

International Journal of Pattern Recognition and Artificial Intelligence
© World Scientific Publishing Company

A survey of automatic event detection in multi-camera third generation surveillance systems

Tiziana D'Orazio

*Institute of Intelligent Systems for Automation - C.N.R.
via Amendola 122/D-I Bari, 70126, ITALY
dorazio@ba.issia.cnr.it*

Cataldo Guaragnella

*Politecnico di Bari, Via Orabona 4,
Bari, 70126, ITALY
guaragnella@poliba.it*

Third generation surveillance systems are largely requested for intelligent surveillance of different scenarios such as public areas, urban traffic control, smart homes and so on. They are based on multiple cameras and processing modules that integrate data coming from a large surveillance space. The semantic interpretation of data from a multi view context is a challenging task and requires the development of image processing methodologies that could support applications in extensive and real time contexts. This paper presents a survey of automatic event detection functionalities that have been developed for third generation surveillance systems with a particular emphasis on open problems that limit the application of computer vision methodologies to commercial multi-camera systems.

Keywords: Multi-view; video analysis; image processing; low-level and high-level feature extraction; modelling of feature dynamics; event detection.

1. Introduction

The term third generation surveillance systems was introduced at the beginning of 2000 to refer to systems that provide full digital solutions to the design of surveillance systems, starting at the sensor level, up to the presentation of visual information to the operators¹. These systems represent the evolution of the first generation surveillance systems (analogue CCTV Systems from multiple remote locations which present video data to the human operators) and the second generation surveillance systems (automated visual surveillance which combines advances in digital video communications and image processing techniques with CCTV systems). Both the drop in prices of technological equipments and the increasing demand of security and safety systems, have brought to a large diffusion of surveillance cameras especially in wide public environments, such as airports, metro and railway stations, malls, parking lots, museum and so on. The large number of cameras used to cover these areas provides huge data that actually do not contain any ac-

2 *T. D'Orazio*

tionable information, since monitoring is tiring, expensive and ineffective². The development of an additional layer of intelligent processing is the central point of third generation surveillance systems in order to automatically detect "interesting" events in the constant flow of video and provide alarms to the surveillance staff that evaluates and responds to the events.

The point on which the scientific community has been debating in the last years, is the concept of "interesting" and the related computer vision methodologies that can be involved to highlight these "interesting" events from a large flow of noisy data. Some functionalities, already available to automatically detect simple events, have been applied to second generation surveillance systems. In the context of smart homes, motion detection capabilities are enough to guarantee the requested security level³. Surveillance systems that connect cameras via wireless video servers to Home PCs offer simple motion detection capabilities and are on sale at hardware and consumer electronics stores for few hundreds of dollars. Abandoned/removed object detection approaches are available that use both fixed or moving cameras with an environment model^{4,5}. People accessing to forbidden areas can be detected by using calibrated cameras and motion detection approaches⁶. Some functionalities are already available for transportation infrastructures, such as the detection of traffic flows⁷, the detection of vehicles stopping in forbidden areas⁸, and so on.

When the concept of "interesting" events becomes more complex, the automatic detection requires a high level semantic interpretation that is not always easy to perform. For example, if the surveillance system has to detect not only abandoned/removed objects but also to recognize the person who carried the object, the task becomes more challenging since it is necessary to track people in the environment, estimate their positions, evaluate interactions with objects. Let's suppose that a surveillance system is asked to recognize people loitering. In this case it is necessary to track people among multi cameras, evaluate their trajectories, compare them with trajectories of the normal flow of people, and in case provide an alarm. Surveillance systems can also be asked to detect panic situations in flows of people, or to recognize brawls in the crowd. These systems should be able to analyze the motion parameters of moving areas and also to discern anomalous situations in long term observations. Eventually, cameras controlling access points can be asked to recognize that people enter with their personal passes and do not use other people passes. In these cases, biometric analysis are required. These are just few examples of the possible applications that can be developed to support intelligent surveillance systems.

From the computer vision point of view many methodologies have been already developed to solve specific tasks, but their applications in multi view and real contexts are not always immediate. The semantic interpretation becomes challenging when the events happen in different areas of the environments and requires the observation and recognition of the same object in different cameras that due to different positions, optical characteristics, fields of view can perceive the same object

in a very dissimilar color and shape.

Although many efforts have been done in this direction, much work has to be done in order to have architectures and methodologies that could support applications in extensive and real time contexts. Survey on remote surveillance systems for public safety have been published in the last years^{9,10,11,12} and tried to cover all the aspects of architectures, technologies and applications. Some recent reviews are specific on selected topics such as action recognition, activity recognition from 3D Data, or wireless video architectures^{13,14,15}. Aim of this paper is to review the literature of the last years but from a different perspective: the central idea is to understand the type of events that can be recognized by a multi-camera surveillance system. Without the presumption of being exhaustive, we have selected from the large number of publications on this subject, those papers which have demonstrated an advantage gained from the multi-camera framework, for the recognition of events occurring in large areas. As the application contexts are different we cannot provide comparisons of the relative merits of the different approaches but only an overview of the field and an insight in the methodologies that are mature for the development of some simple surveillance functionalities. We have organized the selected literature according to the types of events and the camera architectures that can be used. At the end of this review we summarize the image processing methodologies in relation with the camera network topologies, and we highlight open problems and limitations in order to give clear advices to interested readers on the points that require more research efforts.

The remaining of this paper is organized as follows. The literature has been divided in three main areas: surveillance with passive cameras, surveillance with active cameras, and video anomalies detection (see figure 1).

In particular in the passive cameras group (section 2) we have identified different application areas: vehicle tracking is essential for transportation applications such as monitoring traffic parameters or preventing accidents for safety issues. People tracking across a multi camera system is the preliminary step for any further analysis of behavior understanding. Human action recognition techniques starting from the silhouettes of segmented people from different cameras try to detect robust features that allow action recognition. Finally we have considered the problem of detecting the camera network topology when large networks of un-calibrated cameras are used. The use of active cameras (section 3) in camera networks has the great advantage of reducing sensing resources to monitor the same area but poses more complex problems to coordinate data among cameras with different and time varying characteristics. In section 4 we have considered those applications in which anomalies are detected without object detection and tracking and are based on motion detection and recognition of abnormal situations that could take place both in unusual times and locations. A brief survey of public datasets, which are used to compare performances of different methodologies, is reported in section 5.

2. Passive Cameras

When passive cameras are used, according to the applications and the surveillance tasks, different camera network topology overlaps can be chosen (see figure 2). Totally overlapping cameras (see fig. 2(a)) provide a complete coverage of the observed scene and can be used for specific surveillance tasks in which the exact pose of the observed objects is necessary. Non Overlapping and Partially overlapping cameras (see fig. 2(b) and (c)) are the most common configurations for large area surveillance tasks. Partially Overlapping or not Overlapping Top view cameras (see fig. 2(d)) are specific for traffic surveillance applications as the resolution is not enough to identify and track smaller objects.

Many surveillance systems impose the constraint of having cameras calibrated in the same world reference system in order to have the correspondences between the observed objects in the image planes and their real positions in the scene. Anyway camera calibration is not always possible especially when large camera networks cover wide areas, therefore discovering the camera topology becomes a prerequisite task for any surveillance application.

An important point that surveillance systems have to consider and that is strictly related to the camera network topology is the ability to maintain as separated different objects moving close in the scene. This task depends on the dimension and the shape of the observed objects and their relative positions in the image planes. As a successive step, the surveillance systems have to recognize the same objects in different view, task that becomes more complex if the overlapping areas among different cameras are reduced in the camera network topology.

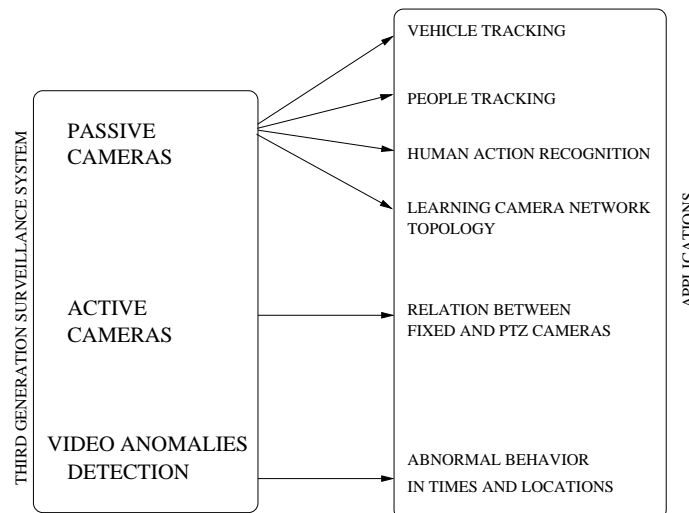


Fig. 1. Third generation surveillance systems

In the following subparagraphs we have done a distinction between systems that monitor rigid objects (vehicle tracking) and non rigid objects (people tracking), as the above problems can be faced in different ways. In traffic contexts, shape, motion and trajectory constraints are imposed to simplify the multi-view associations among vehicles, while in people surveillance applications, people shape is greatly variable, motion is neither predictable nor constrained in determined paths.

