Detection and Classification of Human Movements in Video Scenes

A. G. Hochuli, L. E. S. Oliveira, A. S. Britto Jr., and A. L. Koerich

Postgraduate Program in Computer Science (PPGIa)
Pontifical Catholic University of Parana (PUCPR)
{hochuli,soares,alceu,alekoe}@ppgia.pucpr.br
WWW home page: www.ppgia.pucpr.br

Abstract. This paper presents a novel approach for classification of human movements in videos scenes. It consists in detecting, segmenting and tracking foreground objects in video scenes to further classify their movements as conventional or non-conventional. From each tracked object in the scene are extracted features such as position, speed, changes in direction and temporal consistency of the bounding box dimension. These features make up feature vectors that are stored together with labels assigned by a human supervisor which categorize the movement. At the classification step, an instance-based learning algorithm is used to classify the object movements as conventional or non-conventional. For this purpose, feature vectors generated from objects in motion are matched against the reference feature vectors previously labeled. Experimental results on video clips from two different databases (Parking Lot and CAVIAR) have shown that the proposed approach is able to detect non-conventional human movements with accuracies between 77% and 82%.

1 Introduction

The classification of events in video scenes is a relative new research area in computer science and it has been growing more and more due to the broad applicability in real-life. One of the main reasons is the growing interest and use of video-based security systems, known as CCTV. However, the majority of the CCTV systems currently available in the market have limited functionality which comprises capture, storing and visualization of video gathered from one or more cameras. Some CCTV systems already include motion detection algorithms and are able to constrain the recording of videos only when variations in the scene foreground are detected. The main utility of such systems is the recording of conventional and non-conventional events for further consultation and analysis. In other words, such systems do not have any embedded intelligence which is able to provide a classification of the events. They do not have mechanisms to warn operators when non-conventional events are occurring. Such an attribute would be very helpful to prevent and detect in an active fashion the occurrence of non-conventional events.

Besides the need of a more efficient tool in the security area, the detection of non-conventional events in video scenes could be used in other contexts, such as: to detect when an elderly people has an accident inside his/her house [1, 2], non-conventional activities in an office, transit infractions [3]. Therefore, a non-conventional event can be viewed as an action that does not belong to the context.

The research in this area has been focused on two main streams: state-space modeling and template matching [4]. In the former, most of the approaches employ Markov Process and state transition functions [1], hidden Markov models [3] and hierarchical hidden Markov models [2] to model categories of non-conventional events inside pre-defined environments. Essentially an event is characterized by a sequence of actions modeled by a graph called model. When an event presents a sequence which is not modeled, it is considered as non-conventional. The main disadvantage of the model-based approaches is that its use in a novel environment requires a remodeling. The latter uses an approach based on the movement trajectory prototypes [5]. Such prototypes are in fact statistical information about the motion in the time and space domains, such as object centroid position, object edges, variation in velocity and direction. Based on such information, the algorithm computes the trajectory and matches it against other previously known trajectories. In a similar manner, but with lower complexity, Niu et al. [6] use only the object position and velocity to design curves which describe the motion. The use of people gait, represented through an histogram, was used to classify non-conventional situations inside a house [7]. Classification is carried out through a regular histogram. Besides this approach is base on the object features and not on the object motion, the author points out that the variation of distance between the objects and cameras is a serious drawback that may produce errors in the histogram projections. Therefore, one of the challenges in the automatic analysis of video scenes is the adaptability to different environments as well as a real-time classification of the events.

In this paper we present a novel approach which has a relative ability to be adapted to different application environments and which is able to detect non-conventional human movements in video scenes. Such an approach has a calibration period and further it extracts features from the foreground objects in motion through the scene. A non-parametric learning algorithm is used to classify the object motion as conventional or non-conventional. The proposed approach has four basic steps: detection and segmentation of foreground objects, tracking of foreground objects, features extraction from their motion, and movement classification as conventional or non-conventional event.

