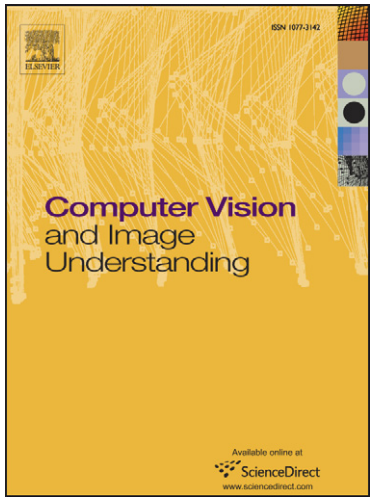


HECOL: Homography and Epipolar-based Consistent Labeling for Outdoor Park Surveillance

Simone Calderara, Andrea Prati, Rita Cucchiara

PII: S1077-3142(08)00012-X
DOI: 10.1016/j.cviu.2007.07.006
Reference: YCVIU 1422

To appear in: *Computer Vision and Image Understanding*

Received Date: 7 October 2006
Revised Date: 15 June 2007
Accepted Date: 7 July 2007

Please cite this article as: S. Calderara, A. Prati, R. Cucchiara, HECOL: Homography and Epipolar-based Consistent Labeling for Outdoor Park Surveillance, *Computer Vision and Image Understanding* (2008), doi: 10.1016/j.cviu.2007.07.006

HECOL: Homography and Epipolar-based Consistent Labeling for Outdoor Park Surveillance

Simone Calderara [a], Andrea Prati [b,*], Rita Cucchiara [a]

[a] *Dipartimento di Ingegneria dell'Informazione, University of Modena and Reggio Emilia, Via Vignolese, 905 - 41100 Modena, Italy*

[b] *Dipartimento di Scienze e Metodi dell'Ingegneria, University of Modena and Reggio Emilia, Via Amendola, 2 - 42100 Reggio Emilia, Italy*

Abstract

Outdoor surveillance is one of the most attractive application of video processing and analysis. Robust algorithms must be defined and tuned to cope with the non-idealities of outdoor scenes. For instance, in a public park, an automatic video surveillance system must discriminate between shadows, reflections, waving trees, people standing still or moving, and other objects. Visual knowledge coming from multiple cameras can disambiguate cluttered and occluded targets by providing a continuous *consistent labeling* of tracked objects among the different views.

This work proposes a new approach for coping with this problem in multi-camera systems with overlapped Fields of View (FoVs). The presence of overlapped zones allows the definition of a geometry-based approach to reconstruct correspondences between FoVs, using only homography and epipolar lines (hereinafter HECOL: Homography and Epipolar-based COnsistent Labeling) computed automatically with a training phase.

We also propose a complete system that provides segmentation and tracking of people in each camera module. Segmentation is performed by means of the SAKBOT (Statistical and Knowledge Based Object Tracker) approach, suitably modified to cope with multi-modal backgrounds, reflections and other artefacts, typical of outdoor scenes. The extracted objects are tracked using a statistical appearance model robust against occlusions and segmentation errors.

The main novelty of this paper is the approach to consistent labeling. A specific Camera Transition Graph is adopted to efficiently select the possible correspondence hypotheses between labels. A Bayesian MAP optimization assigns consistent labels to objects detected by several points of views: the object axis is computed from the shape tracked in each camera module and homography and epipolar lines allow a correct axis warping in other image planes. Both forward and backward probability contributions from the two different warping directions make the approach robust against segmentation errors, and capable of disambiguating groups of people.

The system has been tested in a real setup of a urban public park, within the Italian LAICA (Laboratory of Ambient Intelligence for a friendly city) project. The experiments show how the system can correctly track and label objects in a distributed system with real

time performance. Comparisons with simpler consistent labeling methods and extensive outdoor experiments with ground truth demonstrate the accuracy and robustness of the proposed approach.

