An Integrated Surveillance System for Outdoor Security

C. Micheloni, E.Salvador, F. Bigaran and G.L.Foresti
Department of Mathematics and Computer Science, University of Udine
Via delle Scienze 206, 33100 Udine, ITALY

Abstract

An integrated system for the detection, active tracking and recognition of people in wide outdoor environments will be hereafter discussed. Specifically, a static sensor with a wide view is used to detect people inside the environment and to classify their behaviours. As outcome of anomalous activities, an active camera is selected to focus its attention on a particular person. Here, techniques for face detection are employed to determine a region of interest where to extract features used by a tracking algorithm for an autonomous gaze of a PTZ camera. Finally, a face recognition phase will be considered to recognize the person of interest. Results will show how the integrated system is able to detect, track and recognise people inside a tough environment such as a parking lot.

1. Introduction

One of the objectives of advanced video surveillance applications is the automatic identification of people for security purposes. An effective solution for enhancing monitoring and control capabilities of remote human operators without requiring the participants cooperation or knowledge can be provided by person identification systems based on the analysis of face images. Recognition of faces from a video sequence in an uncontrolled environment is one of the most challenging problems in face recognition [1]. Due to the fact that video acquisition usually occurs outdoors or indoors but with bad conditions for video capture and to the fact that subjects are not cooperative, large illumination and pose variations in the face images may be encountered. In addition, partial occlusion and disguise are possible. Moreover, typically the face image sizes are small and make not only the recognition more difficult, but also affect the accuracy of face detection, as well as the accurate extraction of features that are often needed in recognition methods.
This paper describes work aimed at performing automatic person identification through face recognition in a realistic outdoor surveillance scenario. In particular, the proposed integrated multisensor system focuses on the challenging problem of acquiring good quality face patterns from a distance for an accurate discrimination. A static camera is used to identify the presence of people and analyse their behaviour and an active pan-tilt-zoom (PTZ) camera is exploited to provide useful image resolution for recognition purposes. Relatively few efforts can be found which have addressed the issue of acquiring data at multiple resolutions for extracting higher quality, relevant information. In [2], Trivedi et al. propose a framework for analysing human activities in indoor intelligent spaces using networks of static and active cameras. Modules for multicamera-based person tracking, event detection, focus of attention and streaming face recognition are discussed. In the work of Hampapur et al. [3], controlled pan-tilt-zoom cameras driven by a 3D wide baseline stereo tracking system are used to automatically acquire zoomed-in views of a persons head in indoor environments for recognition purposes. In this work, we consider an outdoor surveillance scenario.
The paper is organized as follows. The general architecture of the system is described in Section 2. In Section 3, the main processing steps related to the static sensor are briefly discussed. Section 4 presents the proposed active vision techniques used to track the targets of interest and to acquire close-up views of their faces for recognition purposes. Experimental results are discussed in Section 5 and conclusions are finally drawn in Section 6.

2. System architecture

The architecture of the proposed surveillance system is illustrated in Figure 1. A two-camera system made up of a static and an active camera connected to two processing nodes is adopted. The static camera covers a wide area and is employed to monitor the behaviour of people in the environment. The active camera, characterized by controllable zoom and pointing direction (pan and tilt angles), is used to obtain a closer view of people of interest and higher resolution images of people's faces. An initial camera calibration and registration process exploiting the ground-plane hypothesis [15] allows to establish a correspondence between camera views so that position information of moving objects can be exchanged in a common coordinate frame. In the first analysis phase, people in the field of view of the static camera are detected and tracked over time and the observed trajectories are analysed in order to extract events of