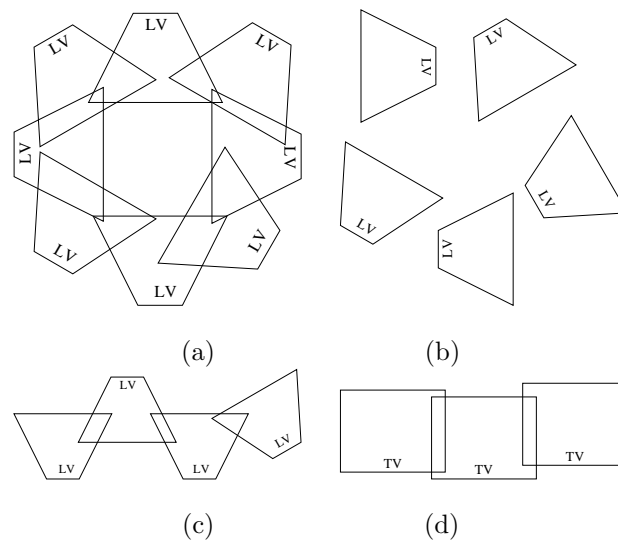


Fig. 2. Types of Camera Overlap: (a) Totally overlapping Lateral View (LV) Cameras; (b) Non Overlapping Lateral View (LV) Cameras; (c) Partially overlapping Lateral View (LV) Cameras; (d) Partially Overlapping Top View (TV) Cameras.

2.1. Vehicle Tracking

Cameras are becoming increasingly ubiquitous features of modern transportation infrastructures due to their low cost, ease of maintenance, and the wealth of information they provide about traffic conditions. Tracking is a central problem in many transportation applications as vehicle trajectories provide data enough for the estimation of many traffic parameters of interest. The complexity of the problem depends on the cameras positions and the relative fields of view. If vehicles occlude each other the detection and tracking problems are challenging since two or more close moving vehicles can be confused with trucks. The multi view approach can help to disambiguate these situations.

Overlapping and Calibrated Cameras: Many methods that use overlapping cameras and camera calibration parameters (such as the projection matrix of each camera) do not perform any matching between views since they use *correspondences between moving objects in overlapping areas* (Ref.^{16,17,18}). Among these,



Fig. 3. The track of the same object in two non overlapping views in Denman et al.¹⁶

some works try to detect moving objects combining the uncertainty in the location estimation of each view and providing a *probability distribution* of detected vehicles^{19,20}. This kind of camera configuration allows also the *3D reconstruction* of the observed scene and many methods propose the construction of 3D models that can be matched with a set of examples stored in a common data base(Ref.^{17,21}).

In Denman et al.¹⁶, a combination of both motion detection and optical flow modules is used to track and monitor vehicles in a short-term parking lane in real-time. Automated alerts of events such as parking time violations, breaching of restricted areas or improper directional flow of traffic can be generated and communicated to attending security personnel. An automated vehicle tracking method that supplies trajectory, orientation, and dimension data about identified vehicles in real-world units is presented in Atev et al.¹⁸. The tracking method is based on the measurement of some corners of the boxes that represent vehicles (typically 6 of the 8 corners are visible). Even during static and dynamic occlusions, at least a few of those corners remain visible and can be easily matched among successive images. Different views resolve the targets at different resolutions, and in turn location estimates on the plane have different variances. In Sankaranarayanan et al.¹⁹ the theory for modelling the relation between the camera-plane geometry to the variance of location estimates on the ground plane is presented. Probability fusion maps are used in Lamosa et al.²⁰ to detect vehicles in traffic scenes. The vehicle images from multiple cameras are inverse perspective-mapped and registered onto a common reference frame, combining the multiple camera information to reduce the impact of occlusions.

A *3D scene reconstruction* is proposed in Smolic et al.¹⁷ that uses a database to extract objects model on which the appropriate textures of the extracted moving objects are superimposed to provide the final 3D scene representation. When cameras are placed to obtain very different views of the plane, the targets are observed at different resolutions, producing location estimations on the plane with different variances. A system that allows dynamic 3-D scene reconstruction from a limited number of input cameras is presented in Muller et al.²¹. A priori knowledge about

the scene is exploited, such as plane background areas and ground plane constraint for foreground objects. Dynamic objects are processed by mapping all textures of an object onto a common synthetic 3-D model. The model is selected from a database by comparing its 2-D projections into the initial views with the original textures.

Overlapping and Non Calibrated Cameras: When cameras are overlapping but not calibrated the geometric correspondence between moving regions is not always possible. For this reason some methods use the *objects' appearance* to match the same physical vehicle in the images provided by different views^{22,23}. In Ferecatu et al.²², during a training phase a set of SIFT key points are extracted to train the canonical correlation analysis transformation, and track cars in a highway by using slightly overlapping top views and non calibrated cameras. When an object enters in one view by analyzing its key features the most probable among the candidates in the second view is evaluated and associated as best match. A graph-based approach is used in Shahri et al.²³ to match candidate objects in different views. The similarity of individual vehicle attributes (color, size and length) as well as the similarity of attributes of all their neighbors are evaluated. The traffic laws and conventions are used to constrain the definition of neighbor.

Partially Overlapping Top View Cameras: If top views are used, occlusions are less probable, but it becomes important to perform data association after mapping objects into the ground plane world coordinates. Tracking a car while it travels within a lane reduces to a simpler *1D tracking problem* when top view cameras are used. In Dixon et al.²⁴ a set of partial object tracks are generated in each cameras and then combined in a complete object track (see figure 4) by using the sequences of world-space position observations, an appearance descriptors denoting the mean color of all the pixels covered by each track, and its start/end frame on which a temporal model is built.

Non Overlapping Cameras: The main difficulty in tracking objects with non overlapping cameras arises from the fact that an object may disappear from one camera and reappear later with a different appearance. Common approaches are based on *trajectory analysis* and *trajectory grouping* together with *motion prediction* among distant cameras. In Kim et al.²⁵ the problem of tracking an unknown number of objects in a system of uncalibrated cameras sparsely distributed without overlapping fields of view is considered. The method exploits the statistics on overall traffic and the probabilistic dependence of a path in one camera view on the previous path in another camera view. The dependency and the frequency of allowable paths are represented in a graph model. An approach for activity analysis in multiple synchronized but uncalibrated static camera views is presented in Wang et al.²⁶. Under a probabilistic model, the approach jointly learns the distribution of an activity in the feature spaces of different camera views. Then, it accomplishes the following tasks: 1) grouping trajectories, which belong to the same activity but may be in different camera views, into one cluster; 2) modelling paths commonly taken by objects across multiple camera views; and 3) detecting abnormal activities.

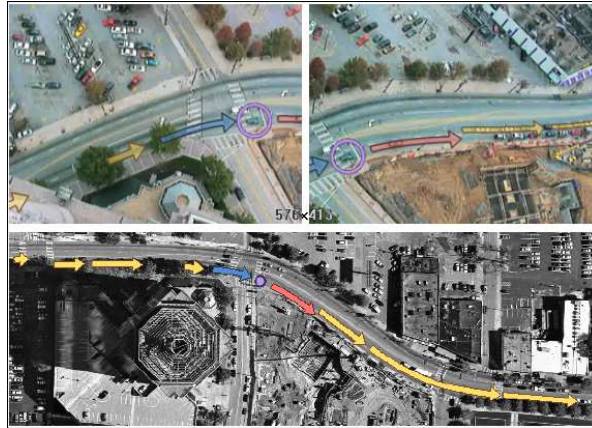


Fig. 4. The top images show the partial object tracks for one vehicle in two cameras; the bottom image shows the complete object track from all seven cameras that were generated for the same vehicle (Dixon et al.²⁴)

2.2. People Tracking

People tracking is certainly a more complex task than vehicle tracking since people are non rigid objects whose shapes greatly change also among different frames of the same camera. These variations increase when cameras with different fields of views observe people having different orientations.

Overlapping and Calibrated Cameras: Many methods that consider overlapping and calibrated cameras, use the *relative geometrical information* of the objects positions in different views to associate the object tracks among different cameras^{27,28,29} and attempt to solve occlusions in crowded scenes (see Ref. ^{30–47}).

Feature-based approaches use also *feature analysis* to discriminate among different couples of candidates in the association of object tracks (Ref.^{48,49,50}).

Geometrical approaches have the great advantage of being light to implement but are strictly dependent on the information that the camera is able to extract in each view. Occlusions in one view can be solved if objects appear separated in at least one of the other views³⁷. A method for multi-camera image tracking in the context of image surveillance with static overlapping and calibrated cameras is presented in Black et al⁰⁶.^{30,31}. Moving objects are detected by using background subtraction and viewpoint correspondence, between detected objects, is established by using the ground plane homography constraint. The Kalman Filter, using a constant velocity motion model, is then used to facilitate the 3D object tracking. According to the reported results, the system is robust in handling dynamic, static and partial occlusions, but it suffers when the constant velocity model is not valid (i.e. objects that accelerate or decelerate), or the trajectory changes during the period of occlusions. In Dai et al.³², the problem of people association and consistent labelling through exploring geometrical correspondences of objects is considered.

The cameras are geometrically related through joint combinations of multi-camera calibration, ground plane homography constraint, and field-of-view lines. In Fleuret et al.³³ the problem of keeping track of people who occlude each other using a set of calibrated and overlapping cameras is addressed. A mathematical framework combines the probabilities of occupancy of the ground plane at individual time steps with dynamic programming to track people over time. Basic color and motion models are used to estimate the optimal individual trajectories. The results show that the system can track up to six people in a room for several minutes by using only four cameras, in spite of severe occlusions and lighting variations. Target information from multiple views are fused in Hu et al.³⁴, and a co-training strategy is applied to generate a representative set of training bags from all views.