1 Introduction

Computer vision approaches for people detection and tracking have been deeply studied in the last decade. Indoor and outdoor applications for video surveillance have been developed both in research projects (such as Pfinder [1], VSAM [2] or W4 [3]) and, more recently, in commercial systems. Typical approaches tackle the problem of people surveillance with the aim of detecting the presence and position of people in the image only. Occlusions and people overlaps suggested the adoption of probabilistic techniques to allow robust tracking of the person's position, especially indoors and with limited fields of view (FoVs). For instance, two recent works using probabilistic state definition and particle filtering are BramBLe [4] and the model presented by Lanz in [5]. These methods predict and update the status vector (position and speed) of the objects, but do not take into account the aspect of the objects itself, i.e. they do not provide people segmentation. In many surveillance applications, instead, the silhouette, shape and visual aspect are very important. For instance, for abnormal behaviour analysis, gait and posture must be classified. For people identification, biometric and soft biometric features (face, clothing, etc.) must be extracted. For video surveillance in public places, people obscuration should be provided for, by eliminating in the video the shape of people [6, 7]. For these purposes, people segmentation based on background suppression and deterministic tracking based on appearance are normally adopted [8,9]. This work addresses the aforementioned contexts, applied in outdoor surveillance of public parks.

In large outdoor environments, multi-camera systems are required. Distributed video-surveillance systems exploit multiple video streams to enhance the observation ability. Hence, the problem of tracking is extended from single to multiple cameras: people's shape and status must be consistent not only in a single view, but also in space (i.e., observed by multiple views). This problem is known as *consistent labeling*, since identification labels must be consistent in time and space.

Often cameras' FoVs are disjoint, due to installation and cost constraints. In this case, the consistent labeling should be based on appearance only. If cameras' FoVs are overlapped, consistent labeling can exploit geometry-based computer vision. This could be done with a precise system calibration and 3D reconstruction could

* Corresponding author,
Email address: prati.andrea@unimore.it (Andrea Prati).

be used to solve any ambiguity. However, this is not often feasible, in particular if the cameras are pre-installed and intrinsic and extrinsic parameters are not available. Thus, partial calibration or self-calibration methods can be adopted to extract only some of the geometrical constraints, e.g. to compute the ground-plane homography.

This paper proposes a novel method, called HECOL (Homography and Epipolar-based COnsistent Labeling), to provide consistent labeling of people segmented in large areas covered by multiple overlapped cameras. The method takes into account both geometrical and shape features in a probabilistic framework. Homography and epipolar lines are computed to create relationships between cameras. The multi-camera system is modelled as a Camera Transition Graph (CTG) that defines the possible overlap between cameras in a given setup. When a new object is detected, the exploration of the graph selects a subset of compatible labels which may be assigned to the object in order to limit the search space. An off-line training phase allows to compute the *Entry Edges of Field of View* (or E^2oFoV, in short) that define the area of overlapped FoV between cameras and permit to construct the homography. The learning phase also allows to compute the location of the epipoles of the overlapped cameras with a robust algorithm based on RANSAC optimization.

At runtime, people segmentation and tracking is provided from each camera module. The system exploits a suitable modification of the SAKBOT algorithm [10] that defines a statistical background model, selectively updated according to the knowledge of the moving objects in the scene and which can cope with shadows. The model has been improved in order to deal with specific problems of cluttered outdoor scenes, with multi-modal background distributions. Tracking is provided with a deterministic approach based on pixel appearance, that can cope with multiple and long-lasting occlusions [8].

The main novelty of the paper lies in the phase of consistent labeling that defines a probabilistic framework with forward and backward contributions: it checks the mutual correspondence of people using the axis of the objects precisely warped in the other FoV using epipolar lines. It accounts for the matching of the warped axis and the shapes of people. This makes the method particularly robust against segmentation errors and allows to disambiguate groups of people.

This paper is structured as follows. After a section presenting related works, section 3 describes the overall architecture of distributed surveillance system for people detection and tracking in outdoor scenarios. The Bayesian-competitive method is fully described in section 4. In particular, the construction and use of the Camera Transition Graph are described in subsection 4.1, while the Bayes framework is presented in the remainder of the section. Section 5 shows the full set of experiments carried out for analyzing the performance of the proposed system. Appendix 1 details the self-calibration method to create homography and epipolar lines. Finally, Appendix 2 describes the new modified version of SAKBOT conceived for outdoor cluttered

environments.

