Performance Evaluation of the Covariance Descriptor for Target Detection

Pedro Cortez–Cargill*, Cristobal Undurraga–Rius*, Domingo Mery* and Alvaro Soto*
*Computer Science Department
Pontificia Universidad Católica de Chile
Av.Vicuña Mackenna 4860 (143), Santiago, Chile
Email: pmcortez@uc.cl, caundurr@uc.cl, dmery@ing.puc.cl, asoto@ing.puc.cl

***Abstract*—In computer vision, there has been a strong advance in creating new image descriptors. A descriptor that has recently appeared is the Covariance Descriptor, but there have not been any studies about the different methodologies for its construction. To address this problem we have conducted an analysis on the contribution of diverse features of an image to the descriptor and therefore their contribution to the detection of varied targets, in our case: faces and pedestrians. That is why we have defined a methodology to determinate the performance of the covariance matrix created from different characteristics. Now we are able to determinate the best set of features for face and people detection, for each problem. We have also achieved to establish that not any kind of combination of features can be used because it might not exist a correlation between them. Finally, when an analysis is performed with the best set of features, for the face detection problem we reach a performance of 99%, meanwhile for the pedestrian detection problem we reach a performance of 85%. With this we hope we have built a more solid base when choosing features for this descriptor, allowing to move forward to other topics such as object recognition or tracking.**

***Index Terms*—Region Covariance, target detection.**

I. Introduction

One of the most extraordinary abilities of the human vision is to recognize objects and faces. No matter the angle, size, luminosity or occlusion of the object, the human vision is able, in almost every case, to recognize the object or person. This ability is primordial in many aspects of our lives, for example, without this capacity to recognize faces or facial expression we could not have a satisfactory social life. Given this definition, the next logical step is to design machines or systems that could achieve to imitate this ability automatically, to use them in applications such as vigilance or quality control. Computer vision is a subfield of artificial intelligence. Its main objective is to program machines that could understand or recognize the patterns of a scene or the characteristics of an image. These tasks have been a remarkable challenge that have not yet been achieved. Thanks to the advances in technology and the research conducted in the last few years, have been created many different applications for detection and recognition in varied fields. This includes video-games, driver assistance, video edition, quality control, transit control, vigilance, security, *tracking*, etc. For example: for driver assistance, there are applications that warn the drivers when they are falling sleep, using facial expression recognition [1]; in quality control there are many applications which can determinate if a product is in perfect shape or not, using features such as size, shape, color, etc. of an image [2]; in vigilance and security there are applications which, from a security video, detect strange objects or behaviors (robberies, violence, trespassing, etc.) [3].

Actually, to attain these tasks different techniques are used through which relevant information is obtained of images or videos known as features and descriptors [4]. The features selection is an important step for the detection and recognition of objects. A descriptor must be ideally discriminative, robust, and fast to compute. There is a great variety of descriptors, some of them are focused in being computed faster, meanwhile others, in obtaining as much information as possible. On the other hand, there are algorithms that detect regions of interest and invariants to size, luminosity and perspective; this way only the features of relevant regions are computed, instead of the entire image. This technology is known as *viewpoint invariant segmentation* [5], [6].

In this paper we have defined a methodology which determinates the performances of different covariance matrix built from distinct sets of features. With this we are able to define which ones are the best for detection of objects, faces and pedestrians. First, we obtain a set of images, where we select a specific target that we want to detect. Next, we find, in a search image, the region with the smallest distance to the target region initially selected. This way we define an acceptation threshold, that decides if the object is in the search image or not. We obtain the performance for each set of features used in the creation of the covariance descriptor. Finally when an analysis with the best set of features is performed we get for the face detection problem, a performance of 99%, meanwhile for the pedestrian detection problem, we get a performance of 85%.

The rest of the paper is organized as follows. In section 2 we describe the state of art of the problem; in section 3 we explain the mathematical basics, the hypothesis and the implementation of the problem; in section 4, we present the proposed methodology and the results obtained; and finally

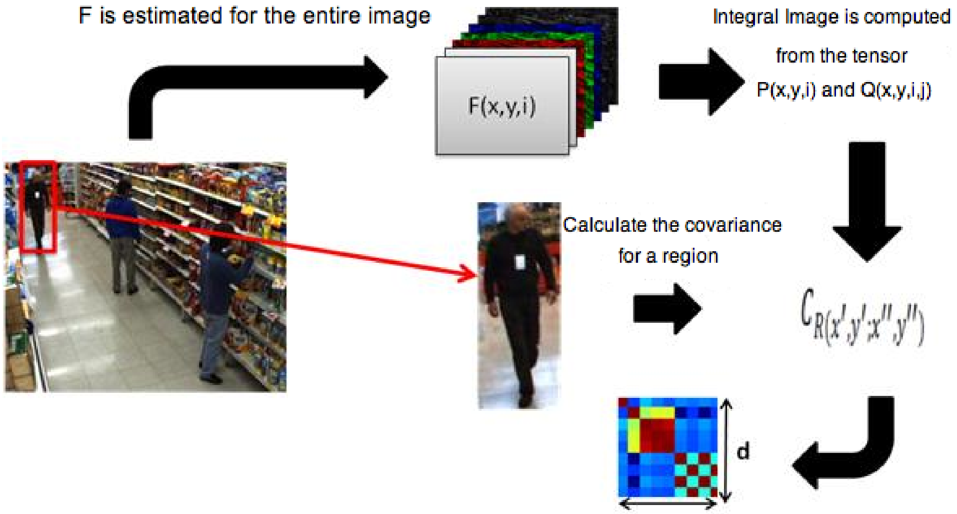


Fig. 1. Computation of the covariance descriptor for a region.

in section 5 we present our conclusions.