0-7803-9385-6/05/$20.00 ©2005 IEEE

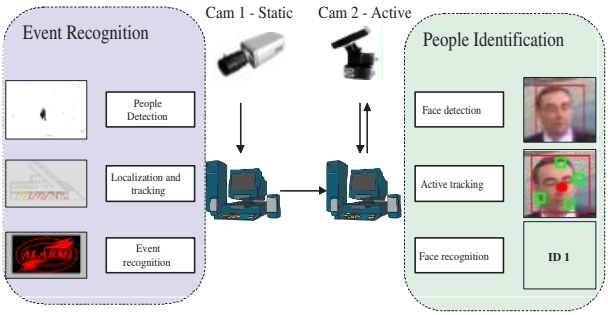

Figure 1: System architecture.

interest for the classification of people's behaviour.
Relying on the results of the first stage of analysis, the active PTZ camera allows an event-based attention focusing. When an anomalous behaviour is signalled, the PTZ camera is used to focus the system's attention on the people involved, with the objective of recognizing their identity. To this end, face detection, tracking and recognition techniques are employed. Face detection and tracking allow the extraction of image regions corresponding to faces and of features inside these regions that are exploited to gaze the PTZ camera at the target. The sequence of face images is then analysed by the face recognition module.

3. People detection and event recognition

The processing node connected to the static sensor performs first of all all the low-level image processing operations needed to identify moving objects by means of change detection and to extract features for object classification and tracking.
Adaptive background modelling, image differencing and adaptive thresholding are employed to extract blobs of the moving objects in the scene [14]. Blob attributes (dimensions, area, centroid's coordinates) are then extracted to distinguish people from other object classes, such as vehicles, and to establish a correspondence between instances of the same object in successive frames. This allows to keep track of the object's position over time [4]. Thanks to camera calibration, the position and motion parameters of each detected object on a 2D top view map of the monitored environment are also computed [15] through a perspective transformation matrix.
To classify people's behaviour and to detect anomalous events, computed trajectories are analysed. A "middle-level" analysis is first of all performed to process trajectory information in such a way that useful information can be provided to an high-level event recognition system for the identification of anomalous behaviours. A trajectory clustering and analysis algorithm is adopted [16]. The algorithm is dynamic and works on the fly. Trajectories with common features are grouped in clusters and probabilistic predictions on the future movements of the object are used to improve the detection.
In order to task the active camera, events are defined by a set of k classified objects over a sequence of n consecutive frames. An off-line Event Database is built with the models of normal and anomalous events. For each event class, a set of features characterizing the observed trajectory is extracted and stored in the database. An adaptive high order neural tree (AHNT) is then applied for recognizing events. More details about the event recognition algorithm can be found in [4]. The detection of an anomalous event raises an alarm which is communicated to the second processing node, connected to the active sensor. Information about the position of the area of interest in the monitored scene is also sent to the active camera which will be controlled to focus the system's attention on it.

Figure 2: Example of feature extraction within the ROI for a detected face.

4. Focus of attention for person recognition

The proposed focus of attention is the result of the cooperation between a face detection module and an active tracking module. In particular, when the PTZ camera is tasked for target tracking, the face detection module is exploited to better drive the gaze control. Initially, the PTZ control system detects motion in the area subject to suspects behaviours in order to facilitate a first face detection. Once a face is detected (or a group of faces in case of a group of people), the active tracking module switches his operative mode from survey to focusing. In this phase, trackable features are extracted within the bounding box containing the face and are tracked to estimate the motion of the pattern of interest (see Figure 2). The active tracking proceeds then autonomously to track the features by rejecting those not accurately tracked and by adding new features as substitutes until a new face pattern is trustworthy detected. Moreover, the area covered by the extracted and tracked features is exploited to control the zoom parameter. As the area becomes

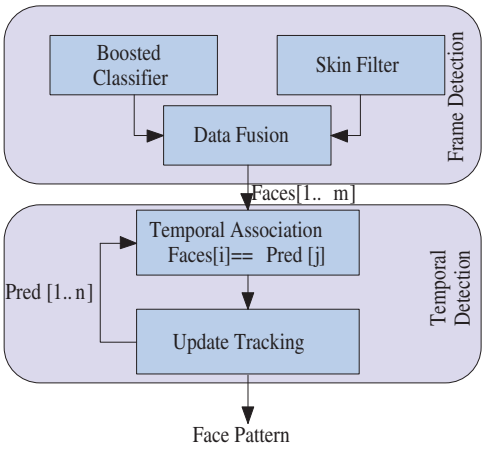

Figure 3: Block diagram of the proposed face detection and tracking module.

wider a zoom out operation is performed while a zoom in is carried out when the area shrinks. Each processing step is described in more detail in the following sections.