2 Related Works

Reference Paper	Color Information	Features	Geometric Information	Calib.
Kang *et al.* (2003) **[11]**	Polar representation	—	Homography	Yes
Stauffer *et al.* (2003) **[12]**	—	—	Homography and TCM	No
Orwell *et al.* (1999) **[13]**	Statistical color mod. & matching	—	—	No
Chang *et al.* (2001) **[14]**	Color appearance and HPCA	Object apparent height	Epipolar geometry	No
Yue *et al.* (2004) **[15]**	—	—	Homography	Yes
Khan and Shah (2003) **[16]**	—	—	EoFoV	No
Dockstader & Tekalp (2001) **[17]**	—	Person's body model	3D projection	Yes
Tsutsui *et al.* (1998) **[18]**	—	—	3D projection	Yes
Krumm *et al.* (2000) **[19]**	Color histogram	—	Stereo match	Yes
Zhou and Aggarwal (2005) **[20]**	—	—	3D trajectory projection and EKF	Yes
Black and Ellis (2005) **[21]**	—	—	3D location	Yes
Nummiaro *et al.* (2003) **[22]**, Tan and Ranganath (2003) **[23]**	Color appearance	Face database	—	No
Mittal and Davis (2001) **[24]**, Mittal and Davis (2003) **[25]**	Color Region	—	Epipolar geometry and 3D projection	Yes
`HECOL` (2007)	—	—	Epipolar geometry and homography	No

Table 1
Related works on multi-camera object matching and multiview tracking data fusion (TCM - Tracking Correspondence Model, HPCA - Hierarchical Principal Component Analysis, EKF - Extended Kalman Filter)

The approaches to consistent labeling can be generally classified into three main categories: appearance-based, geometry-based and mixed-approach. Table 1 summarizes the most relevant techniques presented in literature in recent years, describes their use of color and/or geometric information, and reports the features adopted.

Appearance-based approaches base the matching essentially on the color of the

objects. Several methods (e.g., [13]) exploit some metrics computed from people's colors histogram acquired on different views to label corresponding objects consistently. Using color information only can lead to errors when camera sensors acquire images under different light conditions. Color calibration can be performed to reduce errors, but each camera's color space can evolve unpredictably in general outdoor surveillance applications. To reinforce matching based on colors, general features are used. For instance, in [22] and [23], matching reliability is improved by the adoption of a database containing people's observed faces. This method works only if the camera's FoV assures a frontal view of the people's faces. In these cases, information on faces are very useful in identifying the same person viewed by different cameras but, in order to obtain a reliable match, faces's snapshots must be sufficiently detailed. This constraint on resolution is not feasible in most surveillance systems, especially in outdoor environments, limiting the actual applicability of the method.

Geometry-based approaches exploit geometrical relations and constraints between different views to establish the consistent labeling. In this process, camera calibration parameters can be employed or not. Exploiting calibration, the relationship between overlapped cameras can be easily modelled in the 3D space and warping techniques can be applied with high accuracy. In [15], Yue *et al.* propose a homographic mapping between a portion of the image plane and a specific 3D hyperspace (i.e., the ground plane) and provide consistency based on the people's positions. Although this approach performs robustly, there is no need of a complete calibration to detect corresponding planes and establish the mapping. Additionally, using only a two-dimensional planar warping, people close to each other are detected as a single object. Zhou and Aggarwal in [20] recently proposed the adoption of an extended Kalman filter operating on people's 3D trajectory to efficiently match moving objects on different views and to deal with occlusions exploiting filter prediction. In this approach, authors use a full camera calibration process to re-project the trajectory on the 3D space. The use of trajectories in a predictive framework can efficiently deal with occlusions caused by static elements but can mislabel corresponding objects in cluttered situations where trajectories overlap. Mittal and Davis in [25] proposed color region matching along epipolar lines to obtain a 3D re-projection of the objects viewed simultaneously from at least two cameras. This kind of re-projection produces a mapping similar to an approximated bird-eye view of the scene, where matching can be accomplished by means of data point clustering. Obviously, this kind of matching can not be computed in absence of calibration parameters. Conversely, a fully uncalibrated approach, based on the image projections of overlapped cameras' field of view lines, has been initially proposed by Khan and Shah in [16]: the lines delimiting the overlapping zones in the FoVs of the cameras are computed in a training phase with a single person moving in the scene. At run time, when one or more people have a camera handoff, the distances from the lines are used to disambiguate objects, assuring label consistency. Even though this approach represents an innovation in the use of image plane geometry relations, it achieves low accuracy when several people cross the FoV lines

