b$ el fondo de la imagen. Con base en lo anterior, es importante notar que se requiere del parámetro $b$ para dar una estimación más precisa del contraste.
Fondo de la imagen

En [6] se hace una propuesta para obtener una aproximación al fondo de la imagen utilizando transformaciones morfológicas. Las curvas generadas por tal operador tienen la característica de tocar los mínimos regionales, además de poder controlar la profundidad con el parámetro de tamaño $\lambda$. La ecuación para calcular el fondo de la imagen es:

$$b = e_1[\overline{\gamma}(f)]$$

(12)

Donde $b$ representa el fondo de la imagen, $e_1$ es el erosionado tamaño 1 de la apertura por reconstrucción $\overline{\gamma}(f)$ tamaño $\lambda$ y $f$ representa a la imagen de entrada. En la figura 1 se ilustra la idea de detección del fondo de la imagen usando la ecuación 12. Sustituyendo la ecuación 12 en la 11 se obtiene:

$$C[f(x)] = k \log L + e_1[\overline{\gamma}(f)](x), \quad L > 0$$

(13)

Debido a que el valor máximo de intensidad en las imágenes procesadas en este trabajo es 255, se propone detectar los valores del contraste de la imagen procesada en el intervalo $0 < C \leq 255$. De esta forma, el valor de $k$ en la ecuación 13 se obtendrá de la siguiente manera:

$$k[f(x)] = \frac{255 - e_1[\overline{\gamma}(f)](x)}{\log(255)}$$

(14)

4 Índice de contraste

La ecuación 13 se usa en esta sección para introducir un método para cuantificar el contraste usando operadores morfológicos. Considere una imagen procesada por un operador y un elemento de estructura cuadrado $\mu B_x$ con su origen en el centro. Para cada conjunto de pixeles cubiertos por el elemento de estructura se consideran los elementos máximo y mínimo, esto es $I_{\max}(\mu B_x)$ e $I_{\min}(\mu B_x)$, de tal manera que una aproximación al parámetro $L$ en la ecuación 13 puede expresarse como:

$$L = \frac{I_{\max}(\mu B_x)}{I_{\min}(\mu B_x)}$$

(15)

con $I_{\max}(\mu B_x) \neq 0$ e $I_{\min}(\mu B_x) \neq 0$. Nótese que el parámetro $L$ es detectado localmente. Sustituyendo la expresión 15 en la 13 se obtiene la siguiente ecuación:

$$C_{\mu,\lambda}(f(x)) = k \log \left( \frac{I_{\max}(\mu B_x)}{I_{\min}(\mu B_x)} \right) + e_1[\overline{\gamma}(f)](x)$$

(16)

Con $I_{\max}(\mu B_x) \neq 0$ e $I_{\min}(\mu B_x) \neq 0$

Debe notarse que los valores de $I_{\max}(\mu B_x)$ e $I_{\min}(\mu B_x)$ pueden ser sustituidos por $\delta_{\mu B}(f)(x)$ y $\epsilon_{\mu B}(f)(x)$ respectivamente en el punto $x$. Esto se debe a que en los filtros de orden, la erosión y dilatación morfológicas se obtienen de esta manera [8], por lo tanto $I_{\max}(\mu B_x) = \delta_{\mu B}(f)(x)$ e $I_{\min}(\mu B_x) = \epsilon_{\mu B}(f)(x)$, entonces

$$C_{\mu,\lambda}(f(x)) = k \log \left( \frac{\delta_{\mu B}(f)(x)}{\epsilon_{\mu B}(f)(x)} \right) + e_1[\overline{\gamma}(f)](x)$$

(17)

Con $\epsilon_{\mu B}(f) \neq 0$ y $\delta_{\mu B}(f) \neq 0$
El operador gradiente ha sido definido en la ecuación 6, considerando que
\[ \log(\delta_{\mu}(f)) = \delta_{\mu}(\log(f)) \text{ y } \log(\varepsilon_{\lambda}(f)) = \varepsilon_{\lambda}(\log(f)) \] (20),
la ecuación 17 se escribe como:
\[ C_{\mu,\lambda}(f)(x) = k \ grad_{\mu}(\log(f))(x) + \varepsilon_{\lambda}[\bar{\gamma}_{\lambda}(f)](x) \] (18)
Con \( f(x) \neq 0 \). Para tener un índice global del contraste de la imagen, se considera la suma de los valores de \( C_{\mu,\lambda}(f)(x) \) y se denota como \( \Theta_{\mu,\lambda} \):
\[ \Theta_{\mu,\lambda} = k \ \sum \ \Grad_{\mu}(\log(f))(x_{i,j}) + \sum \ \varepsilon_{\lambda}[\bar{\gamma}_{\lambda}(f)](x_{i,j}) \] (19)
con \( f(x_{i,j}) \neq 0 \)
Donde \( f(x_{i,j}) \) representa el valor de la intensidad en niveles de gris en el punto \( x_{i,j} \), mientras que \( m \) y \( n \) denotan las dimensiones de la imagen. El parámetro \( k \) se especificó en la ecuación 14. Posteriormente, la ecuación 19 se divide por el volumen de la imagen original \( Vol[f] = \sum_{0 \leq i,j \leq n} f(x_{i,j}) \); con esto se evita trabajar con números grandes del contraste global. De esta manera, el índice de contraste \( X_{\mu,\lambda} \) se obtiene a partir de la ecuación 20.
\[ X_{\mu,\lambda} = \frac{\Theta_{\mu,\lambda}}{Vol[f]} \] (20)

5 Experimentos
Se utilizará la Fig. 2 para determinar la imagen con mejor contraste. La imagen original se encuentra en la Fig. 2(a), mientras que las imágenes en las figuras 2(b)-2(l) se obtuvieron a través del mapeo de contraste definido en la ecuación 8 considerando los valores mostrados en la tabla 1.

Se consideran los siguientes valores para los parámetros del mapeo de constante: \( \mu_1=12 \) y \( \mu_2=9 \), en este ejemplo la elección de los valores de \( \mu_1 \) y \( \mu_2 \) son grandes y diferentes para poder observar mejor el desempeño del mapeo de contraste, los valores de \( \alpha, \beta \) se obtuvieron de la siguiente manera: se considera un intervalo de valores entre 0 y 255, como \( \alpha \) y \( \beta \) deben tomar valores entre 0 y 1, se aplicó una regla de 3, considerando que el valor de 1 corresponde a 255, además, se tomaron incrementos de \( \alpha \) y \( \beta \) de 10 unidades, también observe que \( \alpha \) y \( \beta \) toman valores diferentes para cada mapeo de contraste.

Una vez determinados los valores que se involucran en cada mapeo, la imagen de la Fig. 2(a) se modifica, y se obtiene el conjunto de imágenes que se muestran en las figuras 2(b)-2(l). Posteriormente se calcula el índice de contraste \( X_{\mu,\lambda} \) para las imágenes de las figuras 2(b)-2(l), los valores respectivos se muestran en la tabla 2.

<table>
<thead>
<tr>
<th>Imagen</th>
<th>( \alpha )</th>
<th>( \beta )</th>
<th>( \mu_1 )</th>
<th>( \mu_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 (b)</td>
<td>0</td>
<td>0.039</td>
<td>12</td>
<td>9</td>
</tr>
<tr>
<td>2 (c)</td>
<td>0.039</td>
<td>0.078</td>
<td>12</td>
<td>9</td>
</tr>
<tr>
<td>2 (d)</td>
<td>0.078</td>
<td>0.117</td>
<td>12</td>
<td>9</td>
</tr>
<tr>
<td>2 (e)</td>
<td>0.117</td>
<td>0.156</td>
<td>12</td>
<td>9</td>
</tr>
<tr>
<td>2 (f)</td>
<td>0.156</td>
<td>0.196</td>
<td>12</td>
<td>9</td>
</tr>
<tr>
<td>2 (g)</td>
<td>0.196</td>
<td>0.235</td>
<td>12</td>
<td>9</td>
</tr>
<tr>
<td>2 (h)</td>
<td>0.313</td>
<td>0.352</td>
<td>12</td>
<td>9</td>
</tr>
<tr>
<td>2 (i)</td>
<td>0.431</td>
<td>0.470</td>
<td>12</td>
<td>9</td>
</tr>
<tr>
<td>2 (j)</td>
<td>0.549</td>
<td>0.588</td>
<td>12</td>
<td>9</td>
</tr>
<tr>
<td>2 (k)</td>
<td>0.666</td>
<td>0.705</td>
<td>12</td>
<td>9</td>
</tr>
<tr>
<td>2 (l)</td>
<td>0.823</td>
<td>0.862</td>
<td>12</td>
<td>9</td>
</tr>
</tbody>
</table>
Los valores de la tabla 2 se compararán con otros valores obtenidos por medio de dos métodos para cuantificar el contraste, el primer método fue introducido en [12], mientras que el segundo método fue propuesto en ([14, 15]).