In Mishra et al.³⁵ a 3D surveillance system using multiple cameras surrounding the scene is presented. The cameras are fully calibrated and assumed to remain fixed in their positions. Object detection and interpretation are performed completely in 3D space. Using depth information obtained by a stereo approach, persons can be separated from the background and their postures identified by matching with 3D model templates. The problem of people occlusion or group of people is not considered.

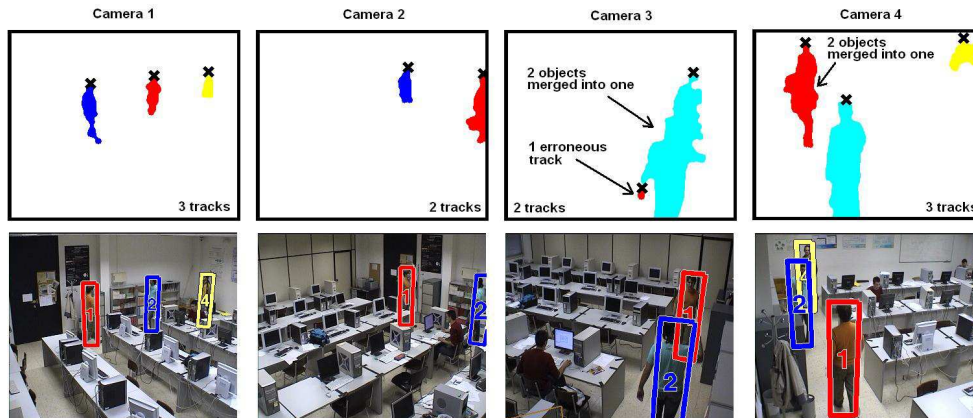


Fig. 5. First row: objects detected by each camera, with their 2D head position indicated with a cross. Second row: final estimation of 3D objects. Results of Mohedano et al.⁴⁹

Many methods use geometry constraints only to reduce the complexity of the matching process and use feature analysis to discriminate among different couples of candidates. In Mohedano et al.⁴⁹ 2D tracking systems running independently in each camera, are combined using Bayesian association of the monocular tracks. The correspondence between actual objects and 2D tracks is established according to the projection of the 3D object position onto the camera image plane and the color histogram coherence reasoning (see figure 5). Also in Mohammadi et al.⁴⁸, a

single-view tracking algorithm on each camera is performed initially, and then a consistent object labelling algorithm is applied on all views. Corresponding objects are extracted through a Homography transform from one view to the other and vice versa. For each region a set of descriptors, such as gravity mass center, histogram and texture, is extracted to find the best match between different views based on region descriptors similarity. The results demonstrate that the method is able to deal with multiple objects and occlusion, appearance and disappearance of objects are resolved using information from all views. The method proposed in Song et al.⁵¹ investigates distributed scene analysis using concepts of consensus from multi-agent systems. While each camera estimates certain parameters independently, these parameters are subsequently shared with neighboring cameras to arrive at a final estimate.

A complete approach to reconstruct 3D shapes in a dynamic event from silhouettes extracted from multiple videos recorded using a geometrically calibrated camera network is presented in Guan et al.⁵². It is based on a probabilistic volumetric framework for automatic 3D dynamic scene reconstruction and automatically learns the appearance of the dynamic objects, tracks the motions and detects surveillance events such as entering/leaving the scene.

Overlapping and Non Calibrated Cameras Methods that use overlapping and non calibrated cameras generally use the information coming from overlapping areas to learn *color similarity* and *color transformation* among different views⁵³. An objects detection algorithm for color dynamic images from overlapping cameras is proposed in Hatakeyama et al.⁵⁴ for a surveillance system under low illumination. It provides an automatic calculation of a fuzzy corresponding map and color similarity for low luminance conditions, detects little chromatic regions in CCD camera images under low illumination and presents regions with a possibility of occlusion situation. The method requires an initial preprocessing for the color similarity calculation among different views.

Also the person descriptor proposed in Quinn et al.⁵⁵ is based on color information but the tracked person is segmented into head, torso, leg and feet regions, and described by MPEG-7 Color Layout Descriptor and quantized histograms in the HSV color space. The multi view model is composed of 2D models of the person as viewed at different angles.

Non Overlapping Cameras When multiple non overlapping cameras are used, the surveillance task has to reconstruct the paths taken by all objects despite the fact that a moving object can be temporarily out of view of any cameras. People that move in large environments covered by non overlapping cameras can be difficult to observe since their paths can be unpredictable (they can stop, or change the direction in sudden ways) and close people can be detected by a unique blob, which makes difficult their characterization with color and shape appearance.

In the literature, different *re-identification methods* have been developed, some on them focusing on the matching between trajectories⁵⁶, others focusing solely on the appearance of the body. The latter are referred to as appearance-based

methods, and can be grouped in two sets. The first group is composed by the single-shot methods (Ref.^{57–67}), that model a person analyzing a single image. They are applied when tracking information is absent. The second group encloses the multiple-shot approaches; they employ multiple images of a person (usually obtained via tracking) to build a signature (Ref.^{68–71}). Some works try to learn the camera network topology in order to simplify the people association problem by predicting the relation between the exit location and time from one camera and the entry location and time into neighbor cameras^{72–77} (see figure 6).

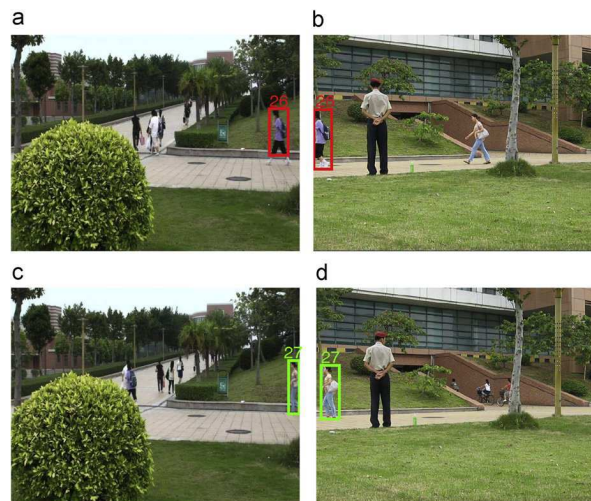


Fig. 6. Examples of correctly detected and matched objects in outdoor scene in Lian et al.⁷⁴

Color is the most commonly feature used in the appearance based approaches, sometimes encoded in the form of histograms or cumulative histograms. Different channels and their combination are used; RGB, HSV, YUV, YCbCr are the color space widely applied for the re-identification task⁵⁷. In order to represent patches of the silhouette other color descriptors are used, such as dominant color descriptors which compute the most recurrent color values, or stable color region descriptors which extract the homogeneous color by grouping neighboring color blobs⁶⁹. When cameras have a good resolution also texture information represent key features to characterize the signature. The most used textural features are Gabor filters, HAAR-like features, or DCT coefficients, as well as ratio of colors, ratio of oriented gradients and ratio of saliency maps describe texture variation between different patches⁶⁰. Interest points such as SIFT features or Hessian Affine invariant operators together with other shape information are matched to compare images from different cameras^{63,64}. The use of a cascade of grids of region descriptors has been demonstrated to outperform other single descriptor approaches⁶¹. The descriptors which embed information from multiple images per person, show that the presence

of several occurrences of an individual is very informative for re-identification⁶⁸. Major details on appearance based re-identification approaches can be found in recent reviews^{78,79}.

Alternative approaches to people tracking across multiple un-calibrated cameras use *gait analysis* as a new biometrics technique. In Ref.^{80,81,82} gait analysis is proposed as a solution for subjects identification across a network of cameras. The completely unobtrusiveness without any subject cooperation or contact for data acquisition make gait particularly attractive for identification purposes in camera handover. A 2D markerless view-independent gait analysis algorithm has been presented (see figure 7): the method does not need camera calibration or prior knowledge of subject pose. Since the choice of the cameras characteristics is a key-point for the development of a smart surveillance system, the performance of the proposed approach has been measured with respect to different video properties. Tests on synthetic and on real video sequences allowed performance evaluation of the proposed approach with respect to different spatial resolution, frame-rate, data compression and image quality. The obtained results show that gait analysis can be efficiently used for view-independent subjects identification with commercially available video cameras.

