Improving Tracking Algorithms Using Saliency

Cristobal Undurraga[1] and Domingo Mery[2]

Computer Science Department, Pontificia Universidad Católica de Chile
Av.Vicuña Mackenna 4860, Santiago, Chile
Email: caundurr@uc.cl[1], dmery@ing.puc.cl[2]

Abstract. One of the challenges of computer vision is to improve the automatic systems for the recognition and tracking of objects in a set of images. One approach that has recently gained importance is based on extracting descriptors, such as the covariance descriptor, because they manage to remain invariant in the regions of these images despite changes of translation, rotation and scale. In this work we propose, using the Covariance Descriptor, a novel saliency system able to find the most relevant regions in an image, which can be used for recognition and tracking objects. Our method is based on the amount of information from each point in the image, and allows us to adapt the regions to maximize the difference of information between the region and its environment. The results show that this tool's improvements can boost trackers precision up to 90% (with initial precision of 50%) without compromising the recall.

Keywords: Saliency, Edge Detector, Tracking, Object Recognition, Covariance Descriptor.

1 Introduction

When recognizing objects in an image there are several approaches to define them. It can be done by: points of interest, descriptors with relevant information [9]; bags of words, areas that define the object [10]; features of a region, variance of the features of the region [11]; local appearence, attention operators based on symetry [7]; among others. One of the methods that has shown good results in object recognition and tracking is based on regions characterized by the *covariance descriptor* proposed by Porikli et al. [11]. This descriptor represents a region or window by a covariance matrix formed from the image's features. Our approach was inspired by the good results obtained in different applications that use this descriptor [2], [12], [14].

One of the problems from tracking algorithms that use one of the descriptors previously described, is chosing the correct window that gives sgnificant information for the recognition. To determine this region, one has to take into consideration the final use of the system. By example, for people tracking, one would choose a face detector such as Voila-Jones [13] or a people detector such as Felzenszwalb [5]. Several trackers and detectors use a rectangular region,

from which arises the problem that within the chosen region is the described object and also the background of the image. This causes that a large amount of information of the region is not from the object, causing a low performance of the tracker. Therefore if the object has little significant infomation and the background has many, the trackers will get confused by considering that the background is more important than the object.

We searched to solve this problem through the quantification of the image's information. This would be acheived by adapting the target region to reduce the noise caused by the background, maximizing the information's contrast between the window and its neighborhood. Currently, to determine if an area or a point contains relevant information, we use saliency systems. Itti et al. [6] present a model for saliency detection, which searches for saliencies in three diferents layers: a color layer, an intensity layer and an orientation layer. Then it linearly combines the zones found by the three layers to obtain the saliency map of the image. On the other hand, Achanta et al [1] present a method to determine salient regions in images using low-level features of luminance and color.

Our saliency method, unlike the method of Itti et al. [6], searches for saliency zones by integrating the color, intensity and orientation layers, and evaluating their covariance on a point through its neighborhood. With this information, we quantify the amount of variation in a pixel allowing us to form a system of saliency. From this we retrieve the saliency map and we define a better region to initialize the tracker, obtaining a high improvement in the precision, from a 50% to a 90%.

In this paper, we present a novel system for the improvement of recognition and tracking algorithms through the quantification of a pixel's information. It's simple and fast. Using the properties of the covariance descriptor to establish the variance of diferents features in a pixel, we found the areas containing the largest amount of information. This article is organized as follows: in section 2 we present the mathematical bases, the hypothesis and the implementation of the problem; in section 3, we present the methodology and the results; finally, in section 4 we present the conclusions and future works.

2 Proposed Method

2.1 Covariance Descriptor

The covariance descriptor for one point proposed by Porikli et al. [11], is formally defined as:

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

where: I is an image which can be in RGB, grayscale, infrared, etc.; x and y are ther coordinates of a pixel; F is a $W \times H \times d$ matrix, where W is the width of the image, H the height and d is the number of features used; and ϕ_i is the function that relates the image to the i-th feature; i.e. the function to get the i-th feature

from de image I. The proposed method uses a 11 characteristic tensor F , which is defined by:

$$F(x, y, i) = [x\ y\ R\ G\ B\ |I_x|\ |I_y|\ \sqrt{|I_x|^2 + |I_y|^2}\ |I_{xx}|\ |I_{yy}|\ \sqrt{|I_{xx}|^2 + |I_{yy}|^2}] \quad (2)$$

For further analysis in the selection of features for the covariance descriptor and the method of calculation of the covariance matrix, we encourage the reader to take a look on the work of Cortez et al. [4].