simultaneously, or in presence of segmentation errors. It addresses neither the problem of the disambiguation of groups nor that of simultaneous detections of new objects. Stauffer and Tieu, in [12], propose an interesting method for building a graph representing the topology of a network of overlapped cameras directly from tracking data. Although the cameras' registration stage is very interesting and partially similar to the one proposed in this paper, the matching stage relies only on the homographies and the objects' position on the ground plane showing its weakness in the case of noisy tracking data such as partially extracted objects or grouped objects.

Mixed approaches combine information about the geometry with information provided by the visual appearance. Different techniques are adopted to fuse information, based on probabilistic information fusion [11] or on Bayesian Belief Networks (BBN) [14] [17]. The combination of both elements embodies advantages of previous approaches but also maintains some drawbacks. Dockstader *et al.*, for instance, in [17] use a Bayesian network to simultaneously perform spatial and temporal data fusion from multiple cameras by exploiting two Kalman-based predictors, one operating on the image plane's data and another working on observation on a 3D Cartesian space. To obtain the latter data elements, a three-view triangulation and 3D re-projection is performed using camera calibration parameters. In [14], Chang *et al.* combine the contributions from colors, features, and geometrical elements. Even though this approach seems to be very effective it still performs only one-to-one matching. Therefore it proves to fail in the case of segmentation errors and does not treat the problem of groups of people. Color histograms are used by Krumm *et al.* in [19], together with stereo matching techniques. Normalized histograms are only partially robust to viewpoint changes but suffer color delocalization. Another main drawback of the use of histograms is the definition of the number of bins to obtain a discriminant representation when color distributions cluster on some particular values because of both poor lighting conditions and acquiring sensor quality. To overcome data delocalization problems, mixed color-spatial representations were proposed, such as the correlograms used by Ho *et al.* in [26]. Kang *et al.* in [11] propose a polar representation to correct color delocalization. Although this representation can be very suitable for the shape of a single person, it cannot be adopted in the case of groups of people.

Another class of approaches presented in literature deals with multi-view geometry to analyze and impose continuity in the objects' trajectory across camera streams. This can be realized exploiting the homography mapping combined with an estimation technique, such as the use of 3D Kalman filters ([21], [18]). However, this class of approaches cannot solve the grouping problem since the trajectory could not be accurate if people enter in one view as already grouped.

The HECOL approach described in this paper belongs to the class of pure geometry-based approaches, but is unique since it does not require camera calibration and exploits both planar homography and epipolar geometry by automatic training in an

off-line phase. Actually, people shape is exploited, not in terms of color or texture, but in terms of shape used to disambiguate groups. Indeed, the proposed method has the important property of identifying people in groups.

Group detection has been extensively studied recently as the starting point for high-level reasoning on behavior analysis. Most single view approaches rely on the tracking system for detecting and identifying people grouping in the scene and, subsequently, correctly reacquiring targets after the splitting events. This task can be accomplished either observing target trajectories as in [27], or focusing the attention on target appearance [28, 29]. Single view tracking-based approaches cannot deal with people entering as already grouped in the scene, since the no information on the target is available prior to the grouping event. To overcome this problem, projection histograms [3] have been exploited to identify and count people in groups by searching for the projection of the heads; this approach obtains good results only in camera setups with pan angle close to zero degrees; this constraint remains unavoidable since perspective deformations can negatively affect the histogram shape. Information redundancy coming from the adoption of overlapped cameras can be very useful in characterizing groups. In [30] Cupillard *et al.* perform group detection in a system with overlapped cameras by representing moving objects via a local graph model on each view; then they fuse graphs together exploiting a set of fixed rules. Although this approach performs very robust also in cluttered scenes, such as metro stations, it still relies the matching on some 3D additional information not directly inferable from the image plane.