This paper is organized as follows: Section 2 presents an overview of the proposed approach as well as the main details of the detection, segmentation and tracking algorithms. Section 3 presents the feature extraction while the classification of human movements is discussed in Section 4. Experimental results achieved in video clips from two databases are presented in Section 5. Conclusions and perspective of future work are stated in the last section.

2 System Overview

Through a video camera placed in an strategic point in the environment, video is captured and its frames are processed. First, there is a step to detect and segment the objects in motion, or foreground objects, which aim is to look at the video frames for the regions where the objects of interest may be present. These regions are tracked at the subsequent frames. Only the regions where the objects of interest may appear are tracked. From such objects of interest are extracted some features, not from the objects, but features from the object motion. Features like position, velocity, x, y coordinates, direction variation, and temporal consistency of the bounding box dimension are extracted to make up feature vectors. Such feature vectors are matched against other feature vectors which have been previously labeled and stored in a database. In this step it is employed a temporal window and the dissimilarities between the feature vectors represent mean values for the temporal windows. Using a majority voting rule, the motion of the objects of interest is classified as conventional and non-conventional. Figure 1 presents an overview of the proposed approach and the relationship between the main modules.

The main idea of the proposed approach is that such an strategy could be applied to detect some types of human movements without much effort to be adapted to the environment, since we do not take into account specific information from the objects or scene, but from the object motion. First the solution is adapted to environments where the flow of people in the camera view is moderate, since our research is focused on the movement classification and therefore we do not pay attention to situations where overlapping or occlusion may happen.

2.1 Detection and Segmentation of the Foreground Objects

Several approaches to detect motion have been proposed in the last years [8]. However, the main limitation of such techniques refers to the presence of noise due to the variations in the scene illumination, shadows, or spurious generated by video compression algorithms. In this case, the most elementary techniques based on the background subtraction yields to the detection of several false foreground regions. To minimize the impact of the noise the strategy proposed by Stauffer and Grimson [9] employs Gaussian functions to classify the pixels as belonging to the background or to the foreground. At each frame, the pixels are matched against a mixture of Gaussian distributions according to its variance, standard deviation and weight. All the pixels that could be absorbed by a Gaussian distribution are considered as belonging to the background. If there is no Gaussian distribution that can absorb the pixel, then it is considered as a foreground pixel.

Gaussian distributions are able to absorb continuous motion and this is one of the greatest merit of this approach. If there is at the scene an object executing a periodic motion, the blades of a fan for example, after a small time such a motion is absorbed by a Gaussian distribution and considered as belonging to the background.

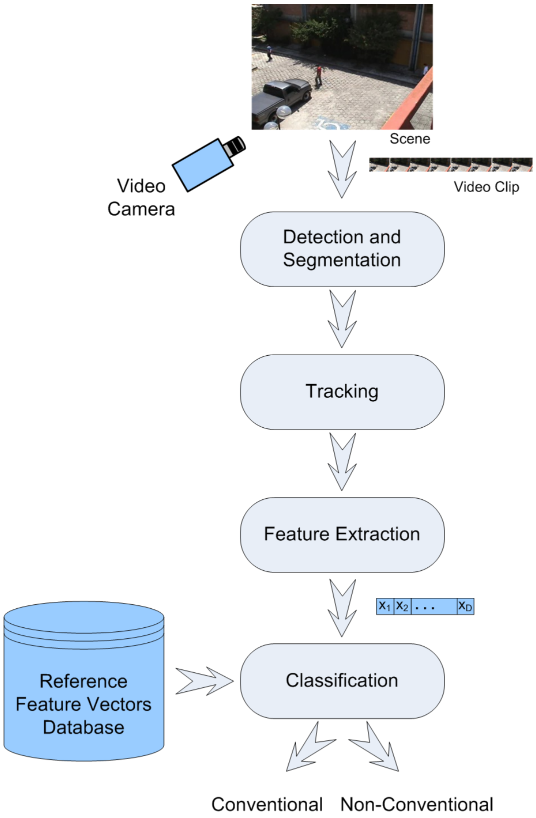

Fig. 1. Overview of the proposed approach to detect non-conventional human movements in video scenes.

