TACKLING EVENT DETECTION IN THE CONTEXT OF VIDEO SURVEILLANCE

Răducu DUMITRESCU[1], Diana GRAMA[2], Bogdan IONESCU[3]

Rezumat. *În acest articol discutăm despre problematica detecţiei automate a evenimentelor în contextul sistemelor de supraveghere video. O primă etapă de analiză o constituie estimarea fundalului. În acest sens, am testat trei abordări diferite, astfel: diferenţa cadrelor succesive, media "alunecătoare" şi o estimare a filtrării mediane. Aceste tehnici furnizează informaţii despre schimbările survenite de la o imagine la alta şi sunt folosite mai departe pentru detecţia prezenţei umane în scenă. Aceasta este realizată folosind o abordare orientată pe contur. Contururile obiectelor sunt extrase din regiunile ce se modifică şi parametrizate. Silueta unei persoane va furniza o semnătură particulară a acestor parametri. Rezultatele experimentale realizate dovedesc potenţialul acestei metode pentru detecţia evenimentelor din scenă. Totuşi, acestea sunt nişte rezultate preliminare, reprezentând primele noastre rezultate în această direcţie.*

Abstract. *In this paper we address the problem of event detection in the context of video surveillance systems. First we deal with background extraction. Three methods are being tested, namely: frame differencing, running average and an estimate of median filtering technique. This provides information about changing contents. Further, we use this information to address human presence detection in the scene. This is carried out thought a contour-based approach. Contours are extracted from moving regions and parameterized. Human silhouettes show particular signatures of these parameters. Experimental results prove the potential of this approach to event detection. However, these are our first preliminary results to this application.*

Keywords: background estimation, human detection, video surveillance, event detection.

1. Introduction

One of the first image-processing systems has been successfully used in the years after 1920 to improve images submitted by transoceanic cable between London and New York. Although these techniques have been improved continuously, their true potential was revealed by using numerical computer. Technological progress in electronics, optics or computer engineering have increased processing power while lowering costs of the equipments and thus accelerating the introduction of digital image processing in more and more fields of activity.

[1] Eng., Faculty of Electronics, Telecommunication and Information Technology, University Politehnica of Bucharest (e-mail: raducu.dumitrescu@gmail.com).
[2] Eng. Faculty of Electronics, Telecommunication and Information Technology, University Politehnica of Bucharest, (diana_g0812@yahoo.com).
[3] Lect. dr. eng., LAPI – The Image Processing and Analysis Laboratory, Faculty of Electronics, Telecommunication and Information Technology, University Politehnica of Bucharest, (bionescu@alpha.imag.pub.ro).

Nowadays, if we attempt to define this new domain in the context of the actual technological evolution, one may say that "image processing holds the possibility of developing the ultimate machine that could perform the visual functions of all living beings" [26].

These "possibilities" are used successfully in various applications of great interest, such as medical imaging to support and improve medical diagnosis, remote sensing to support military or civil applications, astronomy, biology, criminology, biometric systems, and so on. One area of wide interest, which makes the subject of this paper, is *video surveillance*. Intelligent video surveillance systems are a very paying industry, constantly expanding, supported on one side of the technological progress of the data acquisitions and transmission protocols and by the fast development of urban infrastructure. The existing solutions aim at replacing the human operator in various tasks, to increase productivity, reduce human and material losses, law enforcement, accident prevention, etc.

2. Previous Work

In this paper we address two common video surveillance issues. First, we deal with *automatic extraction of changing contents,* which is related to background extraction techniques. Secondly, we address the problem of *detecting human presence* in the scene and discuss a contour-based approach.

2.1. Existing background extraction techniques

The extraction of changing contents in video sequences may be done by one of the following methods: frame differencing [19] [20], background subtraction and optical flow [17] [18] (which additionally provides motion information).

One of the most efficient techniques and commonly adopted with the existing video surveillance systems performing in real-time, is background subtraction. It consists in lowering a reference image, denoted background, from the current frame or in a certain time window. The content of this image should not change during the video. Background is subtracted from each current frame and the image resulting from this operation is binarized through a thresholding approach. This leads to a binary mask. After improving the object shape in the binary image, typically done by morphological operations [10], the result is the retrieval of changing regions, denoted generically foreground. Ideally, this corresponds to moving objects from the scene.

