Improving the Segmentation Stage of a Pedestrian Tracking Video-based System by means of Evolution Strategies*

O. Pérez, M.A. Patricio, J. García, and J.M. Molina

Universidad Carlos III de Madrid. Computer Department,
Avenida de la Universidad Carlos III, 22 Colmenarejo 28270. Madrid. Spain.
{opconcha,mpatrici,jgherrer}@inf.uc3m.es, molina@ia.uc3m.es,
WWW home page: http://www.giaa.inf.uc3m.es/

Abstract. Pedestrian tracking video-based systems present particular problems such as the multi fragmentation or low level of compactness of the resultant blobs due to the human shape or movements. This paper shows how to improve the segmentation stage of a video surveillance system by adding morphological post-processing operations so that the subsequent blocks increase their performance. The adjustment of the parameters that regulate the new morphological processes is tuned by means of Evolution Strategies. Finally, the paper proposes a group of metrics to assess the global performance of the surveillance system. After the evaluation over a high number of video sequences, the results show that the shape of the tracks match up more accurately with the parts of interests. Thus, the improvement of segmentation stage facilitates the subsequent stages so that global performance of the surveillance system increases.

1 Introduction

Surveillance systems are usually made up by several interconnected processing blocks or stages that form a high-level representation of the sensed world. The optimization of a video surveillance system consists of improving the particular performance of a stage of the system by adding new computations and adjusting the parameters which run this stage, so that the whole system increases its global performance. In [1], the authors showed how to construct and tune a multi-stage video surveillance system to obtain a good performance in the tracking of aircraft and vehicles moving in an airport surface [2].

In this work, we adapt the system to track people based on the same architecture of the tracking system for surface surveillance in airports. The first new problem that arises from this adaptation is that the parts of interest or blobs appear more fragmented as people are less compact (especially for the extremities) than aircraft or vehicles [3]. Second, one of the main drawbacks of outdoor motion estimation is shadows [4] - [6], which attached to the moving

* Funded by CICYT (TIC2002-04491-C02-02)

people make the system deform the real target. Moreover, shadows might [7] interferer in subsequent stages of the tracking system so that it would be desirable to remove them in a previous phase. The purpose of this study is to improve the performance of this segmentation stage by adding a new block so that the system detects more compact people-shaped blobs and eliminates if possible the shadows so that the total performance of the whole system increases. The parameters that regulate this new block will be searched and tuned by an Evolution Strategy (ES), which has proved to be valid for this type of problems in works like [1]. The main goal of this work is to present based upon the idea that most morphological image analysis tasks, can be reframed as image filtering problems and that ES can be used to discover optimal filters which solve such problems. This paper also introduces the fitness function that assesses the performance of the segmentation block [8, 9]. Finally, we must test the improvement on the complete system for which we need an evaluation function. Although there is a large literature and previous works on metrics for performance evaluation [1], [10]-[12], this paper shows an original proposal based on a minimal *ground truth* record and it is able to evaluate a large number of video samples to obtain significants statistical results.

This paper attempts to address these points. First, section 2 presents a study of the segmentation stage in our video surveillance system. Section 3 details the main problems that we face on tracking people and the solutions adopted. Then, the evaluation function to assess the global performance of the system is presented in Section 4. The details of the experiments and the final conclusions are given in Sections 6.

2 Segmentation Stage

Automated visual surveillance aims to provide an attention-focussing filter to enable an operator to make an optimum decision whenever an unusual event occurs. This is achieved by directing the operators attention only to those events classified as unusual. A generic video processing framework for automated visual surveillance system [13, 14] is depicted in Figure 1. Although some stages require interchange information with others, this framework provides a good structure for the comprehension of our work.