4.1. Face detection and tracking

The proposed face detector integrates different modules in order to improve the overall detection performance by means of tracking capabilities and information fusion techniques. As illustrated in Figure 3, it is composed by a face pattern detector, a skin filter, a module to fuse skin and face pattern information, and a tracking module. The video stream is processed and a set of face patterns associated to each person in the scene is extracted.

The face pattern detection module is based on the boosted cascade classifier proposed by Viola and Jones [5] and Lienhart and Maydt [6]. In the adopted scheme, every stage of the cascade is built of basic classifiers boosted using Gentle Adaboost [12] and the input of the basic classifiers are Haar-like features. Since such method results in a large number of false positives (i.e. detection of non-face patterns), additional techniques have been integrated to reduce this type of errors.

A first improvement has been obtained by considering the fact that colour can provide very useful information for face detection in complex environments. A skin filter [7], based on the YC_bC_r colour space for computational efficiency, has been therefore developed to extract image regions corresponding to skin. Only face patterns including a skin region are validated. Although some non-skin patterns are inevitably mismatched with skin, the skin filter allows to reduce the overall false positives. Moreover, as illustrated in Figure 4, inside the face ROI, obtained by enlarging the output of the face pattern detector (face bounding box) of

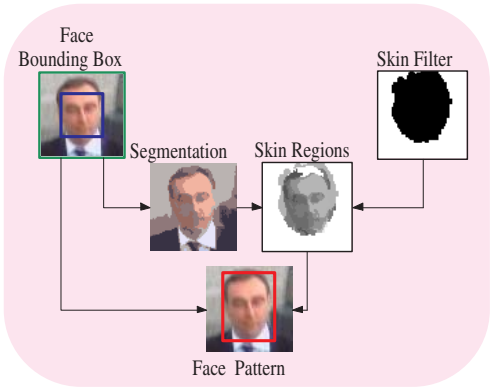

Figure 4: Block diagram of the proposed data fusion module.

1.5 times, a pyramidal segmentation is executed. The resulting regions are validated by the skin filter to obtain the skin regions. If the majority of a skin region's area is within the face bounding box, then such region is retained and considered to expand the bounding box in order to include the region entirely. This procedure allows to increase the robustness of the final detection.

A second improvement has been obtained by means of tracking face patterns over time. A camshift technique, initialised on the previous face pattern, is employed to predict the position of the face pattern in the current frame. This allows to speed up the detection in the current image and to remedy to miss detections. When no face pattern is detected by the face detector for a predefined number of frames, the face is considered as lost and a new track is initiated as soon as a new face pattern is detected.

4.2. Active tracking

The active tracking module receives face patterns from the face detector and extracts a set of features for each pattern inside the corresponding bounding box to determine a fixation point. To this end, the work of Tordoff and Murray [13] has been considered. They solve the problem of tracking a fixation point, typically the centre of mass of an object, by adopting an affine projection. In particular,it is possible to consider the following affine projection:

$$\mathbf{x} = MX + \mathbf{t} \tag{1}$$

where $\mathbf{x}$ is the fixation point, M is the affine transform, X is the real position of the point and $\mathbf{t}$ is a translation vector. The main issue in adopting this technique concerns the selection of the point to use for the affine transform computation. We have adopted a feature clustering technique able to classify a tracked feature as belonging to different moving objects or to the background. In the context of a