a) Cuantificación de contraste basada en el análisis del histograma. En [12] la medida de contraste se obtiene al aplicar la ecuación 21, seguida de una cuantificación del ancho del histograma considerando el segundo momento alrededor del nivel de contraste 0. El segundo momento M2 está dado por la ecuación 22, la cual se muestra a continuación.

\[
C = \frac{L_{\text{max}} - L_{\text{min}}}{L_{\text{max}} + L_{\text{min}}} \tag{21}
\]

\[
M2 = \frac{\sum_{i=1}^{N} C_i^2 p(C_i)}{N} \tag{22}
\]

<table>
<thead>
<tr>
<th>Imagen</th>
<th>(X_{\mu, \lambda})</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 (b)</td>
<td>0.6183</td>
</tr>
<tr>
<td>2 (c)</td>
<td>0.5956</td>
</tr>
<tr>
<td>2 (d)</td>
<td>0.6070</td>
</tr>
<tr>
<td>2 (e)</td>
<td>0.6191</td>
</tr>
<tr>
<td>2 (f)</td>
<td>0.6242</td>
</tr>
<tr>
<td>2 (g)</td>
<td>0.6279</td>
</tr>
<tr>
<td>2 (h)</td>
<td>0.6341</td>
</tr>
<tr>
<td>2 (i)</td>
<td>0.6422</td>
</tr>
<tr>
<td>2 (j)</td>
<td>0.6427</td>
</tr>
<tr>
<td>2 (k)</td>
<td>0.6281</td>
</tr>
<tr>
<td>2 (l)</td>
<td>0.6001</td>
</tr>
</tbody>
</table>

Fig 2. (a) Imagen original, (b)-(l) Imágenes obtenidas al aplicar el mapeo de contraste definido en la ecuación 8 utilizando los parámetros que se muestran en la Tabla 1
donde $L_{\text{max}}$ y $L_{\text{min}}$ son las máxima y mínima luminancias en la región analizada, $p(C_j)$ es el número normalizado de ocurrencias del pixel con contraste $C_j$.

iii) El contraste total es el promedio de los contrastes globales obtenidos para la imagen original y las imágenes submuestreadas.

La siguiente expresión resume los pasos anteriores ([14,15]):

$$c = \frac{\sum_{\forall \text{pixel}} \frac{\sum_{8-\text{vecinos}} |p_i - p_j|}{8}}{\sum_{\forall \text{nivel}} \# \text{pixeles} / \# \text{niveles}}$$

(23)

El resultado de aplicar la ecuación 23 al conjunto de imágenes de la Fig. 2 se muestra en la tabla 4. Para comparar los resultados de las tablas 2, 3 y 4 se presentan sus gráficas respectivas en la Fig. 3. Observe en las gráficas de la Fig. 3 que la imagen con un mejor contraste utilizando el índice $X_{\mu,\lambda}$ corresponde a la imagen 2(j), mientras que las medidas M2 y C revelan que la imagen 2(k) es la imagen con mejor contraste.

### Tabla 3. Valores del parámetro M2 para las imágenes de la Fig. 2

<table>
<thead>
<tr>
<th>Imagen</th>
<th>M2</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 (b)</td>
<td>6.39e-5</td>
</tr>
<tr>
<td>2 (c)</td>
<td>5.35e-5</td>
</tr>
<tr>
<td>2 (d)</td>
<td>6.57e-5</td>
</tr>
<tr>
<td>2 (e)</td>
<td>6.95e-5</td>
</tr>
<tr>
<td>2 (f)</td>
<td>7.2e-5</td>
</tr>
<tr>
<td>2 (g)</td>
<td>7.41e-5</td>
</tr>
<tr>
<td>2 (h)</td>
<td>7.63e-5</td>
</tr>
<tr>
<td>2 (i)</td>
<td>8.08e-5</td>
</tr>
<tr>
<td>2 (j)</td>
<td>8.32e-5</td>
</tr>
<tr>
<td>2 (k)</td>
<td>8.40e-5</td>
</tr>
<tr>
<td>2 (l)</td>
<td>8.2e-5</td>
</tr>
</tbody>
</table>

b) Cuantificación del contraste basado en el promedio de diferencias. Los pasos a seguir para medir el contraste en imágenes digitales usando el método propuesto por [14,15] se presentan a continuación:

i) Obtener una pirámide de imágenes submuestreadas. En nuestro caso particular, las imágenes utilizadas fueron de los siguientes tamaños: 205×205, 154×154 y 103×103, partiendo de una imagen original de tamaño 256×256.

ii) Para cada imagen del inciso i) se calcula la suma del valor absoluto de la diferencia de cada pixel de la imagen con sus 8 vecinos. La suma final se divide por el tamaño de la imagen procesada. El resultado en este paso es el contraste global de la imagen analizada.

### Tabla 4. Medida del contraste usando la ecuación 23

<table>
<thead>
<tr>
<th>Imagen</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 (b)</td>
<td>9.46</td>
</tr>
<tr>
<td>2 (c)</td>
<td>10.37</td>
</tr>
<tr>
<td>2 (d)</td>
<td>10.89</td>
</tr>
<tr>
<td>2 (e)</td>
<td>11.93</td>
</tr>
<tr>
<td>2 (f)</td>
<td>12.07</td>
</tr>
<tr>
<td>2 (g)</td>
<td>12.48</td>
</tr>
<tr>
<td>2 (h)</td>
<td>13.03</td>
</tr>
<tr>
<td>2 (i)</td>
<td>13.66</td>
</tr>
<tr>
<td>2 (j)</td>
<td>14.15</td>
</tr>
<tr>
<td>2 (k)</td>
<td>14.23</td>
</tr>
<tr>
<td>2 (l)</td>
<td>13.22</td>
</tr>
</tbody>
</table>
Establecer de manera perceptual cual contraste es mejor entre las imágenes 2(j) y 2(k) no es sencillo. Sin embargo, analizando las gráficas en la Fig. 3 se puede determinar lo siguiente. En la Fig. 3(a) se encuentra que la imagen 2(j) posee importantes cambios de intensidad entre los contornos y las regiones alrededor de ellos, ya que tiene el valor más alto de contraste. Note que la imagen original en 2(a) es modificada por un mapeo, el cual realiza el contraste a través de los valores de la apertura, del cierre y de la imagen original de acuerdo al criterio de proximidad. El comportamiento de la curva en la gráfica 3(a) indica que existen cambios importantes en la intensidad de los pixeles a medida que se realza el contraste en la imagen procesada y se obtiene que las dos imágenes con mejor contraste son las imágenes en las Figs. 2(j) y 2(k). Posteriormente debido al comportamiento del mapeo de contraste se van fusionando las regiones por lo que la cantidad de contornos disminuye. Debido a este comportamiento del mapeo de contraste, las propuestas dadas en ([12 ,14, 15]) detectan que la mejor imagen es aquella donde existen esas regiones planas y que tienen importantes
cambios en las intensidades de los píxeles de la imagen. Esta situación se ilustra en la Fig. 4, donde pueden observarse como resaltan mejor los contornos de la imagen que fue detectada por el índice $X_{\mu,\lambda}$, Fig. 2(j), M2 y C, Fig. 2(k). Note que los contornos son más gruesos en la Fig. 4(b) que en la Fig. 4(a).

De esta manera, al observar los contornos de las imágenes en la Fig. 4, se concluye que los contornos están mejor preservados en la Fig. 4(a) que en la Fig. 4(b). Esta diferencia es un criterio que permite determinar que la imagen en la Fig. 2(j), presenta un mejor contraste.

6 Conclusiones

En este trabajo se presentó un índice de contraste basado en la ley de Weber, el cual se implementó a través de operadores morfológicos de uso común. Dicho índice involucra a los contornos y al fondo de la imagen. Los contornos se obtuvieron a través del gradiente morfológico, mientras que el fondo de la imagen se detectó mediante la erosión de la apertura por reconstrucción. Finalmente la propuesta dada en este trabajo se comparó con dos medidas de contraste reportadas en la literatura. El resultado de la comparación es que no es simple decir cual imagen es mejor visualmente, ya que ambas imágenes detectadas por los dos métodos presentan un buen realce en el contraste, sin embargo, al visualizar los contornos es posible tener un criterio para elegir la imagen con mejor contraste, y que en el ejemplo dado corresponde a la imagen detectada por el índice de contraste propuesto en este trabajo.

Agradecimientos

Los autores Angélica Jiménez Sánchez e Israel Santillán agradecen a CONACyT por la beca de doctorado otorgada.

Referencias

Índice de contraste morfológico basado en el análisis de los contornos y el fondo de la imagen


Angélica R. Jiménez Sánchez realizó sus estudios de licenciatura y maestría en Matemáticas aplicadas en la Universidad Autónoma de Querétaro (UAQ) en 2005 y 2007, actualmente realiza sus estudios de doctorado en esta misma institución. Es profesora-investigadora de la UAQ. Sus intereses en la investigación incluyen a la morfología matemática y visión por computadora.