A relevant problem in computer vision is the detection and tracking of moving objects in video sequences. The detection of moving objects can be difficult for several reasons. We need to account for possible motion of the camera, changes in illumination of a scene and shadows, objects such as waving trees, objects that come to a stop and move again such as vehicles at a traffic light, etc. Once the moving objects have been identified, tracking them through the video sequence can also be difficult, especially when the objects being tracked are occluded by buildings or moved in and out of the frame due to the motion of the camera.

By Segmentation Stage, we mean the task of detecting regions that correspond to moving objects such as people and vehicles in video. This is the first basic step of almost every vision system since it provides a focus of attention and

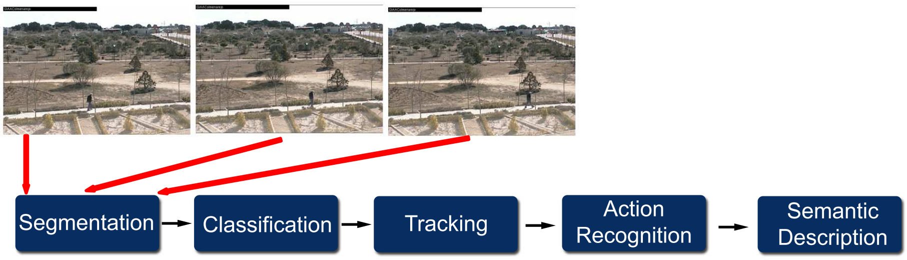


Fig. 1. A generic video processing framework for automated visual surveillance system.

simplifies the processing on subsequent analysis steps. As we have said above, due to dynamic changes in natural scenes such as sudden illumination, shadows, weather changes, motion detection is a difficult task to process reliably.

Frequently used techniques for moving object detection are background subtraction, statistical methods, temporal differencing and optical flow. Most of the moving object detection techniques are pixel based [16, 17]. Background subtraction techniques attempt to detect moving regions by subtracting the current image pixel-by-pixel from a reference background image that is created by averaging images over time in an initialization period. The pixels whose difference exceeds a threshold are classified as foreground. Although background subtraction techniques perform well at extracting most of the relevant pixels of moving regions, they are usually sensitive to dynamic changes when, for instance, repetitive motions (tree leaves moving in windy day, see Figure 2), or sudden illumination changes occur.

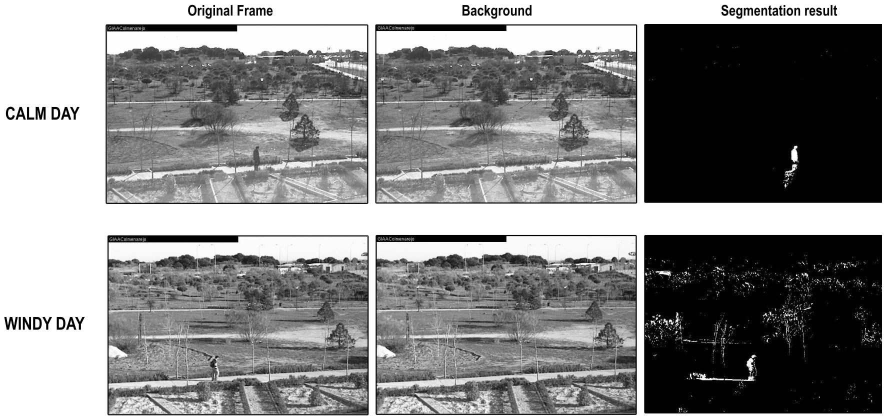


Fig. 2. Different segmentation results obtained in different condition. The first row shows the excellent segmentation results in a calm day. However in the second row, due to the tree leaves in a windy day, we observe brightness changes almost everywhere in the image. Thus, the segmentation stage obtains worse performances.

More advanced methods that make use of the statistical characteristics of individual pixels have been developed to overcome the shortcomings of basic background subtraction methods. These statistical methods are mainly inspired by the background subtraction methods in terms of keeping and dynamically updating statistics of the pixels that belong to the background image process. Foreground pixels are identified by comparing each pixels statistics with that of the background model. This approach is becoming more popular due to its reliability in scenes that contain noise, illumination changes and shadow [14, 18]. Temporal differencing attempts to detect moving regions by making use of the pixel-by-pixel difference of consecutive frames (two or three) in a video sequence [19].

