Shape Based People Detection for Visual Surveillance Systems

M. Leo, P. Spagnolo, G. Attolico, and A. Distante

Istituto di Studi sui Sistemi Intelligenti per l'Automazione - C.N.R.
Via Amendola 166/5, 70126 Bari, ITALY
{leo,spagnolo,attolico,distante}@ba.issia.cnr.it

Abstract. People detection in outdoor environments is one of the most important problems in the context of video surveillance. In this work we propose an example-based learning technique to detect people in dynamic scenes. A classification based on people shape and not on image content has been applied. First, motion information and background subtraction have been used for highlighting objects of interest, then geometric and statistical information have been extracted from horizontal and vertical projections of detected objects to represent people shape. Finally, a supervised three layer neural network has been used to properly classify objects. Experiments have been performed on real image sequences acquired in a parking area. The results have shown that the proposed method is robust, reliable, fast and it can be easily adapted for the detection of any other moving object in the scene.

1 Introduction

The people detection problem has been largely studied in literature due to many promising applications in several areas, above all visual surveillance. The availability of new technologies and the corresponding decrease of their costs cause more attentions to developing real-time surveillance system. Surveillance systems have been already installed in many locations such as highways, streets, stores, homes and office, but often have been used only as record supports to control the situations. A visual system that automatically detects abnormal situations and alerts for possible anomalous behaviors would be greatly appreciated. The first step is then to built a visual system able to detect moving objects, recognize people and individuate illegal behaviors. In this paper, we focus on the problem of people detection in outdoor environments with a static TV camera. Our application context is a park where people and cars go through. In this context the main aim is to detect the presence of people in order to recognize their gestures and individuate illegal actions such as thefts. We propose a new technique that combines high detection rate with fast computational time in order to process image sequences in real time. Detecting people in images is more challenging than detecting other objects because people can be articulated in their shape and can assume a variety of postures, so it is nontrivial to define a single

J. Kittler and M.S. Nixon (Eds.): AVBPA 2003, LNCS 2688, pp. 285-293, 2003.

model that captures all these possibilities. Moreover, people wear different dresses of different colors and then the interclass variation in the people class can be very high, making difficult the recognition by using color-based and fine edge-based techniques.

In literature two main categories of approaches for moving object classification have been developed: Shape-based classification and Motion-based classification. Motion based classification is performed extracting information on periodic properties of motion. In [1] time-frequency analysis was applied to detect and characterize the periodic motion and in [2] residual optical flow was used to analyze rigidity and periodicity of moving entities. Shape-based classification has been performed extracting different descriptions of moving regions. The descriptions can be based on intensity gradient image [12], image blob dispersedness, area, ratio of the blob bounding box [3,4], mean and standard deviation of silhouette projection histogram [5], wavelet representation [6,7,8]. The above approaches, namely shape-based classification and motion-based classification, have also been effectively combined for moving objects classification[9]. Usually, motion based classification techniques require a lot of time for parameters computation and they cannot perform in real time. Moreover they require that the orientation and the apparent size of the segmented objects do not change during several periods and this constraint cannot be applied in a context of video surveillance where there are no rules in the movement. The shape based classification techniques, instead, do not need temporal information (they work on a single frame) and they can be implemented in real time because the parameters calculation is more fast and simple.

In this work, a shape-based classification technique that uses statistical and geometric information extracted from the horizontal and vertical projection of the binary silhouette of the moving object has been adopted. The use of horizontal and vertical projections has several advantages: reduces the problem of noises in images still maintaining all the information to distinguish a moving object from others in contexts of video surveillance; it can be performed in a very small amount of time even without specialized hardware or algorithmic dodges. In this work, seven parameters that incorporate geometric and statistical information of the moving object have been extracted from the projections of binary shapes and have been provided as input to a neural classifier.

The paper is organized as follows: in section 2 a brief overview of the approach for highlighting objects of interest in a static background is described; in section 3 the feature extraction step, and the feature learning step are detailed. Finally, the experimental results obtained on real image sequences are reported in section 4.

