ICCTA 2004
14th International Conference on
Computer Theory and Application
Alexandria, Egypt, 28-30 September 2004

Knowledge Based Video Surveillance System

Alberto Amato[1], Vincenzo Di Lecce[2], Vincenzo Piuri[3]
[1] II Engineering Faculty of Politecnico di Bari - Viale del Turismo 8 - 74100 Taranto – Italy
[2] DIASS - Politecnico di Bari - Viale del Turismo 8 - 74100 Taranto – Italy – corresponding author
[3] DTI - University of Milan - Via Bramante 65, 26013 Crema (CR) - Italy

Abstract –In this work, the authors propose a semantic event detection system based on a neural classifier and suitable for video surveillance.
Actual video surveillance systems are composed by a distributed network of active video sensors that present their output to a human operator. The main trouble is that human operators are often unable to monitor the output of a large number of video sensors. The goal of the proposed system is to automatically collect real time information to improve the situational awareness of security providers and decision makers.
Focus of the system is the evolution from classical "frame difference" paradigm to the new "known scene → no alarm" and "unknown scene → alarm" paradigm, where the meaning of scene is related to spatial-temporal events.
The proposed system is able to detect mobile objects in the scene and to class their movements as allowed or disallowed, in order to raise an alarm, also using a video camera mounted on a motorized pan scanner.
The results show that the system is able to compensate background changes due to camera motion and to detect mobile object in the scene.

Keyword – alarm detection, neural classifier, mobile camera.

I. INTRODUCTION

More and more cameras are on the market at a good price. This cost effectiveness let a video surveillance system be easily adopted. This fact has led to the spread of video surveillance systems composed by many video sensors. On the other hand, the great amount of visual material (to see) and the limit of the human attention in front of a monitor have led users to employ video data only "after the fact" as a forensic tool. In order to overcome this situation, many researchers are developing automatic methods to interpret the content of surveillance video sequences because, a continuous 24-hour monitoring of surveillance video is necessary to alert security officers to a disallowed act in progress, in this way they have enough time to avoid it.

The application of video surveillance has a wide range of purposes, from traffic monitoring [1] to human activity understanding [2]. For this reason, the difference in behavior of an automatic system depends on the particular application. For example, the surveillance system of the banks has the task of detecting an intrusion, while highway surveillance has to compute statistics about the traffic and to raise alarms in case of an anomalous situation.

In order to pay attention to a wide area, video surveillance systems use different kinds of camera such as: fixed cameras, omnidirectional cameras or mobile cameras. Detection and tracking of moving object are important tasks for visual based surveillance systems. Especially, it is highly difficult to detect moving objects from image sequences obtained by a moving camera because of the apparent motion of a static background [3, 4]. A large number of methods that face this problem have been proposed in literature; some are based on the direct detection of camera parameters with sensors [5], optical flow estimation [6] and the registration of the statistic background using a geometric transformation [7]. Once moving objects in the scene are successfully detected and tracked, information such as position, trajectories, etc., can be used to make assumptions of object behavior. Once the system owns a symbolic description of object behavior, alarm can be related to a particular event. In this way, the operator is lightened from the burden of discerning between standard or abnormal situations [8].

The natural evolution of the video surveillance systems is in the increase of background knowledge using neural networks, fuzzy logic and/or artificial intelligence approach.

Under this hypothesis the surveillance systems can evolve from classical "frame difference" to "known scene → no alarm" and "unknown scene → alarm", where the meaning of scene is related to spatial-temporal events.

Therefore no difference can exist from fixed to motorized pan scanner video camera.

In this paper, a neural network is used to generate a background knowledge represented by a variable number of points (centroides) in the frames space (each frame is represented by a n-dimensional vector also called signature). In the intrusion detection mode the distance from the actual frame among all computed centroides is calculated. If this distance exceeds a threshold an alarm will be released. The proposed system introduces a new method to solve the problem of the apparent background motion, in CCTV panning camera systems.

In the next sections, after a brief related work, the frame signature extraction method is presented, in section IV there is a brief discussion about the neural network utilized for background knowledge, section V describes the experiments conducted and the results obtained in this work and finally the conclusions and the final remarks are in section VI.

II. RELATED WORK

The main purpose of video surveillance is to guarantee realtime monitoring. The primary research issue of video surveillance is the automated motion detection. Finally, alerting responses to indicate the alarm conditions need to be presented to a decision maker. Thus a change detection model plays an important rule in a video surveillance system [9, 10, 11].