2.1 Background subtraction

In our system we have implemented a background subtraction approach [20]. The segmentation algorithm is based on the detection of targets contrasting with local background, whose statistics are estimated and updated in an auxiliary image, Background. Then, the pixel level detector is able to extract moving features from this static background, simply comparing the difference with a threshold:

$$Detection(x,y) = [Im(x,y) - Back(x,y)] > THRESHOLD * \sigma \tag{1}$$

where σ represents the standard deviation of pixel intensity. A low threshold would mean a higher sensitivity value, leading to many false detections and higher probability of detection and not corrupting target shape quality. This is one of the key parameters of the system. The background statistics (mean and variance) for each pixel are estimated, from the sequence of previous images, with a simple iterative process and weights to give higher importance to the most recent frames. Besides, in order to prevent targets from corrupting background statistics, the update is just performed for pixels not too near of a tracked target, using the tracking information in the detector. So, the statistics for k-th frame are updated as:

$$\begin{array}{l} Back(x,y,k) = \alpha Im(x,y,k) + (1-\alpha)Back(x,y,k-1) \\ \sigma^2(x,y,k) = \alpha[Im(x,y,k) - Back(x,y,k-1)]^2 + (1-\alpha)\sigma^2(x,y,k-1) \end{array} \tag{2}$$

being x and y pixels out of predicted tracks.

In Figure 3 some segmentation results are depicted following this approach.

2.2 Morphological post-processing

As we can see in Figure 3, the last step (labelled as 'Segmentation result') obtains excellent results. However, it seems obvious that we can improve the segmentation stage. A pedestrian zoom views of the Figure 3 are depicted in Figure 4. The white pixels make up the pedestrian and set up the foreground pixel map,

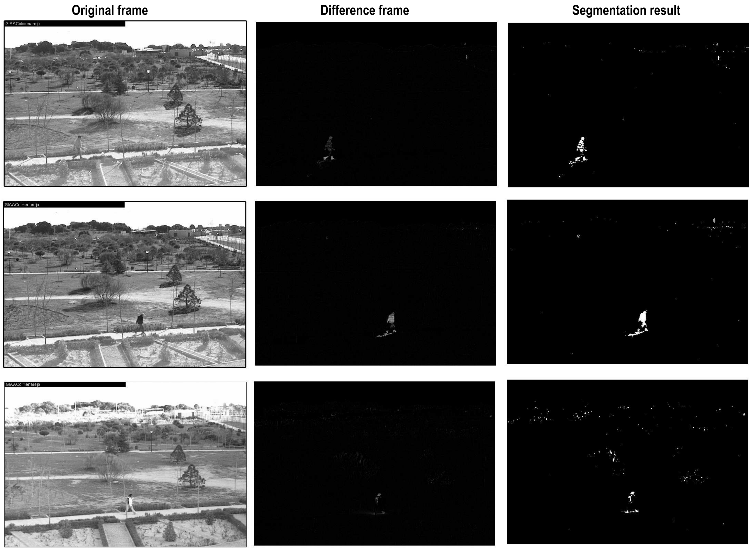


Fig. 3. Instances of the segmentation stage. Although the results are good enough (third column), notice that in the third row, the object detected is rather difficult to track.

in which there are unconnected and missing areas. Furthermore, in all over Figure 3 there is a lot of noise which can confuse later processing. The goal of the segmentation stage is not only to produce foreground pixel maps as accurately as possible, e.g. by removing the special types of noise, but rather to make the pedestrians segmentation more visible and easier to process in the classification stage (see Figure 1).

