CIHSPS 2005 – IEEE International Conference on
Computational Intelligence for Homeland Security and Personal Safety
Orlando, FL, USA, 31 March – 1 April 2005

Neural Network Based Video Surveillance System

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

Abstract *– Video surveillance systems are usually composed of a network of active video sensors that continuously capture the scenes and present them to a human operator for analysis and event detection. Unfortunately human operators are often unable to monitor the video streams coming from a large number of video sensors.*
In this paper a semantic event detection system based on a neural classifier is presented to screen continuous video streams and detect relevant events, specifically for video surveillance. The goal of the proposed system is to automatically collect real-time information to improve the awareness of security personnel and decision makers.
Our research is focused on the use of the "known scene → no alarm / unknown scene → alarm" paradigm, where the meaning of scene is related to spatial-temporal events, instead of the classical "frame difference" paradigm. The proposed system is able to detect mobile objects in the scene and to classify their movements (as allowed or disallowed) so as to raise an alarm whenever unacceptable movements are detected. This ability is supported also for video cameras mounted on a motorized pan scanner: experiments showed that the system is able to compensate the background changes due to the camera motion.

Keywords *– Alarm detection, neural classifier, mobile camera.*

I. INTRODUCTION

Since video cameras are nowadays on the market at a good price, video surveillance systems become more and more popular. Video surveillance is in fact adopted in a wide range of application cases, from traffic monitoring [1] to human activity understanding [2].

In particular, the use of many video sensors to ensure a high coverage is economically viable. This leads to capture a great amount of visual material to be monitored and screened for event detection. However, the limited abilities of the human personnel in performing this task in real time, on multiple screens, and with a constant level of attention may reduce in the practice the actual usability and effectiveness of multi-sensor real-time video surveillance, often confining video data to a forensic tool for after-crime inspection.

To overcome this drawback many researchers are developing automatic methods to analyze and understand the content of surveillance video sequences: the real-time screening of continuous video streams will allow for immediately alerting the security officers whenever a disallowed event occurs so that they can take the appropriate actions. These techniques are unfortunately strictly related to the specific behavior and the detailed operation of the video surveillance system and, thus, to the specific application. For example, the surveillance system in a bank is expected to detect intrusions, while in highway applications it computes statistics about the traffic and raises alarms in the case of anomalous traffic flows.

To monitor a wide area video surveillance systems use different kinds of cameras, such as fixed, omni-directional, and mobile cameras. Detection and tracking of moving object are the fundamental tasks performed on the image streams flowing from these sensors. Heterogeneity and camera mobility must be taken into account to create flexible and effective systems. In particular, camera movements must be carefully compensated in the image analysis in order to identify correctly the events occurred in the scenes. It is especially difficult to detect moving objects in image sequences captured by a moving camera because of the apparent motion of a static background [3, 4]. Several methods, proposed in literature, deal with removal of the background apparent motion, for example those based on the direct identification of the camera parameters with suitable sensors [5], the 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, kinematical information (e.g., position, trajectories) can be extracted and used to make assumptions on the object behavior. By exploiting the symbolic description of the object behavior, the system can associate an alarm to each specific event. In this way, the operator is lightened from the burden of monitoring continuously all scenes and can focus his attention on the possibly abnormal situations [8].

An adaptable and flexible approach is desirable to achieve high accuracy and robustness analyzing video streams, detecting objects and tracing them. The classical "frame difference" techniques are not sufficiently able to reach the desired quality since they are essentially focused on the local information flow that they observe. A semantic-based approach is therefore valuable to incorporate and exploit the background knowledge into the analysis and detection process. The approach considered in this paper is therefore based "known scene → no alarm / unknown scene → alarm" strategy, where the meaning of the scene is related to the spatial-temporal events identified in the streams. This approach, implemented by using computational intelligence techniques, allows for overcoming the differences between fixed and motorized pan scanner video camera.