Although the principle of the technique is simple, the background estimation remains a challenge due to implementation practical aspects, e.g. slow or sudden change of scene illumination, camera movement (caused by wind or vibrations produced by cars), changes in the background geometry (parked cars), real-time

capabilities, etc. Background estimation must be robust to face the challenges above, but sensitive enough to detect all moving objects in the frame.

According to [22] existing background subtraction techniques can be classified into three main categories: basic background modeling, statistical background modeling and background estimation. Basic background modeling use in general average [23] or median approaches [15][16][24], or some estimates, e.g. running average, approximation of median, etc. Statistical background modeling use rather mathematical modeling than considering background an image itself, e.g. single Gaussian distribution [2], Mixture of Gaussians [13] or Kernel Density Estimation [14]. Finally, background estimation techniques use filtering approaches inspired by signal processing, e.g. Kalman filtering, Wiener filtering.

2.2. Existing human detection approaches

Once we retrieve changing contents from the video flow, one may address the classification of this content. The most common application is to detect human presence in the scene and determine its behavior. The relevant literature concerning human detection can be divided into techniques which require background subtraction (see the previous section) and techniques that can detect humans directly.

In order to detect humans, background estimation is followed by a human model construction which uses different features. For example, foreground object classification can be based on the object's shape as in [1], or using a mixture of texture and contour features, that attempts to locate head, hands and feet to identify human model [2]. The method in [1][1] uses a shape-based approach for classification of objects following background subtraction based on frame differencing. The goal is to detect the humans for threat assessment. The target intruder is classified as human, animal or vehicle based on the shape of its boundary contour. The similarity between contours is measured using the L2 norm. In [2] it is proposed a real-time system (called Pfinder) for detecting and tracking humans. The background model uses a Gaussian distribution in the YUV space at each pixel, and the background model is continually updated. The person is modeled using multiple blobs with spatial and colors components and the corresponding Gaussian distributions. Person blob models are initialized using a contour detection step which attempts to locate the head, hands and feet. This system is geared toward finding a single human, and makes several domain-specific assumptions and works in real-time.

On the other hand, direct human detection techniques operate on different types of information extracted from image or video, e.g. motion information and shape [3], periodic motion [4] or shape templates [5]. In [3] detection of humans is performed, directly, from static images or from video flow using a classifier

trained on human shape and motion features. The method restricts itself to the case of pedestrians, i.e. humans are always in upright walking poses. Another example is the approach in [4] which focuses on detecting periodic motions and is applicable to the detection of characteristic periodic biological motion patterns, such as walking. The system is capable of detecting periodic human motion, but it also has knowledge of the period which is useful for extracting more information about gait, such as stride length. The system performance is real-time. [5] deals with the challenging scenario of a moving camera mounted on a vehicle. Shape-based template matching is performed based on the Chamfer distance. A hierarchical tree of templates is constructed from a set of templates, which allows for efficient matching. The method also includes a Kalman filter based tracker for taking advantage of the temporal information for filling in missed detections.

3. The Proposed Background Extraction Approaches

In this paper we have tested and compared the results of three background extraction approaches which are presented in the sequel.

Frame differencing. Frame differencing is the simplest technique for the detection of changing content. The current frame is subtracted from the previous frame and if the absolute difference is great than threshold *Th* then the pixel is consider as part of a moving object, thus:

$$M(x,y)_t = \begin{cases} 1, & if \left| I(x,y)_t - I(x,y)_{t-1} \right| > Th \\ 0, & else \end{cases} \quad (1)$$

where $M(x,y)_t$ is a binary image, $M(x,y)_t=1$ for a moving pixel and 0 otherwise and $I(x,y)_t$ represents the image at time index *t*.

In this case the background is always approximated with previous frame. The method diagram is presented in Fig. 1.

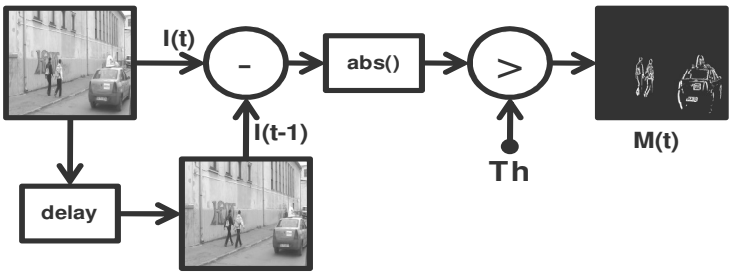

Fig.1. Block diagram for frame differencing.