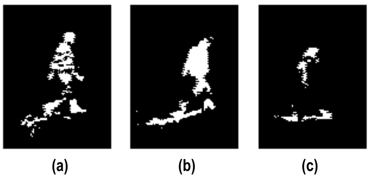


Fig. 4. A pedestrian zoom views of Figure 3. It is clear that we can improve the segmentation stage. In the images appear unconnected and missing areas.

In order to improve segmentation results, morphological operators have been implemented. The field of mathematical morphology contributes a wide range of operators to image processing, all based around a few simple mathematical concepts from set theory. Morphology is a broad set of image processing operations that process images based on shapes. Morphological operations apply a structuring element to an input image, creating an output image of the same size. The most basic morphological operations are dilation and erosion. In a morphological operation, the value of each pixel in the output image is based on a comparison between the corresponding pixel in the input image and its neighbors. By choosing the size and shape of the neighborhood, a morphological operation can be tuned to be sensitive to specific shapes in the input image. In our case, an erosion operator has been chosen as first post-processing step in order to remove the noise. Then, we apply a delation operator to improve the size and shape of the pedestrian.

Now, our problem is concerned with the selection of the size of the suitable structuring element and the number of iterations of the erode and dilate operations. We define the rectangular size of structuring elements and the number of iteration of erosion and dilate process by the next parameters: HORIZONTAL-SIZE-ERODE, VERTICAL-SIZE-ERODE, HORIZONTAL-SIZE-DILATE, VERTICAL-SIZE-DILATE, ITERATIONS-NUMBER-ERODE and ITERATIONS-NUMBER-DILATE. Besides, we have to establish another parameter involving in the segmentation stage: the THRESHOLD in Equation 1. The election of the values of these parameters makes a big difference in the performance of the system. Thus, in the next section, we show how to use Evolution Strategies in order to optimize these parameters.

3 Optimizing morphological parameters by means of Evolution Strategies

Evolutionary Computation (EC) comprises several robust search mechanisms based on underlying biological metaphor. Having been established as a valid approach to problems requiring efficient and effective search, EC are increasingly finding widespread application in business, scientific and engineering circles. Not much work has been applied in automatic visual surveillance systems using EC. Perhaps the main trouble is related with the enormous amount of data to process. In [21] genetic programming is used to segment video sequences. Hwang [22] shows a genetic algorithm which uses both spatial and temporal information to segment and track moving objects in video sequences. In [1], an Evolution Strategy (ES) for optimizing the parameters regulating a video-based tracking system is presented.

We have implemented ES for improving the segmentation stage by adjusting the parameters listed above. Regarding the operators, the type of crossover used in this work is the discrete one and the replacement scheme which is used to select the individuals for the next generation is $(\mu + \lambda) - ES$.

In an ES, the fitness is a function that gives a *score* to the outcome of the system and its design is probably the most critical task concerning both the domain problem and the ES itself. In fact, it must be based on the foreground pixel map's features and most of the parameters within this domain algorithm could affect the outcome of the segmentation stage.

After the morphological post-processing of an image, its foreground pixel map consist of several blobs (i.e. coherent connected regions). In order to simplify the process, we represents the blobs by its bounding rectangle. Let NB be the Number of Blobs in a foreground pixel map. In our experimentation we have been working with videos where there is only a pedestrian, and therefore we expect to found a short number of blobs in our ideal segmentation stage.

Let Im and $\widehat{Im}$ be the image before and after the morphological post-processing, respectively. We define $Im(x, y)$ and $\widehat{Im}(x, y)$ as `true`, if and only if the pixel (x, y) belongs to a moving object, respectively. We define the Density ratio, $D(B)$, of a blob, B, as:

$$D(B) = \frac{1}{n} \quad Card\{Im(x, y) \quad \wedge \quad \widehat{Im}(x, y)\}; \qquad \forall (x, y) \in B. \tag{3}$$

where n is the number of pixels in the blob B and *card* stands for the cardinality (i.e. number of pixels) of a set. The operator $'\wedge'$ (*and*) is applied to assess which part of the processed image contains detected pixels in the original image.