2.2 Saliency Model

In order to establish the amount of information contained in a pixel, we build the matrix F from (1). The idea is to get the amount of information for a pixel, that's why we define the region of the descriptor as the neighborhood of the pixel. But we need a metric to evaluate the covariance matrix. In our experiments we tested with the largest singular value, with the infinity norm, the determinant and the logarithm of the absolute value of the determinant. The latter gave the best results. Therefore, we define the magnitude of the obtained matrix C_R as the logarithm of the absolute value of the determinant of the matrix. Thus, we define the amount of information I for a pixel (x, y) with a neigborhood N as:

$$S(x, y) = \log(|\det(C_{R(N)})| + 1) \quad (3)$$

2.3 Saliency Region Detector

With the saliency map already obtained, we determined the window where the higher amount of information was concentrated. For this we created an algorithm that reduces the size of a window to maximize the information within it. For a fast calculation we used the same method to calculate the covariance matrix: first, we created the integral matrix of the saliency map I_S; and then, we calculated the information in a region with:

$$I_{S(R)} = I_S(x, y) + I_S(x', y') - I_S(x', y) - I_S(x, y') \quad (4)$$

We defined a line as a rectangle with a side of one pixel long. Then we set the window as the entire image and begun to reduce it. We set a stopping point: defining what percentage of the image's information we wanted to be inside the window. Then, for each side, we calculated how much information they gave. The one that gave less information was reduced, and so on until the region contained the percentage of information previously defined.

2.4 Effectiveness Score

To evaluate tracking algorithms there are two widely used metrics: precision and recall. But having two scores that are almost as important, is a problem. Here is where an other score is needed, that combines both metrics, as it does the

F-score. However, for a further analysis on the tracked path, the precision is more important than the recall, so a $F_{0.5}$-score is advised. Using a variable parameter tends to cause conflict because of the variability of the results when the parameter is altered. So to measure the performance of the tracking algorithms we propose a new score called effective score defined as:

$$E - score = \sqrt[3]{precision^2 * recall} \tag{5}$$

This new score gives us a powerfull tool for choosing the best percentage of information taking into account the precision and the recall. This score was not used to compare us with other methods, but to determine the best percentage for a set of images.

2.5 Automatic determination of the percentage of information

We have discussed an algorithm to choose a better initial region, however, we have to set the parameter of the percentage of information. If we want this to work automatically then we have to establish a method that sets the value of the parameter. For this task we have chosen a bayesian network where the set of variables are: the percentage of information (A_i); the training videos (B_k); the most similar training video (C); and the success for tracking (E). From the joint distribution we have:

$$\arg\max_i P(E|A_iC) \tag{6}$$

After a few arithmetics operations and considering the law of total probabilty we have:

$$\arg\max_i \sum_{k=1}^{m} [P(E|A_iB_k)P(B_k)] \sum_{j=1}^{n} [P(E|A_jC)] \tag{7}$$

where: A_i is a given percentage; and B_k is a video of the training set, where $P(B_k)$ is the probability of the test region to be like the training region given theirs covariance matrices similarities.

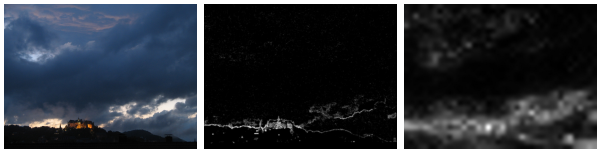
3 Experiments and Results

The aim of the experiments described below is to show the main aspects of our method and then to show a successful application for the improvement of monitoring systems.

3.1 Saliency Region Detector

The goal of this algorithm is to determine if a point is salient or not. For each point of the image, we assign the square region of five pixels as its neighborhood.

From this we obtain a map of saliency using the variation of the features that form the covariance matrix. In comparison with other algorithms, our map is much more visually understandable, since the saliency is for each point and not for an area (Figure 1).



(a) (b) (c)

Fig. 1. Example of the saliency map obtained through the Covariance matrix: (a) original image; (b) our proposed saliency map; (c) Itti's saliency map [6].

In reality, indoor backgrounds are complex and have too much information. That's why we use a similar process, minimizing the information within the window. Thus, we leave most of the background information outside, which produces noise and errors when tracking. By eliminating the sides that contain higher amount of information we reduce the window, which decreases the information within it and maximises the contrast of information between the region and its neighborhood (Figure 2).

(a) (b) (c)

Fig. 2. Example of windows obtained removing sides that contain higher amount of information and leaving 66 percent of it within de window: (a) region obtained from full image; (b) region obtained from image cropped whith a pedestrian detector [5]; (c) region obtained from image cropped whith a face detector [13].