II. State of Art

In the object detection field there are many different approaches, one of them is based on features. In this approach we indentify two main tasks. The first one is to extract the features which should give as much information as they can from the object, region or image. The second task is the detection of the object or region through a good classification of the features previously extracted.

The features extraction methods might be classified in basically two main groups, based in their representation of the object. The first group is one that from a detection algorithm for relevant points obtains a set of representative regions, for example: edge and corner detection proposed by Harris et al. [7]; saliency and scale detection [8] or affine invariant regions detection [5] proposed by Kadir et al. More recent methods are no longer using image intensity as a descriptor and are starting to use edges and image gradients in a special context and in different scales such as: SIFT descriptor, proposed by Lowe [9]; shape context descriptors, proposed by Belongie et al. [10]. All these methods base their detection on establishing correspondences between the relevant points obtained from the target image and those obtained from the search image. Many of these algorithms are not robust enough for face and pedestrian detection since they are not invariant to certain transformations such as scale and illumination changes, two big problems to resolve. The most robust of them has shown to be SIFT, which is robust to plane transformations, which is not the case for face and pedestrian detections. The second group is the one that finds an object descriptor inside of a window. The image is heavily analyzed searching for correspondences between an origin window and a search window. Recent studies use as object descriptor: intensity templates proposed by Rowley et al. [11] and Sung et al. [12]; Haar-Wavelets based descriptors, which are a set of base functions that encode visual patterns such as those proposed by Papageorgiu et al. [13]. These methods have resulted quite robust for face detection as the deformations are few and well known. We can see that the latter has been completely demonstrated in this context [14], [15], [16]. But the problem of detecting deformable elements has not been resolved yet. This is why we have seen the need to dig a little deeper.

Recently Porikli et al. proposed [17] a simple solution that integrates multiples features, which are simple and easy to calculate; such as gradient, color, position or intensity, it can even integrate features from infrared or thermal cameras. This descriptor belongs to the second group of methods previously described, where the region or window is represented by the covariance matrix formed from the features of the image. The covariance matrix has been used in various applications and some improvements and additions have been proposed like: Tuzel et al. [18] and Yao et al. [19] proposed to utilize the covariance descriptor plus a LogiBoost classifier for pedestrian detection; Hu et al. [20] proposed to use the particle filter for object *tracking*, weighting the particles with metrics for the covariance matrix; Meer et al. proposed [21] an algorithm to track objects using the covariance descriptor and Lie algebra to create an actualization model. All these innovative contributions try to improve the covariance descriptor, but none of them try to relate the features selection with the problem to solve. In our work we contribute with statistic data about what kind of features are more useful given a specific problem. This way we obtain the real importance of features selection to the descriptor.

III. Proposed Method

A. Theoretical Framework

The covariance descriptor proposed by Porikli et al. [17], is formally defined as:

$$F(x,y,i) = \phi_i(I,x,y) \tag{1}$$

Where I is an image (which can be RGB, black and white, infrared, etc.), F is a $W \times H \times d$ matrix, where W is the image width, H the image height, d is the number of features used and ϕ_i is a function that relates the image with the i-th features, i.e. the function that obtain the i-th feature from the image I. It is important to note that the features are obtained at pixel level (Figure 1).

The goal is to represent the target from the covariance matrix of the F matrix, built from these features. The covariance is the statistical measure of the variation or relation between two random variables, this can be negative, zero or positive, depending the relation between them. In our case the random variables represent the features. The diagonal of the matrix represents the variance of each feature, meanwhile the rest represents the correlation between features.

The use of the covariance matrix as a descriptor has multiple advantages: 1) it unifies both spatial and statistical information of the target; 2) it provides an elegant solution to merge different features and modalities; 3) it has a low dimensionality; 4) it is capable of comparing regions; without being restrained to a fixed window size because it does not matter the region size, the descriptor is of size $d \times d$; 5) it might be easily computed, for all regions.

Regardless the benefits that the representation of a region as a covariance matrix brings, the calculation for any window or region given the image, is computationally prohibitive using conventional methods. Tuzel et al. [22] proposed a computational superior method to calculate the covariance matrix for any rectangular window or region of an image from the formulation of the integral image. The concept of the integral image was initially introduced by Viola et al. [23], for fast computation of Haar features.