This technique is very sensitive to the threshold value and cannot detect the entire shape of a moving object with quasi-uniform intensity. Main advantages are the

reduced computational load, little memory space needed and it is highly adaptable to changes in background.

Running average. It is a fast algorithm that constructs the background as an estimate of the average of the previous *N* frames. It estimates the background from only the current frame at time index *t*, $I(x,y)_t$, and the previous background $B(x,y)_{t-1}$ at time index *t-1*, thus:

$$B(x,y)_t = \alpha \times I(x,y)_t + (1-\alpha) \times B(x,y)_{t-1} \tag{2}$$

where α is the learning ratio which determined the speed of adaptation to illumination variations (a common value is around 0.05). The method's diagram is presented in Fig. 2.

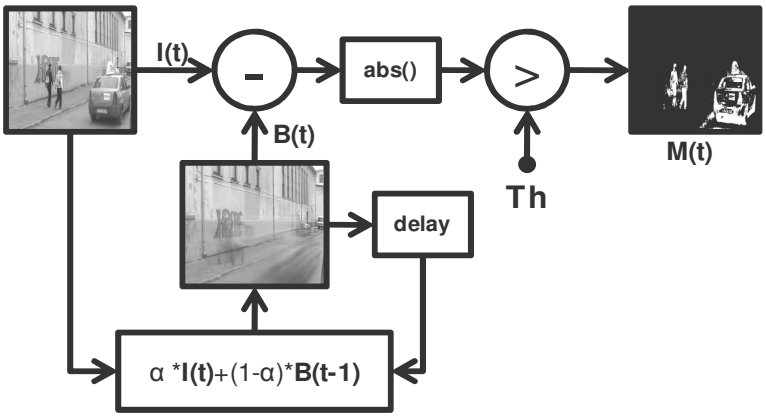

Fig.2. Block diagram for running average.

Once the background is estimated, changing content (foreground) is determined using the same approach as for frame differencing, thus computing the binary image $M(x,y)_t$:

$$M(x,y)_t = \begin{cases} 1, & if \left| I(x,y)_t - B(x,y)_t \right| > Th \\ 0, & else \end{cases} \tag{3}$$

where *Th* is a threshold.

Approximation of median filtering. If a pixel in the current frame has a value greater than the corresponding background pixel, the background pixel is incremented by 1. Otherwise, if the current pixel is less than the background pixel, the background is decremented by one. In this way, the background eventually converges to an estimate of the median, where half the input pixels are greater than the background, and half are less than the background (convergence time will vary based on frame rate and amount movement in the scene.). The following equations describe this process:

$$B(x,y)_t = B(x,y)_{t-1} + \mathrm{sgn}(I(x,y)_t - B(x,y)_{t-1}) \tag{4}$$

$$M(x,y)_t = \begin{cases} 1, & if \left| I(x,y)_t - B(x,y)_t \right| > Th \\ 0, & else \end{cases} \tag{5}$$

where $B(x,y)_t$ is the background estimated at time index t, $sgn()$ represents the signum function defined such: $sgn(x)=-1$ if $x<0$, $sgn(x)=1$ if $x>0$, and $sgn(x)=0$ if $x=0$, $I(x,y)_t$ is the current frame at time index t, $M(x,y)_t$ is the binary image corresponding to changing content (value 1) and Th a threshold. The method diagram is presented in Fig. 3.