Marchesotti et al. [12] present an algorithm to track moving objects in a scene for dynamic occlusions between moving objects, especially tracking pedestrians moving in indoor environments.

The tracking method is based on joint application of Kalman filtering and correlation-based shape matching techniques. A technique based on contour analysis and shape description is used to count people in the scene, shape matching algorithm is based on the maximization of a correlation function changing according to the shape pose parameters.

A semi-automatic video surveillance approach has been presented in [13] to overcome the drawbacks of traditional video surveillance systems and an alarm generator has been implemented in order to pre-filter the data. A symbolic representation of events has been produced to generate alarms and to store them in a database for off-line consulting.

Event detection is considered in Stringa et al. [14]. The purpose of event detection is to send an alarm signal to the human operators, utilizing the fact that a pixel belongs to a moving object if it is different from both of its background and the previous acquired images. A sequence of interesting video shots is retrieved by using a feature vector to detect an event.

Isaac and Medioni [15] have suggested a graph representation of the moving regions extracted from a video acquired by a moving airborne platform. It allows dynamic inference of moving objects in order to perform robust tracking. It also provides a confidence measure characterizing the reliability of each extracted trajectory.

In [16], the authors show how an adaptive algorithm is a good possibility to robustly detect the illumination changes in outdoor video-surveillance images. This is achieved by using a background updating module together with a segmentation algorithm, especially for sudden changes.

Our research in this paper intends to provide a system that is able to detect mobile objects in the scene and to class their movements as allowed or disallowed, in order to raise an alarm, also using a video camera mounted on a motorized pan scanner.

III. FRAME SIGNATURE EXTRACTION

In this work, each frame is processed in order to extract a low level visual feature called angular spectrum [17] that can be stored in a data vector. This data vector contains data easier than those of the original image to handle. However it is able to represent the image in all the processes of video analysis.

The distance between two signatures is considered as related to image similarity and then it is used in the clustering phase as similarity assessment in order to compute the related centroids.

These centroids, obtained using a neural network clustering algorithm, represent the knowledge of the system.

The main considerations about this method are based on the fact that the image visual properties are mainly related to the largest image components, and among their features the shape, the texture and the orientation play a major role.

In many cases shapes can also be defined in terms of presence and distribution of oriented sub-components. A thick object has most of its lines arranged along its primary direction, while a thin elongated object has a peak in the line orientation distribution. Therefore the orientation of objects within an image is a key attribute in the definition of the images.

On the basis of this assumption, a metric for image classification has been defined. It is based on orientation, which is quantified by signatures formed from angular spectra of image components, in two-dimensional space.

The problem of finding the distribution of image lines direction has been approached analyzing its Fourier transform. A pre-processing phase is applied to the image before computing the Fourier transform: a 2D Hamming function transforms the signal describing the image into a periodic signal that can be represented by a finite number of frequency components. The visual effect achieved by the application of a Hamming function is comparable to a low pass filter whose effect is negligible in the centre of the image, and increases towards the borders. The borders reduce their contribution to the image frequency spectrum, focusing the analysis on the central part of the image.

The low frequencies correspond to the large components within the image, which are usually relevant to interpret the image and can be assumed to be more frequently localized in central areas (e.g., foreground objects) or to be spread in all the image (e.g., recurrent shapes, landscapes). Their contribution to the image is almost completely preserved. High frequencies correspond to small details and fine-grained textures; a reduction of these components means focusing on the foreground component, located in the image central area, and ignoring the details of the peripheral contour.

In the proposed application, for each image, a signature based on 7*10 values, is generated. The polar representation of the image spectrum is split in 10 sectors; in each sector 7 harmonics are evaluated and recorded in a histogram.

In order to evaluate the distance between two angular spectrum signatures, a weighted Euclidean distance has been used:

$$3)\quad dw(n,m) = w_{n,m} * d(n,m)$$

where the weight $w_{n,m}$ depends on the distance between the largest term of the reference image signature and the largest term in the first band (terms 1:10) of the examined signature.

IV. THE NEURAL NETWORK FOR BACKGROUND KNOWLEDGE

It is not easy to define a suitable training set for a neural network system that has to deal with real-world data that are complex and variable. In this work, a large number of video sequences, in which there are no alarms, have been processed using the technique showed in the previous paragraph. Once that each frame has been processed in order to obtain a signature, a neural network clustering algorithm is used to organize the signatures into clusters.