As stated in the introduction, one of the novel features introduced in our approach is the addition of the exploitation of epipolar geometry to enhance consistent labeling. In stereo systems, epipolar geometry is used to reduce complexity of the stereo matching problem and can be computed even in total absence of camera calibration. Most computation techniques exploit the generalized Longuet-Higgins equation, proposed by Luong and Faugeras in [31], which relates point correspondences in overlapped cameras' image planes through the fundamental matrix. Hartley *et al.* [32] have demonstrated the stability of the linear method using eight pairs of corresponding points. Non-linear techniques exist which reduce the number of correspondence pairs to seven [33]. In real uncontrolled environments, finding a large number of corresponding points is a difficult task to accomplish without introducing outliers that affects negatively the accuracy of the estimated fundamental matrix. Faugeras *et al.* in [34] introduce a technique for computing fundamental matrix from at least two corresponding planes exploiting the relation between homographies and fundamental matrices. As a particular case, this method can be applied even when only one homographic transformation between planes in the stereo cameras is known, leading to a full recovery of epipolar geometry from the knowledge of only two corresponding points not lying on the plane involved by the homographic transformation. This method can produce an imprecise estimate of the epipole location if the extracted corresponding points are not accurate enough.

Our approach exploits this technique in the learning phase and employs the highest points of the tracked heads seen by different cameras as corresponding points. In fact, head points are easier to extract than feet points since they do not suffer problems due to projected shadows. Moreover, to improve the accuracy of the computed fundamental matrix we propose to exploit an off-line RANSAC optimization technique [35] to reject false match and to reduce outlier influence in the estimation of the epipole location.

The geometric information learnt during the training stage are then robustly combined in a statistical framework using Bayes rule to obtain a reliable match even in presence of uncertainties on the geometric measure itself i.e. when there are segmentation errors. The use of Bayesian frameworks has been widely adopted to statistically link data acquired by multiple sensors [36]. In a network of disjoint cameras, Bayes rule provides a global optimization in order to select the best data association from objects' appearance and trajectories, applied both in traffic [37] and in people surveillance [38]. The general framework presented in [36] states for the need to perform a global association of objects at system level, considering all the hypotheses of link between two objects in any camera. The global hypothesis space is a high dimensional space and finding the best hypotheses may be computationally infeasible for real time application. In order to reduce the complexity of the MAP search, Kettnaker *et al.* in [39] propose to use a linear programming technique. However, in the case of overlapped cameras this high dimensionality can be reduced by performing data warping among cameras directly, without any prediction stage and with no need to globally find the best hypotheses. In fact, if the track is always visible at least in one camera, the pair-wise approach ensures the system consistency.

3 System Overview

The system defined at Imagelab (`http://imagelab.ing.unimo.it`) is a complex set of modules as depicted in the diagram of Fig. 1. The acquisition platform is composed of several cameras, both fixed and PTZ (Pan-Tilt-Zoom). Each of the modules connected with a fixed camera segments and tracks moving people.

Moving people segmentation is achieved by using the background suppression approach called SAKBOT (Statistical And Knowledge-Based Object Tracker) presented in [10]. The background pixels are defined by two models: the first statistical model updates the pixels at each frame using the temporal median function over the previous n sampled pixels; the second model exploits the knowledge of previous background and of the corresponding moving objects. Specifically, the pixels belonging to the current moving objects are not used for updating the model in order to prevent the gradual inclusion of slowly-moving objects into the background. Instead, pixels detected as foreground at previous steps but classified as