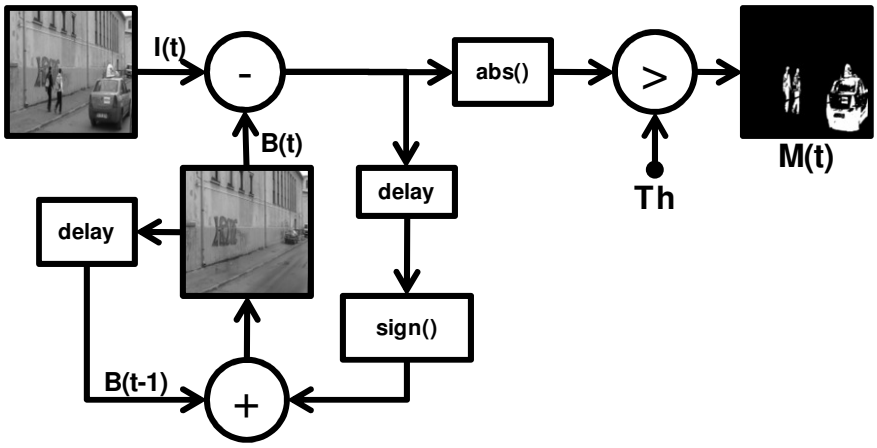

Fig.3. Block diagram from approximation of median filter.

This method has the advantage of providing the accuracy of some higher-complexity methods but with a computational complexity comparable to frame differencing.

4. The Proposed Human Detection Approach

We propose a human detection method which is based on the analysis of image contour structural features and background extraction. Basically, the implemented technique consists of two main processing steps, namely: background estimation and object contour parameterization. The algorithm is presented in Fig. 4 and each step is discussed with the following.

Pre-processing. The first step consists of converting true color images (16 million color palette) to grayscale (256 values), thus:

$$P(x,y) = 0.2989 \cdot R(x,y) + 0.587 \cdot G(x,y) + 0.114 \cdot B(x,y) \tag{6}$$

where (x,y) are the coordinates of the current pixel, (R,G,B) represent the Red, Green and Blue components and P is the resulting gray level.

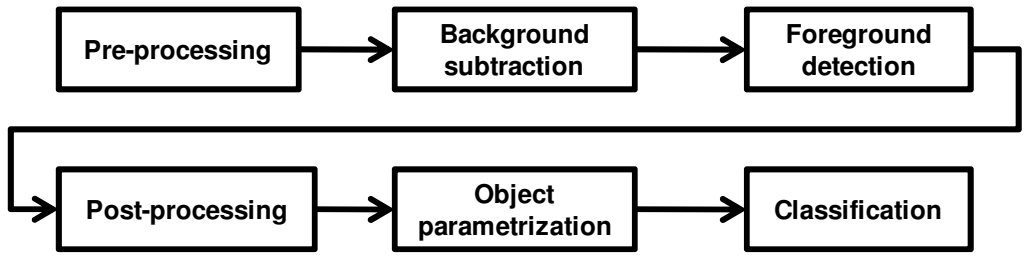

Fig. 4. The proposed human detection approach.

Next, the image noise was filtered using median filtering techniques [7] to preserve as much as possible edge transitions in the image. Additionally, contrast was enhanced with histogram equalization due to its efficiency [8] [9]. All these pre-processes were adopted to improve contour/edge information and thus strengthening human silhouettes in the scene.

Background detection. This step aims at recovering changing content from the video sequence, as we assume that target people are in motion. We use a median filtering technique (see equation 5, more details on background extraction are provided in Section 3).

Post-processing. Experimental tests show that regions obtained after background extraction are not suited for the classification as they are, e.g. false regions are always present, contours are often non-uniform, etc. To enhance their appearance we have adopted several morphological operations. Morphological operators are shape oriented mathematical operations that simplify image data, preserving their essential shape characteristics and eliminating irrelevancies [10]. Mathematical morphology provides a number of important image processing operations, including erosion, dilation, opening and closing. All these morphological operators take two pieces of data as input: the input image and the structuring element. The structuring element consists of a pattern specified as the coordinates of a number of discrete points relative to some origin. It basically determines the precise details of the effect of the operator on the image.

In the post-processing we have adopted the following operations: image closing and opening, gap filling and edge detector. By image closing we fill gulfs, channels and lakes smaller than the structuring element. On the contrary, image opening removes capes, isthmus and islands smaller than the structuring element [11]. Applied one after another it allows smoothing and enhancing object geometry.

Additionally, objects that were on the boundary of the image and objects whose area is smaller than a threshold (experimentally determined) are removed. Finally, we extract edges of the objects, i.e. the exterior contour. The result of the post processing is an image which contains only the contour of the objects that were

candidates for being classified as human silhouettes (examples are provided in Section 5).