Let $AR(B)$ the Aspect Ratio of a blob, B. A blob is represented by its bounding rectangle. We define $AR(B)$ as:

$$AR(B) = \frac{width(B)}{height(B)} \tag{4}$$

where *width* and *height* stands for the bounding rectangle's width and height of a blob, respectively. Since, in our system, pedestrians are the object that we have to track, in contrast of shadows or noise, we expect to get a small value for the AR(B) ratio in every blob.

At last, the fitness function that we have to minimize is:

$$fitness \; = \; \alpha NB \; + \; \beta \sum_{\forall B \in \widehat{Im}} AR(B) \; - \; \gamma \sum_{\forall B \in \widehat{Im}} D(B) \tag{5}$$

where α, β and γ are normalization coefficients.

4 Evaluation System

The main requirement for surveillance systems is the capability for tracking objects of interests in operational conditions, with satisfactory levels of accuracy and robustness. The difficult task is the definition of an automatic, objective and detailed procedure able to capture the global quality of a given system in order to support design decisions based on performance assessment. There are many studies that evaluate video surveillance systems against the ground truth or with synthetic images. Our contribution is a new methodology to compute detailed evaluations based on a minimal amount of hand-made reference data and a large quantity of samples. The result is a robust assessment based on a statistical analysis of a significant number of video sequences. Thus, our work used the proposed evaluation system to assess the surveillance system and check the increase of the total performance.

4.1 Basis of the Evaluation System

The system requires as reference a function *f(x,y)* (it could be a function defined on parts) that describes as well as possible the mean track followed by the targets we want to track. In this case, we have recorded a set of video sequences of people walking along a footpath. Our set of samples was divided into two groups: (1) 50 video sequences of people moving from right to left along a footpath, and (2) 50 video sequences of people moving from left to right along a footpath.

Thus, the subsequent assessment was separated into two steps and we obtained two sets of results for each one of the video sequences.

The function *f(x,y)* that approximates the objects' trajectory was very simple in both cases. It was a straight line that was considered the ground truth for the system.

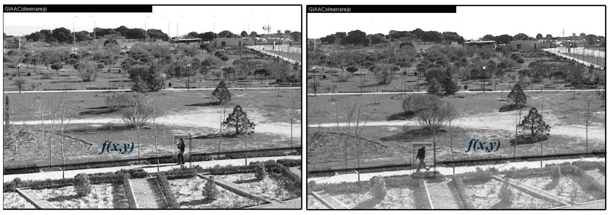


Fig. 5. Video shot samples from the two sets of sequences and the function f(x,y) that approximates the trajectory of each pedestrian.

4.2 Evaluation Metrics

This section explains the core of the evaluation system and how it worked in our experiments. The evaluation system collected the tracks that were given by the tracking system for each frame in all the video sequences. Then, a distance to the reference function *f(x,y)* was computed so that only the tracks whose distance was less than a given margin were considered for the subsequent assessment. The set of metrics considered for this particular problem are listed below:

Absolute Area. It is computed by calculating the area of the detected track.

Transversal Error. It is defined as the distance between the center of the track and the segment which is considered as ground truth in this moment.

Continuity Faults. This metric checks if a current track existed the previous moment or did not. If the track did not exist, it means that this track was lost by the tracking system and recovered in a subsequent frame. This behavior must be computed as continuity fault. This continuity metric is a counter where one unit is added each time a continuity fault occurs.

Changes of Direction. This metric marks when a track changes its direction. This metric is also a counter where one unit is added each time a change of direction occurs.

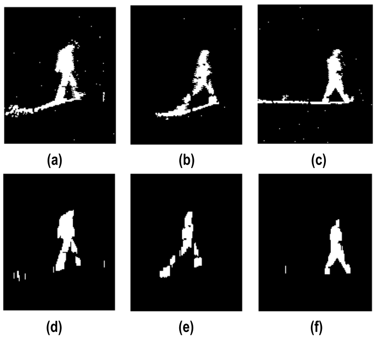


Fig. 6. Segmentation results before (from (a) to (c)) and after (from (d) to (f)) the morphological post-processing.