2 Object Detection and Binary Shape Extraction

The first problem that a visual system has to solve to recognize any object is to extract from a complex image the objects of interest that are candidate to be matched with a searched model. In our contest the objects of interest are both moving objects and static objects that differ from a background model. For this reason we have implemented an algorithm for objects detection that uses an adaptive background subtraction scheme. The key idea is to maintain a statistical model of the background: for each pixel a running average and a form of standard deviation is maintained. So,

in the test phase, a pixel is labeled as foreground if its intensity value differs from the running average two times more than the standard deviation [3]. Any background subtraction approach is sensitive to variations of the illumination conditions. To solve this problem, it is necessary to frequently update the information about the running average and the standard deviation of all background pixel values. For the updating, an exponential filter has been used; the implemented updating equations are described in [10]. After these steps, a reliable model of the background is available at each frame, so it is possible to exactly extract the objects of interest. But the results of the object detection subsystem described above cannot be used directly by the object recognition system since they present an undesirable drawback: the shadows. Each foreground object contains its own shadow, because also this area effectively differs from the background. It's necessary to remove that, because the structure of the object shape radically changes. Starting from the observation that shadows move with their own objects, but that they have not a fixed texture, the removing algorithm proposed in [10] has been implemented. Firstly, the image is segmented calculating the photometric gain for each pixel, then segments that present the same correlation calculated in the background image and in the current image are eliminated. Finally, an additional step is made in order to further simplify the following classification phase. All the moving blob with the area lower than an appropriate threshold are removed. This step allows to concentrate the attention only on the object of interest as a car or a person. At this point each detected object is represented by a binary shape that can be provided as input to the classifier. In presence of many objects, each of them is considered separately by the moving detection algorithm. Each blob is detached from the remaining and analyzed at different times. For each frame the algorithm extracts a number of binary images equals to the number of different objects detected in the scene.

3 Feature Extraction and Learning

After the detection of the objects of interest, it is necessary to extract some attributes for modeling and recognize automatically their shapes (*pattern recognition problem*). The problem of pattern recognition can be decomposed in a *features extraction problem* and in a *classification problem*. In the first case a set of coefficients (*pattern*), that allows to describe the input information in a significant way, has to be found. The extracted characteristics are provided as input to the classifier then the objective is to reduce the number of coefficients still preserving the relevant information of the input shape. In order to satisfy these two contrasting requirements, in this work, each binary image has been processed as follows: the horizontal and vertical projections of the whole image are computed; from these projections the most important geometric and statistic properties have been extracted. In the first step the system generates a new representation of the binary image of the target. Each item of this new representation is the sum of the black points in the rows of the binary image for the vertical projection, and the sum of the black points in the columns for the horizontal projection, that is:

$$Hor\,\mathrm{Pr}oj(j) = \sum_{i=1}^{N} I(i,j) \qquad Ver\,\mathrm{Pr}oj(i) = \sum_{j=1}^{M} I(i,j) \tag{1}$$

where N is the number of rows and M is the number of columns in the binary image I. In this way the information is collected in a set of coefficients equals to the sum of the rows and columns in the image, obtaining a substantial reduction of the representation coefficients. The initial coefficients of the binary image representation were MxN whereas, after the projections their number becomes M+N. In addition, since the target is a small part of the whole image, many coefficients of the projections are zero. These coefficients are removed by the system and the number of coefficients decreases further. Starting from the remaining coefficients the system evaluates seven parameters that compose the *pattern* associated with each object of interest. The parameters P1…P7 are :

1. Max value of the horizontal projection :

$$P1 = \max_{j=1\ldots M}\left\{ Hor\,\mathrm{Pr}oj(j) = \sum_{i=1}^{N} I(i,j) \right\} \tag{2}$$

2. Max value of the vertical projection:

$$P2 = \max_{i=1\ldots N}\left\{ Ver\,\mathrm{Pr}oj(i) = \sum_{j=1}^{M} I(i,j) \right\} \tag{3}$$

3. Sum of the coefficients of the horizontal projection (or of the vertical projection):

$$P3 = \sum_{j=1}^{M} Hor\,\mathrm{Pr}oj(j) = \sum_{J=1}^{M}\sum_{i=1}^{N} I(i,j) = \sum_{i=1}^{N} Ver\,\mathrm{Pr}oj(i) = \sum_{i=1}^{N}\sum_{J=1}^{M} I(i,j) \tag{4}$$

4. Mean value of the horizontal projection:

$$P4 = \sum_{j=1}^{M} Hor\,\mathrm{Pr}oj(j) / K \tag{5}$$

where K ≤ M is the number of non zero items in the horizontal projection.