In this paper we tackle therefore the real-time analysis of video streams for surveillance applications, namely detection and identification of human presence and motion, by using a semantic-based approach to create a modeling and analysis framework which is independent from the specific application case and to abstract from the characteristics of the cameras (fixed, mobile, or panning devices). Real-time operation is taken into account as a fundamental goal to allow the effective use in continuous monitoring. Computational intelligence technologies are adopted to introduce flexibility and accuracy, without impairing system performance; namely, neural networks are used to hold the knowledge about the background so as to achieve camera independence and solve, in an innovative way, the problem of the apparent background motion at low computational cost. Each image frame is represented by a n-dimensional vector, called signature, which characterizes the content of the frame itself in the semantic frame space. For intrusion detection an alarm is raised when the distance of the current frame signature from the ones previously collected on that area exceeds a given threshold.

Section II will briefly review the related works; in Section III the overall structure of the proposed system is presented. In Section IV the frame signature extraction method is proposed, while Section V deals with the use of neural technologies to create the knowledge about the background scene. Section VI presents the experiments and the achieved results.

II. RELATED WORK

The main purpose of video surveillance is to guarantee real-time monitoring to detect the presence and the motion of persons in critical areas. The primary research issue in video surveillance is therefore automated motion detection, so that the alerting responses pointing out possibly critical conditions can be presented to the final human decision maker. Consequently, the change detection model plays an important rule in a video surveillance system [9, 10, 11].

An algorithm was presented in [12] to track moving objects in the scene by considering dynamic occlusions among moving objects; this is especially useful to track pedestrians moving in indoor environments. The tracking method is based on 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; the shape matching algorithm is based on the maximization of a correlation function that changes according to shape pose parameters.

A semi-automatic video surveillance approach was presented in [13] to reduce the number of alarms presented to the user. A symbolic representation of events was introduced to generate and manipulate the alarms at a high-abstraction level as well as to store them in a database for subsequent off-line analysis. An alarm generator was implemented to pre-filter the incoming data.

Event detection is another important function in surveillance system directed to send an alarm signal to the human operators. The approach proposed in [14] exploits the fact that a pixel belongs to a moving object if it is different both from the background and the previous acquired images. To detect an event a sequence of relevant video shots is retrieved by using a feature vector.

A graph representation was suggested in [15] to represent the moving regions which were extracted from a video acquired by a moving airborne platform. This approach allows dynamic inference of moving objects to achieve robust tracking and provides a confidence measure characterizing the reliability of each trajectory.

Robust detection of illumination changes in outdoor video-surveillance images was effectively achieved in [16] by using an adaptive algorithm. This approach dynamically learns and updates the background by using a segmentation algorithm, especially for sudden changes.

III. THE REAL-TIME, KNOWLED□E-BASED HETERO□ENEOUS-VIDEOCAMERA SURVEILLANCE SYSTEM

Our research was directed to create a system being able to detect mobile objects in the scene observed with various types of video cameras (including the ones mounted on motorized pan scanners) and to classify their movements as allowed or disallowed so that a suited alarm can be raised.

In Figure 1 a block diagram of our system is introduced. The system is composed by the following blocks:

1. *Moving Camera*. As input sensor, in this work we tested a fixed camera (AXIS 205) and a moving camera (AXIS 2130).
2. *Knowledge Collector*. This component, implemented by using a RPCL neural network, captures the background knowledge in the video sequences (i.e., the centroids of the frames) and provides information to remove the background from the subsequent real-life images. To reduce the amount of data to be handled a low-level visual feature (called string signature [17]) is extracted from each frame and stored in a data vector. This data vector contains data which can be handled in an easier way than those of the original image, even though it completely represents the image in the subsequent analysis.
3. *Alarm Detector*. This module analyzes the scene and raises alarms when appropriate. This process consists of the following steps:
   a. the camera sends a frame;
   b. the signature is extracted from the frame;
   c. the frame is classified as “known” or “unknown” by using the knowledge stored in the neural network;
   d. if the scene is classified as “unknown” then the system raises an alarm.