Clustering is an important technique used in discovering inherent structure present in the set of objects. Clustering algorithms attempt to organize unlabeled pattern vectors into clusters or “natural groups” such that, points within a cluster are more similar to each other than to points belonging to different clusters. Neural network have been employed in many clustering problems [18].

A neural network is constituted by a number of nodes connected by links, which are characterized by the connection of strengths (weights). The nodes are organized in layers (one input, one or more hidden and one output) and an activation function determines the transformation they perform. During the training phase the weights are adjusted in order to simulate the input-output relation that is present in the learning database. For this reason the choice of the database for tuning the network and for testing its performance is surely a crucial point [19].

Other important aspects of a neural network are: the number of nodes and connections, the number of hidden layers, the selection of the activation function and the number of iterations. It is known that the property of generalizing about the “knowledge”, acquired in the training phase, tends to diminish when the dimension of the neural network increases. Indeed, increasing the number of nodes and connections, it is possible to observe an increase of the performance referred to the learning set, in a certain sense it “stores” the pattern encountered in the training phase. This behavior may be useful in some applications such as recognition of well determined patterns, but it does not occur in case of alarm classification, in which the generalization property is essential. So, a minimization on the number of nodes and connections must be considered only designing a neural network for this kind of application.

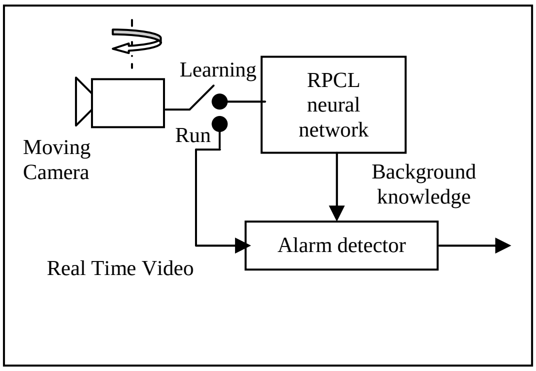

Figure 1 - a schematic view of the proposed system

In the proposed application a Rival Penalize Competitive Learning (RPCL) neural network composed of 200 elements has been utilized. Its learning coefficients are $\alpha_f(0)=0,2$ and $\alpha_r(0)=0,06$. Using about 5000 frames (500-700 frames per turn) and after 35 iterations, only 132 elements survive. The weight vectors of these elements have been adopted as centroids.

Figure 1 shows a schematic view of the proposed system. The training set is a set of frames in which there is no alarm. All these frames are processed in order to obtain their signatures as shown in the previous section. The so obtained signatures are clusterized using a RPCL neural network [20] owning the capability to detect the number of the classes in the dataset automatically.

The so obtained centroides (background knowledge) are used to compute the minimum distance to the current frame signature, and to raise an alarm if overcoming an experimental threshold.

It is well known that a perfect background knowledge is hard to come by in real world application. The typical situation in neural network design is that in which only partial domain knowledge is available. Though insufficient to determine the final neural network structure completely, such partial knowledge can be used to prime neural network design.

This fact justifies the experimented false alarm raised by the system in the first phase of working. In order to improve the performance, it is possible to include new frames (those that raise false alarms) in the training set.

One of the strength points of this system is his applicability to moving cameras.

This kind of video devices are very useful when a wide area range must be monitored. Nevertheless, the use of these devices leads to a new difficulty, indeed, the system must distinguish the real motion in the scene between the apparent motion introduced by the moving camera.

The proposed system resolves this problem using the fusion of the image retrieval approach and the neural network clustering. Indeed, each frame belongs to a cluster because his content is similar to that of the other images in that cluster regardless the motion condition introduced by the moving camera.

For example, this system can be suitable for the management of a crowded public space. In this kind of application one of the main concerns is the timely detection of excessive crowding level. In order to resolve this problem using the proposed system, the neural network should be trained using video sequence in which there is not an excessive crowding level. Thanks to the system architecture, the way in which the video is recorded (using a still camera or a moving camera) has no effects on the results.

V. EXPERIMENTS AND RESULTS

The proposed system has been tested with a moving camera rotating at 360 degrees and at a variable angular speed.

In order to populate the training set, several recording sessions have been done in various hours of the day in order to consider the daily light variation of the scene.