Jorge D. Mendiola Santibañez recibió el grado de licenciado en electrónica por parte de la Benemérita Universidad Autónoma de Puebla, México y el grado de maestría en electrónica por parte del INAOE, México. El grado de doctor fue otorgado por la Universidad Autónoma de Querétaro (UAQ), México. Actualmente es profesor-investigador en la UPQ y es miembro del sistema nacional de investigadores (SNI), México. Sus intereses en la investigación incluyen a la morfología matemática y visión por computadora.
Gilberto Herrera Ruiz recibió el grado de doctor en ingeniería de la academia Húngara de ciencias en 1992. Sus intereses en la investigación incluyen el procesamiento de imágenes, sistemas de manufactura, máquinas y herramientas, así como tecnologías para invernaderos. Actualmente es profesor-investigador en la Universidad Autónoma de Querétaro y es miembro del sistema nacional de investigadores (SNI) México.

Israel Santillán realizó sus estudios de licenciatura en el Instituto Tecnológico de Estudios Superiores de Monterrey y la maestría en la Facultad de Ingeniería de la Universidad Autónoma de Querétaro, México. Actualmente realiza sus estudios de doctorado en esta misma institución. Es profesor investigador de la UPQ y es miembro del sistema nacional de investigadores (SNI), México. Sus intereses en la investigación incluyen a la morfología matemática, visión por computadora.

Artículo recibido el 23/08/2010; aceptado el 16/08/2011.
Diseño óptimo de transformadores de Hilbert
sin multiplicadores con base en el uso de un subfiltro simple

David E. Troncoso Romero, Miriam G. Cruz Jiménez y Gordana Jovanovic Dolecek
Departamento de Electrónica, Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE),
Puebla,
México
{dtroncoso, miriam_gcj, gordana}@inaoep.mx

Resumen. Los transformadores de Hilbert altamente selectivos pueden ser diseñados eficientemente mediante el método de Transformación en Frecuencia (Frequency Transformation, FT), donde un bloque básico, formado con dos subfiltros idénticos, es implementado repetidamente. El número de bloques utilizados se obtiene de la longitud de un filtro prototipo. Recientemente se ha utilizado la técnica Segmentación-Intercalamiento (Pipelining-Interleaving, PI) para evitar el uso repetitivo del bloque básico, reduciendo el número de coeficientes requeridos. Sin embargo, el diseño del subfiltro y del filtro prototipo está basado en una búsqueda heurística. En este artículo se presenta el método óptimo para diseñar el subfiltro y el filtro prototipo, minimizando el número de coeficientes. Además, se propone una estructura alternativa que permite utilizar únicamente un subfiltro dentro del bloque básico. Como resultado, el número total de coeficientes es disminuido. Se demuestra con un par de ejemplos que el método de diseño es óptimo, simple y eficiente.

Palabras Clave. Filtros digitales, transformador de Hilbert.

Optimal Design of Multiplierless Hilbert Transformer based on the Use of a Simple Subfilter

Abstract. Very sharp Hilbert transformers can be efficiently designed by using the Frequency Transformation (FT) method, where a basic building block, formed with two identical subfilters, is repeatedly implemented. The number of the building blocks used is obtained from the length of a prototype filter. Recently, the Pipelining-Interleaving (PI) technique has been applied to avoid the repetitive use of the basic building block, reducing the number of required coefficients. However, the design of the subfilter and the prototype filter is based on a heuristic search. In this paper, we present an optimal method to design the subfilter and prototype filter minimizing the number of coefficients. Additionally, an alternative structure, which permits to use a unique subfilter inside the basic building block, is presented. As a result, the total number of coefficients is decreased. Two examples show that the proposed design method is optimal, simple, and efficient.

Keywords. Digital filters, Hilbert transformer.

1 Introducción

Un transformador de Hilbert es un filtro que introduce un desplazamiento de fase de π/2 en la señal de entrada. Su respuesta en frecuencia ideal está dada por la siguiente ecuación,

\[
H(e^{j\omega}) = \begin{cases} 
-j, & 0 < \omega < \pi, \\
j, & -\pi < \omega < 0.
\end{cases}
\]  

(1)

Este filtro es utilizado en telecomunicaciones, procesamiento de voz e imágenes médicas, entre otros [6, 7, 8].

Los transformadores de Hilbert pueden ser diseñados como filtros con Respuesta al Impulso Infinita (Infinite Impulse Response, IIR) o como filtros con Respuesta al Impulso Finita (Finite Impulse Response, FIR). Estos últimos pueden tener fase lineal exacta, y su estabilidad está garantizada. Sin embargo, ellos tienen una complejidad computacional más alta en comparación con los filtros IIR para una misma especificación. Esta complejidad aumenta conforme la banda de transición se hace más angosta [1].
Diferentes técnicas han sido propuestas para llevar a cabo el diseño eficiente de transformadores de Hilbert FIR de baja complejidad con banda de transición angosta [2, 5, 6, 7, 8].

Un método efectivo es el basado en un filtro de media banda formado con base en la técnica de Enmascaramiento de Respuesta en Frecuencia (Frequency-Response Masking, FRM) [7]. Una modificación a ese método utiliza una nueva estructura basada en FRM, la cual se basa en un subfiltro de corrección de la respuesta en frecuencia [8]. En esa propuesta todos los subfiltros son diseñados simultáneamente bajo el mismo problema de optimización. Por otra parte, se tiene el método basado en el diseño de filtros FIR de banda ancha y fase lineal con una respuesta al impulso Polinomial-Sinusoidal por Segmentos (Piecewise Polynomial-Sinusoidal, PPS) [6]. El método de Transformación en Frecuencia (Frequency Transformation, FT) [9], fue utilizado para diseñar transformadores de Hilbert FIR con base en la interconexión en cascada de múltiples copias de un bloque básico simple [5]. Este bloque básico es obtenido con la conexión en cascada de dos subfiltros transformadores de Hilbert idénticos de baja complejidad.

Debido a que los multiplicadores son elementos muy caros en un filtro digital, el diseño sin multiplicadores es preferible. Recientemente un método simple y eficiente ha sido propuesto para diseñar transformadores de Hilbert sin multiplicadores con banda de transición muy angosta, con base en el método FT [2]. En esa propuesta la implementación repetida del mismo bloque básico fue evitada empleando la técnica Segmentación-Intercalamiento (Pipelining-Interleaving, PI) [3]. Además se utilizó el método de Eliminación de Subexpresiones Comunes (Common Subexpression Elimination, CSE) [10] para reducir el número de sumadores en los subfiltros transformadores de Hilbert que componen al bloque básico.

La aplicación de la técnica propuesta por [2] da como resultado transformadores de Hilbert FIR altamente selectivos con un número reducido de sumadores. No obstante, el diseño del subfiltro y del filtro prototipo está basado en una búsqueda heurística, por lo que el resultado no es óptimo. Adicionalmente, la estructura y el algoritmo de diseño propuestos en ese artículo se basan en la implementación del bloque básico como la conexión en cascada de dos transformadores de Hilbert idénticos de bajo orden. Si bien esto permite que el aumento de la frecuencia de reloj de los transformadores de Hilbert que conforman dicho bloque no sea demasiado alto, el número total de sumadores requeridos en el diseño no es completamente reducido. Esto provoca que el filtro ocupe más espacio y más recursos de hardware.

En este artículo se presentan dos contribuciones que permiten una mayor reducción en el número de coeficientes requeridos. En primer lugar, se presenta el método óptimo para diseñar el subfiltro y el filtro prototipo, de tal manera que el número de coeficientes sea minimizado. En segundo lugar, se propone modificar la estructura presentada por [2], utilizando la técnica PI no solamente en la interconexión de bloques básicos sino también dentro del propio bloque, de tal manera que se pueda evitar el uso repetido de los dos subfiltros transformadores de Hilbert que lo forman. Como resultado, se obtiene una reducción adicional en el número total de coeficientes.

La organización de este artículo es como sigue. La Sección 2 presenta la explicación del método de Transformación en Frecuencia. El método de diseño propuesto es detallado en la Sección 3. Por último, la discusión de resultados es presentada en la Sección 4.

2 El método de transformación en frecuencia

El método de Transformación en Frecuencia de [5] permite diseñar un transformador de Hilbert con rizos muy pequeños y banda de transición muy angosta utilizando repetidamente un subfiltro simple. El número de veces que este subfiltro es utilizado está en función de la longitud de un filtro prototipo. Ambos, el subfiltro y el filtro prototipo, son transformadores de Hilbert FIR basados en filtros Tipo-III y Tipo-IV. Un filtro Tipo-III tiene respuesta al impulso antisimétrica y su longitud debe ser impar. Similarmente, un filtro Tipo-IV
también tiene antisimetría en su respuesta al impulso, pero su longitud es siempre par.