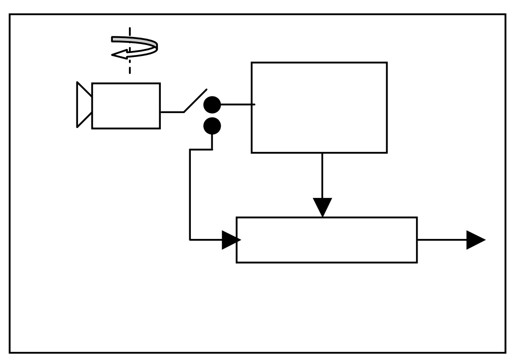

Figure 1 - a schematic view of the proposed system

IV. FRAME SI□NATURE EXTRACTION

A low-level visual feature (called string signature [17]) is extracted from each frame of the video to represent concisely the image contained in the frame itself. This should be enough significant and representative of the image content so that it can be used in the video analysis process and, especially, in motion detection and image retrieval.

It is well known that edges convey essential information in a picture. Their accurate detection is therefore fundamental. Discovering edges in an image is in fact the first step to recognize geometric shapes within the image itself. Edge detection, ideally, identifies all the curves outlining the objects in an image (often called shape or contour). The contour is in fact frequently used in image retrieval.

In this paper an innovative contour-based system is used to simplify and compact video frames. This allows for extracting a string-based image description, consisting of the local contour orientation, suited for motion detection and image retrieval from a database.

To detect the motion as well as to search for images in the video database the distance among signatures is related to the similarity among the corresponding images. Therefore the signatures can be used in the clustering phase as similarity assessment to compute the centroids of the classes, which represent the background knowledge acquired by the system in the video analysis.

While investigating the typical characteristics of images in the perspective of retrieval, the following properties were observed in [18] and assumed in this paper as hypotheses to study the opportunities for creating efficient image retrieval systems:

- all objects in an image are characterized by edges (limiting a shape or constituting a boundary in the cases which are most useful for the image understanding);
- the fundamental objects characterizing the semantic content of an image are in foreground.

In our extensive experiments [17] we also noted a further typical aspect concerning the main local directions of the edges in an image, which we consider as an additional assumption:

- every edge in an image can approximately described by using a sequence of prevalent directions chosen in the set composed only of the four principal 2D directions (namely, horizontal, vertical, and the two diagonal ones) and by the lack of direction. These principal directions can be identified by the symbols "—", "|", "/", "\", and "o" (or "null"), respectively.

Since the global structure of an image can be described approximately by the prevalent directions of the edges, the image can be therefore concisely represented as the string of symbols associated to the prevalent directions of the edges contained in the image. This string is obtained by suitably concatenating the symbols that represent the local directions of the edges themselves. Since the string of symbols characterizes the directions of the edges, it also constitutes a signature for the whole image (called symbols string signature).

To create this signature we first need to identify the main edges in the image, for example by using the Canny's algorithm [19]. The edged image (i.e., the image in which only edges are left) is divided in windows by using an *n*x*n* regular grid. In each window the prevalent direction θ of the edges is computed according to the method presented in [20]. This approach is based on the fact that the gradient of a line in any of its points is a vector and, thus, has both magnitude and direction. For an edge in the window analysis, in the discrete case, the magnitude $|\Delta\square|$ and the local edge direction θ in a point of the edge can be approximated by:

$$|\Delta G| = \left(|\Delta_H| + |\Delta_V|\right)/2$$

$$\theta = tan^{-1}\left(\Delta_V / \Delta_H\right) + \pi/2$$

where Δ_H and Δ_V are the 3x3 Sobel's operator in the horizontal and vertical directions in the considered point. The prevalent direction θ is a real number $(0 \le \theta \le \pi)$, measured counter-clockwise, being the horizontal direction zero.

For each image window the prevalent directions histogram is created as follows. In the abscissa we consider the 5 discretized prevalent directions aforementioned in the assumption 3. For each of these directions we count the number of occurrences having magnitude $|\Delta\square|$ not smaller than a threshold *t* in the image pixels (being *t* a non critical parameter). The prevalent direction in the window is the discretized prevalent direction corresponding to the maximum value in the prevalent direction histogram.