3.2 Improvement of Tracking Algorithms

The application to which we tested our saliency model was tracking algorithms. We experimented with twenty videos obtained from a supermarket, hoping to verify if there was an improvement by applying it at the beginning of a tracking system. In scenarios like the supermarket, where everything is done to attract

Fig. 3. Results obtained on images with lots of information in the background. First row corresponds to the reference result obtained using the ONBNN [3]. The second row is the result where the initial region was modified with our algorithm.

the customer's attention, the saliency algorithms select the background as points of interest and not the person, who tends to have more uniform colors.

As tracking algortihms we used two state of the art methods: the on-line naive bayes nearest neighbor (ONBNN) for covariance descriptors [3], and the TLD real-time algorithm [8]. As initial region we used two different methods: ground truth of a person, to analyse how important is a small amount of background for a tracking system; and a people detector [5], to determine the benefits of the method in a full tracking system. Finally, as saliency algorithms we used three methods: our method, center surround by Achante et al. [1] and Itti et Al algorithm [6].

The results show that saliency algorithms increase the precision of tracking algorithms, however, in some cases, they tend to sacrifice the recall (Table 1). Our method gets higher results in precision without compromising to much recall, or even increasing it. This allows us to obtain better F-Score and E-Score and through an analysis of these scores, we found that better results for ground truth are found using a 70% of the information and for felzenszwalb are found using a 40% of the information. Using better regions, we can prevent that a region gets stuck in the initial position because it doesn't get confused with the background (Figure 3).

Finally, using our bayesian network and choosing the precision as the success of the tracker, we can improve the precision of the ONBNN tracker from a 57.1% to a 98.2% but decreasing the recall. Inversely, if we choose the recall as the success of the tracker we can improve it from a 45.9% to a 56.5% while also increasing the precision to 85.2 %. However, using a Score as the success, we increase the precision to an average of 97.7% while increasing the recall to an average of 49.8% (Table 2). This shows that scores give us powerful information for the improvement of tracking algoritms. They improve the precision and the recall at the same time, reaching high precision levels.

Table 1. Average percentage from twenty videos with initial region ground truth and a people detector, after using a saliency model to adapt the region (100 % means no saliency algorithm was used and *a*uto the autoselection of percentage using the E-Score as the tracking success).

	(%)	Ground Truth						Felzenszwalb					
Tracker		ONBNN			LTD			ONBNN			LTD		
Saliency		Us	C.S.	Itti's	Us	C.S.	Itti's	Us	C.S.	Itti's	Us	C.S.	Itti's
Precision	10	**95.6**	85.9	5.9	46.2	39.6	0.0	**84.7**	82.6	8.7	12.8	15.7	0.0
	20	**92.9**	91.8	10.6	43.7	51.0	0.0	90.4	**91.1**	14.0	25.2	19.9	0.0
	30	**94.3**	92.6	24.7	64.4	52.7	5.9	**88.0**	79.0	22.5	25.8	27.1	0.0
	40	89.0	**92.5**	44.8	68.7	51.5	2.2	**86.6**	78.1	31.8	26.6	58.6	0.0
	50	**88.5**	87.5	53.4	67.1	68.4	18.0	74.4	**79.2**	45.5	23.3	63.4	0.0
	60	85.2	**87.0**	58.7	65.6	67.5	22.3	**72.0**	66.1	50.0	22.6	65.1	0.0
	70	**87.7**	76.7	73.5	62.5	61.3	27.2	70.3	**71.9**	67.7	42.2	69.2	0.0
	80	76.3	74.6	**82.5**	54.7	61.5	28.8	65.4	67.8	**79.7**	57.9	64.2	0.0
	90	78.1	76.1	**84.5**	59.0	59.1	31.9	65.4	62.8	**88.8**	64.8	57.7	0.0
	100		57.1			51.3			60.4			59.2	
	Auto	94.4	-	-	79.6	-	-	85.7	-	-	72.9	-	-
Recall	10	**21.2**	16.0	0.0	15.4	9.8	0.0	**18.3**	16.7	0.0	4.8	3.4	0.0
	20	**26.4**	21.7	0.7	16.0	17.5	0.0	**26.2**	24.5	0.1	9.6	6.1	0.0
	30	**34.0**	30.4	1.2	25.5	19.2	0.0	**31.9**	30.1	0.5	9.4	10.6	0.0
	40	**38.7**	37.6	2.0	30.2	24.5	0.1	**37.0**	34.7	0.9	13.2	18.6	0.0
	50	**44.1**	39.2	3.4	30.9	28.0	1.4	38.4	**38.7**	1.8	13.7	26.7	0.0
	60	43.8	**45.2**	6.3	34.3	36.8	2.4	**40.6**	36.5	3.1	19.6	28.8	0.0
	70	**47.9**	43.5	8.1	33.5	33.0	3.1	**41.7**	41.7	5.5	28.9	32.8	0.0
	80	45.3	**46.7**	12.7	34.8	32.7	3.1	41.5	**43.3**	8.2	41.1	33.4	0.0
	90	**49.1**	48.2	16.2	34.4	32.7	3.7	**44.2**	41.3	12.8	52.3	32.5	0.0
	100		45.9			38.1			45.6			48.2	
	Auto	50.4	-	-	40.2	-	-	52.4	-	-	42.2	-	-

4 Conclusions

To improve the tracking systems we have developed a novel saliency model that uses the covariance descriptor. This saliency system allows us to determine whether an object will be easy or difficult to follow in a video, given that the background contains more- or less- information than the object, and to extract enough information to improve recognition and tracking systems.