El filtro prototipo debe ser un filtro FIR Tipo-IV, es decir, con longitud par dada como $L_P = 2N$ y respuesta al impulso antisimétrica de la forma $p(2N - 1 - n) = -p(n)$. Su respuesta en frecuencia está expresada como

$$P(e^{j\Omega}) = e^{-j(2N-1)\Omega/2} P_0(\Omega),$$

donde $P_0(\Omega)$, el término de fase-cero, está dado por

$$P_0(\Omega) = j \cdot \sin\left(\frac{\Omega}{2}\right) \sum_{n=0}^{N-1} \tilde{d}(n) \cos(\Omega n),$$

con $\Omega$ denotando el dominio de frecuencias del filtro prototipo. Los coeficientes $\tilde{d}(n)$ pueden ser obtenidos de la respuesta al impulso $p(n)$ [1].

Consideremos la equivalencia

$$\cos(\Omega n) = T_n(\cos(\Omega))$$

con $T_n(x)$ siendo el polinomio de Chebyshev de $n$-ésimo grado.

El término de fase-cero puede entonces ser reescrito como

$$P_0(\Omega) = j \cdot \sin\left(\frac{\Omega}{2}\right) \sum_{n=0}^{N-1} \alpha(n) \left[\cos(\Omega n)\right]^n,$$

donde los coeficientes $\alpha(n)$ están relacionados con los coeficientes $\tilde{d}(n)$ a través de los polinomios de Chebyshev. Con base en la siguiente equivalencia

$$\cos(2x) = 1 - \sin^2(x) = 1 + \left(j \cdot \sin(x)\right)^2,$$

el término de fase-cero puede ser expresado por

$$P_0(\Omega) = j \cdot \sin\left(\frac{\Omega}{2}\right) \sum_{n=0}^{N-1} \alpha(n) \left[1 + \left(j \cdot \sin\left(\frac{\Omega}{2}\right)\right)\right]^n,$$

(6)

El subfiltro puede ser un filtro Tipo-III o Tipo-IV. Consideremos el caso de un filtro Tipo-III, con longitud impar dada como $LG = 2M + 1$ y respuesta al impulso antisimétrica de la forma $g(2M - n) = -g(n)$. Su respuesta en frecuencia está expresada como

$$G(e^{j\omega}) = e^{-j(2M+1)/2} G_0(\omega),$$

donde $G_0(\omega)$ es el término de fase-cero, dado por

$$G_0(\omega) = j \cdot \sum_{n=1}^{M} c(n) \sin(\omega n).$$

(8)

Los coeficientes $c(n)$ pueden ser obtenidos a partir de $g(n)$ [1]. Obsérvese que el término $G_0(\omega)$ puede ser puesto en (6) mediante la sustitución dada por la siguiente expresión,

$$j \cdot \sin\left(\frac{\omega}{2}\right) = j \cdot \sum_{n=1}^{M} c(n) \sin(\omega n),$$

(9)

resultando en

$$H_0(\omega) = j \cdot \sum_{n=1}^{M} c(n) \sin(\omega n) *$$

$$\sum_{n=0}^{N-1} \alpha(n) \left[1 + 2 \left(j \cdot \sum_{n=1}^{M} c(n) \sin(\omega n)\right)\right]^n,$$

(10)

donde $H_0(\omega)$ es el término de fase-cero del filtro total. Por lo tanto, la transformación en frecuencia se obtiene de (9) y está dada por

$$\Omega = 2 \sin^{-1}\left[\sum_{n=1}^{M} c(n) \sin(\omega n)\right],$$

(11)

La ecuación (11) implica que la magnitud de la respuesta en frecuencia del filtro prototipo se preserva, pero su dominio de frecuencias es cambiado por el subfiltro.

La función de transferencia del transformador de Hilbert total, está dada como

$$H(z) = z^{-M(2N-1)} H_0(z),$$

(12)

donde

$$H_0(z) = G(z) \sum_{n=0}^{N-1} z^{-2M(2N-1)-n} \alpha(n) \left[H_1(z)\right]^n,$$

(13a)
\[ H(z) = z^{2M} + 2G(z), \]  

con \( G(z) \) siendo la función de transferencia del subfiltro.

Consideremos la especificación del transformador de Hilbert deseado dada su respuesta en frecuencia de la siguiente manera,

\[ (1-\delta) \leq |H_0(\omega)| \leq (1+\delta), \quad \text{para} \quad \alpha_1 \leq \omega \leq \pi - \alpha_1 \]  

(13b)

Donde \( \omega_L \) es la frecuencia límite inferior de la banda de paso y \( \delta \) es el rizo permitido. Entonces el filtro prototipo debe cumplir la especificación dada en términos de la frecuencia transformada por

\[ (1-\delta) \leq |P_0(\Omega)| \leq (1+\delta), \quad \text{para} \quad \Omega_L \leq \Omega \leq \pi \]  

(14)

Con \( \Omega_L \) siendo la frecuencia límite inferior del filtro prototipo. El subfiltro debe cumplir simultáneamente la especificación dada por

\[ \nu_a = \frac{1}{2} + \frac{1}{2} \sin \left( \frac{\alpha_1}{2} \right), \]  

(16a)

\[ \delta_a = \frac{1}{2} - \frac{1}{2} \sin \left( \frac{\alpha_1}{2} \right) \]  

(16b)

Nótese que para poder diseñar ambos, el filtro prototipo y el subfiltro, es necesario determinar el valor \( \Omega_L \). La Figura 1 presenta una estructura eficiente [2], donde el uso repetitivo del bloque básico \( H_1(z) \), dado en (13b), se evitó con base en la técnica PI. El número de coeficientes requeridos se redujo, ya que el bloque básico se implementa una sola vez en forma expandida, \( H_1(z^K) \). \( K \) es el factor de expansión y también representa la cantidad de aumento en la frecuencia de reloj del filtro \( H_1(z^K) \). No obstante, el valor \( \Omega_L \) fue obtenido en forma heurística [2].

3 Método propuesto

A continuación se presentará, en la Sección 3.1, el método de optimización para obtener el valor \( \Omega_L \). En la Sección 3.2 se dará una modificación a la estructura de la Figura 1, que permite una mayor reducción en el número de coeficientes. La Sección 3.3 presentará un ejemplo detallado con los pasos de diseño propuestos.

3.1 Minimización del número de coeficientes requeridos

El número total de coeficientes requeridos en la estructura de la Figura 1 puede ser aproximado por

\[ N_{coef} = \beta(L_3 + 1) + L_p / 2, \]  

(17)

Donde \( L_3 \) es la longitud del subfiltro, \( L_p \) es la longitud del filtro prototipo y \( \beta = 3/4 \) si se desea utilizar un subfiltro Tipo-III o \( \beta = 3/2 \) si el subfiltro es Tipo-IV.

La longitud del filtro prototipo puede ser estimada eficientemente con base en los resultados de [8], de la siguiente manera,
Diseño óptimo de transformadores de Hilbert sin multiplicadores con base en el uso...

\[ L_p = ([0.002655 \log_2(\delta)]^3 + 0.031843 \log_2(\delta))^2 - 0.554993 \log_2(\delta) - 0.049788)/(\Omega_L / 2\pi) + 1. \] (18)

De modo similar, la longitud del subfiltro puede ser estimada utilizando la misma fórmula, solamente sustituyendo \( \omega_L \) en lugar de \( \Omega_L \) y \( \delta_G \) en lugar de \( \delta \), con \( \delta_G \) dado en (16c). Sustituyendo (18) en (17) y calculando \( \delta_G \) de (16c), se llega a

\[ N_{\text{coef}} = \beta [([0.002655 \log_2(0.5 - 0.5 \sin(\Omega_L / 2))]^3 + 0.031843 \log_2(0.5 - 0.5 \sin(\Omega_L / 2))]^2 - 0.554993 \times \log_2([0.5 - 0.5 \sin(\Omega_L / 2)] - 0.049788)/(\omega_L / 2\pi)) + 2\beta + \frac{1}{2} \{(0.002655 \log_2(\delta)]^3 + 0.031843 \log_2(\delta)]^2 - 0.554993 \log_2(\delta) - 0.049788)/(\Omega_L / 2\pi) + \frac{1}{2}. \] (19)

Observamos que las variables en (19) son \( \beta \), \( \delta \), \( \omega_L \) y \( \Omega_L \). Sin embargo, \( \beta \), \( \delta \) y \( \omega_L \) son valores conocidos que pueden ser sustituidos previamente. Por lo tanto, el número total de coeficientes queda explícitamente en función de \( \Omega_L \). Por lo tanto, el problema de optimización se expresa como

\[
\min_{\Omega_L} N_{\text{coef}}
\]

tal que \( 0 < \Omega_L < \pi \), (20)

con \( N_{\text{coef}} \) dado en (19). En esta propuesta se ha utilizado la función de MATLAB fminbnd para resolver (20).