Let be P a $W \times H \times d$ matrix, the first order tensor of the integral image

$$P(x',y',i) = \sum_{x<x',y<y'} F(x,y,i) \quad i = 1 \dots d \tag{2}$$

Let be Q a $W \times H \times d \times d$ matrix, the second order tensor of the integral image

$$Q(x',y',i,j) = \sum_{x<x',y<y'} F(x,y,i)F(x,y,j) \tag{3}$$

$$i,j = 1 \dots d$$

Now, let be

$$P_{x,y} = \left[\begin{array}{ccc} P(x,y,1) & \dots & P(x,y,d) \end{array} \right]^T \tag{4}$$

$$Q_{x,y} = \begin{pmatrix} Q(x,y,1,1) & \dots & Q(x,y,1,d) \\ \vdots & \ddots & \vdots \\ Q(x,y,d,1) & \dots & Q(x,y,d,d) \end{pmatrix} \tag{5}$$

It should be noted that the matrix $Q_{x,y}$ is symmetric and to calculate P and Q it takes $d + (d^2 + d)/2$ steps. The complexity of calculating the integral image is $O(d^2WH)$. Figure 2(a) shows that the covariance matrix of a point (x,y) represents the region from the origin to the given point. Figure 2(b) shows graphically that the covariance matrix for any region of the image is calculated as:

$$R_Q = Q_{x',y'} + Q_{x'',y''} - Q_{x'',y'} - Q_{x',y''} \tag{6}$$

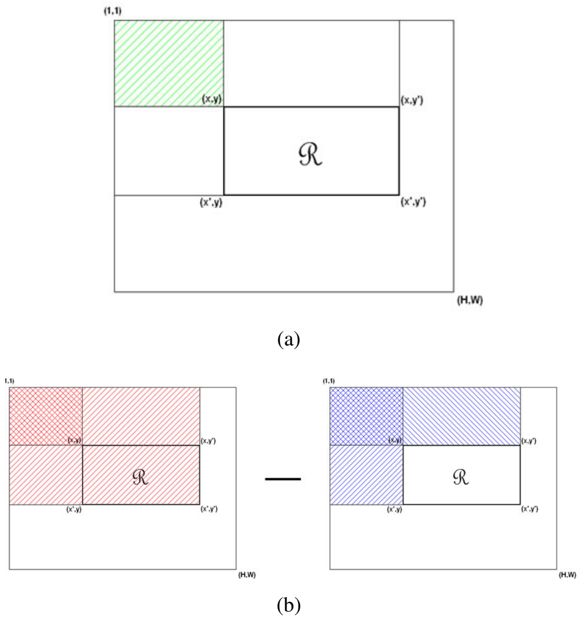
$$R_P = P_{x',y'} + P_{x'',y''} - P_{x'',y'} - P_{x',y''} \tag{7}$$

$$C_{R(x',y';x'',y'')} = \frac{1}{n-1}[R_Q - \frac{1}{n}R_P R_P^T] \tag{8}$$

Where $n = (x'' - x')(y'' - y')$. This way, after building the first order tensor P and the second order tensor Q, the covariance of any region can be computed in $O(d^2)$.

The covariance descriptor does not lie on Euclidean space, therefore we cannot use the classics algorithms of machine intelligence, such as nearest neighbors, Mahalanobis distance, etc. On the other hand, the covariance matrixes are symmetric positive-definite matrixes, which are included in the Lie algebra or the *Riemannian Manifolds* [18]. The Riemannian Manifolds is a mathematical space that on a small enough scale resembles the Euclicean Space. It is a real differentiable manifold in which each tangent space is equipped with an inner product in a way which varies smoothly from point to point. This allows one to define various notions such as angles, length of curves, areas (or volumes), curvatures, gradients, etc. and so generalize the Riemannian Manifolds as a Euclidean space [24].

In our research, to compare two regions from the covariance matrixes, we will use a metric proposed by Forstner et al.



(a)

(b)

Fig. 2. Graphical representation of the calculation of the covariance matrix: (a) given (x, y) point; (b) region R given from points (x, y) and (x', y').

[25]. Which is defined as:

$$\rho(C_1, C_2) = \sqrt{\sum_{i=1}^{n} \ln^2 \lambda_i(C_1, C_2)} \tag{9}$$

Where $\lambda_i(C_1, C_2)_{i=1..n}$ are the generalized eigenvalues of C_1 and C_2 such that,

$$\lambda_i C_1 x_i - C_2 x_i = 0 \quad i = 1 \ldots d \tag{10}$$

This metric satisfies the axioms of the symmetric positive-definite matrixes C_1 and C_2:

$$\rho(C_1, C_2) \geq 0 \tag{11}$$

$$\rho(C_1, C_2) = 0 \Rightarrow C_1 = C_2 \tag{12}$$

$$\rho(C_1, C_2) = \rho(C_2, C_1) \tag{13}$$

$$\rho(C_1, C_2) + \rho(C_1, C_3) \geq \rho(C_2, C_3) \tag{14}$$

To build the matrixes, we will use different color spaces which provide powerful information about the target. There are various color spaces and new spaces can be invented with transformations of the already existing ones. The most common space is RGB (Red, Green, Blue) from which other spaces arise. The CMY color space, which is used on television and there is not a simple conversion between these two. Other spaces are HSL and HSV (Hue, Saturation, Lightness, Value) which are obtain from RGB.