Object parameterization. Having determined all these contours one have to establish whether a foreground object contains a human or not. To do so, we needed a procedure to characterize the human contour, to uniquely describe it. For this purpose, the object parameterization was introduced in the processing chain.

We attempt to characterize each contour property with several numeric parameters and therefore transposing the classification problem to the classification of some feature vectors.

First, we determine the gravity center of the object and then we define as its signature the sequence of Euclidean distances computed between the gravity center and each point from the object's contour. For instance, if d_1 represents the distance between the center of gravity, denoted $G(a,b)$, and a point from the contour, $P(x,y)$, then it is given by:

$$d_1 = d(P,G) = \sqrt{(a-x)^2 + (b-y)^2} \tag{7}$$

The algorithm is illustrated in Fig. 5.

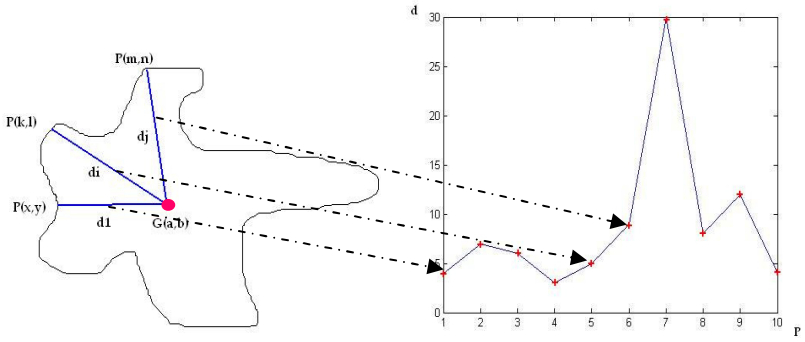

Fig. 5. Example of contour signature (right graph, where *d* represent Euclidean distances and *P* are exterior contour points).

5. Experimental Results

The proposed approaches have been tested on several video sequences recorded from different locations and from different perspectives, summing up to 1 hour of footage.

Each sequence was manually labeled in order to constitute a ground truth. In the following we present some of the experimental results.

5.1. Background extraction results

Fig. 6 shows the comparative results obtained by applying the three background estimation methods discussed in Section 3. One may observe in Fig. 6.a. that frame differencing method tends to detect only the outer edges of moving objects. This is an unwanted effect when segmenting objects with non-textured surface. On the other hand, moving-average method and the median filter technique achieve good results. Being recurrent methods they tend to introduce "ghosts" into the background, which are generated exclusively by moving objects. However, for running average, this phenomenon can be controlled with the parameter α (see Fig. 7). A major disadvantage of the recursive methods is the reduced degree of background adaptation to changes. If some objects are moving very slowly or even stall, they will be considered as background and kept during a long period of time. In what concerns the computational complexity, all three methods are similar, nevertheless, as expected, frame differencing is faster in the detriment of the quality of the resulting background.

Fig.6. Foreground detection examples:
a. Frame differencing.
b. Running average (α = 0.01).
c. Median filtering.

Fig.7. Background estimation examples (frame 177)
a. Frame differencing; **b.** Running average(α = 0.001); **c.** Median filter.

Fig. 7 presents several background estimation examples with all three methods, thus: frame differencing (Fig. 7.a.), moving-average method (Fig. 7.b.) and the median filter method (Figure 7.c.). Considering the test database, we may conclude that median filter is the most reliable method, providing the smallest number of artifacts and the most proper background.

5.2. Human detection results

Several examples are depicted in Fig. 8, 9 and 10. Fig. 8 illustrates an example of image processing chain. As a result, we obtain the contours that are to be classified. We apply first image closing (see Fig. 8.a.) that emphasize the objects, followed by gap filling (see Fig. 8.c). As we are interested in getting the complete contours of the objects, we decide to eliminate the boundary objects. Due to previous processing steps, the car from the image touches the bottom boundary of the image and therefore it is removed (see Fig. 8.d.). We can now detect the contour of the remaining objects, depicted in Fig. 8.e. The goal is to detect human silhouettes; therefore, small objects are not of interest, so we eliminate them. The result of this last post-processing step is presented in Fig. 8.f.

Fig. 9 presents several final examples of significant foreground objects (found inside the different colored bounding boxes). The contours of these objects are used further to compute the objects' signatures.