However, for objects that present a slow motion, only the edges are highlighted. The central parts of the object are absorbed quickly, resulting in a set of sparse points of the object. To reconstruct the object without changing its size like as a morphological operation, a local background subtraction is carried out on these regions. A 3x3 window is applied at each pixel that is not absorbed by a Gaussian distribution, and inside such a window the pixels are subtracted from a fixed background. Thus if the pixel belongs to an object, the neighbor pixels that were previously absorbed by the Gaussian distribution will be highlighted. In this step, we can retrieve the pixels of object that was absorbed by gaussian function, but using a simple background subtraction these pixel are highlighted again.

To eliminate the remaining noise is applied a 3x3 median filter. The partial result is a set of pixels from the object in motion, possibly with non-connected pixels. A contour detection algorithm based on polygonal approximation is used to assure that these pixels make up a single object. In such a way, what was before a set of pixels is now a single object called blob which has all its pixels connected. Figures 2 and 3 show in a simplified way the detection and segmentation of foreground objects. Once a blob is identified, it must be tracked while it is presented in the camera field of view.

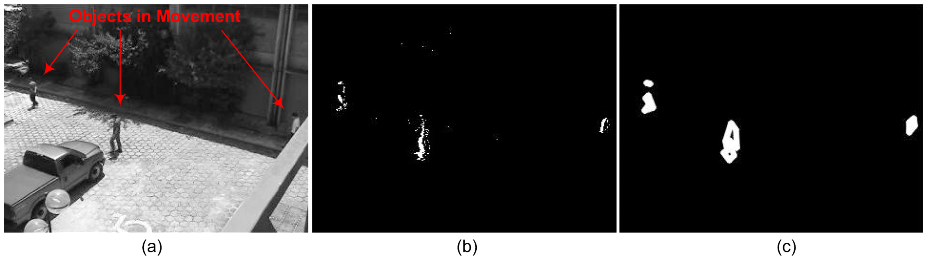

Fig. 2. An example of motion detection and segmentation on a video clip from Parking Lot Database: (a) original video frame with objects in motion, (b) motion segmentation by Gaussian distributions, (c) resulting *blobs* after applying filters, background subtraction and contour detection.

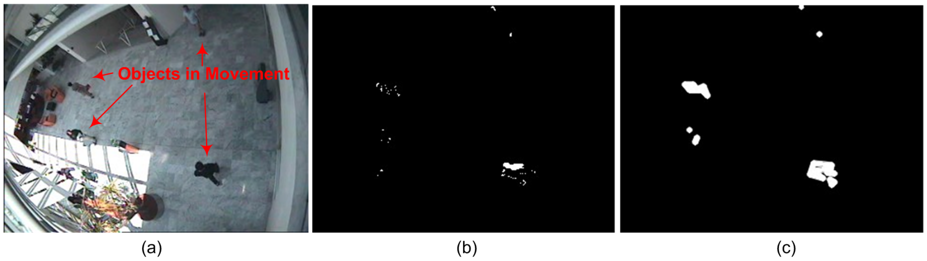

Fig. 3. Motion detection and segmentation in a video clip from CAVIAR Database: (a) original video frame with objects in motion, (b) motion segmentation by Gaussian distributions, (c) resulting *blobs* after filtering, background subtraction and contour detection.

2.2 Tracking Objects

The tracking consists in evaluating the trajectory of the object in movement while it remains in the scene. To eliminate objects that are not interesting under the point of view of the tracking, it is applied a size filter which discards blobs that are not consistent with the width and height of the objects of interest. The idea of using filters to eliminate undesirable regions was proposed by Lei and Xu [10], where the filtering take into account the velocity and direction of the motion applied to a cost function. The tracking of the objects in the scene and the prediction of its position in the scene is done by an approach proposed by Latecki and Miezianko [11] with some modifications. Suppose an object O^i in the frame F^n, where O^i denotes a tracking object. In the next frame F^{n+1}, given j regions of motion, R^j, we have to know which R^j represents the object O^i from the preceding frame. The following cost function is used:

$$Cost = (w_P * d_P) + (w_S * d_S) + (w_D * d_D) + d_T \tag{1}$$

where w_P, w_S, and w_D are weights that sum to one, d_P is the Euclidean distance in pixels between the object centers, d_S is the size difference between the bounding boxes of the region of motion, d_D is the difference in direction between the object position estimated by the Lucas-Kanade algorithm [12] and the last known center of the object in the preceding frames and the difference between the center of the region of movement and the center of the object, and d_T is the difference of the time to live (TTL) of the object. These parameters are better described as follows.