Gevers et al. [26] proposed the new spaces $c_1c_2c_3$ and $l_1l_2l_3$. They also proposed the space $m_1m_2m_3$ which is defined as the relation to a neighbor pixel. We propose to use the mean from the neighbor pixels. Be $R^V G^V B^V$ the means from the neighbor pixels and $R^X G^X B^X$ the values from the evaluated pixel. These are defined in Table I, in Table II we can observe their invariants.

TABLE I
ECUATIONS TO BUILD THE NEW COLOR SPACES.

Color Table	
c_1	$\tan^{-1}(\frac{R}{max(G,B)})$
c_2	$\tan^{-1}(\frac{G}{max(R,B)})$
c_3	$\tan^{-1}(\frac{B}{max(R,G)})$
l_1	$\frac{(R-G)^2}{(R-G)^2+(R-B)^2+(G-B)^2}$
l_2	$\frac{(R-B)^2}{(R-G)^2+(R-B)^2+(G-B)^2}$
l_3	$\frac{(G-B)^2}{(R-G)^2+(R-B)^2+(G-B)^2}$
m_1	$\frac{R^X G^V}{R^V G^X}$
m_2	$\frac{R^X B^V}{R^V B^X}$
m_3	$\frac{G^X B^V}{G^V B^X}$

The goal of this work is to use all these seven color spaces, in different problems.

B. Hypothesis

The problem defined in this work is to find a descriptor efficient enough, fast to compute and with high degree of invariance towards different conditions. The problem arises because most invariant descriptors are bigger and therefore have a larger computacional cost.

This way, we wish to demostrate that for different detection problems of a target we need distinct features to build the covariance descriptor. The descriptor itself is invariant to certain changes of ilumination and scale, but depends heavily on the selected features.

Finally our hypothesis is to demostrate that for different problems, the performance increases when using varied features to build the covariance matrix.

C. Implementation

To achieve the goals previously defined, in a first part, we will focus on implementing in a satisfactory way the covariance descriptor proposed by Porikli et al. [17]. This includes the implementation of: the new method for the covariance matrix calculation for any region proposed by Porikli and Tuzel [22]; the distance between covariance matrixes proposed by Fröstner et al. [25]; and a search algorithm using windows inside an image (Figure 1). All this

TABLE II
RESUME OF COLOR SPACES AND THEIR INVARIANTS TO DIFFERENT CONDITIONS (+ DENOTE INVARIANT - DENOTE SENSITIVE TO THE CONDITION) [26].

	viewing direction	surface orientation	highlights	illumination direction	illumination intensity	illumination color	inter reflection
I	-	-	-	-	-	-	-
RGB	-	-	-	-	-	-	-
rgb	+	+	-	+	+	-	-
S	+	+	-	+	+	-	-
$c_1c_2c_3$	+	+	-	+	+	-	-
H	+	+	+	+	+	-	-
$l_1l_2l_3$	+	+	+	+	+	-	-
$m_1m_2m_3$	+	+	-	+	+	+	+

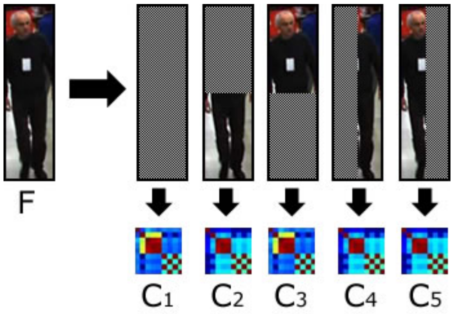
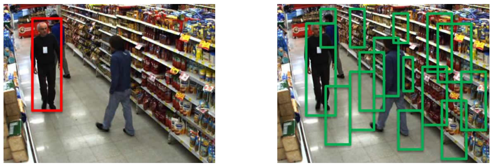


Fig. 3. Occlusion problem can be managed by assigning the distance between descriptors as the shortest distance between the descriptors of each sub-region.



(a) Origin Image

(b) Search Image

Fig. 4. Representation of the origin and search image: (a) Origin image and the target in red. (b) Search image and all the search regions in green.

implementation was realized with the *MATLAB* program.

In order to implement the covariance descriptor of a region, first we make the F matrix from (1), next we obtain the first and second tensor from (2) and (3). Finally we obtain the covariance matrix for any region from (8). The idea is to obtain the region that minimizes the distance between the covariance descriptor, from the origin image and the search image.

In our research, we use two different metrics to measure the similarity between two covariance descriptors. The first one is the metric based on the generalized eigenvalues of two covariance matrices, as defined in (9). The second is a metric that uses the comparison of various subsets of covariance matrices through distance (9). This latter is used to detect pedestrians because the idea is to reduce the occlusion by assigning the distance between descriptors as the shortest distance between the descriptors of each sub-region (Figure 3).

$$\rho(O,T) = \min_j \rho(C_j^O, C_j^T) \qquad (15)$$

where C_j^O is the covariance matrix of the sub-region of the origin region O and C_j^T is the covariance matrix of the sub-region of the search region T.