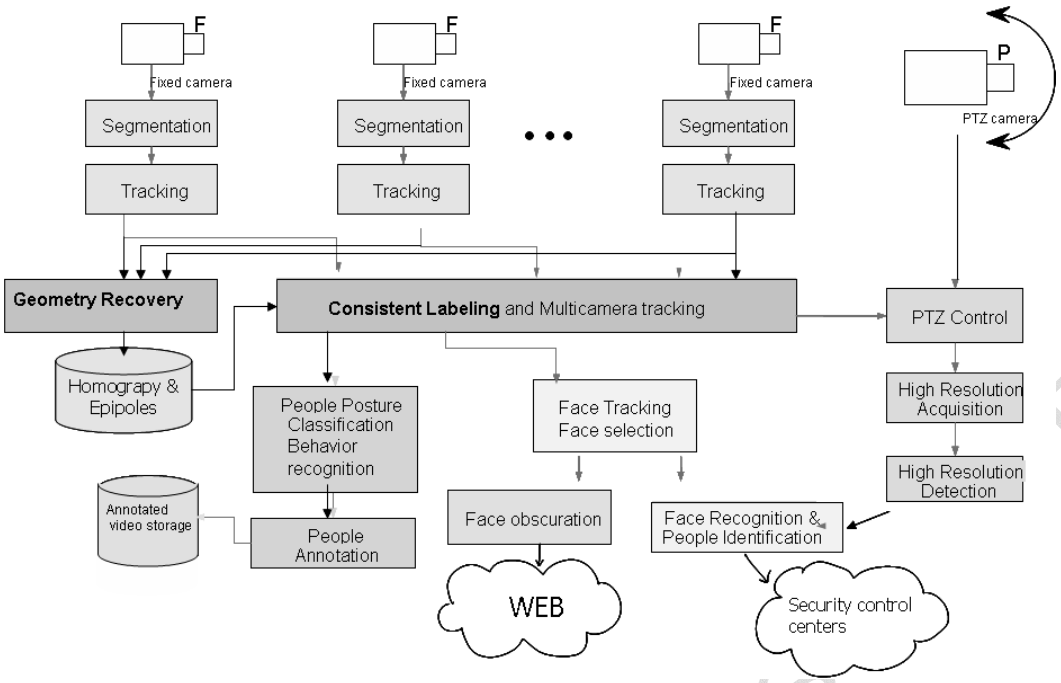


Fig. 1. Diagram of the outdoor surveillance system developed at Imagelab.

noise or shadows are included in the statistical model. This approach is critical when a stopped object starts to move, generating two “foreground” objects, one real and one apparent (also called “ghost”). To avoid deadlock situations for the ghosts, a specific ghost suppression algorithm has been conceived. Further details can be found in [10].

Despite its effectiveness in most situations, the SAKBOT system still suffers of some drawbacks in actual complex situations, such as those common in actual park scenarios. In particular, the adoption of the median is motivated by its reasonable approximation of the mode of a pdf. However, in the case of multi-modal distributions, this approach is likely to fail. For this reason, we modified the SAKBOT approach to properly work in cluttered outdoor environments. The modified approach contains proper techniques for background bootstrapping, ghost suppression, and object validation, and has been compared with state-of-the-art background modelling techniques, such as the mixture of Gaussian of Stauffer and Grimson [40]. Additional details on these improvements are reported in Appendix 1.

Moving objects detected by SAKBOT are then classified as person or non-person according with their geometrical shape and size (scaled according to their position on the ground plane). The objects detected as moving and validated as people are then tracked in each single view by means of an appearance-based algorithm. The algorithm uses a classical predict-update approach. It takes into account not only the status vector containing position and speed, but also the memory appearance model and the probabilistic mask of the shape [8]. The former, also called dynamic template in [3], is the adaptive update of each pixel in the color space. The latter is

a mask whose values, ranging between 0 and 1, can be viewed as the probability for that pixel to belong to that object. These models are used to define a MAP (Maximum A Posteriori) classifier that searches the most probable position of each person in the scene. The tracking algorithm is a suitable modification of a work, previously proposed by Senior [41], that includes a specific module for coping with large and long-lasting occlusions. Occlusions are classified into three categories: self-occlusions (or apparent occlusions), object occlusions, and people occlusions. Occlusion handling is very robust and has been tested in many applications. It can keep the shape of the tracked objects very precisely and has been exploited for people posture classification [42]. Details on the tracking algorithm can be found in [8].

As shown in Fig. 1, the outputs of each camera module are processed by the Bayesian-competitive consistent labeling module. This is indeed the main novelty of this paper and we will devote to it most of our discussion in Section 4. This module aims at assigning the same label to the different instances of the same person viewed by different cameras.

The global label assignment assured by the consistent labeling is the fundamental step for the subsequent, higher-level tasks shown in Fig. 1. For instance, a PTZ camera can be instructed to zoom on a specific person, given his/her position in the scene and calibrated cameras. This can allow the system to take high-resolution images to be further processed for face recognition and people identification. Another important task accomplished thanks to consistent labeling is the acquisition, for each tracked person, of multiple snapshots from multiple viewpoints. These snapshots can be stored and used for subsequent retrieval of information. An example of output and stored images is shown in Fig. 2.