Pan&Tilt camera movement such a method can be used to find the most reliable cluster defining the background displacement [11]. However, in the case of a zoom operation, it is not possible to identify a unique cluster containing the features of the background to represent its motion.
Therefore, to distinguish the clusters we need first to compute the affine transform for each tracked object. Such a computation is performed over all the features belonging to an object. In particular, let $\hat{A}$ be the computed affine transform and f_i^{t-1}, f_i^t the position of a generic feature respectively at time instant $t-1$ and t. At this point, the effective displacement of the feature i is $\mathbf{d}_i = f_i^t - f_i^{t-1}$, while the estimated one for the effects of the affine transform is as follows:

$$\tilde{\mathbf{d}}_i = f_i^t - \tilde{f}_i^{t-1} = f_i^t - \hat{A} f_i^{t-1} \quad (2)$$

where $\tilde{f}_i^{t-1}$ represents the position of the feature i after the compensation of the camera motion by means of the affine transform $\hat{A}$.
Let TFS_{obj} be the set of features extracted from a window (i.e. *fovea*) centred on the object of interest. Then, the clusters computation is performed according to the following rule:

$$C_{obj}(\tilde{\mathbf{d}}) = \left\{ f_i \in TFS_{obj} | \left\| \tilde{\mathbf{d}}_i - \tilde{\mathbf{d}} \right\|_2 \leq r_{tol} \right\} \quad (3)$$

where $C_{obj}(\tilde{\mathbf{d}})$ is the cluster having all the features i whose displacement $\tilde{\mathbf{d}}_i$ is such that the norm between it and the vector $\tilde{\mathbf{d}}$ is lower than a defined threshold r_{tol}. Once the computation of all the clusters is done, we can easily find the background cluster and therefore all the features that have been erroneously extracted from the background in the previous feature extraction step. After having deleted all the features either belonging to the background or to a cluster with cardinality lower than three, we can apply the technique proposed by Tordoff-Murray over each cluster to determine the fixation point.
Let $\mathbf{g}_i$ be the fixation point we need to compute for each object i, then we need to solve the following equation:

$$\mathbf{g}'_{\tilde{\mathbf{d}}_k} = N\mathbf{g}_i + \mathbf{r} \quad (4)$$

where $\mathbf{g}'_{\tilde{\mathbf{d}}_k}$ is the new position of the fixation point for the set of features belonging to the cluster $C_{obj}(\tilde{\mathbf{d}}_k)$, computed from the position of the old fixation point $\mathbf{g}$ of the same object.

4.3. Face recognition

The objective of the face recognition module is to determine the identity of the detected person's face by comparing it with face images of known identity stored in a database (the reader is referred to [8, 9] for a comprehensive critical survey of face recognition techniques).

In this work, a simple approach based on the Fisherfaces technique proposed by Belhumeur et al. [10] has been tested. Face recognition is performed on each face image extracted by the previously described face detection module. Each image is downsampled to 25×25 pixels for computational efficiency. In order to deal with uncertainty in the face detection process and, in particular, to reduce the effect of misalignment of detected test faces with respect to trained classes, the face pattern is translated of some pixels (up to 3 pixels in the current implementation) in every direction in search of the position which provides the highest similarity with one of the reference classes. Single-frame recognition results are then accumulated over a sequence of L frames before taking a decision on the recognized identity. A weighted majority decision rule is applied to the stream of recognized identities to improve recognition rates. For each single-frame recognition decision, the corresponding weight is proportional to the computed similarity between test image and reference class in the database.