5. Mean value of the vertical projection:

$$P5 = \sum_{i=1}^{N} Ver\,\mathrm{Pr}oj(i) / H \tag{6}$$

where H ≤ N is the number of non zero items in the vertical projection.

6. Standard deviation of the horizontal projection :

$$P6 = \sum_{j=1}^{M} \left| Hor\,\mathrm{Pr}oj(j) - (\sum_{j=1}^{M} Hor\,\mathrm{Pr}oj(j) / K) \right| \tag{7}$$

7. Standard deviation of the vertical projection :

$$P7 = \sum_{i=1}^{N} \left| Ver\,\mathrm{Pr}oj(i) - (\sum_{i=1}^{N} Ver\,\mathrm{Pr}oj(i) / H) \right| \tag{8}$$

The parameters P1,P2 and P3 provide geometric information: the max value of the horizontal projection is the height of the object; the max value of the vertical projection is the width of the object and P3 is its area.

The parameters P4, P5, P6 and P7 provides, instead, statistical information: P4 and P5 are normalized with the number of coefficients in the respective projections and they supply information on the centroids of the object. In the same way P6 and P7 are normalized with the number of coefficients in the respective projections and they detain the information of the object shape. This method is invariant to the object translation in the scene. In this way it overcomes the problems of using sliding windows [16] in order to search the target. The possibility to easily adapt this method to detect other moving objects in the scene is another advantage of the proposed method. The figure 1 shows the scheme of the whole people detection system. Figure 2 shows the horizontal and vertical projections of two images containing a person and a car.

The extracted parameters are provided as input to the classifier. The classification is performed by a three layer neural network trained, with the Back propagation algorithm, on a set of positive examples relative to selected people images and negative examples relative to others moving objects in the scene.

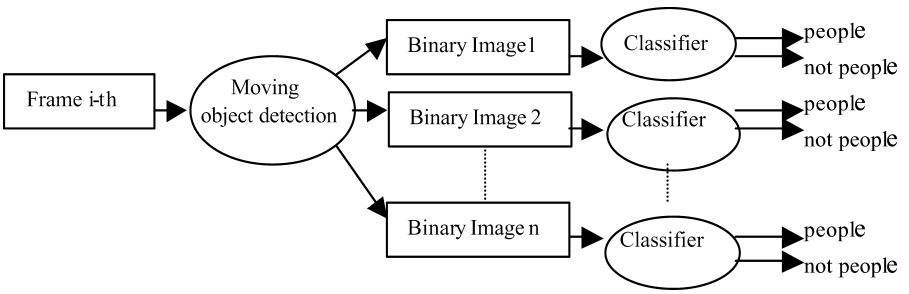

Fig. 1. The people detection system

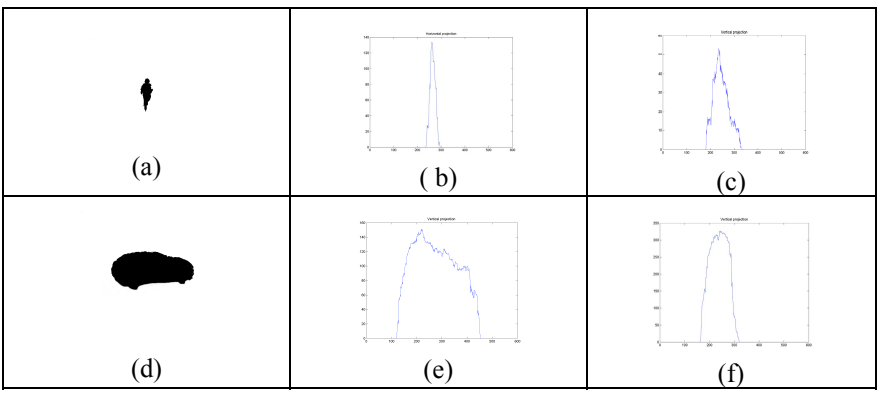

Fig. 2. Two examples of feature extraction by projection: (a) (d) two binary images containing a person and a car - (b)(e) horizontal projections (c) (f) vertical projections