4 Bayesian-competitive Consistent Labeling

The HECOL approach is based on two separate processes, as depicted in Fig. 3. The first process (left part of Fig. 3) is performed off-line and aims to compute overlapping constraints among camera views to create the homography and to extract epipolar geometry. This process runs for each pair of overlapped cameras and it is described in the Appendix 2. Three tasks are accomplished during this process:

- the computation of the *Entry Edges of Field of View* (E^2oFoV, Appendix 2.A), which ensure consistency of the correspondent extracted lines and permit to robustly compute inter-camera image homographies (Appendix 2.B);
- the recovery of the epipole location exploiting parallax property of perspective images and RANSAC optimization technique (Appendix 2.C);
- the synthesis of the Camera Transition Graph (CTG) which represents the consistency constraints of the whole multi-camera system (described later in section

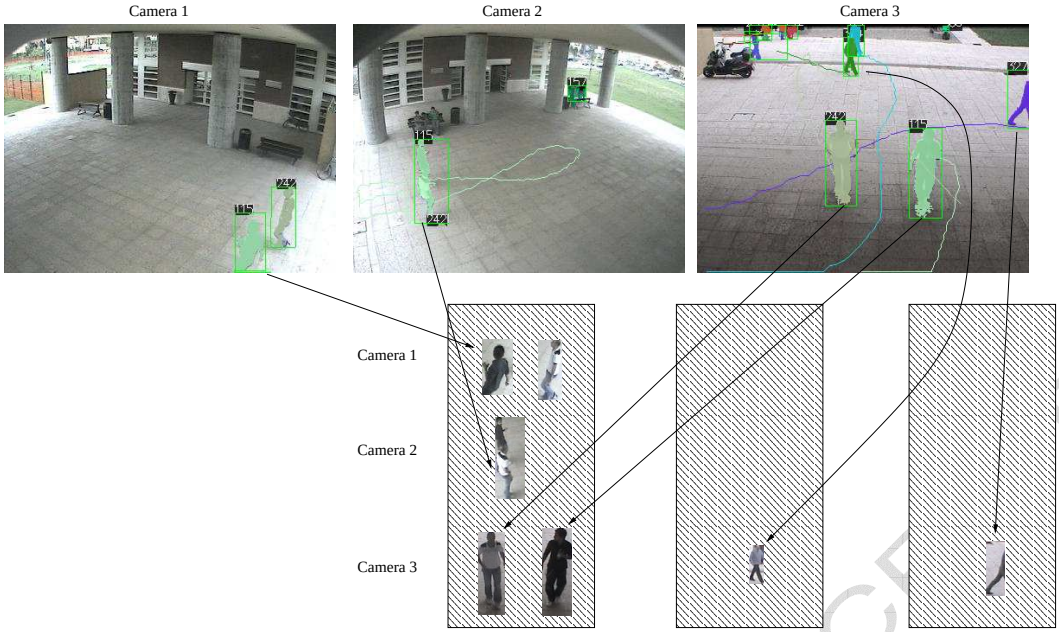


Fig. 2. Example of the logging system.