It is noteworthy that the angular spectrum signature has a good reaction to the light change, indeed, as figure 2 shows, a 40% variation of the brightness in a frame corresponds to a 7% variation in the relative angular spectrum signature.

Figure 3 shows the results of an experiment in which a video sequence of 5860 frames has been used to test the system. In this video there is a boy walking in a forbidden area around the frame number 2000.

The first frame that the alarm detector does not recognize is that number 2021. Figure 4 shows the first frame with alarm and the last frame without alarm.

The same procedure has been applied to twenty-five videos. These videos have been recorded in various situation such as: parking area, laboratory, classroom and home.

These experiments made with other video sequences have proved the quality of the results shown in the proposed example.

Figure 5 shows the trend of the probability to find an alarm in comparison to the dimension of the object expressed in percentage of the occupied image.

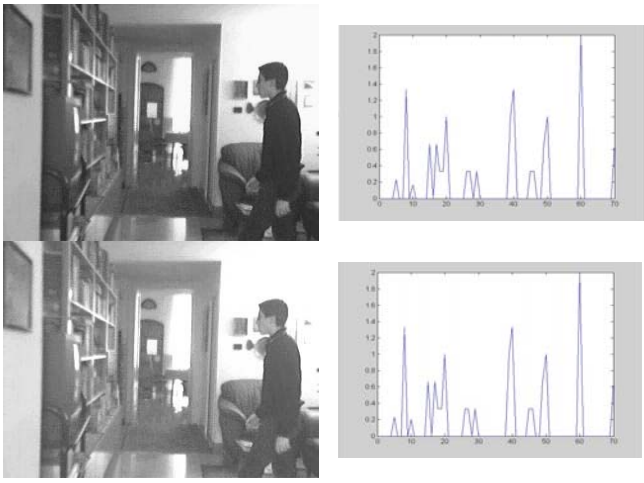

Figure 2 – this figure shows that a 40% variation of the brightness in a frame corresponds to a 7% variation in the relative angular spectrum signature

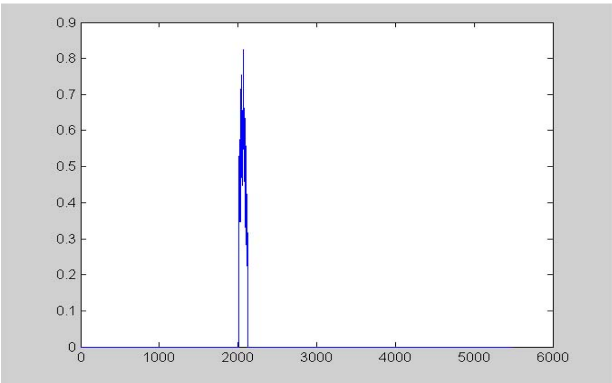

Figure 3 – This figure shows the output of the neural classifier. An alarm is visible round the frame 2000

Figure 4.a - The last allowed frame of the video sequence

Figure 4.b - The first disallowed frame of the video sequence

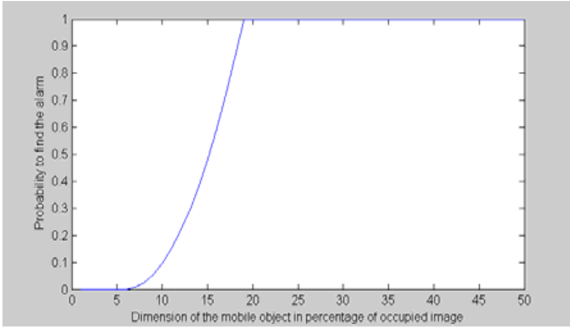

Figure 5 – This figure shows the trend of the probability to find an alarm in comparison to the dimension of the object expressed in percentage of occupied image

VI. CONCLUSIONS

In this work, the authors propose a video surveillance system that uses a neural network to generate a background knowledge.

Each video-frame is used to compute a signature according to the well known image retrieval approach.

The used signature shows a low sensitivity to the brightness variation (a 40% variation of the brightness in a frame corresponds to a 7% variation in the relative angular spectrum signature) and it is suitable for creating very homogenous clusters. This characteristic is very important, because, video surveillance systems usually videotape 24 hours each day, so the daily brightness variation produces a variation in the scene that must not raise an alarm.