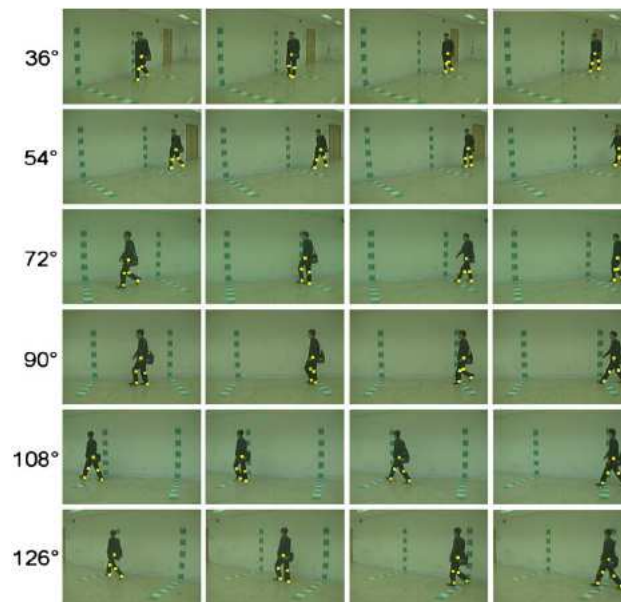


Fig. 7. Markerless joints extraction in different view points in Goffredo et al.⁸⁰

2.3. Human Action Recognition

Recognition of human action from multiple views is a popular task in the computer vision community since it has applications in video surveillance and monitoring, but also in other contexts such as human computer interactions or augmented reality. Many works have been published on this subject, and all of them suppose to have the segmented images of the human body from different cameras (generally overlapping cameras) and focus their attention on the *selection of robust features* for action recognition in critical situations and varying conditions (Ref.^{83–87}). The source of variability is related to several factors: the variation of speed, viewpoint, size and shape of the performer, but also the difficulty of interpret the beginning of the action, and finally the motion of human body that is not rigid in nature. A recent review on activity recognition has been published in Holte et al.⁸⁸, which presents also qualitative comparisons of a few promising approaches on publicly available datasets. We demand to this review for major details and references on this subject. In this section we consider just a few works that have applied different feature extraction techniques to give a general idea of the considered problems.

In Ahmad et al.⁸³ combined local-global (CLG) optical flow is used to extract motion flow feature and invariant moments with flow deviations are used to extract the global shape flow feature from the image sequences (see figure 8). Actions are modelled by using a set of multidimensional HMMs for multiple views using the combined features, which enforce robust view-invariant operation. Different human actions in daily life are successfully recognized in indoor and outdoor environments using a maximum likelihood estimation approach. Multiview approaches for auto-

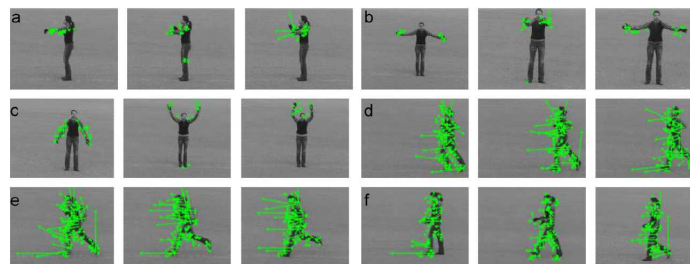


Fig. 8. CLG optical flows overlapping on the image of several actions in Ahmand et al.⁸³: (a) Boxing; (b) hand clapping; (c) hand waving; (d) jogging; (e) running; (f) walking.

matic detection of a falling person are presented in Ref.^{84,85}. In Thome et al.⁸⁴, the posture classification is performed by a fusion unit, merging the decision provided by the independently processing cameras in a fuzzy logic context. In each view, the fall detection is optimized in a given plane by performing a metric image rectification. A theoretical analysis of the chosen descriptor defines the optimal camera placement for detecting people falling in unspecified situations, and proves that

two cameras are sufficient in practice. Variation in the area and orientation of the human silhouettes is used in Mirmahboub et al.⁸⁵ to distinguish falls from other activities. In Gkalelis et al.⁸⁶, a view independent human movement representation and recognition method is presented. The binary masks of a multi-view posture image are first vectorized, concatenated and the view correspondence problem between train and test samples is solved using the circular shift invariance property of the discrete Fourier transform (DFT) magnitudes. The projective depths of selected points of the human skeleton are used in Ashraf et al.⁸⁹ to identify similar motion from varying viewpoints. The method requires the representation of human body as a set of points to decompose a body posture into a set of perspective depths. In Luo et al.⁸⁷, each camera processes its data locally by extracting sparse 3D features which characterize the observed motion and sends limited information to the base station. The information transmitted from local cameras are fused by a Naive Bayesian technique that integrates the prediction results from the SVM of each camera.

However, although the considered approaches show promising results, action recognition has some shortcomings. First of all, the quality of the segmentation and occlusions are crucial for the outcomes of feature-based and shape-based techniques. Second, the number of cameras will influence the level of details, and then the possible actions that can be recognized. The recent availability of low cost RGB-depth cameras offers new opportunities to explore 3D computer vision techniques in surveillance scenarios at limited costs. But these sensors are usually limited to a range of up to about 6-7 meters, and the estimated data can become distorted by scattered light from reflective surfaces.

2.4. Learning Camera Network Topology

When large camera network are distributed over wide areas, discovering the relationship between cameras becomes an important issue to develop intelligent surveillance system. *Tracking known moving objects* and establishing object correspondence across multiple cameras is proposed in Nam et al.⁹⁰ for representing the spatio-temporal topology of the camera network with overlapping and non-overlapping fields of view (FOVs). To track people successfully in multiple camera views, a Merge-Split (MS) approach for object occlusion in a single camera and a grid-based approach for extracting the accurate object feature are used. In addition, the appearance of people and the transition time between entry and exit zones is considered for tracking objects across blind regions of multiple cameras with non-overlapping FOVs. A technique for the registration of multiple surveillance cameras is presented in Konw et al.⁹¹, which tries to recover the relative pose of several stationary cameras that observe one or more objects in motion. In order to find correspondences between two images, trajectories are extracted after tracking of moving object from each cameras and then matched. When more than one object are present in the scene all the possible combinations between couples of

trajectories have to be evaluated as possible match. The method does not consider the people appearance for the matching process, but only the distances among all the possible trajectories are evaluated. In Kassebaum et al.⁹² a 3-D target-based localization solution for smart camera networks is presented. The method requires only small, pairwise view overlaps, making it more suitable for larger networks deployed for human or environmental monitoring. A 3-D target is used to provide all feature points needed for camera localization reducing the computation and communication costs and making the the algorithm suitable for battery-powered, processing-constrained smart camera platform.

3. Active Cameras

Sensing with a mix of static and mobile cameras has great advantages since the sensing resources may be allocated dynamically reducing the number of cameras required in order to monitor the same area. The advantage of using Pan Tilt and Zoom cameras is that it is possible to control the camera parameters to improve tracking and recognition performances. A common application is to have a fixed master camera, controlling a wide area, whose tracking results are used to guide the PTZ control of the mobile slave camera which zooms on the object of interest to obtain high resolution images. However, PTZ camera networks pose much more complex problems to solve than classical stationary camera networks. Assuming that all the cameras observe a planar scene, the image relationships between the camera image planes undergo a planar *time-variant homography*.

In Del Bimbo et al.⁹³ a framework exploiting a PTZ camera network to achieve high accuracy in the task of relating the feet position of a person in the image of the master camera, to his head position in the image of the slave camera is presented (see figure 9). The method exploits a prebuilt map of visual 2D landmarks of the wide area to support multi-view image matching. The landmarks are extracted from a finite number of images taken from a non calibrated PTZ camera, in order to cover the entire field of regard. Each image in the map also keeps the camera parameters at which the image has been taken. At run-time, features that are detected in the current PTZ camera view are matched to those of the base set in the map. The matches are used to localize the camera with respect to the scene and hence estimate the position of the target body parts.

An architecture for a multi-camera, multi-resolution surveillance system is described in Bellotto et al.⁹⁴. The aim is to support a set of distributed static and pan-tilt-zoom (PTZ) cameras and visual tracking algorithms, together with a central supervisor unit. Each camera (and possibly pan-tilt device) has a dedicated process and processor. Asynchronous interprocess communications and archiving of data are achieved via a central repository, implemented using an SQL database. The goal of the system, as regulated by a supervisor process, is to keep track of all targets in a scene using the overview camera, and to acquire high-resolution, stabilized images of the faces of all agents in the scene using zoom cameras in closed loop tracking

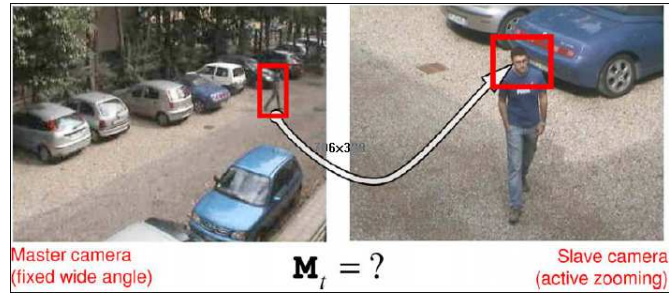


Fig. 9. Estimation of the time-variant transformation M_t that maps feet to head from the master to the slave view in Del Bimbo et al.⁹³. Left: Wide angle master camera view in which a person is detected. Right: Close up view from the slave camera.

mode. In Scotti et al.⁹⁵ an omnidirectional imaging device is used in conjunction with a pan tilt zoom (PTZ) camera leading to a sensor that is able to automatically track at a higher zoom level any moving object within the guarded area. A joint calibration strategy is necessary for the functioning of the coupled camera system.

4. Video Anomaly Detection

In this category, all those applications in which anomalies are detected without or before object detections and tracking, are considered. These paradigms are based on *motion detection and abnormal behavior recognition* that could take place both in unusual times and locations. Starting from these identified anomalies, region of interest can be extracted and object detection and tracking could take place in a successive phase.