### 3.2 Estructura propuesta

De la Figura 1b se observa que el bloque básico \( H_1(z) \) está formado por dos subfiltros conectados en cascada. Esta conexión puede ser evitada utilizando también la técnica PI (Pipelining-Interleaving). Como resultado, se obtiene la estructura de la Figura 2. Esta modificación trae como consecuencia una disminución en el número de coeficientes requeridos, ya que el subfiltro se implementa solamente una vez dentro de \( H_1(z) \). El número total de coeficientes requeridos en la estructura propuesta se determina utilizando (19), solamente sustituyendo \( \beta = 1/2 \) si el subfiltro es Tipo-III o \( \beta = 1 \) si es Tipo-IV. Por lo tanto, el valor \( \Omega_L \) óptimo se obtiene resolviendo (20) con los valores adecuados de \( \beta \).

Debe notarse que con la estructura propuesta se requiere que el bloque básico \( H_1(z^2) \) sea controlado con una mayor frecuencia de reloj. El aumento en la frecuencia de reloj está dado, para la estructura de la Figura 1b y para la estructura propuesta, como

\[
K = \gamma \left[ \frac{L_p}{2} - 1 \right]
\]

(21)
donde \( \gamma = 1 \) para la estructura de la Figura 1b y \( \gamma = 2 \) para la estructura propuesta. Sustituyendo (18) en (21), \( K \) se puede estimar en función de \( \Omega_L \). Derivando esta función con respecto a \( \Omega_L \) e igualando la derivada a cero podemos obtener el valor en el que \( K \) es mínimo. El mínimo \( K \) existe en un valor \( \Omega_L \) infinito. Esto significa que cuando hay restricción en el aumento de la frecuencia de reloj de \( H_1(z^2) \), \( \Omega_L \) no debe ser menor a un mínimo valor dado por

\[
\Omega_L = \gamma \pi \cdot \frac{[0.002655 \log_2(\delta)]^3 + 0.031843 \log_2(\delta)]^2 - 0.554993 \log_2(\delta) - 0.049788}{(K + 0.5\gamma)}.
\]

(22)

Nótese de (21) que en la estructura propuesta el valor \( K \) siempre es un número par, dado que \( \gamma = 2 \) en ese caso.

**Fig. 2.** Estructura propuesta para el bloque básico \( H_1(z) \)

### 3.3 Procedimiento de diseño

El procedimiento de diseño de un transformador de Hilbert con especificación dada en (14) se da en los siguientes pasos.

**Paso 1.** Resolver (20), sustituyendo los valores adecuados de \( \beta \), \( \delta \) y \( \omega_L \), para obtener \( \Omega_L \). Si hay...
A continuación se presenta un ejemplo de diseño.

**Ejemplo 1.** Diseñe un transformador de Hilbert que satisfaga (14), con $\delta = 0.004$ y $\omega_p = 0.01\pi$. Considere como máximo un aumento en la frecuencia de reloj del bloque $H_r(z^K)$ dado por $K = 9$.

**Paso 1.** Se utiliza $\beta = 3/4$ debido a que $K$ es impar. Esto significa que la estructura del bloque básico será la de la Figura 1b, con un subfiltro Tipo-III. Los valores $\delta = 0.004$, $\beta = 3/4$ y $\omega_p = 0.01\pi$ se sustituyen en (19). Se resuelve (20) utilizando la función fminbnd de MATLAB y se obtiene $\Omega_L = 0.139\pi$. Debido a que hay una restricción en el aumento de la frecuencia de reloj, también se calcula $\Omega_L$ utilizando (22), donde se sustituyen $K = 9$ y $\gamma = 1$. Se obtiene $\Omega_L = 0.1503\pi$. De ambos valores obtenidos, se elige $\Omega_L = 0.1503\pi$, por ser el más alto de los dos.

**Paso 2.** Se estima la longitud del filtro prototipo, $L_p$, sustituyendo $\Omega_L = 0.1503\pi$ en (18). Se obtiene $L_p = 20$. Para estimar la longitud del subfiltro, $L_G$, se necesita calcular $\delta_G$ usando (16c). Se obtiene $\delta_G = 0.3831$. Entonces se sustituye $\delta_G$ y $\omega_L = 0.01\pi$ en (18). El valor estimado es $L_G = 38.37$. Los filtros cumplen las especificaciones requeridas usando $L_p = 20$ y $L_G = 23$.


**Paso 4.** Los coeficientes del subfiltro son simplificados usando el método de [10]. Los coeficientes de la estructura son obtenidos de su relación con $p(n)$ (ver (2)-(4)) y representados en forma Canónica de Dígitos con Signo (Canonical Signed Digit, CSD).

<table>
<thead>
<tr>
<th>$g(n)$</th>
<th>$-g(L_G-1-n)$</th>
<th>$x_1 = z^6 - 2^{-2}$</th>
<th>$x_2 = z^6 + 2^{-2}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$g(0)$</td>
<td>$-2^{-3}x_2 - 2^{-3}x_1$</td>
<td>$g(4) = -2^{-3}$</td>
<td>$g(8) = -2^{-3}$</td>
</tr>
<tr>
<td>$g(2)$</td>
<td>$-2^{-3}x_1$</td>
<td>$g(6) = -2^{-3}x_2$</td>
<td>$g(10) = -2^{-3}x_2 - 2^{-3}x_1$</td>
</tr>
</tbody>
</table>

A continuación se presenta un ejemplo de diseño.

Diseñar el filtro prototipo tal que satisfaga (15) y el subfiltro tal que satisfaga (16). Estimar las longitudes iniciales $L_p$ y $L_G$ usando (18).

**Paso 2.** Diseñar el filtro prototipo tal que satisfaga (15) y el subfiltro tal que satisfaga (16). Estimar las longitudes iniciales $L_p$ y $L_G$ usando (18).

**Paso 3.** Redondear los coeficientes del filtro prototipo de la siguiente manera,

$$p(n) = 2^{-b_p} \cdot \text{round}[p(n)/2^{-b_p}]$$  \hspace{1cm} (23)

donde $p(n)$ es la respuesta al impulso con coeficientes redondeados, $B_p$ es la longitud de palabra del filtro prototipo y $\text{round}[x]$ indica la operación de redondeo hacia el entero más cercano a $x$. Los coeficientes del subfiltro se deben redondear usando (23), simplemente sustituyendo $p(n)$, $B_p$ y $p(n)$ respectivamente, con $g(n)$ siendo la respuesta al impulso con coeficientes redondeados y $B_G$ la longitud de palabra del subfiltro. $B_p$ puede ser estimada como sigue [4],

$$B_p = -\log_2 \left( \frac{\delta}{2 \sqrt{(2L_p-1)/3}} \right)$$  \hspace{1cm} (24)

$B_G$ puede estimarse usando (24), solamente sustituyendo $\delta_G$ y $L_G$ en lugar de $\delta$ y $L_p$, respectivamente. Los valores $L_p$ y $L_G$ deben ser los obtenidos finalmente en el Paso 2 (no los valores estimados). Si los filtros no satisfacen la especificación correspondiente, incrementar $B_p$ y $B_G$.

**Paso 4.** Aplicar el algoritmo de [10] para eliminar las subexpresiones comunes en el subfiltro. Además obtener los coeficientes $a(n)$ a partir de $p(n)$ y representarlos en forma Canónica de Dígitos con Signo (Canonical Signed Digit, CSD).

A continuación se presenta un ejemplo de diseño.

**Ejemplo 1.** Diseñe un transformador de Hilbert que satisfaga (14), con $\delta = 0.004$ y $\omega_p = 0.01\pi$. Considere como máximo un aumento en
Diseño óptimo de transformadores de Hilbert sin multiplicadores con base en el uso...

4 Discusión de resultados

Con el propósito de comparar el método propuesto, a continuación se realizan dos transformadores de Hilbert tomados de las referencias [6, 7, 8]. Ambas estructuras del bloque básico $H_1(z)$, la de la Figura 1b y la de la Figura 2, son utilizadas en los diseños propuestos.

**Ejemplo 2.** Diseñe un transformador de Hilbert que satisfaga (14), con la siguiente especificación [7]: $\delta = 0.0001$ y $\omega_P = 0.00625\pi$. La Tabla 3 muestra las características del filtro prototipo y del subfiltro para ambos casos, cuando se usa el bloque básico de la Figura 1b y cuando se utiliza el de la Figura 2. La Tabla 4 exhibe la comparación contra los diseños de [2, 7] en términos del número de sumadores y del número de multiplicadores requeridos. De la Tabla 4 puede verse que utilizando el método de optimización propuesto con la estructura de la Figura 1b para el bloque básico $H_1(z)$, la reducción en el número de sumadores es apenas del 1.24% con respecto al resultado de [2]. Por otra parte, cuando se utiliza la estructura propuesta para el bloque básico, se obtiene una reducción en el número de sumadores de 24.84% con respecto a [2]. Nótese que en este caso el número de sumadores es 5.5% menor que el número de sumadores utilizado en [7]. Además, el filtro propuesto no requiere multiplicadores.