Obtaining the object's signature is the final processing step. The data resulted from this step of the algorithm is used in the classification process. Although not identical, the human silhouette can be differenced from other objects using the computed signatures, as it has particular features. Several examples depicted in Fig. 10 prove the potential of the proposed contour signature in retrieving human silhouette. However, false detections may occur.

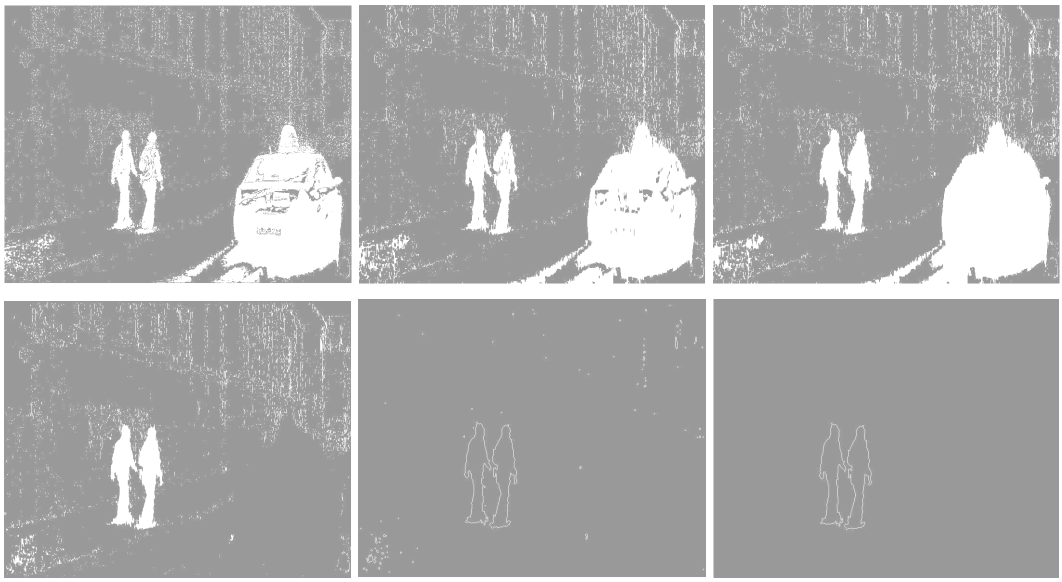

Fig. 8. Post-processing steps example
a. Before post-processing; **b.** After image closing; **c.** After gap filling.
d. After boundary object removal; **e.** After edge detection; **f.** After small object removal.

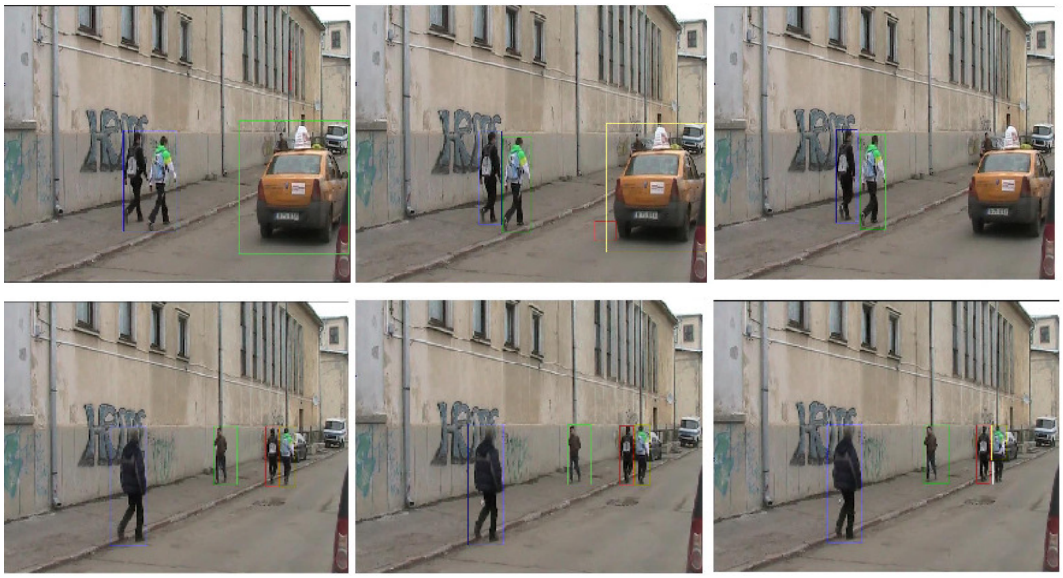

Fig. 9. Results of the detected objects in different frames (see color boxes).