To each image window we associate the symbol corresponding to the discretized prevalent direction computed in this way. This matrix of symbols describes concisely the image. The symbols string signature is finally obtained by concatenating the symbols associated to each image window by increasing row index and, within each row, by increasing column index.

To evaluate the similarity between two images, I_1 and I_2, the distance between the two corresponding symbols string signatures, S_1 and S_2, must be defined. According to [21] the transform list is adopted to support this distance measurement. Any string S_1 can be transformed into any other string S_2 by adding, deleting, or replacing one symbol at a time. The transform list stores the sequence of strings (by starting from S_1) that are produced by a sequence of applications of these operators to produce the string S_2. The length of the transform list is the number of operators that have been applied in the sequence (i.e., the number of strings minus 1). A minimal transformation list is one having the minimum length among all possible sequences of operators and starts from S_1 to generate S_2, i.e., one that needs the minimum number of operators to complete the desired transformation. The distance between two strings S_1 and S_2 is the length of a minimal transform list.

V. THE NEURAL NETWORK FOR BACK□ROUND KNOWLED□E IDENTIFICATION

To capture the background knowledge in the video sequences (i.e., the centroids of the frames) and to remove it from the subsequent real-life images, a neural network approach was adopted.

A suitable training set was defined: a large number of video sequences in which there were no alarms were processed by using the technique discussed in the previous section. The signature was extracted for each of these frames.

A neural network clustering algorithm was used to organize the signatures into clusters [22]. Clustering is an important technique used to discover the inherent structure present in the set of the envisioned objects. In the case of alarm classification the generalization property of neural networks is essential to deal with small variations of the background. In our research a Rival Penalize Competitive Learning (RPCL) neural network [23] was used to compute the centroids.

During the operating life the distance of the current frame is measured in relation to the centroids, which constitute the background knowledge. The centroid having the minimum distance from the current frame signature is the one to which the current image is more similar. If this distance is higher than an experimentally-defined threshold, the frame is considered to be not sufficiently similar to the ones represented by that centroid; therefore an alarm is raised since something suspicious may have been occurred in the scene.

It is well known that perfect background knowledge is hard to obtain in real world applications. The typical situation is therefore that only a partial knowledge of the domain is available. Though often not completely sufficient to determine the final neural network structure, such a partial knowledge can be used to prime the neural network design. This fact justifies the experimented false alarm raised by the system at the beginning of its operating life on real-world video sequences. To improve the quality of the operation, the frames generating these false alarms during the initial life can be included later in the training set and network configuration updated by incorporating such new knowledge.

The proposed approach and system is very robust and effective in dealing also with 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 in frame processing and image understanding: the system must distinguish the real motion in the scene from the apparent motion introduced by the moving camera. The proposed system solves this problem by merging the image retrieval approach with the neural network clustering. In fact, each frame belongs to a cluster because its content is similar to that of the other images in such a cluster, regardless from the possible motion introduced by the moving camera.

Consequently, our system is very suited for monitoring a crowded public space. In this kind of application one of the main concerns is in fact the timely detection of excessive crowding level. To tackle this problem by using the proposed system, the neural network capturing the background knowledge must be trained by using video sequences in which there is not an excessive crowding level. Thanks to the system architecture, the way in which the video is recorded (i.e., by using a fixed camera or a moving camera) has no effects on the final results.

VI. EXPERIMENTS AND RESULTS

The proposed system was implemented and tested with fixed cameras and with moving cameras, which is capable of rotating for 360 degrees and at a variable angular speed. In this paper we report specifically the results for the more complex case of moving cameras.

To populate the training set, several recording sessions were performed in various hours of the day so that to incorporate the daily light variation of the scene. The string signature is quite robust to the light change: it is shown in figure 2 that 40% variation of the brightness in a frame corresponds to only 7% variation in the relative string signature. This characteristic is very important because video surveillance systems usually videotape 24 hours each day: even though the daily brightness variation produces a variation in the scene, an alarm should not be raised only for this reason.