In Saligrama et al.⁹⁶ a behavior subtraction approach has been presented that is analogous to background subtraction. However, unlike background subtraction, which operates on photometric quantities, behavior subtraction operates on dynamic features; photometric stationarity is replaced by dynamic stationarity thus treating regular motion as a background activity that needs to be ignored. To afford the main issue of how to ensure that activity seen at a specific location(s) in one camera is the same activity seen in other cameras, the authors consider the property of geometric invariance. Even if the geometric properties change based on orientation and zoom levels, the motion labels are busy for the same duration and at the same physical times. In Ermis et al.⁹⁷ an algorithm for matching camera regions in a heterogeneous camera network with overlapping fields of view is presented. It does not require any knowledge of any camera parameters such as location, orientation, epipolar geometry, etc. Anomalies can be detected by learning a nominal distribution from training data and declaring as anomalies those test points which are least likely under the nominal distribution. The method works directly on the models of corresponding pixels that observe the same location and it can be used also when cameras observe the scene from significantly different ori-

entations with different zoom levels. Another approach that avoids object tracking under challenging surveillance conditions is presented in Loy et al.⁹⁸. Specifically, since a complex scene naturally consists of multiple local scene regions that encompass distinctive activities, each camera view is first decomposed automatically into regions, across which different spatio-temporal activity patterns are observed. Then, a Cross Canonical Correlation Analysis that learns activity correlations by exploiting the underlying spatial and temporal correlation of regional activities in a holistic manner is applied. An approach to automated surveillance site activity segmentation that does not require any object based detection and tracking tasks is presented in Xu et al.⁹⁹. The output of a synchronized camera network, is subject to a specialized feature extraction procedure, which is conducted in an appearance-based image subspace. A data clustering method is used to generate video clusters that represents a fine summarization of the dynamic scene contents. The temporal distribution of these videos reflects the underlying dynamics or the peace of the scene, ie. the status and manner how crowds move over the time in the environment. In Loy et al.¹⁰⁰ anomalies are associated with deviations in the expected temporal dynamics embedded in complex behaviors (e.g. atypical duration and irregular temporal order). The same framework has been used to detect abnormal correlations among objects which could span across large spatial and temporal visual context. A similar approach is used in de Leo et al.¹⁰¹, where an action that occurs only once is considered an anomaly with respect to those classified as normal after a grouping phase that finds common patterns.

5. Data sets

Many researchers and organizations have done lots of work on the issues described before. However it is very difficult to make comparisons among different approaches for several reasons: the application contexts, the main objectives, the typologies of events, the experimental setups are in many case far from each other, then quantitative evaluation of the performances and in-depth comparisons of the relative merits are quite difficult. In the last decade, the scientific community made huge efforts to define public benchmark data sets to test and compare performances among different approaches. Some of these data sets are more realistic than others and provide multi camera real scenarios of both indoor and outdoor contexts. Among the others we cite: PETS datasets¹⁰³ (from 2000 onwards) provide multi-camera indoors and outdoor scenarios for multiple objects tracking; i-Lids¹⁰⁴ provides, among the others, videos with abandoned baggages and parked vehicles; CAVIAR¹⁰⁵, VIPER¹⁰⁶, ViSOR¹⁰² provide data sets for people re-identification, MuHAVi¹⁰⁷ provides a multi-camera human action video data set, ISSIA soccer data set¹⁰⁸ for synchronized multi-camera players and ball tracking, and so on. It is noteworthy to mention also the recent increase of public challenges in many important international conferences such as ICPR, CVPR, which allow precise quantitative comparisons and ranking of various algorithms, with respect to accurate ground-truths available for

all the frames of all the videos. Anyway, although these attempts put in evidence that some methodologies perform better than others in those specific contexts, it remains difficult to assess the real level of performances of the published algorithms in different scenarios. Many approaches use ad-hoc methodologies to afford the problems deriving by the specific camera resolutions, scenarios, lighting conditions that public data sets impose. Pre-processing techniques, properly devised for these contexts, hardly maintain the same performances when applied in real life scenarios with many variabilities both in the kind of events and in the environmental situations. This consideration is confirmed by the fact that even though the development of smart surveillance systems is being discussed by the scientific community since the beginning of 2000, commercial surveillance systems are not available yet, unless for simple functionalities such as motion detection in predefined areas, cross-line detection with top view cameras, cars parked in emergency lane, and so on.

6. Discussion

Surveillance system may approach the problem in two different ways: by detecting, tracking and recognizing moving objects in order to understand their behaviors, or detecting anomalies without any object detection process. The application of these two alternative approaches depends on the surveillance context. A few applications on traffic contexts have been found in the last year for the second class of approaches demonstrating a new and increasing interest on this methodology, while the remaining large research activities use strategies based on object detection and tracking. The main problem of this kind of approaches is the integration of information coming from different cameras, and the consistent labelling of the same objects when observed from different points of view. In this paper we distinguish between rigid objects, such as vehicles, and non rigid objects such as people.

In figure 10 we summarize the image processing methodologies that have been applied in different surveillance contexts. Their applications in multi camera networks are strictly dependent on the camera network topologies. When cameras are overlapping and calibrated, stereo approaches are largely preferred³⁵. Some works use depth information for 3D reconstruction of moving vehicles and matching of extracted objects with models in a database. In a similar way, using the depth information obtained from stereo approaches, person postures can be identified by matching with 3D model templates³³. Other approaches use the depth information only to solve object occlusions when close vehicles or group of people appear as a unique blob in the image.

If image plane transforms are known, homographic projections allow the detection of moving object positions in a 2D world plane⁴⁸. Many mathematical frameworks propose to estimate probabilities of occupancy of the ground plane to track people over time. Geometry constraints are used to reduce the complexity of the matching process demanding to other feature analysis the discrimination among different couples of candidates.

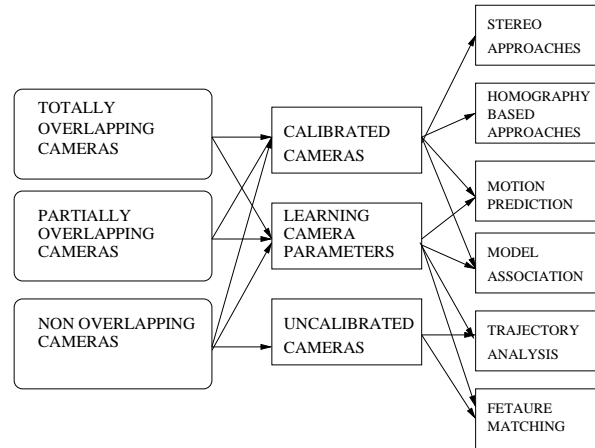


Fig. 10. Overview of image processing methodologies used with different camera set up in multi-camera intelligent surveillance systems

When cameras are not calibrated it is still possible to apply a learning phase during which some parameters on cameras can be extracted. A new trend of research use the knowledge of people walking along known trajectory to extract both the camera network topology and color brightness transformation among neighbor cameras^{72,74}. In these cases the camera calibration task is replaced by a training process that requires the knowledge of kinematic parameters of moving objects.

A much more difficult task results when moving object have to be tracked among views coming from not overlapping cameras. Vehicles that move along streets have the great advantage of having a predictable trajectory and a shape that cannot considerably change when observed from different views. For this reason, many surveillance systems for traffic applications are based on geometrical constraints, and kinematic evaluations to predict trajectories, to evaluate the frequency of allowable path, to estimate the arrival and departure locations²⁴. A few of them use also color and shape information to match different tracks among cameras. Also the fixed vehicle dimensions allow a prediction of the occupancy of each track among different views.

People tracking is more complex with non overlapping cameras. People are not rigid objects whose shape can greatly change also inside the same view. The appearance can vary according to the parts of the dresses that are visible in each view. The trajectory cannot be predictable since people can suddenly change their motion directions. Last but not least point is the problem of people aggregation that causes multiple blobs containing more persons. Here, maintaining the same identity of individuals along time becomes a problem of person re-identification that is typically addressed with similarity search as in multimedia retrieval. If the camera fields of view are not so far and the lighting conditions are not so different,

it is plausible that a person exiting a camera field of view will soon appear in the field of view of a neighbor camera and will generate similar color distribution. But when cameras are distant the approaches have to search the best association based on similarity among robust features that can be extracted in each view on single tracks. Some works use the appearance of objects, others match trajectories, other use gait recognition approaches. Among the most used features to characterize the object appearance there are color features that are relatively robust towards the size and orientation changes but suffer to compensate inter-camera distortions as well as illumination changes. Among the most used color features there are color histograms in different color spaces (RGB, HSV, and so on), color templates (in images with the same resolution), color layout features representing the spatial distribution of colors in an image, and color signature^{57,69}. In order to make these features more robust, some methods apply the feature extraction process to the entire object shape, others consider three main body regions (head, torso and legs) and in the same way, some of them apply the extraction process to a single image of the candidate others to multiple images for each individual. Alternative approaches use gait analysis as a new biometric techniques^{80,7,82}. The analysis and grouping of trajectories belonging to the same activity has been used as well to cluster and model paths of objects across different cameras and associate tracks without any camera calibration.

When people are well segmented and separated in a multi-view camera system (generally overlapping), many works focus their attention on the selection of robust features for action recognition, optical flow analysis, posture classification and so on^{83,87}.

Considering the analysis of the related literature, lets go back to the point on which we started our discussion: what are the interesting events that, at the moment, an intelligent multi-camera surveillance systems can detect? In the context of vehicular traffic monitoring, the involved image processing methodologies are mature for being implemented in real-world settings and the constraints on the domain can greatly simplify the detection and tracking tasks. Commercial systems that provide license plate recognition, detection of cars stopped in emergency lane, automatic traffic analysis, and so on, are already available on the market. On the other hand, much remains to be done in the area of people behavior analysis and modelling, especially in large public environments. For some specific events such as motion detection in predefined areas, cross line detection by top-view cameras, overcrowding, abandoned packages detection, possible applications are feasible in well known contexts and using proper camera setups (fixed and calibrated cameras).