Next, to detect the target on the search image, we use a brute-force search algorithm (Figure 4), because having calculated the first and second order tensor of the image, we can calculate the covariance descriptor in $O(d^2)$. This way, in the search image, we compare the origin descriptor with 300 new regions or windows randomly chosen through the all image. From these 300 windows, which are of the same size than the origin window, we select the 4 with shorter distance to the origin region. For each of these 4 regions we compare with 30 concentric regions but with different size. From the 120 regions obtained we select the one with shortest distance to the origin region.

Finally, we used eight different matrixes F (based in the proposed colors ont the previous section), to build the covariance descriptors. The matrixes F defined can be observed in Table III. Where R, G and B are Red, Green and Blue; $|I_x|$ is the Intensity first derivative in the x direction, $|I_y|$ is the intensity first derivative in the y direction; $|I_{xx}|$ is the intensity second derivative in the x direction; $|I_{yy}|$ is the intensity second derivative in the y direction; $tan^{-1}(\frac{I_x}{I_y})$ corresponds to the edge orientations.

IV. EXPERIMENTS AND RESULTS

A. Methodology

Before starting to measure the descriptor performance, the test methodology must be defined. First we select the target from the origin image, we will call it "origin region". Next, in other image, we search for the most similar meaning the region of shortest distance to the origin region (Figure 4). We will call this region "search region".

To determine whether an image has or not the origin region, we define a factor k, which establishes an acceptation limit

TABLE III
FEATURES TO FORM THE MATRIXES F.

Matrixes F - Features	
F_1	$\begin{bmatrix} x & y & R & G & B & \|I_x\| & \|I_y\| & \|I_{xx}\| & \|I_{yy}\| \end{bmatrix}$
F_2	$\begin{bmatrix} x & y & H & S & L & \|L_x\| & \|L_y\| & \|L_{xx}\| & \|L_{yy}\| \end{bmatrix}$
F_3	$\begin{bmatrix} x & y & H & S & V & \|V_x\| & \|V_y\| & \|V_{xx}\| & \|V_{yy}\| \end{bmatrix}$
F_4	$\begin{bmatrix} x & y & R & G & B & \|I_x\| & \|I_y\| & \sqrt{\|I_x\|^2 + \|I_y\|^2} & \|I_{xx}\| & \|I_{yy}\| & \tan^{-1}(\frac{I_x}{I_y}) \end{bmatrix}$
F_5	$\begin{bmatrix} x & y & \|I_x\| & \|I_y\| & \sqrt{\|I_x\|^2 + \|I_y\|^2} & \|I_{xx}\| & \|I_{yy}\| & \tan^{-1}(\frac{I_x}{I_y}) \end{bmatrix}$
F_6	$\begin{bmatrix} x & y & c_1 & c_2 & c_3 & \|I_x\| & \|I_y\| & \|I_{xx}\| & \|I_{yy}\| \end{bmatrix}$
F_7	$\begin{bmatrix} x & y & l_1 & l_2 & l_3 & \|I_x\| & \|I_y\| & \|I_{xx}\| & \|I_{yy}\| \end{bmatrix}$
F_8	$\begin{bmatrix} x & y & m_1 & m_2 & m_3 & \|I_x\| & \|I_y\| & \|I_{xx}\| & \|I_{yy}\| \end{bmatrix}$
F_9	$\begin{bmatrix} x & y & R & G & B & H & S & L & c_1 & c_2 & c_3 & l_1 & l_2 & l_3 & \|I_x\| & \|I_y\| & \sqrt{\|I_x\|^2 + \|I_y\|^2} & \|I_{xx}\| & \|I_{yy}\| & \tan^{-1}(\frac{I_x}{I_y}) \end{bmatrix}$

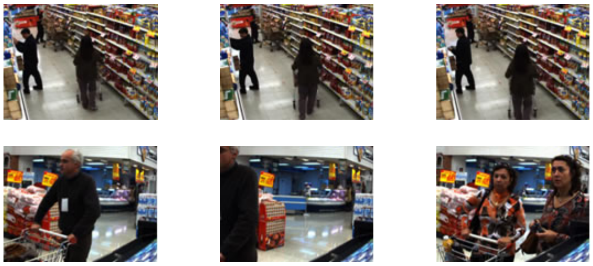


Fig. 5. Video 1 (up): pedestrian or object detection. Video 2 (bottom): face or object detection.

or threshold of the measured distance between the covariance descriptors. Therefore, if the distance between descriptors is greater than k, the search region is not similar enough to the origin region. Meanwhile if it is shorter than k, then the image has the target on it. On the other hand if classifying we obtain a shorter distance than the k factor, but the region is misplaced, we will considered this case as a false positive.