La Tabla 5 muestra las características del filtro prototipo y del subfiltro para ambos casos, cuando se usa el bloque básico de la Figura 1b y cuando se utiliza el de la Figura 2. La Tabla 6 presenta la comparación contra los diseños de [2, 6, 8] en términos del número de sumadores y del número de multiplicadores requeridos. Es posible ver de la Tabla 6 que el filtro resultante diseñado con el método de optimización propuesto, usando la estructura de [2] para el bloque básico (ver Figura 1b), utiliza 9.82% menos sumadores que el diseño realizado en [2]. Por otra parte, utilizando la estructura propuesta en el bloque básico, la reducción en el número de sumadores es de 38.17%.

**Ejemplo 3.** Diseñe un transformador de Hilbert que satisfaga (14), con la siguiente especificación [6, 8]: $\delta = 0.0001$ y $\omega_P = 0.00125\pi$. 

**Tabla 2. Coeficientes de la estructura del ejemplo 1**

| $\alpha(0)$ | -2^1 + 2^2 - 2^3 + 2^-1 + 2^-2 |
| $\alpha(1)$ | -2^2 + 2^-1 - 2^3 + 2^-2 - 2^-1 |
| $\alpha(2)$ | -2^-2 + 2^-3 |
| $\alpha(3)$ | -2^-3 + 2^-1 - 2^-2 |
| $\alpha(4)$ | -2^3 + 2^-1 + 2^-4 |
| $\alpha(5)$ | -2^1 + 2^-1 - 2^2 - 2^-1 + 2^-2 |
| $\alpha(6)$ | 2^1 - 2^1 - 2^3 - 2^-1 |
| $\alpha(7)$ | 2^2 - 2^-3 - 2^-1 |
| $\alpha(8)$ | -2^1 + 2^-1 - 2^3 - 2^-2 |
| $\alpha(9)$ | -2^2 - 2^-2 |

**Fig. 3. Estructura basada en la técnica PI para el ejemplo 1**

**Fig. 4. Respuesta en magnitud del transformador de Hilbert del Ejemplo 1; (a) Detalle de banda de paso, (b) Detalle de banda de transición**
Obsérvese que aunque el número de sumadores utilizados es mayor que el de [8], ese diseño requiere 107 multiplicadores. Adicionalmente, el diseño de [6] requiere 48 multiplicadores. El filtro propuesto, por otra parte, no necesita multiplicadores.

5 Conclusiones

En este artículo se ha presentado un método eficiente para diseñar transformadores de Hilbert con especificaciones muy estrictas y con la minimización del número estimado de los coeficientes. La estructura de diseño está basada en la técnica de Transformación en Frecuencia (Frequency Transformation, FT), donde se utiliza repetidamente un bloque básico compuesto de dos subfiltros idénticos conectados en cascada.

Se ha tomado ventaja de la técnica de Segmentación-Intercalamiento (Pipelining-Interleaving, PI) para evitar el uso repetitivo del mismo bloque básico, y también para evitar el uso repetitivo del subfiltro dentro de dicho bloque. Adicionalmente se ha aplicado redondeo a los coeficientes del filtro para que los multiplicadores sean sustituidos por sumadores y corrimientos. Debido a que la técnica PI incrementa la frecuencia de reloj del filtro, se ha introducido un criterio para obtener el diseño óptimo del filtro sin
Diseño óptimo de transformadores de Hilbert sin multiplicadores con base en el uso...

...de cierto límite dado para el aumento de la frecuencia de reloj.

Como los ejemplos proporcionados lo demuestran, se ha obtenido una reducción en el número total de sumadores utilizados en el filtro. Por lo tanto, el método propuesto da como resultado diseños con menor complejidad, en comparación con algunos otros métodos recientemente propuestos en la literatura.

Referencias


David Ernesto Troncoso Romero recibió el grado de Ingeniero en Electrónica en el Instituto Tecnológico de Minatitlán, Minatitlán, Veracruz, México, y el grado de Maestro en Ciencias en el Departamento de Electrónica del Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE), Puebla, México, en el 2006 y el 2008, respectivamente. Actualmente es estudiante de doctorado en el Departamento de Electrónica del INAOE. Sus áreas de interés actuales están en el diseño de filtros digitales de baja complejidad con Respuesta al Impulso Finita (FIR).

Miriam Guadalupe Cruz Jiménez se graduó en el Instituto Tecnológico de Minatitlán, Minatitlán, Veracruz, México, como Ingeniero en Electrónica, en el 2006. Recibió el grado de Maestría en el Departamento de Electrónica del Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE), Puebla, México, en el 2010.

Gordana Jovanovic Dolecek recibió el grado de Ingeniería Eléctrica por la Universidad de Sarajevo, el grado de Maestría en Ciencias por la Universidad de Belgrado y el Doctorado por la Facultad de Ingeniería Eléctrica de la Universidad de Sarajevo. Fue profesora en la Facultad de Ingeniería Eléctrica de la Universidad de Sarajevo hasta 1993. De 1993 a 1995 estuvo en el Instituto Mihailo Pupin de Belgrado, Yugoslavia. En 1995 se unió al Departamento de Electrónica del Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE), Puebla, México, donde ella es actualmente profesora titular e investigadora. Durante el periodo 2001-2002 y en el 2006 estuvo en el Departamento de Ingeniería Eléctrica y de Cómputo de la Universidad de Santa Bárbara como investigadora visitante. Durante el periodo 2008-2009 estuvo en la Universidad de San...
Diego como investigadora visitante en año sabático. Es autora de tres libros, editora de uno y autora de más de 250 artículos. Es Senior Member de IEEE, miembro de la Academia Mexicana de Ciencias y miembro del Sistema Nacional de Investigadores (SNI), México.

Artículo recibido el 29/11/2010; aceptado el 07/03/2011.
Abstract. We present methods for image annotation and retrieval based on semantic cohesion among terms. On the one hand, we propose a region labeling technique that assigns an image the label that maximizes an estimate of semantic cohesion among candidate labels associated to regions in segmented images. On the other hand, we propose document representation techniques based on semantic cohesion among multimodal terms that compose images. We report experimental results that show the effectiveness of the proposed techniques. Additionally, we describe an extension of a benchmark collection for evaluation of the proposed techniques.

Keywords. Automatic image annotation, region labeling, multimedia image retrieval, ground truth data creation.

1 Introduction

Image retrieval has been an active research area for more than two decades now [1, 2]. Despite a substantial advance achieved in this field, most of the reported work focuses on methods that consider a single modality (i.e., either image or text), thus limiting the effectiveness and applicability of two groups of methods: one based on texts and the other based on images. On the one hand, text-based methods are unable to retrieve images that are visually similar to the query image; on the other hand, image-based techniques cannot retrieve relevant images to queries that involve non-visual information (e.g., places, events, or dates).

Due to the above limitations, in the last few years there has been an increasing interest of the scientific community in the development of retrieval techniques that incorporate both visual and textual information [2, 8]. Nevertheless, current multimodal techniques still deal with both sources of information separately, eventually applying a standard information fusion technique [2]; therefore, such methods do not exploit the association among multiple modalities to obtain better representations of images. Furthermore, in many databases, images are not accompanied with any textual information, which further complicates the application of multimodal retrieval.
methods. In the latter collections, automatic image annotation (AIA) methods are used for assigning text to images because manual labeling of images is a time-consuming and labor extensive task [1].

Both tasks, image annotation and image retrieval, are closely interrelated and hence can be studied jointly. Accordingly, in this paper we faced such problems with the goal of improving the performance of current techniques and overcoming some of their limitations. More specifically, we focused on the region-level AIA task with the goal of giving support to multimodal image retrieval methods that attempt to exploit the redundancy and complementariness of information as provided by labels and text.

We propose methods for the annotation and retrieval of images that are based on semantic cohesion modeling; where semantic cohesion is defined as the degree of affinity of terms in a document according to their meaning or their use in the context given by other terms that occur in the same document. Intuitively, the greater the degree of semantic cohesion among terms, the higher is the probability that such terms are used together in similar contexts. The rest of this paper summarizes our research and outlines the main findings of our work, for further details we refer the reader to the thesis [3] and the representative publications derived from it [4, 7]. Before presenting our methods, we describe the extension we made to a benchmark collection for allowing the evaluation of region-level AIA methods and of image retrieval methods that consider AIA labels.