Table 1. The test sequences and their details

Sequence number	Number of Frames	Number of Binary Images generated	Number of Binary Images with a walking person	Number of Binary Images with a group of walking people	Number of Binary Images with a car	Number of Binary Images with groups of cars and walking people
1	894	1294	1109	185	0	0
2	387	387	0	0	387	0
3	1058	1854	1468	386	0	0
4	51	51	0	0	0	51

Table 2. The classification results

Sequence Number	Positive Response of the Neural Network (people detected)	Negative Response of the Neural Network (people not detected)
1	1294	0
2	0	387
3	1854	0
4	0	51

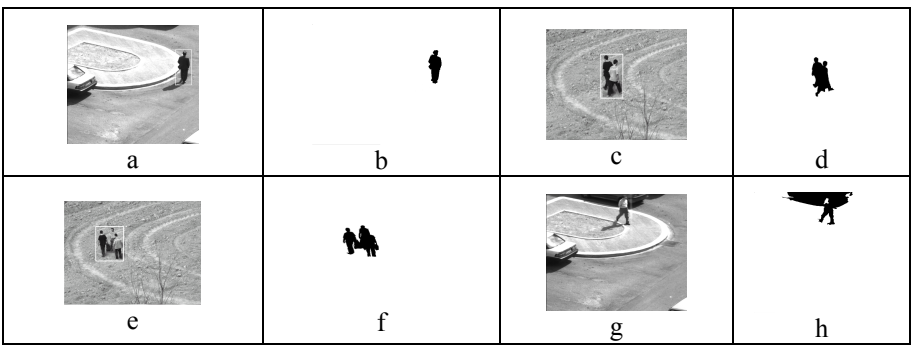

Fig. 4. a-f) Three examples of test images correctly processed; g-h) An example of test images wrongly processed. The people and the car move together and their binary shape are spoiled

4 Experimental Setup and Results

The experiments have been performed on real image sequences acquired with a static TV camera Dalsa CA-D6 with 528 X 512 pixels, connected to a Stream Store able to store up two hours of recorded image. The frame rate is 30Hz. The processing is performed with a Pentium IV, with 1,5 GH and 128 Mb of RAM. The sequences have been acquired in a parking area with the TV camera placed on the sixth floor (22 meters) of a neighbouring building. The parking is very haunted and it's very likely to

have in the same frame many moving targets. Often this is a problem for the algorithms of people recognition because the target are shapeless. The proposed system overcomes partially this problem: when the moving objects are detached, they are analyzed separately but, contrariwise, if different blobs of the targets are connected, the system processes the whole binary shape. Each frame can produce, thus, one or more binary images belonging to the four different categories listed below:

1. one walking person
2. group of walking people
3. one other moving object (probably a vehicle)
4. groups of moving objects and walking people

The neural classifier has been trained with binary images containing a single person or a single car and it has been tested on four different sequences. Table 1 shows the details of each sequence.

In the third column the number of binary images extracted from all the frames of the sequence is shown. Notice that from each frame the number of binary images extracted is equal to the number of moving targets not connected in the scene. The last four columns show the types of images extracted from each sequence. The first and third sequences contain respectively two and three persons walking in different directions, the second contains only a car, and the fourth contains a car and a person connected together for all the time (they move at the same time). In the third sequence there are some frames where the blobs of the three people are connected and thus the binary image contains the whole shape. The classification results are shown in the table 2.

The system is able to detect people also when they are grouped or they assume unexpected postures or a partial occlusion occurs or if some objects carried on modify the usual human body shape. The figure 4 shows some examples of images extracted from the test sequences 1,3 and 4 and the relative binary images extracted by the algorithm described in section 2. In a), c) and e) people are correctly classified and included in a white rectangle. In g) people and car move together and the people is not recognized.