To obtain the results we used two 640×480 videos, rate at 30 frames per second, of a local supermarket (Santiago, Chile) with Point Grey cameras (Figure 5). The first video (corridor video) is use to detect pedestrians or objects, meanwhile the second video (register video) is used for face or object detection. This differentiation is made because the first video has not enough resolution for face detection. Finally, from these videos we obtain two set with 200 images each, where 100 images have the target on it and other 100 images do not.

B. Results

The following results describe the covariance descriptor performance from the different F matrixes previously defined, for each of the image sets.

The F_2 and F_3 features perform exactly the same. The F_8 feature did not provide enough information and therefore we could not calculate accurately the covariance descriptor. The F_9 feature is a set of all the color spaces (except the $m_1 m_2 m_3$) plus the intensity derivatives and the edge orientation; this way we can observe if the color spaces give more information or correlation all together rather than separated. On the other hand, to compare the results all the distances calculated were normalized. The results are summarized on Table IV and Figures 6(a) and 6(b) show the ROC curves for different k values.

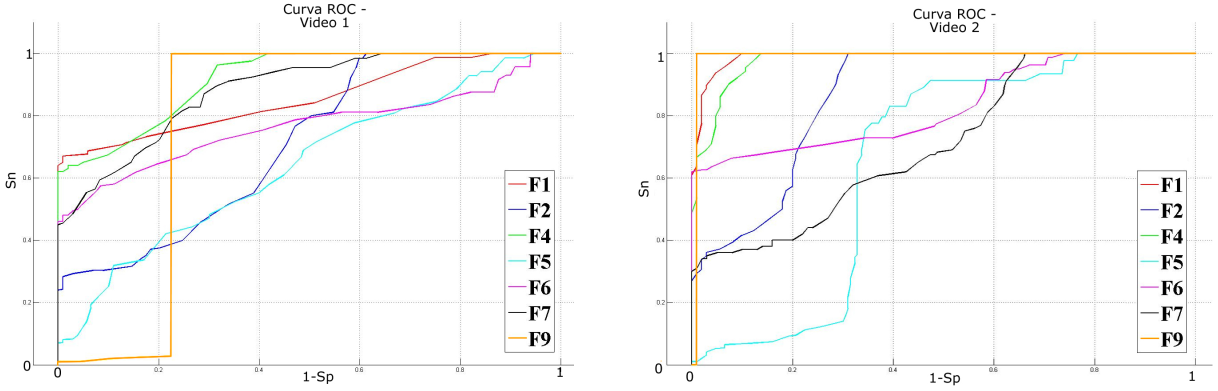
TABLE IV
PERFROMANCE FOR ALL MATRICES F USING THE BEST FACTOR k.

Performance		
Feature	Video1	Video2
F_1	94%	80%
F_2	78%	64%
F_4	92%	83%
F_5	68%	63%
F_6	81%	76%
F_7	66%	78%
F_9	**99%**	**85%**

C. Analysis

From the results we can state that, for the face detection problem the best sets of features are F_9 with 99% efficiency and F_1 with 94% efficiency, meanwhile for the pedestrian detection problem the best sets are F_9 with 85% efficiency and F_4 with 83% efficiency. Also noteworthy is the importance of color spaces regardless of the problem, especially RGB, since the matrix F_5 that does not include any color space had, in both cases, the worst performances.

The set of features F_9 has a higher performance and therefore gives more information or correlation than each set F by itself. This result was expected because using this set we use all the correlation possible between features. Unfortunately, using a matrix of many dimensions makes the computations of tensors P and Q for large images totally



(a) Image Set: Pedestrians

(b) Image Set: Faces

Fig. 6. ROC curves for features F_1 to F_7

contrary to what we seek.

On the other hand, the set of features F_7 shows a high performance, but there was no great discernment between images that had or nor the target, producing a large number of false positives. The sets of features F_1 and F_4 are very similar, but for the pedestrian detection problem the feature $\tan^{-1}(\frac{I_x}{I_y})$ gives more relevant information than for the face detection problem. However, F_4 has higher dimensionality resulting in longer computing times for the tensors. We can see more examples in Figures 7, 8 y 9

All the test were performed in a computer with a Intel Core 2 Duo procesor and 2 Gb Ram. This way, the execution times were of the order of 6 to 7 seconds by processed image using MATLAB. These times are directly proportional to the image size.

V. Conclusions and future works

From the studies presented we can affirm that the covariance descriptor is robust to illumination and shape changes, but has some deficiencies with scale changes. This can be solved normalizing the covariance matrix. It is worth noting the importance of using a color space as feature, especially RGB because it has a high correlation with the intensity gradients of the image.

The best sets of features, for the objects, faces and pedestrians detection are F_9, F_1 y F_4. We must remember that the set of features F_9 is the one that groups all the features used in the other sets and it has a longer computing time. This shows us the importance of features selection to reduce the computing time, and the correlation between them to improve its performance. Surprisingly, any set of features cannot be used because the covariance matrix built might not have enough correlation between features, for example F_8.