$$d_P = |R_c^j - O_c^i| \tag{2}$$

where R_c^j is the center of the region of motion and O_c^i is the last known center of the object. The value of d_P should not be higher than a threshold of proximity measured in pixels. This proximity threshold varies according to the objects are being tracked, mainly due to the speed of such objects in the scene.

$$d_S = \frac{|R_r^j - O_r^i|}{(R_r^j - O_r^i)} \tag{3}$$

where R_r^j and O_r^i denote the size of the box bounding the region of motion and bounding the object respectively.

$$d_D = |arctan(O_s^i - O_c^i) - arctan(R_c^j - O_c^i)| \tag{4}$$

where O_s^i is the object position estimated by Lucas-Kanade, O_c^i and R_c^j are the last know center of object and the center of region of motion respectively. The value of the angle lies between zero and 2π.

$$d_T = (TTL_{MAX} - O_{TTL}^i) \tag{5}$$

where TTL_{MAX} is the maximum persistence in frames and O_{TTL}^i is the object persistence . If the object is found in the current frame, the value of O_{TTL}^i is set to TTL_{MAX}, otherwise it is decreased by one until O_{TTL}^i becomes equal zero, where the object must be eliminated from the tracking. The TTL_{MAX} was set to 3 times the frames per second rate of the video.

Each object from the preceding frame must be absorbed by the region of motion in the current frame that leads to the lowest cost. The values of the object and bounding box centers assume the values of the regions of motion. If there is a region of motion that was not assigned to any object, then a new object is created with the values of such a region. If there is an object that was not assigned to any region of motion, such an object may be occluded and the *Lucas-Kanade* algorithm will fail to predict the corresponding motion. In this case, the motion of such an object is predicted as:

$$O_s^i = S * O_s^i + (1 - S) * (R_c^j - O_c^i) \tag{6}$$

where S is a fixed value of the speed. The region of motion R_c^j, should be the closest region to the object, respecting the proximity threshold. Then, the new position of the object and his bounding box is computed as:

$$O_c^i = O_c^i + O_s^i \tag{7}$$

$$O_r^i = O_r^i + O_s^i \tag{8}$$

3 Feature Extraction

Given an interval t of the trajectory of an object of interest, features are extracted from motion to make up a feature vector denoted by as V_i. Such a vector is composed by five features:

$$V_i = [v_{speed}, v_{posx,posy}, v_{disx,disy}, v_{sizx,sizy}, v_{dir}] \tag{9}$$

where v_{speed} denotes the speed of the object, $v_{posx,posy}$ denotes the coordinate x, y of the object in the scene, $v_{disx,disy}$ denotes the displacement of the object in x and y, $v_{sizx,sizy}$ denotes the temporal consistency of the bounding box based on the variation of its x and y dimensions, and v_{dir} denotes the variation in the direction of the object. These features are computed as:

$$v_{speed} = \sqrt{(O_{c_{t-1}}^i - O_{c_t}^i)^2}/Q \tag{10}$$

$$v_{disx,disy} = O_{c_{t-1}}^i - O_{c_t}^i \tag{11}$$

$$v_{sizx,sizy} = |O_{r_{t-1}}^i - O_{r_t}^i| \tag{12}$$

$$v_{dir} = arctan(O_{c_{t-2}}^i - O_{c_{t-1}}^i) - arctan(O_{c_{t-1}}^i - O_{c_t}^i) \tag{13}$$

The feature extraction is carried out considering an interval of Q frames. Such a value was defined empirically and set to $Q = 3$. Figure 4 illustrates the feature extraction process from a video and the generation of feature vectors.