5. Experimental results

Experiments have been conducted on a real scenario represented by a parking lot. The 640×480 static camera images have been acquired with a Cohu 3812 while the 768×576 moving camera images have been acquired with a Panasonic NV-GS11 mounted on a PTU 47.60. Thirteen single person and five multiple people test sequences have been processed for a total footage of $3830 frames$[1] with 5643 face patterns. As the processing of the static images resulted in the activation of an alarm the active camera was tasked to redirect its gaze at the person of interest. At the end of the repositioning phase the face detector started to extract face patterns. An average detection rate of 92.4% has been achieved with 1.3% false positive rate. As a second metric to assess the performance of the proposed face detector the euclidean distance between the barycentre of the detected face ROI and its ground truth position has been computed. As can be seen in Figure 5, the accuracy of the detector is always less than 20 pixels with average $\mu = 10$ pixels and standard deviation $\sigma = 5$ pixels.
To give an evaluation of the active tracking module we first analysed it capability of continuously maintain the gaze on the target of interest. It is interesting to notice that the person was tracked continuously also during those frame in which the detection failed. In addition, the fixation point was always kept within the face patterns. To further evaluate the active tracking subsystem, we computed two different metrics. First of all, we measured the error given from the difference between the computed position of the fixation point and the human computed ground truth position (see Figure 6 continuous line). The average error was about

[1] Available at http://avires.dimi.uniud.it/avires/sequences/face/index.html

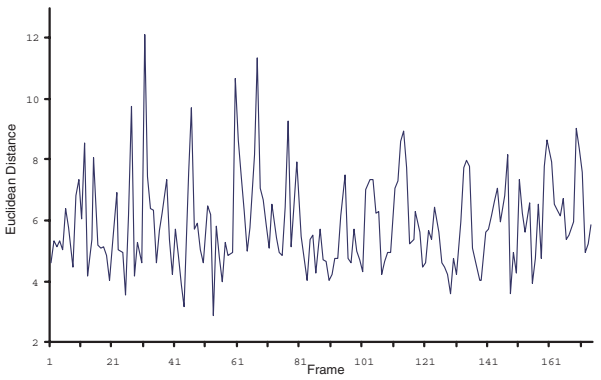

Figure 5: The chart plots the euclidean distance computed between the barycentre of the detected ROI and the ground truth face barycentre.

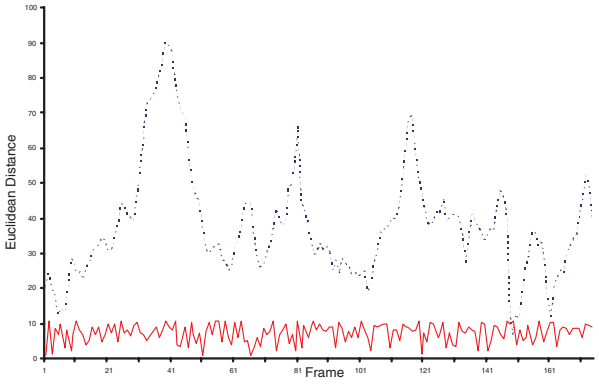

Figure 6: The graph shows the euclidean distance between the position of the fixation point and the centre of the image (dotted line) and the ground truth face barycentre (continuous line).

7.34 pixels with a standard deviation $\sigma = 2.65$ pixels. Since common cameras use to focus the centre of the image, we also computed the average euclidean distance between the fixation point and the centre of the image (see Figure 6 dotted line). The average value was about 38.16 pixels with a standard deviation $\sigma = 16.08$ pixels.
Figure 7 shows some results obtained by processing a test sequence. It is worth noting, in the $2nd$ frame of the $2nd$ row, how the active tracking supplies to a miss detection of the face detector. In the last column frames, an example of how the extracted ROI is used to reject tracked features not belonging to the face pattern is shown. The lower right feature lies outside the ROI and is therefore rejected.

As regards the recognition phase, the system has been trained on a gallery of 5 individuals, with ten face patterns for individual extracted from the recorded videos. Single person videos have been used for testing.The size of face bounding boxes in the probe videos ranges from 28×28 pixels to 74×74 pixels. Partitioning the sequences into non-overlapping segment sequences of L=10 frames, an average correct recognition rate of 64% has been obtained by classifying the original face pattern. By applying the search procedure for the minimum distance bounding box, the recognition rate is improved to 69%. An average recognition rate of 80% and 100%, respectively, has been reached by considering the whole sequences.