5 Methodology and Results

Our general procedure for processing the video sequences was as follows:

1. Take a set of 5 random videos from the two video sequences groups.
2. Use the evolution strategies for adjusting the parameters of the morphological operators added to the segmentation stage. We implemented ES with a size of 10+10 individuals. This population is the minimum that assures the same result as if we had taken a higher number of individuals. The mutation factor of $\triangle\sigma = 0.5$ and the initial seed was fixed at 100.
3. Repeat the experiment with at least three different seeds.
4. If the results are similar, fix the parameters of the morphological algorithms for using them in all videos.
5. Take one video sequences set and the parameters obtained by the evolution strategy. Make the surveillance system work and collect all the people's tracks for each frame of each video sequence.
6. Evaluate these tracks and compare the results (with and without morphological algorithms in the segmentation stage).
7. Repeat the process for the second set of videos sequences from step 5.

In order to compare the effect of the morphological operators, we show some pictures before and after the application of the algorithms (Figure 6). The results of the optimization parameters are shown in Table 1. We can observe that the shadows and noises were removed so that subsequent stages of the surveillance system created more appropriate tracks according to the parts of interests. That is, the results had a real correspondence between the people we were interested in and the resulted tracks of the tracking system. This affected directly the track size, which was smaller as a consequence of the shadow elimination. This effect is displayed in the Table 2.

Finally, the effect on the whole surveillance system is showed in Figure 7. In order to have a more detailed idea of the system performance, the area under

Table 1. Optimization results. Notice that the structuring element shape rewards high and thin objects according to the pedestrians' shape.

HORIZONTAL-SIZE-ERODE	1
VERTICAL-SIZE-ERODE	4
HORIZONTAL-SIZE-DILATE	1
VERTICAL-SIZE-DILATE	4
ITERATIONS-NUMBER-ERODE	2
ITERATIONS-NUMBER-DILATE	2
THRESHOLD	15

Table 2. Numerical statistics of the Absolute Area and Transversal Error.

	Before morphological operators			After morphological operators		
	Mean	Max	Min	Mean	Max	Min
Absolute Area	7033	44890	111	3057.7	25636	203
Transversal Err.	10.5	49.5	0.009	7.8	45.15	0.00055

study is divided into 10 zones. Each zone is defined as a fixed number of pixels of the x-axis, the 10% of the horizontal size of the image. The absolute area and the transversal error show the mean, variance and maximum values for each of these two metrics.

All the metrics presented a remarkable improvement on the behavior of the total surveillance system. The absolute area decreased its mean value from 7033 to 3057.7 (see Table 2 and Figure 7(a) and 7(b)) due to the better adjustment of the tracks to the pedestrian shape. Second, the transversal error improved from a mean value of 10.5 to 7.8, which means that the gravity center of the people's track is closer to the ground truth function $f(x, y)$. Moreover, the last figures show that the number of losses for the tracks and the changes of direction decreased by a factor of 2.