It is important to note that this descriptor has a great future since it can unify both spatial and statistical information. This is why we will continue to reduce the execution time, to build a more complete methodology, to obtain more relevant features for specific problems and finally to reduce the tensors dimensions. All this is a preparation to implement a novel efficient tracking system.

Acknowledgments

This work was partially funded by LACCIR (Latin American and Caribbean Collaborative ICT Research) Virtual Institute grant No. R1208LAC005 (http://www.laccir.org) and Anillos-Chile gran No. ADI-32. Additionally, we would like to thank Fernando Betteley from Cencosud-Chile for allowing us to use the facilities of Supermercados Santa Isabel for the acquisition of videos for the present study.

References

[1] Q. Ji, Z. Zhu, and P. Lan, "Real-time nonintrusive monitoring and prediction of driver fatigue," *IEEE transactions on vehicular technology*, vol. 53, no. 4, 2004.

[2] F. P. Edreschi, D. Mery, F. Mendoza, and J. M. Aguilera, "Classification of potato chips using pattern recognition," *Journal of Food Science*, vol. 69, no. 6, pp. 264–270, 2004.

[3] N. T. Nguyen, H. H. Bui, S. Venkatsh, and G. West, "Recognizing and monitoring high-level behaviors in complex spatial environments," in *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings*, 2003, vol. 2.

[4] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 10, pp. 1615–1630, 2005.

[5] T. Kadir, A. Zisserman, and M. Brady, "An affine invariant salient region detector," *Lecture Notes in Computer Science*, pp. 228–241, 2004.

[6] J. Sivic, F. Schaffalitzky, and A. Zisserman, "Efficient object retrieval from videos," in *12th European Signal Processing Conference (EUSIPCO'04)*, 2004.

[7] C. Harris and M. Stephens, "A combined corner and edge detector," in *Alvey vision conference*, 1988, vol. 15, p. 50.

[8] T. Kadir and M. Brady, "Saliency, scale and image description," *International Journal of Computer Vision*, vol. 45, no. 2, pp. 83–105, 2001.

[9] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[10] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 509–522, 2002.

[11] H. A. Rowley, S. Baluja, and T. Kanade, "Neural network-based face detection," in *1996 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1996. Proceedings CVPR'96*, 1996, pp. 203–208.

[12] K. K. Sung and T. Poggio, "Example-based learning for view-based human face detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 1, pp. 39–51, 1998.

[13] C. Papageorgiou and T. Poggio, "A trainable system for object detection," *International Journal of Computer Vision*, vol. 38, no. 1, pp. 15–33, 2000.

[14] S. Baker, I. Matthews, J. Xiao, R. Gross, T. Kanade, and T. Ishikawa, "Real-time non-rigid driver head tracking for driver mental state estimation," in *11th World Congress on Intelligent Transportation Systems*. 2004, Citeseer.

[15] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 681–685, 2001.

[16] D. DeCarlo and D. Metaxas, "Deformable model-based shape and motion analysis from images usingmotion residual error," in *Computer Vision, 1998. Sixth International Conference on*, 1998, pp. 113–119.

[17] O. Tuzel, F. Porikli, and P. Meer, "Region covariance: A fast descriptor for detection and classification," *Lecture Notes in Computer Science*, vol. 3952, pp. 589, 2006.

[18] O. Tuzel, F. Porikli, and P. Meer, "Pedestrian detection via classification on riemannian manifolds," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 10, pp. 1713–1727, 2008.

[19] J. Yao and J. M. Odobez, "Fast human detection from videos using covariance features," in *ECCV 2008 Visual Surveillance Workshop*, 2008.

[20] H. Hu, J. Qin, Y. Lin, and Y. Xu, "Region covariance based probabilistic tracking," in *Intelligent Control and Automation, 2008. WCICA 2008. 7th World Congress on*, 2008, pp. 575–580.

[21] F. Porikli, O. Tuzel, and P. Meer, "Covariance tracking using model update based on lie algebra," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2006, vol. 1.

[22] F. Porikli and O. Tuzel, "Fast construction of covariance matrices for arbitrary size image windows," in *Proc. Intl. Conf. on Image Processing*, 2006, pp. 1581–1584.

[23] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple," in *Proceedings of CVPR2001*, 2001, vol. 1.

[24] W. Rossmann, *Lie Groups: An introduction Through Linear Goups.*, Oxford Press, 2002.

[25] W. Forstner and B. Moonen, "A metric for covariance matrices," *Qua vadis geodesia*, pp. 113–128, 1999.

[26] T. Gevers and A. W. M. Smeulders, "Color-based object recognition," *Pattern recognition*, vol. 32, no. 3, pp. 453–464, 1999.