The system does not detect a person only when its blob is connected with a blob of a car (sequence 4). In this case the detection based on the relative binary image is hard also for the human eyes and an approach based on the recognition of the human components in the gray levels images [11] appears most appropriate . Notice that, in sequence 4, car and people move at the same time and that their blob are connected: this is a unusual circumstance (only few frames) , specially if the monitored area is wide. In addition it's important to observe that thefts or other illegal actions are done against static objects and when the parking area is not much frequented. Heuristic methods based on temporal consistency, prediction of movements or domain knowledge can be implemented in order to track the people also when the dynamic of the scene inhibits our system. The target classification system proposed is very fast. Each binary image is processed in 0.016 seconds and more binary images can be processed on parallel processors. If there is a number of parallel processors greater than the number of extracted binary images the time elapsed to classify all the detected objects in a frame remains around 0.016 seconds.

5 Conclusions and Future Works

This work deals with the problem of people detection in outdoor environments in the context of video surveillance. An example-based learning technique to detect people in dynamic scenes has been proposed. The classification is purely based on the people shape and not on the image content. Adaptive background subtraction has been used for detecting the objects of interest, then geometric and statistical information extracted from the horizontal and vertical projections are used to represent people shape and, finally, a supervised three layer neural network has been used to classify the extracted patterns. The experiments show that both a single person and a group of people are correctly detected also when other moving objects are in the scene. People are not detected only when their blob is connected with the blob of a moving car that modifies widely the whole binary shape. In this case people detection from the binary shape is hard even for human eyes. In conclusion it possible to assert that the proposed method is robust, reliable, fast and it can be easily adapted for the detection of other moving objects in the scene. Future works will deal with the problem of gesture recognition of the detected people in order to individuate illegal behaviors such as thefts or damaging.

References

[1] R. Cutler, L. Davis: Robust real-time periodic motion detection analysis and applications, IEEE Trans. Pattern Anal. and Mach. Intell. 22 8 (2000), pp. 781-796

[2] A. J. Lipton: Local application of optical flow to analyse rigid versus non rigid motion. In the website http://www.eecs.lehigh.edu/FRAME/Lipton/iccvframe.html

[3] R.T. Collins et al: A system for video surveillance and monitoring: VSAM, final report, CMU-RI-TR-00-12, technical Report, Carnegie Mellon University, 2000

[4] A.J. Lipton, H. Fujiyoshi, R.S. Patil: Moving target classification and tracking from real time video, Proceedings of the IEEE-WACV, 1998, pp. 8-14

[5] Y. Kuno, T. Watanable, Y. Shimosakoda, S. Nakagawa: Automated detection of human for visual surveillance system, Proceedings of ICPR, 1996, pp. 865-869

[6] M. Oren, C. Papageorgiou, P. Sinha, E. Osuna, and T. Poggio "Pedestrian Detection using wavelet templates" In CVPR97, pp. 193-199, 1997

[7] C.Papageorgiou and T.Poggio, A trainable System for Object Detection, International . Journal of Computer Vision,vol.38, no.1, pp. 15-33 (2000)

[8] M. Leo, G. Attolico, A. Branca, A. Distante: Object classification with multiresolution wavelet decomposition , in Proc. of SPIE Aerosense 2002, conference on Wavelet Applications, 1-5 April, 2002, Orlando, Florida, USA

[9] I. Haritaoglu, D. Harwood, L.Davis: W4: real time surveillance of people and their activities, IEEE Trans. Pattern Anal. and Mach. Intell. 22 8 (2000), pp. 809-830

[10] P. Spagnolo, A. Branca, G. Attolico, A. Distante: Fast Background Modeling and Shadow Removing for Outdoor Surveillance, in Proc. of IASTED VIIP 2002, 9-12 September, 2002, Malaga, Spain
[11] A. Mohan, C. Papageorgiou, T. Poggio: Example-based Object Detection in Images by Components, IEEE Trans. Pattern Analysis and Machine Intelligence, Vol.23, No.4, April 2001
[12] L. Zhao, C. Thorpe: Stereo and Neural Network based Pedestrian Detection, IEEE Transaction on Intelligent Transportation Systems, Vol.1, N. 3, September 2000