When interesting events require a deep understanding of people behavior, further progresses in the image processing methodologies are necessary. Intelligent surveillance systems have the necessity to emulate the human ability to detect and recognize people when they are observed from distant point of views, to recognize anomalies that are the result of a complex dynamic interactions, and so on. These tasks could be carried out if robust methodologies for people tracking in crowded

scenes, people re-identification among non overlapping cameras, gesture analysis in low resolution images were available. At the moment, the success of many of these methodologies is still dependent on the quality of preprocessing steps that are far from being completely settled. As a first step, a good segmentation is the fundamental prerequisite for any further analysis. Whatever the context, it is necessary to have moving objects well separated from the remaining background. The presence of ghosts and shadows in the segmentation can greatly modify the object shapes. Just to highlight one of the actual open problem we report in figure 11 some sample images from CAVIAR¹⁰⁵ and TRECVID¹⁰⁹ data sets. Top and bottom images are taken from two different cameras and demonstrate that people appearance could be very different. Humans are able to recognize persons also in these hard conditions with different point of views and low resolution images. In order to emulate this capability robust image processing methodologies are necessary.

In recent years the use of public data sets has promoted comparisons of different methodologies on video sequences providing a number of simple events in quite realistic scenarios. The importance of the availability of datasets shared by the scientific community is quite evident. Algorithms and proposed techniques, able to deal with the several problems rising up in the wide area of video surveillance, such as analysis, tracking and understanding in general, have found a common background for comparison of performances on the same data. More precise and fine tuning of the proposed procedures has come out as the result of this set up. Unfortunately, even the best performing algorithms, working in a satisfactory way on a given dataset, show some difficulties when applied in other datasets or in a real scenario, suggesting the idea that the use of datasets to assess the actual operativeness of the proposed methodologies in real situations of smart surveillance systems is still far from being achievable. A possible way to overcome these limitations could be share with the scientific community all the developed source codes and information to properly operate and reproduce the published results (i.e. giving all the required information about the procedures and the parameter settings to test freely the methodologies in different scenarios). Another possible solution could be set up a web access framework that could be maintained by interested institutions, Universities or research centers, to allow the real time acquisition of videos in real scenarios, (allowing grid computing and cloud storage resources and facilities). In this way, research teams might test their own techniques in the same contexts, such as train stations, airports and so on. Facilitating these realistic comparisons would effectively give a great impulse to the comprehension of the weaknesses and strengths of each methodology and allow researchers to concentrate their attention on hot issues that need to be addressed.

7. Conclusions

In this paper a review on automatic event detection functionalities in third generation surveillance systems has been presented. We have considered only those



Fig. 11. Some sample images from CAVIAR data set (a) and TRECVID data set (b). Top and bottom lines correspond to different cameras.

papers that focus on the application of image processing methodologies to address the multi-view problem finalized to integrate data coming from several cameras in a single processing framework for information exploitation.

The state of art is mature for the case of rigid objects, such as traffic monitoring systems. In this context, well assessed methodologies have been developed and can be applied successfully. On the other side, methodologies for people surveillance are still far from solving the problem, due to the deformable nature of the object in the scene and the large variations of interesting situations that should be detected. From our perspective, in future research, further efforts are needed on the development of robust methodologies for feature matching and object detection among different cameras that could be applied without strict constraints on accurate moving object segmentations or the knowledge of inter/intra-cameras calibration parameters.

8. Acknowledgements

The authors thank the anonymous referees whose valuable suggestions helped to improve the quality of this review. This research was partially funded by PON 01-00980 BAITAH.

References

1. C. Regazzoni, V. Ramesh, G.L. Foresti, *Special Issue on Video Communications, Processing, and Understanding for Third Generation Surveillance Systems*, Proc. IEEE, 2001, Vol. 89, No. 10
2. C. H. M. Donold, *Assessing the human vigilance capacity of control room operators*, Proc. Int. Conf. Humans Interfaces in Control Rooms, Cockpits and Command Centres, 1999, pp. 711.
3. R. Patrick, N. Bourbakis, *Surveillance Systems for Smart Homes: A Comparative Survey*, 21st International Conference on Tools with Artificial Intelligence (ICTAI), 2009, pp. 248-252
4. C.Y. Cho, W.H. Tung, J.S. Wang, *A crowd-filter for detection of abandoned objects*

- in crowded area* 3rd International Conference on Sensing Technology (ICST), 2008, pp. 581-584
5. H. Kong, J.Y. Audiber, J. Ponce, *Detecting Abandoned Objects With a Moving Camera* IEEE Transactions on Image Processing, 2010, Vol. 19 , No. 8, pp. 2201-2210
 6. M. Leo, P. Spagnolo, T. Martiriggiano, A. Caroppo, T. D'Orazio, *A system to automatically monitor forbidden areas* IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS), 2005, pp. 570-575
 7. X. Pan, Y. Guo, A. Men, *Traffic Surveillance System for Vehicle Flow Detection* Second International Conference on Computer Modeling and Simulation (ICCMS) 2010, Vol.1, pp. 314-318
 8. A. Bevilacqua, S. Vaccari, *Real time detection of stopped vehicles in traffic scenes* IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS), 2007, pp. 266-270
 9. M. Valera, S.A. Velastin, *Intelligent distributed surveillance systems: a review* IEE Proceedings - Vision, Image and Signal Processing, 2005, Vol. 152 , No. 2, pp. 192-204
 10. H.M. Dee, S.A. Velastin, *How close are we to solving the problem of automated visual surveillance?* Machine Vision and Applications, 2008, Vol. 19, pp.329-343
 11. N. Haering, P.L. Venetianer, A. Lipton, *The evolution of video surveillance: an overview* Machine Vision and Applications, 2008, Vol.19 pp.279-290
 12. T. D. Raty *Survey on Contemporary Remote Surveillance Systems for Public Safety* IEEE Transactions on Systems, Man, and Cybernetics- Part C: Applications and Reviews, 2010, Vol. 40, No. 5, pp.493-515
 13. D. Weinland, R. Ronfard, E. Boyer *A survey of vision-based methods for action representation, segmentation and recognition* Computer Vision and Image Understanding, 2011, Vol. 115, pp.224-241
 14. J.K. Aggarwal, Lu Xia *Human Activity Recognition From 3D Data: A Review* Pattern Recognition Letters 2014, in press
 15. Y. Ye, S. Ci, A.K: Katsaggelos, Y.W. Liu, Y. Qian, *Wireless Video Surveillance: A Survey*, IEEE Access, 2013, Vol.1, pp. 646-660
 16. S. Denman, C. Fookes, J. Cook, C. Davoren, A. Mamic, G. Farquharson, D. Chen, B. Chen and S. Sridharan *Multi-view Intelligent Vehicle Surveillance System* Proceedings of the IEEE International Conference on Video and Signal Based Surveillance (AVSS), 2006
 17. A. Smolic, K. Mueller, E. Droese, P. Voidt T. Wiegand, *Multiple view video streaming and 3D scene reconstruction for traffic surveillance* Digital Media Processing for Multimedia Interactive Serices, 2003, pp. 427-432
 18. S. Atev and N. Papanikolopoulos, *Multi-View 3D Vehicle Tracking with a Constrained Filter* IEEE International Conference on Robotics and Automation, 2008, pp. 2277-2282
 19. A. C. Sankaranarayanan and R. Chellappa, *Optimal multi-view fusion of object locations*, in Proc. IEEE Workshop Motion Video Comput. (WMVC), 2008.
 20. F. Lamosa, Z.Hu, K. Uchimura *Vehicle detection using probability Fusion Maps generated by multicamera systems* Journal of Information Processing, 2009, Vol. 17, pp. 1-13
 21. K. Muller, A. Smolic, M. Drose, P. Voigt, T. Wiegand, *3-D Reconstruction of a Dynamic Environment With a Fully Calibrated Background for Traffic Scenes* IEEE Trans. On Circuits And Systems for Video Technology, 2005, Vol. 15, No. 4, pp, 538-549
 22. M. Ferecatu, H. Sahbi *Multi-View Object Matching and Tracking Using Canonical Correlation Analysis* IEEE International Conference on Image Processing (ICIP) 2010,