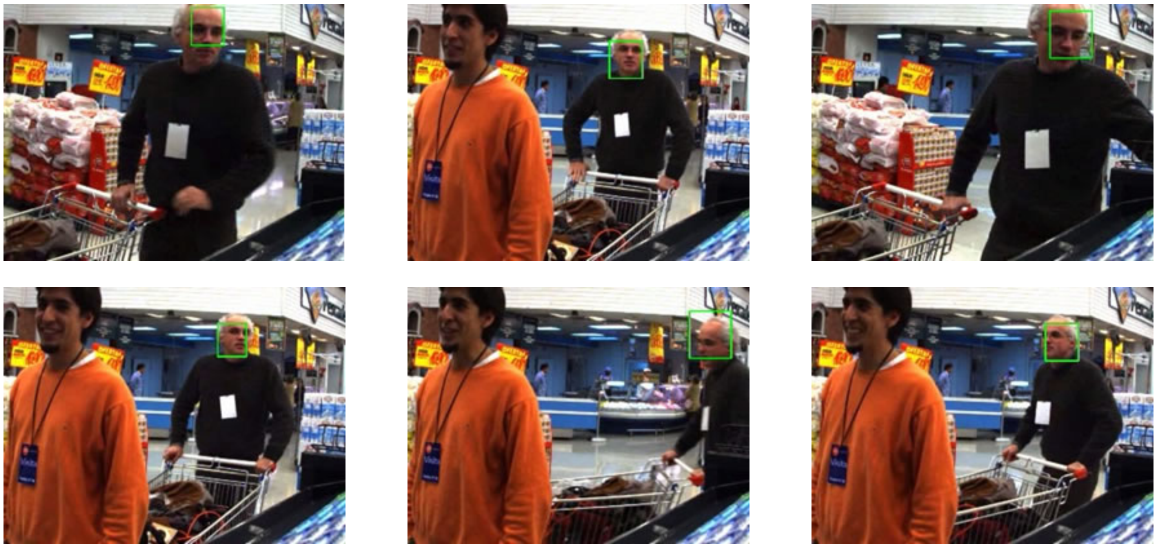


Fig. 7. Face detection for a given region (green rectangle) using F_4.

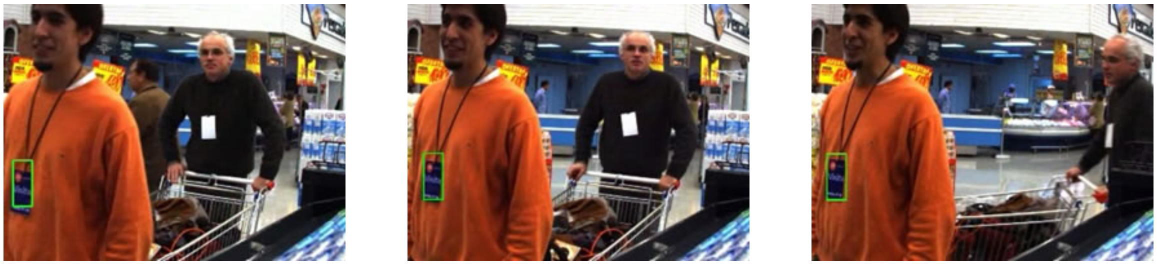


Fig. 8. Object detection for a given region (green rectangle) using F_4.

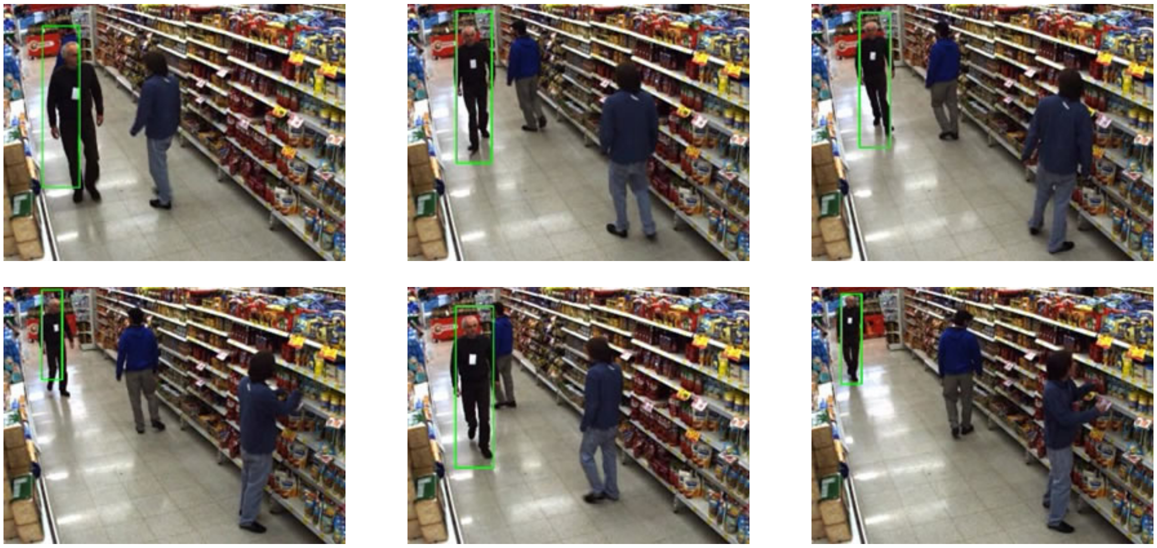


Fig. 9. Pedestrian detection for a given region (green rectangle) using F_4.