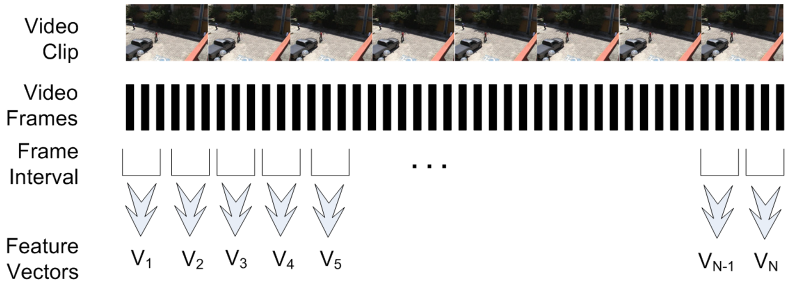

Fig. 4. An overview of the feature extraction process and generation of feature vectors from objects in motion along the scene.

4 Motion Classification

The feature vectors generated from the objects in motion are stored in a database to be further used by a non-parametric classifier. In this paper we have used a instance-based leaning algorithm due to the simplicity and low dimensionality of the feature vectors.

First, a database with reference vectors is generated from the analysis of objects in motion in the video frames. Each reference feature vector has a label assigned to it to indicate if it is representing a conventional (C) or a non-conventional movement (NC). This database is composed by reference feature vectors Z both from conventional and non-conventional movements. At the classification step a temporal window is used to classify segments of the motion almost in real-time. The classification consists in, given an object in motion, a set of features vectors are extracted $\mathcal{V}$. The number of vectors in the $\mathcal{V}$ set varies according to the size of the temporal window. In our case we have defined a temporal windows of size twenty seven frames, that is, the set $\mathcal{V}$ will be composed by nine feature vectors ($27/Q$, where Q is equal 3 which represents the feature extraction interval). The classification process is composed by two stages: first, each $V_i \in \mathcal{V}$ is classified using an instance-based approach, more specifically the k nearest neighbor algorithm (k-NN) [13]; next, the majority voting rule is applied to the feature vectors in $\mathcal{V}$ to come up to a final decision.

For the k-NN algorithm, the Euclidean distance among each feature vector in $\mathcal{V}$ and the Z reference feature vectors stored in the database. The Euclidean distance between a D-dimensional reference feature vector V_z and a testing feature vector V_i is defined as:

$$d(V_z, V_i) = \sqrt{\sum_{d=1}^{D} (V_{zd} - V_{id})^2} \tag{14}$$

The k-closest reference feature vectors will label each feature vector in $\mathcal{V}$ with their labels. After the classification of all feature vectors in $\mathcal{V}$, a final decision on the motion of the object is given by the vote of each member of the set $\mathcal{V}$, and the classification ”conventional” or ”non-conventional” is assigned to the object according to the majority vote. For example, if there are seven feature vectors in $\mathcal{V}$ classified by the k-NN as non-conventional (NC) and two classified as conventional (C), the final decision is to assign the label ”non-conventional” to the object. Figure 5 illustrates the classification process.

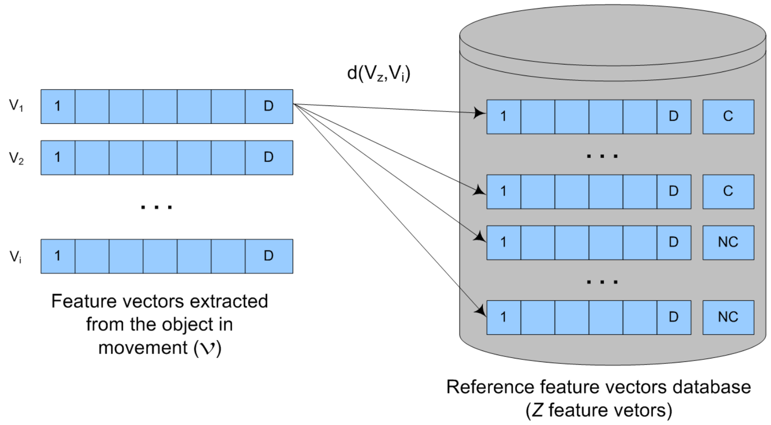
d(V$_z$,V$_i$)
V$_1$
V$_2$
V$_i$
1
D
...
C
NC
Feature vectors extracted from the object in movement ($\mathcal{V}$)
Reference feature vectors database (*Z* feature vetors)

Fig. 5. The classification process: the Euclidean distance between the feature vector extracted from the object in motion and the reference feature vectors stored in the database.