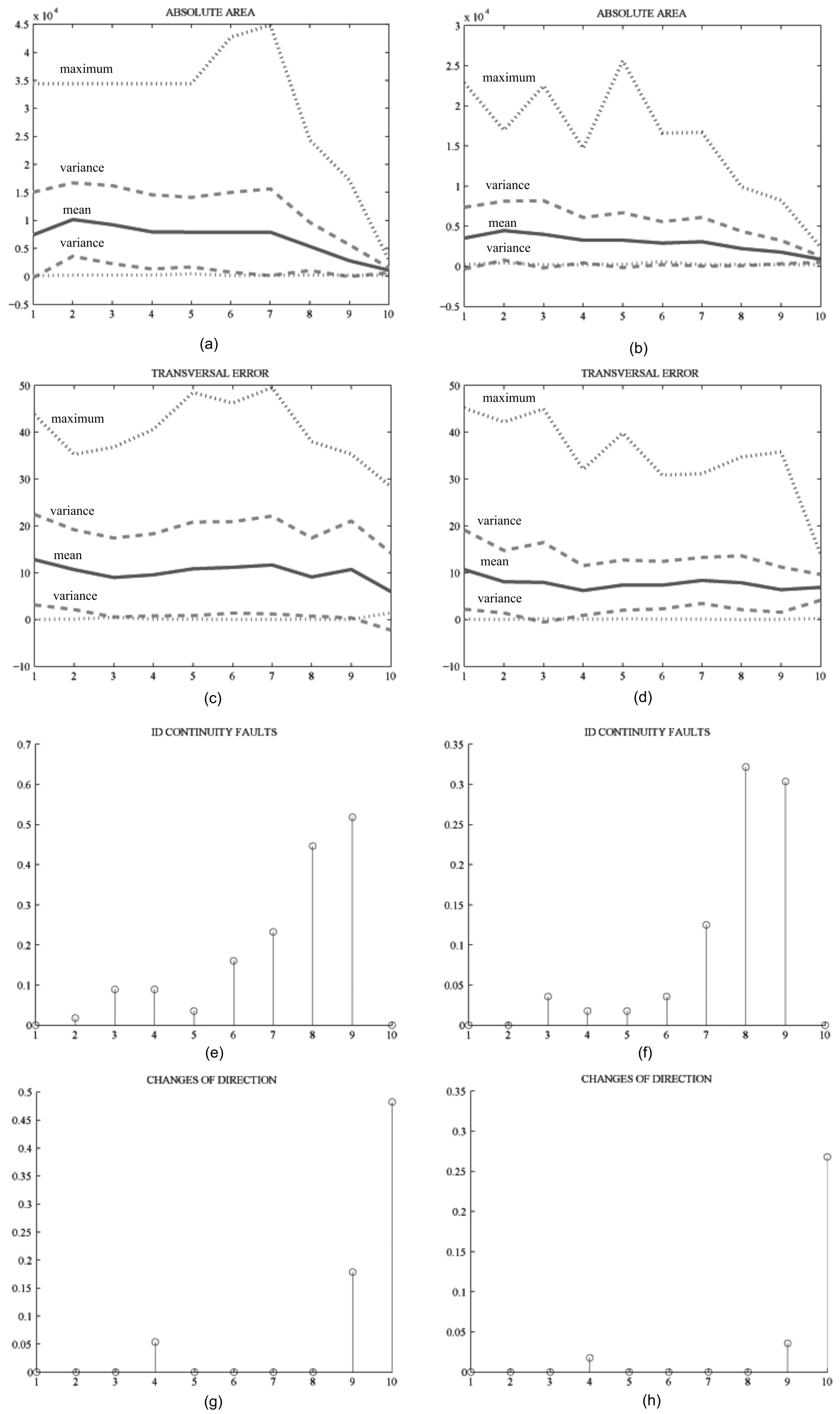
As a final conclusion, we are able to confirm that the improvement in the segmentation stage provides more compact and accurate blobs to the subsequent blocks of the video surveillance system so that the performance of the surveillance system does increased.

References

1. Pérez, O., García, J., Berlanga, A., and Molina, J.M.: Evolving Parameters of Surveillance Video System for Non-Overfitted Learning. Proc. 7^{th} European Workshop on Evolutionary Computation in Image Analysis and Signal Processing. EvoIASP 2005. Lausanne, Switzerland (2005).
2. García, J. A. Besada, J. M. Molina, J. Portillo. Fuzzy data association for image-based tracking in dense scenarios, IEEE International Conference on Fuzzy Systems. Honolulu, Hawaii (2002)
3. Friedman, N. and Russell, S.: Image segmentation in video sequences: A probabilistic approach, in Proceedings of the Thirteenth Annual Conference on Uncertainty

in Artificial Intelligence (UAI 97), Morgan Kaufmann Publishers, Inc., (San Francisco, CA) (1997) 175–181.