These preliminary experimental results have shown cases where the human silhouettes were divided in several pieces or deformed after foreground detection or after post-processing (see Fig. 8 and 10). This is mainly due to the illumination conditions, particular background objects that have similar color as the objects of interest thus making the foreground detection difficult, or false movement

detection due to filming conditions, shadows etc. Some of these issues can be addressed in further work to improve the accuracy of the human detection algorithm.

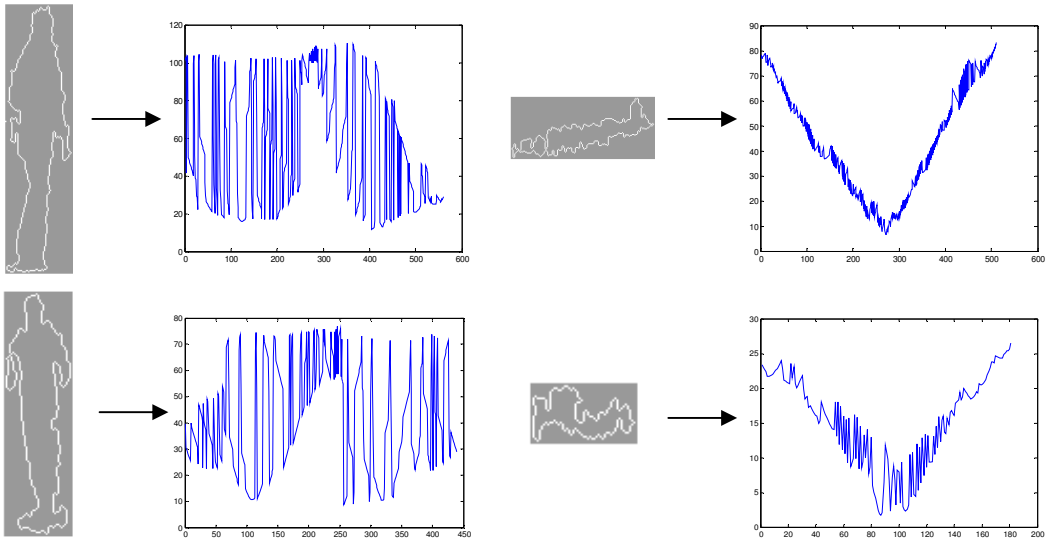

Fig. 10. Signatures from experimental results (shape vs. signature, one may observe that human signatures share some common features).

Conclusions and future work

In this paper we address the issue of event detection in the context of video surveillance systems. First, we deal with background extraction. Several methods are proposed. Secondly, we tackle the detection of human presence in the scene. We use a contour-based classification approach. Experimental tests, carried out on a real video surveillance database prove the potential of this approach to the analysis of human behavior in the scene. However, the work presented in this paper is part of an ongoing project. Further research and development will involve increasing the invariance of the methods, enlarging the feature set and deriving semantic descriptions. Also, the functionality of the methods shall be tested on a real-time environment, e.g. on the video surveillance system from the University Politehnica of Bucharest.

Acknowledgment

The authors would like to thank As. prof. Serban Oprisescu for helping them recording and processing the test videos and Senior Researcher Christoph Rasche for suggesting the contour-based approach.