- pp. 2109-2112
23. H.H. Shahri, G. Namata, S. Navlakha, A. Deshpande, N. Roussopoulos *A Graph-based Approach to Vehicle Tracking in Traffic Camera Video Streams* ACM International Conference Proceeding Series, 2007, Vol. 273, pp. 19-24
 24. M. Dixon, N. Jacobs, R. Pless *An Efficient System for Vehicle Tracking in Multi-Camera Networks* Third ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC), 2009, pp. 1-8
 25. H. Kim, J. Romberg, W. Wolf *Multi-Camera Tracking on a Graph Using Markov Chain Monte Carlo* Third ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC), 2009, pp.1-8
 26. X. Wang, K. Tieu, W. E. L. Grimson, *Correspondence-Free Activity Analysis and Scene Modeling in Multiple Camera Views* IEEE Transaction On Pattern Analysis and Machine Intelligence, 2010, Vol. 32, No. 1, pp. 56-71
 27. J. Yao, J.M Odobez, *Multi-camera multi-person 3d space tracking with mcmc in surveillance scenarios* European Conference on Computer Vision, workshop on Multi Camera and Multi-modal Sensor Fusion Algorithms and Applications (ECCV-M2SFA2), 2008,
 28. E.G. Rieffel, A. Girgensohn, D. Kimber, T. Chen, Q. Liu; *Geometric Tools for Multicamera Surveillance Systems* First ACM/IEEE International Conference on Distributed Smart Cameras, ICDSC,2007, pp. 132-139
 29. A. Kiss, T. Sziranyi, *Localizing people in multi-view environment using height map reconstruction in real-time* Pattern Recognition Letters, 2013, Vol. 34, pp. 2135-2143
 30. J.Black, T. Ellis *Multi camera image tracking* Image and Vision Computing, 2006, Vol. 24, pp. 1256-1267
 31. T. Ellis, J. Black *A Multi-view Surveillance System* IEE Symposium on Intelligence Distributed Surveillance Systems, 2003 , pp. 11/1 - 11/5
 32. X.Dai, S. Payandeh, *Geometry-Based Object Association and Consistent Labeling in Multi-Camera Surveillance*, IEEE Journal on Emerging and Selected Topics in Circuits and Systems, 2013, Vol. 3, No. 2, pp. 175-184
 33. F. Fleuret, J. Berclaz, R. Lengagne, P. Fua, *Multicamera People Tracking with a Probabilistic Occupancy Map* IEEE Trans. on Pattern Analysis AND Machine Intelligence, 20078, V. 30, No. 2, pp. 267-282
 34. K. Hu, X. Zhang, Y. Gu, Y. Wang, *Fusing target information from multiple views for robust visual tracking* , IET Computer Vision, 2014, Vol. 8, No. 2, pp. 86-97
 35. A. K. Mishra, B. Ni, S. Winkler, A. Kassim *3D Surveillance System Using Multiple Cameras* Proceedings of the SPIE, 2007, Vol. 6491,
 36. X.Wang, S. Wang, D. Bi, *Distributed Visual-Target-Surveillance System in Wireless Sensor Networks*, IEEE Transactions on Systems, Man, and CyberneticsPart B: Cybernetics, 2009, Vol. 39, No. 5, pp. 1134-1146
 37. M. Mozerov, A. Amato, X. Roca, J. Gonzalez, *Solving the Multi Object Occlusion Problem in a Multiple Camera Tracking System*, Pattern Recognition and Image Analysis, 2009, Vol. 19, No. 1, pp. 165-171.
 38. J. Shen, W. Yan, P. Miller, H. Zhou *Human Localization in a Cluttered Space Using Multiple Cameras* Seventh IEEE International Conference on Advanced Video and Signal Based Surveillance, 2010, pp. 85-90
 39. S.M. Khan, M. Shah, *A Multiview Approach to Tracking People in Crowded Scenes using a Planar Homography Constraint*, Computer Vision (ECCV), 2006, pp. 133-146.
 40. S.M. Khan, M. Shah, *Tracking multiple Occluding People by Localizing on Multiple Scene Planes*, IEEE Transaction on Pattern Analysis and Machine Intelligence, 2009,

- Vol. 31, No. 3, pp. 505-519
41. B. Kwolek, *Multi camera-based person tracking using region covariance and homography constraint*, 7th IEEE International Conference on Advanced Video and Signal Based (AVSS), 2010, pp. 294-299
 42. D. Arsic, A. Lyutskanov, G. Rigoll, B. Kwolek, *Multi camera person tracking applying a graph-cuts based foreground segmentation in a homography framework* 12th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance, (PETS), 2009, pp. 1-8
 43. J. Orwell, S. Massey, P. Remagnino, D.Greenhill, G.A. Jones, *A Multi- agent framework for visual surveillance*, Proceedings. International Conference on Image Analysis and Processing (ICIP), 1999, pp. 1104-1107
 44. T. D'Orazio, M. Leo, P. Spagnolo, P. L. Mazzeo, N. Mosca, M. Nitti, A. Distanto, *An Investigation into the Feasibility of Real-time Soccer Offside Detection from a Multiple Camera System* IEEE Transactions on Circuit and System for Video Technology, 2009, Vol. 19, n. 12, pp. 1804-1818
 45. T. Zhao, M. Aggarwal, R. Kumar, H. Sawhney *Real-Time Wide Area Multi-Camera Stereo Tracking* IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 2005, Vol. 1, pp. 976-983
 46. T.T. Santos, C. H. Morimoto *Multiple camera people detection and tracking using support integration* Pattern Recognition Letters, 2011, Vol. 32, pp. 47-55
 47. R.Eshel, Y. Moses *Tracking in a Dense Crowd Using Multiple Cameras* International Journal of Computer Vision, 2010, Vol. 88, pp. 129-143
 48. G. Mohammadi, F. Dufaux, M.T. Ha Minh, T. Ebrahimi *Multi-view video segmentation and tracking for video surveillance* SPIE Proc. Mobile Multimedia/Image Processing, Security and Applications, 2009
 49. R. Mohedano, N. Garca *Robust Multi-Camera 3D Tracking from Mono-Camera 2D Tracking using Bayesian Association* IEEE Transactions on Consumer Electronics, 2010, Vol. 56, No. 1
 50. D.T. Lin, K.Y. Huang *Collaborative Pedestrian Tracking with Multiple Cameras: Data Fusion and Visualization* The 2010 International Joint Conference on Neural Networks (IJCNN),pp. 1-8
 51. B. Song, A. Kamal, C. Soto, C. Ding , J.A. Farrell, A.K. Roy-Chowdhury, *Tracking and Activity Recognition Through Consensus in Distributed Camera Networks* IEEE Transaction on Image Processing, 2010, Vol. 19, No. 10, pp. 2564 - 2579
 52. L. Guan, J.S. Franco, M. Pollefeys, *Multi-view Occlusion Reasoning for Probabilistic Silhouette-Based Dynamic Scene Reconstruction*, Int. Journal Computer Vision, 2010, Vol. 90, pp. 283-303
 53. Y. Yun, I. Gu, H. Aghajan, *Multi-View ML Object Tracking With Online Learning on Riemannian Manifolds by Combining Geometric Constraints*, IEEE Journal on Emerging and Selected Topics in Circuits and Systems, 2013 Vol. 3, No. 2, pp. 185-197
 54. Y. Hatakeyama, A. Mitsuta, K. Hirota *Detection algorithm for color dynamic images by multiple surveillance cameras under low luminance conditions based on fuzzy corresponding map* Applied Soft Computing, 2008, Vol. 8, pp. 1344-1353
 55. M.J. Quinn, T. Kuo, B.S. Manjunath *A lightweight multiview tracked person descriptor for camera sensor networks* 15th IEEE International Conference on Image Processing (ICIP), 2008, pp. 1976-1979
 56. V. Kettner, R. Zabih *Bayesian Multi-camera Surveillance* IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1999, pp.253- 259
 57. S. Pedagadi, J. Orwell, S. Velastin, B. Boghossian *Local Fisher Discriminant Anall-*

- ysis for Pedestrian Re-identification* IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 3316-3325
58. M. O. Mehmood, A. Khawaja, *Multi-camera based Human Tracking with Non-Overlapping Fields of View*, International Conference on Image and Graphics (ICIG), 2009, pp. 313-318
 59. Y. Chai, V. Takala, M. Pietikainen, *Matching Groups of People By Covariance Descriptor*, 20th International Conference on Pattern Recognition, 2010, pp. 2744-2747
 60. S. Bak, E. Corvee, F. Bremond, M. Thonnat *Person Re-identification Using Haar-based and DCD-based Signature* Seventh IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 2010
 61. A. Alahi, P. Vanderghenst, M. Bierlaire, M. Kunt, *Cascade of descriptors to detect and track objects across any network of cameras* Computer Vision and Image Understanding, 2010, Vol. 114, pp. 624-640
 62. I.O. de Oliveira, J.L. de Souza Pio, *People Reidentification in a camera network* Eighth IEEE International Conference on Dependable, Autonomic and Secure Computing, 2009, pp. 461-466
 63. K. Jungling, C. Bodensyeiner, M. Arens, *Person re-identification in multi-camera networks*, IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2011, pp. 55-61
 64. N. Gheissari, T. B. Sebastian, P. H. Tu, J. Rittscher, R. Hartley, *Person Reidentification Using Spatiotemporal Appearance* Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 2006, pp. 1528-1535
 65. P.L. Mazzeo, P. Spagnolo and T. D'Orazio *Object Tracking by Non-Overlapping Distributed Camera Networks* Advanced Concepts for Intelligent Vision Systems September (ACIVS), 2009, pp. 516-527
 66. T. D'Orazio, P.L. Mazzeo, P. Spagnolo, *Color Brightness Transfer Function evaluation for non overlapping multi camera tracking* Third ACM/IEEE International Conference on Distributed Smart Cameras, 2009
 67. F. Porikli, A. Divakaran, *Multi-camera calibration, object tracking and query generation*, IEEE International Conference on Multimedia and Expo, 2003, Vol. 1, pp. 653-656
 68. L. Bazzani, M. Cristani, A. Perina, M. Farenzena, V. Murino *Multiple-Shot Person Re-identification by HPE Signature* 20th International Conference on Pattern Recognition (ICPR), 2010, pp. 1413-1416
 69. M. Farenzena, L. Bazzani, A. Perina, V. Murino, M. Cristani *Person Re-Identification by Symmetry-Driven Accumulation of Local Features* IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010, pp. 2360-2367
 70. D. N. T. Cong, L. Khoudour, C. Achard, C. Meurie, O. Lezoray *People re-identification by spectral classification of silhouettes* Signal Processing, 2010, Vol. 90, pp. 2362-2374
 71. O. Hamdoun, F. Moutarde, B. Stanciulescu, B. Steux *Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences* Proceedings of the IEEE Conference on Distributed Smart Cameras, 2008, pp. 1-6
 72. K.W. Chen, C.C. Lai, Y.P. Hung, C.S. Chen, *An adaptive learning method for target tracking across multiple cameras* IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2008, pp. 1-8
 73. B. Song, A.K. Roy-Chowdhury, *Robust Tracking in A Camera Network: A Multi-Objective Optimization Framework* IEEE Journal of Selected Topics in Signal Pro-