6. Summary and Conclusions

An active system for gazing at objects of interest for higher resolution face acquisition has been developed for person identification purposes. Precisely, a static camera is used to monitor the environment in order to detect suspicious events. When a suspect activity arises, an active camera is tasked to acquire face patterns of suspect people. To achieve such results, a robust face detector has been employed to detect face patterns and determine their ROI. These are used by an active tracker to extract and track features that are processed to maintain the fixation point on the target. Finally, a face recognition technique is applied on the face patterns for person identification. When compared with state of the art systems, which are typically tested in relatively controlled environments and are therefore hardly applicable to real world surveillance scenarios, the achieved recognition performance is strongly affected by the tough capture conditions, in terms of illumination, viewpoint, face size, face pose, etc. (see Figure 7). Extensive testing on larger data sets and the use of more reliable techniques based on a spatio-temporal representation of face appearance and dynamics are under investigation and will allow to improve the system performance.

References

[1] A.K. Jain, S. Pankanti, S. Prabhakar, L. Hong, A. Ross, *Biometrics: a grand challenge*, Proc. of 17th Intl. Conf. on Pattern Recognition, ICPR 2004, Aug. 2004, Vol.2, Pages: 935 - 942.

[2] M. M. Trivedi, K. S. Huang, I. Mikic, *Dynamic Context Capture and Distributed Video Arrays for Intelligent Spaces*, IEEE Trans. on Systems, Man and Cybernetics, Part A, Volume: 35, Issue: 1, Jan. 2005. Pages: 145-163.

[3] A. Hampapur, S. Pankanti, A. Senior, Y.-L. Tian, L. Brown, R. Bolle, *Face Cataloger: Multi-Scale Imaging for Relating Identity to Location*, Proc. of IEEE Conf. on Advanced Video and Signal Based Surveillance (AVSS'03), pp. 13-20.