5 Experimental Results

The proposed approach was evaluated in two different databases. The first database consists in CCTV videos where people can execute three types of motion: walking, walking in *zig-zag* and running. These videos where captured in a parking lot through a security camera installed at top of a neighbor building and without any control of illumination and background with a resolution of 720 x 480 pixels, 30 frames per second and compressed using MPEG2. For each kind of motion two video clips with 100 seconds of length, were produced summing up to 200 seconds for each type of motion. For each type of movement, one video clip was used to generate the reference feature vectors (training) and the other was used for testing. The video clip lengths and the number of samples for each type of movement is shown in Table 1.

The main goal of the experiments is to evaluate the accuracy in detecting non-conventional events. Furthermore we are also interested in evaluating the discriminative power of the features. Since there is a low number of features, a force brute strategy was employed to evaluate the feature set. The weights

Table 1. Number of samples generated from the Parking Lot and from CAVIAR Database videos.

	Parking Lot		CAVIAR	
Event	Training	Test	Training	Test
Walking	94	112	57	23
Running	62	31	–	–
Zig-Zag	77	50	–	–
Fighting	–	–	41	16
Total	233	193	98	39

and thresholds described in Section 2.2 were empirically defined on the same video segments used as training. This is known as calibration procedure. The d_P proximity threshold was set to 40, TTL_{MAX} to 45, S to 0.9 and the values of the weights w_P, w_S, w_D to 0.4, 0.1 and 0.5 respectively.

In spite of having three types of motion in the videos, we have considered a two-class problem where "walking" is considered as a conventional event and walking in *zig-zag* and running were considered as non-conventional events. The accuracy is defined as the ratio between the number of events correctly classified and the total number of events.

Among all possible combinations of the features, for the Parking Lot database, the combination of only two features (speed and variation in the direction) has provided the best discrimination between events, see Fig.6. On the other hand the worst accuracy was achieved using only the size of the bounding box. Table 2 presents the confusion matrix for the best combination of features.

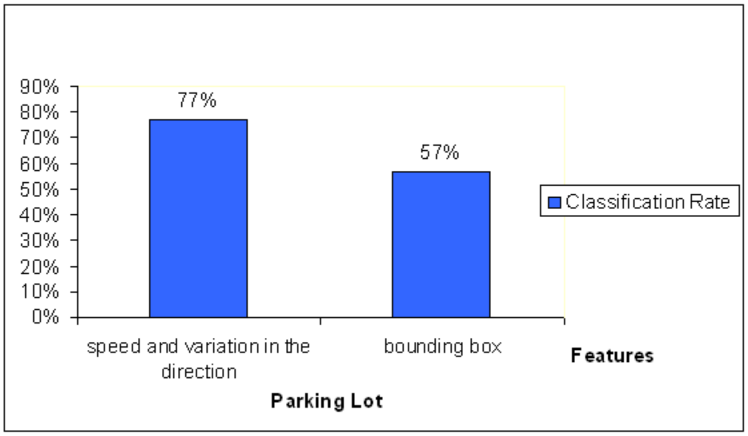

Fig. 6. Results for CAVIAR Database

Table 2. Confusion matrix for the combination of speed (v_{speed}) and variation in the direction (v_{dir}) features.