- cessing, 2008, Vol. 2, No. 4
74. G. Lian, J. Lai, W.S. Zheng *Spatial-temporal consistent labelling of tracked pedestrians across non-overlapping camera views* Pattern Recognition, 2011, Vol. 44, pp. 1121-1136
 75. A. Gilbert, R. Bowden, *Tracking objects across cameras by incrementally learning inter-camera colour calibration and pattern of activity* European Conference on Computer Vision, 2006, pp. 125-136
 76. O. Javed, K. Shafique, Z. Rasheed, M. Shah, *Modelling inter-camera space time and appearance relationships for tracking across non-overlapping views* Computer Vision and Image Understanding, 2008, Vol. 109, pp. 146-162
 77. R. Verrazzi, R. Cucchiara *Event Driven Software Architecture for Multi-camera and Distributed Surveillance Research Systems* IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2010, pp. 1-8
 78. T. D'Orazio, G. Cicirelli, *People re-identification and tracking from multiple cameras: A review* 19th IEEE International Conference on Image Processing (ICIP), 2012, pp.1601-1604
 79. R. Mazzon, S.F. Tahir, A. Cavallaro, *Person re-identification in crowd* Pattern Recognition Letters, 2012, Vol. 14, pp. 1828-1837
 80. M. Goffredo, I. Bouchrika, J.N. Carter, M. S. Nixon *Performance analysis for automated gait extraction and recognition in multi-camera surveillance* Multimedia Tools Application, 2010, Vol. 50, pp. 7594
 81. M. Goffredo, I. Bouchrika, J.N. Carter, M. S. Nixon *Self-Calibrating View-Invariant Gait Biometrics* IEEE Transactions On System, Man, And Cybernetics Part B: Cybernetics, 2010, Vol. 40, No. 4, pp. 997-1008
 82. M. Hu, Y. Wang, Z. Zhang, J.J. Little, D. Huang, *View-Invariant Discriminative Projection for Multi-View Gait-Based Human Identification* IEEE Transaction on Information Forensic and Security, 2013, Vol. 8, No. 12, pp. 2034-2045
 83. M. Ahmad, S.W. Lee *Human action recognition using shape and CLG-motion flow from multi-view image sequences* Pattern Recognition, 2008, Vol. 41, pp. 2237-2252
 84. N. Thome, S. Miguet, S. Ambellouis *A Real-Time, Multiview Fall Detection System: A LHMM-Based Approach* IEEE Trans. On Circuits and Systems for Video Technology, 2008, Vol. 18, No. 11,
 85. B. Mirmahboub, S. Samavi, N. Karimi, S. Shirani *View-Invariant Fall Detection System Based on Silhouette Area and Orientation* IEEE International Conference on Multimedia and Expo, 2012, pp.176-181
 86. N. Gkalelis, N. Nikolaidis, I. Pitas *View independent human movement recognition from multi-view video exploiting a circular invariant posture representation* IEEE International Conference on Multimedia and Expo (ICME) 2009, pp. 394-397
 87. J. Luo, W. Wang, H. Qi, *Feature Extraction and Representation for Distributed Multi-view Human Action Recognition* IEEE Journal on Emerging and Selected Topics in Circuits and Systems, 2013, Vol. 3, No. 2, pp.145-154
 88. M.B. Holte, C. Tran, M. Trivedi, T.B. Moeslund, *Human Pose Estimation and Activity Recognition From Multi-View Videos: Comparative Explorations of Recent Developments*, IEEE journal of selected Topics in Signal Processing, 2012, Vol. 6, No. 5, pp. 538-552
 89. N. Ashraf, C. Sun, H. Foroosh *View invariant action recognition using projective depth* Computer Vision and Image Understanding, 2014, Vol. 123, pp. 41-52
 90. Y. Nam, J. Ryu, Y.J. Choi, and W.D. Cho *Learning Spatio-Temporal Topology of a Multi-Camera Network by Tracking Multiple People* Proceeding of World Academy

- of Science, Engineering and Technology, 2007, Vol. 30, pp. 175-180
91. O. Know, J. Jung, J. Shin, S. Lee, J. Paik *Auto Calibration for Multiple Camera-Based Surveillance System* in Proc. of SPIE, 11th Japan-Koera Joint Workshop on Frontiers of Computer Vision, 2005, Vol. 5960
 92. J. Kassebaum, N. Bulusu, and W.C. Feng, *3-D Target-Based Distributed Smart Camera Network Localization* IEEE Transaction on Image Processing, 2010, Vol. 19, No. 10, pp.2530-2539
 93. A. Del Bimbo, F. Dini, G. Lisanti, F. Pernici *Exploiting distinctive visual landmark maps in pantiltzoom camera networks* Computer Vision and Image Understanding, 2010, Vol. 114, pp. 611623
 94. N. Bellotto, E.Sommerlade, B. Benfold, C. Bibby, I. Reid, D. Roth, C. Fernandez, L. Van Gool and J.i Gonzalez *A Distributed Camera System for Multi-Resolution Surveillance*, Proc. of the 3rd ACM/IEEE Int. Conf. on Distributed Smart Cameras (ICDSC) 2009
 95. G. Scotti, L. Marcenaro, C. Coelho, F. Selvaggi, C.S. Regazzoni, *Dual camera intelligent sensor for high definition 360 degrees surveillance* IEE Proceedings Vision, Image and Signal Processing, 2005, Vol.152 , No. 2, pp. 250-257
 96. V. Saligrama, J. Konrad, and P.M. Jodoin *Video Anomaly Identification*, IEEE Signal Processing Magazine 2010, Vol. 27, No. 5, pp. 18-33
 97. E. B. Ermis, P. Clarot, P.M. Jodoin, V. Saligrama, *Activity Based Matching in Distributed Camera Networks*, IEEE Trans. on Image Processing, 2010, Vol. 19, No. 10, pp. 2595-2613
 98. C.C. Loy,T. Xiang, S. Gong, *Time-Delayed Correlation Analysis for Multi-Camera Activity Understanding*. Int Journal Computer Vision, 2010, Vol. 90, pp. 106129
 99. L. Q. Xu, A. Anjulian, *Relating Pace to Activity Changes in Mono- and Multi- Camera Surveillance Videos*, IEEE Conf. on Advanced Video and Signal Based Surveillance, 2009, pp. 104-109
 100. C. C. Loy, T. Xiang, S.Gong, *Detecting and discriminating behavioural anomalies* Pattern Recognition, 2011, Vol. 44, No. 1, pp. 117-132
 101. C. de Leo, B.S. Majunath, *Multicamera video summarization and anomaly detection from activity motifs* ACM Transactions on Sensor Networks (TOSN), 2014, Vol. 10, n.2
 102. R. Vezzani, R. Cucchiara, *Video Surveillance Online Repository (ViSOR): an integrated framework*, Multimedia Tools and Applications, 2010, Vol. 50, n. 2, Kluwer Academic Press, pp. 359-380
 103. <http://www.cvg.rdg.ac.uk/slides/pets.html>
 104. <http://scienceandresearch.homeoffice.gov.uk/hosdb/>
 105. http://www-prima.inrialpes.fr/PETS04/caviar_data.html
 106. <http://vision.soe.ucsc.edu/node/178>
 107. <http://dipersec.king.ac.uk/MuHAVi-MAS/>
 108. T. D'Orazio, M.Leo, N. Mosca, P. Spagnolo, P.L. Mazzeo, *A semi-automatic system for ground truth generation of soccer video sequences*, 6th IEEE International Conference on Advances Video and Signal Surveillance, 2009
 109. <http://trecvid.nist.gov/>
-



Biographical Sketch and Photo

Tiziana D'Orazio received the Computer Science degree in 1988 from the University of Bari. Since 1997 she has been a researcher at ISSIA-CNR. Her current research interests include pattern recognition, video analysis and computer vision for video surveillance, domotics, intelligent transportation systems, and quality control. She has published over 130 technical papers and book chapters in refereed conferences and journals in the areas of robotics and computer vision.



Cataldo Guaragnella graduated in electronic engineering in 1990 at University of Bari, Italy, and received the Ph.D. degree in Telecommunications by the Politecnico di Bari in 1994. In 1996 he joined the Electrics and Electronics Department of Politecnico di Bari as an assistant professor in Telecommunications. His main research interests include signal and image and video processing/coding, motion estimation in video sequences and multidimensional signal processing.