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

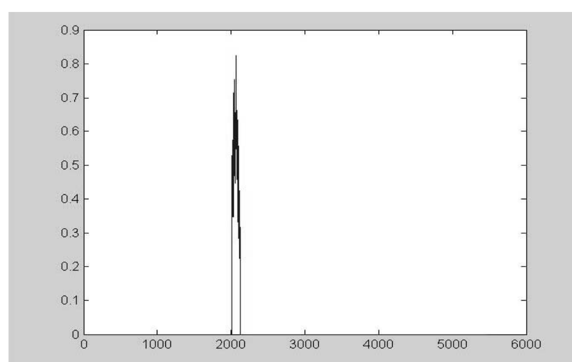

Figure 3 – This figure shows the output of the neural classifier. An alarm is visible round the frame 2000

Figure 3 shows the results of an experiment (composed by 5860 frames) in which a boy is walking in a forbidden area (around the frame number 2000). The first frame that the neural network for alarm detection does not recognize is the number 2021. Figure 4 shows the first frame with alarm and the last frame without alarm.

The same procedure was applied also to other twenty-five videos recorded in various situations (e.g., parking areas, laboratories, classrooms, and home). These experiments confirmed the abilities and the quality of the proposed approach.

Figure 5 shows the trend of the probability to find an alarm in relation to the dimension of the object, given as percentage of the area occupied in the image.

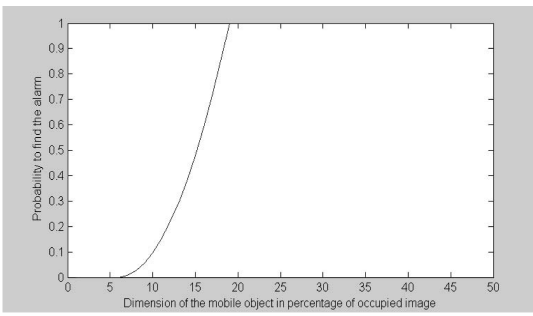

Figure 5 – This figure shows the trend of the probability to find an alarm in comparison to the dimension of the object expressed in percentage of occupied image

VII. CONCLUSIONS

A new approach to design and implement video surveillance systems was presented in this paper. The main goals were to easily and robustly capture the background knowledge in streaming video for video-based monitoring applications, and to make an efficient use of moving cameras.

A neural network was used to understand and learn the background knowledge. For each video frame a signature was computed according to the well-known image retrieval approach. This signature shows a low sensitivity to the brightness variation (40% variation of the brightness in a frame corresponds to only 7% variation in the relative signature) and it is suitable for creating very homogenous clusters. These features are very desirable in real-world video surveillance systems working 24-hour a day, under varying illumination conditions.

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

The preliminary performance evaluation confirms that the proposed approach achieves very appealing performances for video surveillance applications.

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

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

[11] Remagnio, P. Jones, . 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, . Detecting and tracking moving objects for video surveillance, in Proceedings of ICPR (Fort Collins CO. Jun. 23-25, 1999).

[16] Marcenaro, L. ianluca, . 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] A. Amato, V. Di Lecce, "Edge Detection Techniques In Image Retrieval: The Semantic Meaning Of Edge", EC-VIP-MC 2003 4th EURASIP Conference, 2-4 July 2003, Zagreb, Croatia.

[18] K. S. Fu, "Syntactic Pattern Recognition", Prentice-Hall, Englewood Cliffs, New Jersey, 1982.

[19] J. Canny, "A computational approach to edge detection", in *IEEE Trans. Pattern Anal. Mach. Intell*, Vol. 8, Issue 6, 1986.

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

[21] Oommen B. J., Loke R.K.S., "On Using Parametric String Distances and Vector Quantization in Designing syntactic Pattern Recognition Systems", in *IEEE International conference on Computational Cybernetics and Simulation*, Volume 1, 12-15, Oct 1997.

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

[23] 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