Movement	Conventional	Non-Conventional	
	Walking	Running	Zig-Zag
Walking	**90**	10	12
Running	3	**27**	**1**
Zig-Zag	19	**12**	**19**

The second experiment was carried out on some videos from the CAVIAR database [1]. One of the goals of this experiment is to evaluate the adaptability of the proposed approach to different scenarios and well as to different types of non-conventional events. The video clips were filmed with a wide angle camera lens in an entrance lobby. The resolution is half-resolution PAL standard (384 x 288 pixels, 25 frames per second) and compressed using MPEG2. For each kind of movement a number of some videos were used for training while the remaining were used for testing. The videos contain people executing two types of action: walking and fighting. The number of samples for each type of action is shown in Table 1.

Again, the main goal of the experiments is to evaluate the accuracy in detecting non-conventional events. Furthermore we are also interested in evaluating the discriminative power of the features. Among all possible combinations of the features, for the videos from the CAVIAR database, the combination of three features (coordinate, displacement and dimension of the bounding box) has provided the best discrimination between events, while the variation in the direction and bounding box has provided the worst (Fig.7). Table 3 presents the confusion matrix for the best combination of features.

Table 3. Confusion matrix for the combination of coordinate ($v_{posx,posy}$), displacement ($v_{disx,disy}$) and variation in the bounding box ($v_{sizx,sizy}$) features.

Event	Conventional Walking	Non-Conventional Fighting
Walking	**19**	4
Fighting	3	**13**

Above (Fig.8) we change the chosen features between the databases to compare the results. We can observe that is not possible apply the same collection of features into the two databases, but with a simple feature selection, the method is able to choose the better collection of features for the database.

[1] http://homepages.inf.ed.ac.uk/rbf/CAVIAR/

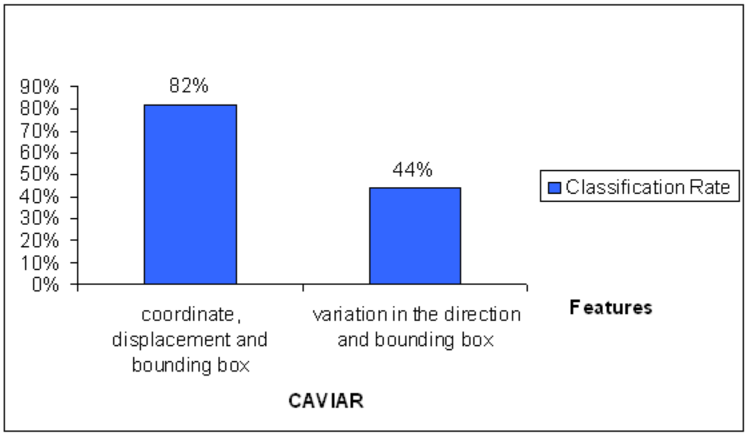

Fig. 7. Results for CAVIAR Database

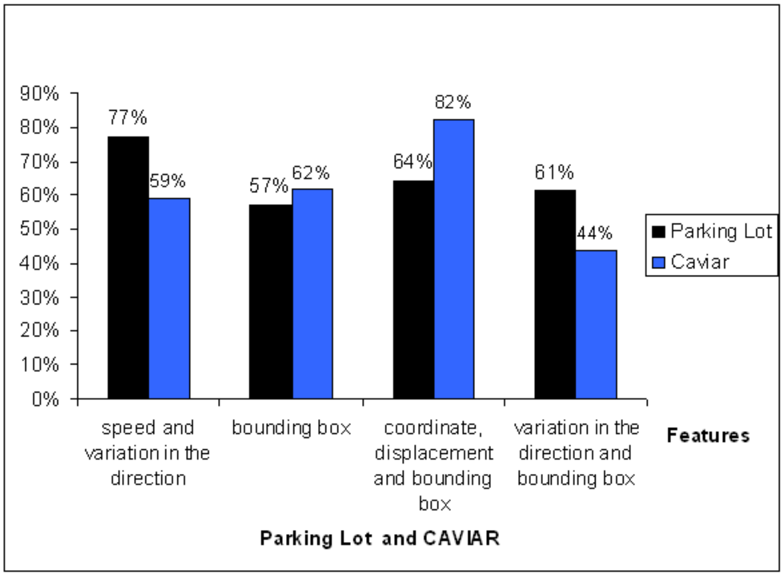

Fig. 8. Features applied into the two database

6 Conclusion

In this paper we have presented a novel approach to non-conventional event detection which is able to classify the movement of objects with relative accuracy. Experimental results on video clips gathered from a CCTV camera and from