REFERENCES

[1] D. J. Lee, P. Zhan, A. Thomas and R. Schoenberger, Shape-Based Human Intrusion Detection, SPIE International Symposium on Defense and Security, Visual Information Processing XIII, 5438:81-91, 2004;

[2] C. R. Wren, A. Azarbayejani, T. Darrell, and A. P. Pentland, Pfinder: real-time tracking of the human body, IEEE Transactions on Pattern Analysis and Machine Intelligence,19(7):780–785, 1997;

[3] P. Viola, M. J. Jones, and D. Snow, Detecting pedestrians using patterns of motion and appearance, IEEE International Conference on Computer Vision, 2:734-731, 2003;

[4] Cutler and L. S. Davis, Robust real-time periodic motion detection, analysis, and applications, IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(8):781-796, 2000;

[5] M. Gavrila and J. Giebel, Shape-based pedestrian detection and tracking, IEEE Intelligent Vehicle Symposium, 1:8-14, 2002;

[6] Neeti A. Ogale, A survey of techniques for human detection from video, http://citeseerx.ist.psu.edu/;

[7] V. V. Bapeswara Rao and K. Sankara Rao, A New Algorithm for Real-Time Median Filtering, IEEE Transactions on Acoustics, Speech, Processing, Vol. 6, p1674-1675, 1986;

[8] W. K. Pratt, Digital Image Processing, (Wiley, California, USA, 2001);

[9] R. C. Gonzalez and R.E.Woods, Digital Image Processing, (Addison-Wesley, Massachusetts, 1993);

[10] R. M. Haralick, S.R. Sternberg, and X. Zhuang, Image Analysis Using Mathematical Morphology, IEEE trans. on Pattern Analysis and Machine Intelligence, Vol.4, p1228-1244, 1986;

[11] Steven W. Smith, The Scientist and Engineer's Guide to Digital Signal Processing, http://www.dspguide.com/;

[12] McKenna, S. J.; Jabri, S. and Duric, Z.; Rosenfeld, A.;Wechsler, H.; Tracking groups of people, CompuVision and Image Understanding, 80, pp 42—56 (2000);

[13] Stauffer, C. and Grimson, W. E. L., Adaptive background mixture models for real-time tracking, Proceedings of CVPR, Jun 1999, pp. 246-252;

[14] Elgamal A.; Duraiswami R.; Harwood D. and Davis L.; Background and foreground modelling using nonparametric kernel density estimation for visual surveillance, Proc of the IEEE, 90, No 7 (July 2002);

[15] Zhou, Q. and Aggarwal, J. K.; Tracking and classifying moving objects from video, Proc of 2nd IEEE Intl Workshop on Performance Evaluation of Tracking and Surveillance (PETS'2001), Kauai, Hawaii, USA (December 2001);

[16] R. Cucchiara, M. Piccardi, and A. Prati, Detecting moving objects, ghosts, and shadows in video streams, IEEE Transactions an Pattern Analysis and Machine Intellignece 25, pp. 1337-1342, Oct 2003;

[17] B.K.P. Horn and B.G. Schunck, Determining optical flow. Artificial Intelligence, vol 17, pp 185-203, 1981;

[18] A. Bruhn, J. Weickert, and C. Schnorr, Lucas/Kanade meets Horn/Schunck: combining local and global optic flow methods, International Journal of Computer Vision,vol.61,no.3,pp. 211–231, 2005;

[19] Alan J. Lipton, Hironobu Fujiyoshi, Raju S. Patil, Moving target classification and tracking from real-time video, submitted to IEEE WACV 98,1998;

[20] L. Wang, W. Hu, and T. Tan. Recent developments in human motion analysis. Pattern Recognition, 36(3):585–601, March 2003;

[21] B.P.L. Lo and S.A. Velastin, Automatic congestion detection system for underground platforms, Proc. of 2001 Int. Symp. on Intell. Multimedia, Video and Speech Processing, pp. 158-161, 2000;

[22] F. El Baf, T. Bouwmans, B. Vachon, Fuzzy Foreground Detection for Infrared Videos, 5th Joint IEEE International Workshop on Object Tracking and Classification in and Beyond the Visible Spectrum, OTCBVS 2008, pages 1-6, Anchorage, Alaska, USA, 27 June 2008;

[23] B. Lee and M. Hedley. Background estimation for video surveillance. Image and Vision Computing New Zealand,2002;

[24] N. McFarlane and C. Schofield, Segmentation and tracking of piglets in images, Mach. Vision Appl. 8 (1995), pp. 187–193,1995;

[25] Hunt B. R., Image processing: the moving horizon, Proceedings of the IEEE, vol. 69, No.5, pp. 499-501, May,1981;

[26] Anil K. Jain, Fundamentals of digital image processing, Prentice Hall, Englewood Cliffs, NJ, 1989.