2 The SAIAPR TC12 Collection

Due to the lack of a suitable database to evaluate the methods we propose, part of our work included the development of a benchmark image collection. Specifically, we proposed an extension of the IAPR TC12 collection, a benchmark data set for the evaluation of image retrieval methods [8]. The extension consisted in manual segmentation and annotation of each image in the IAPR TC12 collection according to predefined rules and by using a hierarchical organization of the vocabulary we defined. The proposed hierarchy is composed of six branches: “Animal”, “Humans”, “Food”, “Man-made”, “Landscape” and “Other”. Summarizing, a total of 20,000 images have been manually segmented and the resultant 99,535 regions were manually labeled by using a vocabulary of 255 labels; the data derived from our extension is publicly available at the official ImageCLEF website. Our extension has increased the number of applications of the collection and its scope in terms of the tasks that can be evaluated with it [4]; furthermore, the extension has been extremely helpful for the evaluation of the methods we developed, and has attracted the interest from the scientific community. A detailed description of our extension to the IAPR TC12 collection can be found in [4].

3 Semantic Cohesion for Image Annotation

For AIA, we propose an energy-based model that attempts to maximize the semantic cohesion among labels that have been assigned to adjacent regions in segmented images [5, 6]. The model seeks to refine the initial labeling as provided by a multiclass classifier trained with purely visual information. The classifier (which can be built by using diverse learning algorithms) provides candidate labels for every region in an image; next, using information about the association between labels (estimated through co-occurrence statistics), the energy-based models select the best combination of labels that should be assigned to the image. Figure 1 illustrates the proposed energy-based model to regions; shaded nodes denote the confidence of classifiers in the candidate labels. We consider dependencies between spatially connected regions.

2 http://imageclef.org/photodata

3 http://imageclef.org/SIAPRdata
We report experimental results obtained with the proposed method over several benchmark image collections of heterogeneous characteristics. Table 1 shows the annotation accuracy obtained by our method and the best reported results for each data set (see caption).

Our experimental results show the usefulness of the proposed method: the multiclass classification approach to AIA proved to be very effective (see OVA-RF); the energy-based model improved the initial labeling for all of the considered collections (see EBM, the difference was statistically significant according the Wilcoxon signed-rank test with 95% of confidence); the proposed method outperformed the best results reported in related works where the authors have used the same collections we did (see Ref).

The main benefits of the proposed method are generality of the approach, easiness of its implementation, its effectiveness and high efficiency. Our work on image annotation with the energy-based model is described in detail in [6].

### 4 Semantic Cohesion for Image Retrieval

For image retrieval, we propose methods based on semantic cohesion among labels and text to represent multimodal documents. Specifically, we propose two forms of representing images based on distributional term representations (DTRs) that have been widely used in computational linguistics [9]. Under the considered DTRs, each multimodal term (i.e., a label term or a word term) is represented by a vector of statistics of occurrence over the documents in a collection or co-occurrences over other terms in a vocabulary.

In this way, the representation of a term will be influenced by the documents in which it occurs (capturing dependencies between terms and documents) for the document-occurrence representation (DOR) or by the terms it mostly co-occurs with (capturing dependencies between terms) for the term co-occurrence representation (TCOR). Once each term in the multimodal vocabulary is represented through DTRs, documents are represented by the weighted sum of the DTRs of terms that appear in the document. Intuitively, each document is represented by the context associated with the terms that occur in the document; where the context is given by other documents in the collection or other terms in the multimodal vocabulary.

<table>
<thead>
<tr>
<th>Data set</th>
<th>Ref.</th>
<th>OVA-RF</th>
<th>EBM</th>
</tr>
</thead>
<tbody>
<tr>
<td>C-AN</td>
<td>45.64%</td>
<td>57.90%</td>
<td>58.97%</td>
</tr>
<tr>
<td>C-AG</td>
<td>50.50%</td>
<td>56.56%</td>
<td>57.23%</td>
</tr>
<tr>
<td>C-BN</td>
<td>39.50%</td>
<td>46.64%</td>
<td>48.74%</td>
</tr>
<tr>
<td>C-BG</td>
<td>43.00%</td>
<td>46.08%</td>
<td>46.87%</td>
</tr>
<tr>
<td>C-CN</td>
<td>42.50%</td>
<td>54.13%</td>
<td>55.59%</td>
</tr>
<tr>
<td>C-CG</td>
<td>47.50%</td>
<td>52.28%</td>
<td>52.71%</td>
</tr>
<tr>
<td>SCEF-I</td>
<td>60.94%</td>
<td>59.99%</td>
<td>60.35%</td>
</tr>
<tr>
<td>SCEF-II</td>
<td>78.73%</td>
<td>81.55%</td>
<td>82.92%</td>
</tr>
<tr>
<td>MSRC-1</td>
<td>93.94%</td>
<td>86.60%</td>
<td>88.82%</td>
</tr>
<tr>
<td>MSRC-2</td>
<td>70.50%</td>
<td>70.86%</td>
<td>76.03%</td>
</tr>
<tr>
<td>VOGEL</td>
<td>71.70%</td>
<td>70.78%</td>
<td>72.54%</td>
</tr>
</tbody>
</table>

We report experimental results obtained with the proposed method over several benchmark image collections of heterogeneous characteristics. Table 1 shows the annotation accuracy obtained by our method and the best reported results for each data set (see caption).
We report experimental results of the developed techniques on the SAIAPR TC12 collection. Table 2 compares the retrieval performance of our proposals: multimodal DOR and multimodal TCOR, with unimodal (text-only and labels-only) and standard multimodal techniques (late fusion, early fusion and intermedia relevance feedback), over two sets of topics (ImageCLEF2007 and ImageCLEF2008).

Experimental results obtained with the standard methods show that the combination of labels and text can be helpful for improving the performance of unimodal strategies significantly. However, the proposed representations achieve better performance than the standard techniques. The difference in performance is statistically significant for multimodal DOR according to the pair-wise t-student test with 95% of confidence. Furthermore, the content of multimodal images is better represented with our techniques, when compared to unimodal or standard multimodal strategies.

In summary, we provide evidence showing that the combination of labels and text can be very helpful for image retrieval, and we prove that the proposed representations provide an effective solution to the multimodal image retrieval task. Our developments on multimodal image retrieval with distributional term representations are explained in detail in [7].

<table>
<thead>
<tr>
<th>Topics</th>
<th>ImageCLEF2007</th>
<th>ImageCLEF2008</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method</td>
<td>MAP</td>
<td>P20</td>
</tr>
<tr>
<td>Multimodal-DOR</td>
<td>0.2141</td>
<td>0.2425</td>
</tr>
<tr>
<td>Multimodal-TCOR</td>
<td>0.1935</td>
<td>0.2392</td>
</tr>
<tr>
<td>Late-fusion</td>
<td>0.1348</td>
<td>0.1858</td>
</tr>
<tr>
<td>Early-fusion</td>
<td>0.189</td>
<td>0.2508</td>
</tr>
<tr>
<td>IM-relevance feedback</td>
<td>0.1659</td>
<td>0.2142</td>
</tr>
<tr>
<td>Labels-only</td>
<td>0.0587</td>
<td>0.1417</td>
</tr>
<tr>
<td>Text-only</td>
<td>0.1241</td>
<td>0.1767</td>
</tr>
</tbody>
</table>

5 Conclusions and Future Work

We provided experimental evidence which shows that the idea of semantic cohesion can be effectively exploited for modeling multimodal information. The proposed methods for image annotation and image retrieval based on such idea obtained superior performance than those reported in related papers; furthermore, our techniques offer additional benefits. Thus, we can conclude that the semantic cohesion modeling, and more specifically, that a modeling based on co-occurrence statistics offers important benefits in terms of effectiveness, efficiency and representation power.

Also, the experimental evidence we provided shows that the combination of labels and text can improve the retrieval performance of unimodal methods, even when standard information fusion techniques are used. Despite the latter is highly intuitive, there are no similar works that attempt to combine text and labels, to the best of our knowledge. As future work, we would like to explore application of the energy-based model to similar problems from structured prediction. We also want to include global image information into the energy function. Concerning image retrieval, we would like to explore the use of multimodal DTRs for combining raw-visual features with textual information and the use of DTRs for generating visual vocabularies for object recognition and image categorization.
Acknowledgements

We are grateful to Nando de Freitas, Aurelio López, Eduardo Morales, J. Francisco Trinidad, and Luis Villaseñor for their valuable comments which helped us to improve our work. Also, we thank the TIA research group at INAOE. This work was supported by CONACyT under Project Grant No. 61335 and Scholarship No. 205834.

References


Hugo Jair Escalante obtained his Ph.D. in Computer Science from the Computational Sciences Department of the National Institute of Astrophysics, Optics and Electronics, where he has been a full-time researcher since January 2012. From August 2010 to December 2011, Dr. Escalante had been an assistant professor at the Graduate Program on Systems Engineering of the UANL. He is a Candidate Researcher of the Mexican National System of Researchers (SNI) and author of more than 30 international papers in journals, books, conference proceedings and workshops. Dr. Escalante obtained the “2010 Best Ph.D. Thesis on AI Award” granted by the Mexican Society of Artificial Intelligence for the thesis reported in this paper. His main research interests are machine learning and its applications in natural language processing and high level computer vision.