We could also improve the initial regions coming from detectors, thus reducing the noise produced by the backgrounds. Although we do not always see improvements in the tracking systems, we could improve the results in cases where they may have failed, keeping the same performance in other cases.

We also propose the use of a bayesian network to efficiently select the best initial region. This allows us to choose if we use, or not, saliency to improve the tracker. We noticed, that bigger regions have less problems to be tracked so, it's more efficient not to use the saliency. However, smaller regions has several

Table 2. Average percentage using the bayesian network to automatically select the best percentage of information. We compare the use of Precision, Recall, F-Score and E-Score as succes of the ONBNN tracker using ground truth for initial region.

	Ground Truth				Felzenszwalb			
	Precision	Recall	F-Score	E-Score	Precision	Recall	F-Score	E-Score
Normal	57.1	45.9	50.0	52.2	60.4	45.6	55.9	54.5
Precision	98.2	29.0	61.3	62.9	92.9	27.8	61.0	61.1
Recall	85.2	56.5	74.1	72.5	81.3	54.4	72.3	70.2
F-Score	95.0	49.2	78.1	75.3	85.7	52.4	74.2	71.8
E-Score	94.4	50.4	77.9	75.4	85.7	52.4	74.2	71.8

problems to be tracked so, it's highly recommended to perform our saliency algorithm to improve the tracking results.

Acknowledgements

The authors would like to thank Fernando Betteley from Cencosud S.A. for facilitate the adquisition of videos in one of Santa Isabel's Supermarkets. Part of this work was done while C.U. was at The National Institute of Astrophysics, Optics and Electronics supported by the LACCIR Short Stays Program. This research was supported in part by LACCIR project #S1009LAC006. This work was supported in part by The School of Engineering, Pontificia Universidad Catolica de Chile, Grant FIA.

References

1. Achanta, R., Estrada, F., Wils, P., Süsstrunk, S.: Salient region detection and segmentation. Computer Vision Systems pp. 66–75 (2008)
2. Batista, J.: A region covariance embedded in a particle filter for multi-objects tracking. Update (2008)
3. Cortez, P., Mery, D., Sucar, L.: Object Tracking Based on Covariance Descriptors and On-Line Naive Bayes Nearest Neighbor Classifier. In: 2010 Fourth Pacific-Rim Symposium on Image and Video Technology. pp. 139–144. IEEE (2010)
4. Cortez, P., Undurraga, C., Mery, D., Soto, A.: Performance evaluation of the Covariance descriptor for target detection. In: 2009 International Conference of the Chilean Computer Science Society. pp. 133–141. IEEE (2009)
5. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE transactions on pattern analysis and machine intelligence 32(9), 1627–45 (Sep 2010)
6. Itti, L., Koch, C., Niebur, E.: A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. Colorectal disease : the official journal of the Association of Coloproctology of Great Britain and Ireland 4(2), 147–149 (Mar 2002)
7. Jugessur, D., Dudek, G.: Local appearance for robust object recognition. cvpr (2000)

8. Kalal, Z., Matas, J., Mikolajczyk, K.: Online learning of robust object detectors during unstable tracking. 2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops pp. 1417–1424 (Sep 2009)
9. Lowe, D.: Object recognition from local scale-invariant features. Proceedings of the Seventh IEEE International Conference on Computer Vision pp. 1150–1157 vol.2 (1999)
10. Nowak, E., Jurie, F., Triggs, B.: Sampling strategies for bag-of-features image classification. ECCV pp. 490–503 (2006)
11. Tuzel, O., Porikli, F., Meer, P.: Region covariance: A fast descriptor for detection and classification. ECCV pp. 589–600 (2006)
12. Tuzel, O., Porikli, F., Meer, P.: Pedestrian detection via classification on Riemannian manifolds. IEEE transactions on pattern analysis and machine intelligence 30(10), 1713–27 (Oct 2008)
13. Viola, P., Jones, M.: Robust real-time object detection. International Journal of Computer Vision (2002)
14. Yao, J., Odobez, J.m., Parc, C.: Fast Human Detection from Videos Using Covariance Features. Learning (2008)