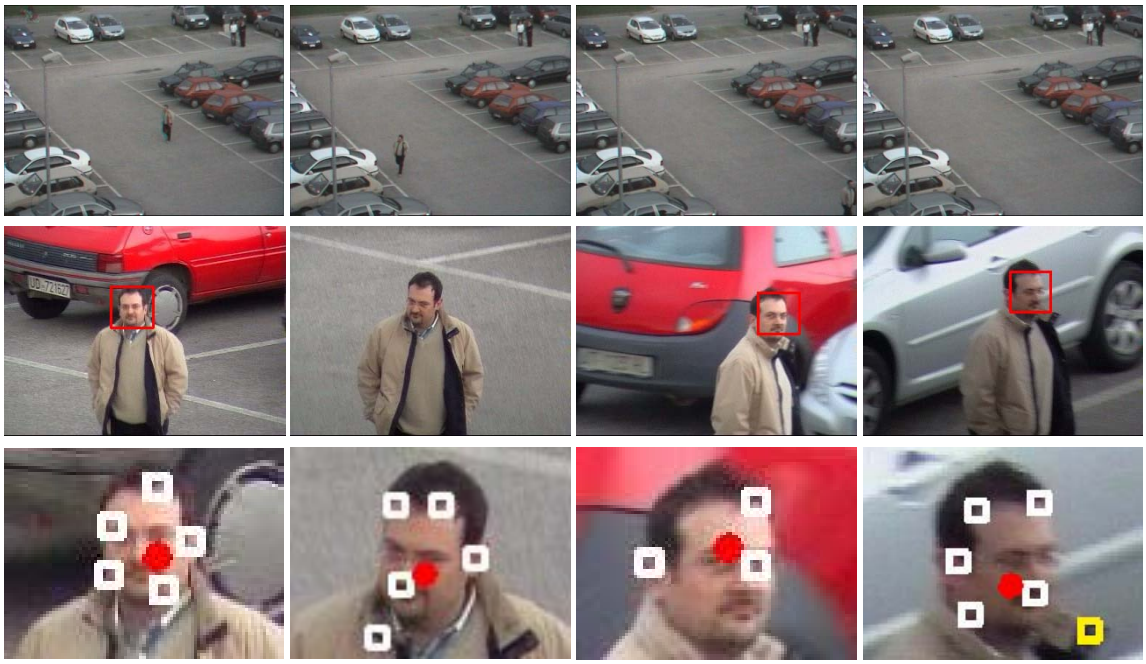

Figure 7: First row shows few frames of a test sequence acquired by the static camera with a wide angle of view. Second row shows the results of the face detector while the third shows the results of the active tracking. In particular, the tracked features (boxes) and the fixation point (circle) are shown. Validated features are drawn in lighter colour with respect to the colour used for rejected features.

[4] G.L. Foresti, C. Micheloni, L. Snidaro, *Event classification for automatic visual-based surveillance of parking lots*, Proc. of IEEE Int. Conf. on Pattern Recognition (ICPR 2004), Volume 3, Pages: 314 - 317.

[5] P. Viola and M. J. Jones. *Rapid Object Detection using a Boosted Cascade of Simple Features*, IEEE CVPR, 2001.

[6] R. Lienhart and J. Maydt. *An Extended Set of Haar-like Features for Rapid Object Detection*, IEEE ICIP 2002, Vol. 1, pp. 900-903, Sep. 2002.

[7] C. Garcia and G. Tziritas, *Face detection using quantized skin colour regions and wavelet packet analysis*, IEEE Transactions of Multimedia, Vol.1, n.3, 1999.

[8] W. Zhao, R. Chellappa, A. Rosenfeld, and P. Phillips, *Face recognition: A literature survey*, ACM Computing Surveys, 35(4), pp. 399-458, 2003.

[9] S. G. Kong, J. Heo, B. R. Abidi, J. Paik and M. A. Abidi, *Recent advances in visual and infrared face recognition - a review*, Computer Vision and Image Understanding, 97(1), pp. 103-135, 2005.

[10] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, *Eigenfaces vs. fisherfaces: Recognition using class specific linear projection*, IEEE Trans. Pattern Analysis Machine Intelligence, 19(7), pp. 711720, 1997.

[11] C. Micheloni, G.L. Foresti, and F. Alberti, *A new feature clustering method for object detection with an active camera*, IEEE International Conference on Image Processing, pages 271–275, Singapore, October 24-27, 2004.

[12] R. Lienhart, A. Kuranov, and V. Pisarevsky, *Empirical analysis of detection cascades of boosted classifiers for rapid object detection*, DAGM03, Magdeburg, Germany, September 2003.

[13] B.J. Tordoff and D.W. Murray, *Reactive control of zoom while tracking using perspective and affine cameras*, IEEE Trans. Pattern Anal. Machine Intell., 26(1):98–112, 2004.

[14] G.L.Foresti, C. Micheloni, L. Snidaro, P. Remagnino and T. Ellis, *Active video-based surveillance system: the low-level image and video processing techniques needed for implementation*, Signal Processing Magazine, IEEE Volume 22, Issue 2, March 2005 Page(s):25 - 37.

[15] C. Micheloni, G.L. Foresti and L. Snidaro, *A Cooperative Multicamera System for Videosurveillance of Parking Lots*, IEE Intelligent Distributed Surveillance Systems, Febrary 26, London, UK, pp. 5/1-5/5,2003.

[16] C.Piciarelli and G.L. Foresti, *Event Recognition by Dynamic Trajectory Analysis and Prediction*, IEE Image for Crime Detection and Prevention, June 2005, Savoy Place, London, UK, pp. 131-134.