Luis Enrique Sucar has a Ph.D. in Computing from Imperial College, London, UK, 1992; a M.Sc. in electrical engineering from Stanford University, California, USA, 1982; and a B.Sc. in electronics and communications engineering from ITESM, Monterrey, Mexico, 1980. He is currently the Director of Research at the National Institute for Astrophysics, Optics, and Electronics (INAOE), Puebla, Mexico. He has been an invited professor at the University of British Columbia, Canada; Imperial College, London; and INRIA, France. He has more than 150 publications in journals, books and conference proceedings, and has directed 16 Ph.D. theses. Dr. Sucar is a Senior Member of IEEE, a member of the Mexican National System of Researchers (SNI), and the Mexican Academy of Sciences. He served as the President of the Mexican AI Society and was a member of the Advisory Board of IJCAI. His main research interests are graphical models and probabilistic reasoning, and their applications in computer vision, robotics, and biomedicine.
Manuel Montes-y-Gómez obtained his Ph.D. in Computer Science from the Computing Research Center of the National Polytechnic Institute of Mexico. Currently, he is a full-time researcher at the Computational Sciences Department of the National Institute Astrophysics, Optics and Electronics, Puebla, Mexico. His research is on automatic text processing. He is author of more than 150 international papers in the fields of information retrieval, question answering, information extraction, and text mining. Dr. Montes has been a visiting professor at the Polytechnic University of Valencia (Spain), the University of Geneva (Italy), and the University of Alabama at Birmingham (USA). In addition, Dr. Montes is a member of the Mexican National System of Researchers (SNI), the Mexican AI Society, the Mexican Association for Natural Language Processing, and the International Web Intelligence Consortium.

Article received on 02/06/2010; accepted on 12/01/2012.
**GENERAL INFORMATION FOR AUTHORS**

**Computación y Sistemas (CyS)** is an open access peer reviewed journal published quarterly. Its goal is to present significant scientific and technological advances in the field of computing science and applications. The topics of interest include, but are not limited to, artificial intelligence, intelligent agents, pattern recognition, natural language processing, multimedia and virtual reality, automata and languages, computer theory, systems theory, logic, applied mathematics, automation, distributed and parallel computing, software engineering, databases, operating systems, computer architectures, and digital signal and image processing. Publication in this journal has no cost for authors.

**Publication norms.** Papers submitted to CyS must contain original research and must not be submitted simultaneously to any other journal, conference, workshop, symposium, or any other forum. If the paper has already been presented, submitted or published elsewhere, the authors must notify the Editor-in-chief. Submissions must comply with the rules specified below.

The journal publishes regular papers and reports on successfully defended PhD thesis. Both types of papers are subjected to the same peer reviewing process.

**General guidelines:**
- **Language:** English or Spanish.
- **Word processors:** MS-Word or TeX.
- **Length:** there is no specific page limit: both short and long papers are welcome, provided the quality and novelty of the paper adequately justifies its length.
- **Page size:** US Letter (8.5” x 11”).
- **Margins:** left and right 2.5 cm; top 3.5 cm, bottom 4 cm.
- **Text:** two columns, single space.
- **Font type:** Arial, 10 pt.

**Structure of the paper:**
- **Title:** In English or Spanish, according to the language of the paper. 14 pt in boldface.
- **Authors:** Given name(s) then family name(s), 10 pt.
- **Authors’ affiliation:** 10 pt.
- **Authors’ email:** 10 pt.
- **Abstract / Resumen:** 100–150 words, in English and in Spanish, 9 pt (if you do not speak Spanish, we will translate it for you).
- **Keywords / Palabras clave:** in boldface, 9 pt, both in English and in Spanish.
- **Translated title:** 12pt in boldface italics.
- **Text:** 10 pt, section titles 12 pt in boldface, subsection titles 11 pt in boldface.
- **Photographs, schemes, drawings, tables and figures:** They must be placed within the text (not on separate sheets) in JPG (photos) or PNG/EPS (drawings) format. Photos must be of high quality in black and white, with good contrast.
- **Formulas:** They should be placed in tables without borders, centered, with numbers justified to the right margin. They must not be presented as images.
- **References:** Alphabetically ordered by last name of the author(s). The guide for bibliographic references and bibliographic citations is available on our web site. TeX users must use bibtex and provide the corresponding .bib file.

**Paper submission:**
- **Submissions must include:**
  - One copy of the paper exactly as it should appear in print (including biographies of the authors and their photos),
  - One copy of the paper without authors’ names, biographies, and self-references (for blind review),
  - A file containing names of authors, institution, affiliation, email, phone, and CONACYT CVU number for those who have it.
- Each of these documents must be provided both in source format (Word or TeX with all necessary files) and in PDF.
- All files must be submitted in one ZIP archive via the web submission system: http://cys.cic.ipn.mx.
INFORMACIÓN GENERAL PARA LOS AUTORES

Computation y Sistemas (CyS) es una revista de acceso abierto con revisión por pares, publicada trimestralmente, cuyo objetivo es dar a conocer internacionalmente avances científicos y tecnológicos significativos de la ciencia de la computación, en áreas tales como: inteligencia artificial, agentes, reconocimiento de patrones, procesamiento de lenguaje natural, multimedia y realidad virtual, lenguajes y autómatas, teoría de la computación, teoría de sistemas, lógica, matemáticas aplicadas, automatización, computación distribuida y paralela, ingeniería de software, bases de datos, sistemas operativos, graficación, arquitectura de computadoras y procesamiento digital de imágenes y señales, entre otras. La publicación no tiene costo para los autores.

Normas de publicación. Los artículos que se someten para revisión deben contener resultados inéditos y originales, no haber sido publicados con anterioridad ni haber sido simultáneamente a otra revista científica, congreso, taller, simposio, etc. Si el artículo ha sido presentado, sometido o publicado en alguna otra fuente, deberá informarse al Editor en jefe. Los artículos deben ajustarse a las especificaciones indicadas más abajo.

La revista publica artículos regulares y reportes de las tesis doctorales sustentadas. Ambos tipos de artículos pasan por el mismo proceso de revisión por pares.

Lineamientos generales:
- **Idioma**: Inglés o español.
- **Procesador de texto**: MS-Word o TeX.
- **Extensión máxima**: No tenemos un límite de páginas específico, aceptamos tanto los artículos largos como cortos, si la calidad y la originalidad del artículo justifique su longitud.
- **Hoja**: Tamaño carta (8.5” x 11”).
- **Márgenes**: Izquierdo y derecho 2.5 cm, superior 3.5 cm, inferior 4 cm.
- **Interlineado**: Sencillo.
- **Texto**: Dos columnas.
- **Tipo de letra**: Arial, 10 pts.

Estructura del manuscrito:
- **Título**: En español o inglés, acorde al texto del artículo. En negritas a 14 pts.
- **Autor(es)**: Nombres seguidos de los apellidos (sin abreviaturas), a 10 pts.
- **Adscripción de autor(es)**: a 10 pts.
- **Correo electrónico de autor(es)**: a 10 pts.
- **Resumen / Abstract**: En español e inglés. Entre 100 y 150 palabras, a 9 pts.
- **Palabras clave / Keywords**: en español e inglés, 9 pts.
- **Título traducido**: Si el original está en español debe traducirse al inglés, o viceversa, en cursiva y negritas a 12 pts después del resumen.
- **Texto**: Arial 10 pts. El título de las secciones debe ir en negrita a 12 pts. y el de las subsecciones en negritas a 11 pts.
- **Figuras**: Deben ser de alta calidad, en blanco y negro, en formato JPG (fotografías) o PNG/EPS (dibujos) e incrustadas en el texto de forma que se puedan manipular independientemente.
- **Fórmulas**: Deberán de presentarse en formato de tabla sin bordes, centradas y la numeración de c/u justificada a la derecha. No deben ser imágenes.
- **Referencias**: Ordenadas alfabéticamente por apellido del autor. Utilizar la guía para referencias bibliográficas y para citas bibliográficas disponible en página web de la revista. Usuarios de TeX deben usar bibtex e incluir el archivo de referencias bibliográficas bib.

Envío de artículos:
- Por favor incluya en su envío:
  - Artículo completo tal cual debe aparecer en la revista (incluyendo las biografías de los autores con fotografías).
  - Copia del artículo sin nombres, biografías o referencia alguna de los autores, para los fines de revisión ciega.
  - Un documento que contenga nombres de los autores, institución, dependencia, correo electrónico, teléfono, y si cuenta con CV en CONACyT favor de indicar el número del mismo.
  - Cada uno de estos documentos se debe proporcionar tanto en el formato fuente (Word o TeX con todos los archivos correspondientes) como en el formato PDF.
  - Prepárelos como un archivo zip y súbálos a través del sitio web: http://cys.cic.ipn.mx.