CAVIAR database have shown the adaptability of the proposed approach to different environments. The proposed approach minimizes the use of contextual, said, information from the scene and from the objects in movement, giving priority to the adaptability to different scenes with a minimal amount of effort. In spite of the preliminary results are very encouraging, since we have achieved correct classification rates varying from 77.20% to 82.05% on video clips captured in different scenes, further improvements are required. Furthermore, broad tests in a variety of environments are also necessary. One of the main sources of errors is the problems related to occlusions. However this problem was not addressed in this work and it will be the subject of our future work.

The use of instance-based learning has lead to satisfactory results and the classification of the events was carried out almost in real-time due to the low dimension of the optimized feature vector as well as a database with few reference feature vectors (223 vectors for the first experiment and 98 vectors for the second experiment) for the Parking Lot database. However, one of the main drawbacks of the proposed approach is the necessity of positive and negative examples, said, examples of conventional an non-conventional events. Our on-going work is now focusing on the use of one-class classifiers which are able to model conventional events only since the capture of non-conventional events in real-life is a time-demanding task.

References

1. Hara, K., Omori, T., Ueno, R.: Detection of unusual human behavior in intelligent house. In: IEEE Workshop on Neural Networks for Signal Processing, Martigny, Switzerland (2002) 697–706
2. Lühr, S., Bui, H.H., Venkatesh, S., West, G.A.W.: Recognition of human activity through hierarchical stochastic learning. In: IEEE Annual Conf on Pervasive Computing and Communications, Fort Worth, USA (2003) 416–421
3. Brand, M., Kettnaker, V.: Discovery and segmentation of activities in video. IEEE Trans. Pattern Anal. Mach. Intell. **22**(8) (2000) 844–851
4. Aggarwal, J.K., Cai, Q.: Human motion analysis: A review. Computer Vision and Image Understanding **73**(3) (1999) 428–440
5. Mecocci, A., Pannozzo, M., Fumarola, A.: Automatic detection of anomalous behavioral events for advanced real-time video surveillance. In: IEEE Intl Symp on Computational Intelligence for Measurement Systems and Applications, Lugano, Switzerland (2003) 187–192
6. Niu, W., Long, J., Han, D., Wang, Y.F.: Human activity detection and recognition for video surveillance. In: IEEE Intl Conf Multimedia and Expo, Taipei, Taiwan, IEEE (2004) 719–722
7. Cucchiara, R., Grana, C., Prati, A., Vezzani, R.: Probabilistic posture classification for human-behavior analysis. IEEE Trans. on Systems, Man, and Cybernetics, Part A **35**(1) (2005) 42–54
8. Hu, W., Tan, T., Wang, L., Maybank, S.J.: A survey on visual surveillance of object motion and behaviors. IEEE Trans. Systems, Man, Cybernetics, Part C **34**(3) (August 2004) 334–352
9. Stauffer, C., Grimson, W.E.L.: Learning patterns of activity using real-time tracking. IEEE Trans. Pattern Anal. Mach. Intell. **22**(8) (2000) 747–757

10. Lei, B., Xu, L.Q.: From pixels to objects and trajectories: A generic real-time outdoor video surveillance system. In: IEE Intl Symp Imaging for Crime Detection and Prevention, London, UK (2005) 117–122
11. Latecki, L.J., Miezianko, R.: Object tracking with dynamic template update and occlusion detection. In: 18th Intl Conf on Pattern Recognition, Washington, USA (2006) 556–560
12. Lucas, B.D., Kanade, T.: An iterative image registration technique with an application to stereo vision. In: 7th Intl Joint Conf Artificial Intelligence, Vancouver, Canada (1981) 674–679
13. Aha, D.W., Kibler, D., Albert, M.K.: Instance-based learning algorithms. Machine Learning **6**(1) (1991) 37–66