The combination of the image retrieval approach with the neural network approach finds a solution to the problem of the apparent motion of a static background in a mobile camera switching from the old concept of classical "frame difference" paradigm to the new "known scene → no alarm" and "unknown scene → alarm" paradigm, where the meaning of scene is related to spatial-temporal events.

The preliminary performance evaluation confirms that the proposed combination reaches satisfactory performances in the video surveillance task.

REFERENCES

[1] D. Koller, K. Daniilidis, H. H. Nagel, Model-based object tracking in monocular sequences of road traffic scenes. *International Journal of Computer Vision*, Vol. 10, 1993, pp. 257-281.

[2] Yuri A. Ivanov and Aaron F. Bobick, Recognition of Multi-Agent Interaction in Video Surveillance, {ICCV} (1), pp. 169-176, 1999.

[3] G. L. Foresti and C. Micheloni, A robust feature tracker for active surveillance of outdoor scenes. *Electronic Letters on Computer Vision and Image Analysis 1(1):21-34, 2003*.

[4] S. Araki, T. Matsuoka et al., Real-time tracking of multiple moving object contours in a moving camera image sequence, IEICE Trans. Inf. & Syst., VOL. E83-D, No. 7, July 2000.

[5] D. Murray and A. Basu, Motion tracking with an active camera, IEEE Trans. Pattern Anal. & Mach. Intell., Vol. 16, no. 5, pp. 449-459, May 1994.

[6] D. Nair and J. Aggarwal, Detecting unexpected moving obstacles that appear in the path of navigating robot, Proc. ICIP'94, vol. 2, pp. 311-315, 1994.

[7] J. Odobez and P. Bouthemy, Detection of multiple moving objects using multiscale MRF with camera motion compensation, Proc. ICIP'94, vol.2, pp. 257-261, 1994.

[8] L. Marchesotti, L. Marcenaro and C. Ragazzoni, A video surveillance architecture for alarm generation and video sequences retrieval, Image Processing. 2002. Proceedings.

[9] Foresti, G. L. Mahoen, P. and Regazzoni, C. Multimedia Video-Based Surveillance System, Requirements, Issues and Solutions. Kluwer Academic Publishers, 2002, USA.

[10] Regazzoni, C. Fabri, G. and Vernazzza, G. Advanced Video-based Surveillance System. Kluwer Academic Publishers, 2002, USA.

[11] Remagnio, P. Jones, G. Paragios, N. and Regazzoni, C. Video-based Surveillance Systems. Computer vision and Distributed processing. Kluwer Academic Publishers, 2002.

[12] Marchesotti, L. Marcenaro, L. Regazzoni C. S. Tracking and counting multiple interacting pedestrian in indoor scenes, in Proceedings of third IEEE international workshop on performance evaluation of tracking and surveillance, (Copenhagen Denmark, June 1, 2002).

[13] Marchesotti, L. Marcenaro, L. Regazzoni, C.S. A video surveillance architecture for alarm generation and video sequences retrieval, in Proceedings of ICIP2002 (Rochester NewYork USA, September 2002).

[14] Stringa, E. Regazzoni, C.S. Content-based retrieval and real time detection from video sequences acquired by surveillance systems, in Proceedings of ICIP98 (Chicago Illinois, Oct. 1998), Vol. III, pp.138-142.

[15] Isaac, C. Medioni, G. Detecting and tracking moving objects for video surveillance, in Proceedings of ICPR (Fort Collins CO. Jun. 23-25, 1999).

[16] Marcenaro, L. Gianluca, G. Regazzoni, C. Adaptive change detection approach for object detection in outdoor scenes under variable speed illumination changes, in Proceedings of Eusipco 2000 (Tampere, Finland).

[17] V. Di Lecce and A. Guerriero, "A Comparative Evaluation of retrieval methods for Duplicate search in Image database", Journal if Visual Languages and Computing, 2001, 12, 105-120.

[18] Lisboa, P.G.J. (1992). Neural Networks: Current applications Chapman and Hall, London

[19] G. Bortolan, R. Degani, JL Willems, ECG Classification with Neural Networks and Cluster Analysis, Computers in Cardiology, 1992.

[20] Xu, A Krzyzakand E. Oja, Rival Penalized Competitive Learning for Clustering Analysis, RBF Net and Curve Detection, IEEE Trans. On Neural Networks Vol. 4, No 4 July 1993, pp 636 – 649