4. Rosin, P.L., and Ellis, T.: Image Difference Threshold Strategies and Shadow Detection, in the 6th BMVC 1995 conf. proc., Birmingham, UK, (1995) 347–356.
5. Jiang, C.: Shadow identification, CVGIP: Image Understanding, **59(2)** (1994) 213–225.
6. Prati, A., Mikic, I., Trivedi, M.M., Cucchiara, R.: Detecting Moving Shadows: Algorithms and Evaluation, IEEE Trans. PAMI **25(7)** (2003) 918–923.
7. Bevilacqua, A.: Effective Shadow Detection in Traffic Monitoring Applications. WSCG 2003, **11(1)**
8. Zhang, Y.J.: Evaluation and comparison of different segmentation algorithms, Pattern Recognition Letters **18** (1997) 963–974.
9. Chabrier, S., Emile, B., Laurent, H., Rosenberger, C.,March, P.: Unsupervised Evaluation of Image Segmentation Application to Multi-spectral Images. 17th International Conference on Pattern Recognition (ICPR'04) **1** (2004) 576–579.
10. Erdem, E., Sankur, B., Tekalp, A.M.: Metrics for performance evaluation of video object segmentation and tracking without ground-truth. ICIP **2** (2001) 69–72.
11. Pokrajac, D. and Latecki, L.J.: Spatiotemporal Blocks-Based Moving Objects Identification and Tracking. IEEE Int. W. Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS). Nice, France. (2003).
12. Black, J., Ellis, T., and Rosin, P.: A Novel Method for Video Tracking Performance Evaluation, Joint IEEE Int. Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS). Nice, France. (2003).
13. Aggarwal, J.K. and Cai, Q.: Human motion analysis: A review. Computer Vision and Image Understanding, **73(3)** (1999) 428-440.
14. Wang, L., Hu, W., and Tan., T.: Recent developments in human motion analysis. Pattern Recognition, **36(3)** (2003) 585-601.
15. Haritaoglu, D.I., Harwood, D., and Davis, L.: W4: Real- Time Surveillance of People and Their Activities, IEEE Trans. Pattern Analysis and Machine Intelligence **22(8)** (2000) 809-830.
16. Remagnino, P., Jones, G.A., Paragios, N., and Regazzoni, C.S.: Video-Based Surveillance Systems. Kluwer Academic Publishers, 2002.
17. Wren, C., Azarbayejani, A., Darrell, T., and Pentland, A.P.: Pfinder: Real-time Tracking of the Human Body, IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI) **19(7)** (1997) 780-785.
18. Stauffer, C., and Grimson, W.: Adaptive background mixture models for realtime tracking. In Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, (1999) 246–252.
19. Lipton, A.J., Fujiyoshi, H., and Patil, R.S.: Moving target classification and tracking from real-time video. In Proc. of Workshop Applications of Computer Vision, (1998) 129-136.
20. Cohen, I., and Medioni, G.: Detecting and Tracking Moving Objects in Video from an Airborne Observer, Proc. IEEE Image Understanding Workshop, (1998) 217–222.
21. Kim, E.Y., Park, S.H., Hwang, S., and Kim, H.J.: Video Sequence Segmentation Using Genetic Algorithms. Pattern Recognition Letter, **23(7)** (2002) 843–863.
22. Hwang, S., Kim, E.Y., Park, S.H., and Kim, H.J.: Object Extraction and Tracking Using Genetic Algorithms, in Proc. IEEE Signal Processing Society ICIP **2** (2001) 383–386.



(a)

(b)

(c)

(d)

(e)

(f)

(g)

(h)

Fig. 7. Metrics for people walking from right to left before and after the morphological process (left and right column respectively)