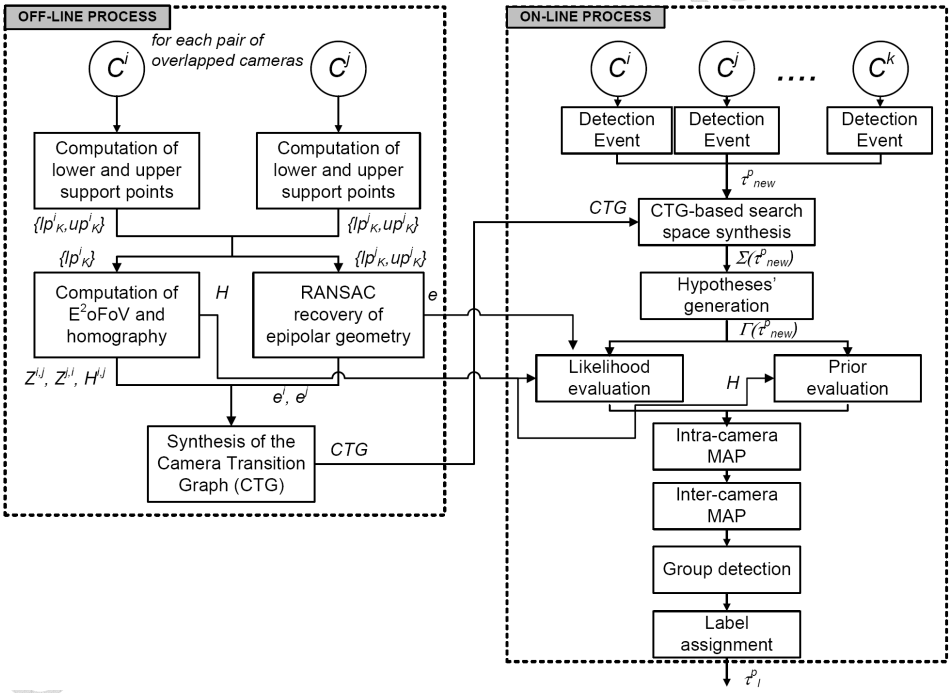


Fig. 3. The HECOL approach for Bayesian-competitive consistent labeling.

4.1).

After this initialization, the main on-line process (right part of Fig. 3) is devoted to manage and maintain system's consistency at each "detection event". Hereinafter, we will refer to the event that generates a new object detected by at least one camera module as a *"detection event"*. This event can be triggered by many causes: a

camera handoff, the entrance of a completely new object, the splitting of a group of people into single people, or a segmentation error. When the new object τ^p_{new} is detected in a camera p, the CTG created in the off-line phase is exploited to synthesize the search space $\Sigma(\tau^p_{new})$. The search space is parsed to generate the hypotheses associated to τ^p_{new}: these hypotheses consist of all the possible combinations of single persons, pairs or groups of people tracked in another camera (and acceptable within CTG constraints) which can be matched with τ^p_{new}. For each hypothesis, the prior and the likelihood are computed in order to obtain *intra-* and *inter-camera* probabilities. Using a MAP approach, the most likely hypothesis is assigned to the new object, enabling detection of groups and final label assignment. This MAP is not based on the estimation of the probability density distribution, but on the calculation of the probability every time with the likelihood function and the prior described above.

4.1 Camera Transition Graph

When a detection event occurs in the image of camera C^i, the multi-camera system must check whether it is a completely new person or it is already present in the FoV of other cameras. This check can be very complex and computationally expensive if there are many camera systems that are detecting many people. To prevent performance from dropping, a CTG graph model (Fig. 4) has been defined to reduce the search space of multi-camera matching. The graph's structure is built using information acquired during the training phase.

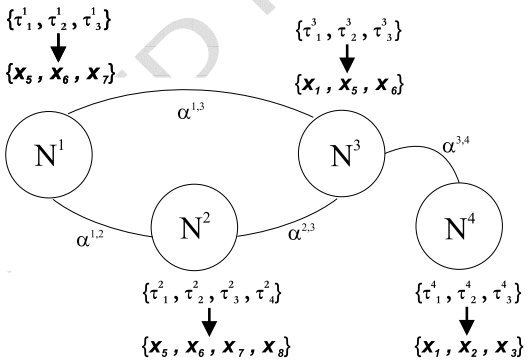


Fig. 4. Example of Camera Transition Graph (CTG) at a time when seven objects are detected and are partially visible from many cameras. For instance, the object with label x_5 is visible from the cameras associated with the nodes N^1, N^2, and N^3.

The CTG is a symmetric graph where each node N^i corresponds to each camera C^i and is associated with the set of objects detected and tracked in its FoV at that time, and each arc $\alpha_{i,j} = \alpha_{j,i}$ indicates the presence of an overlapping zone between C^i and C^j. In the CTG, for each node N^j, we denote with $T^j(t) = \left\{\tau^j_k(t) |\, k = 1, \ldots, K^j(t)\right\}$ the set of objects detected at the time t and with $x^j_k(t)$ their correspondent assigned labels (the instanced values for the variables).

The unary constraint at each node is that two distinct objects must have different labels and they must be conserved during the time. This is the typical tracking problem from a single camera and all unary constraints are checked at each frame by the single camera tracking module. On the other hand, the binary constraint between two nodes is that two instances of the same object viewed by C^i and C^j must have the same label. When a detection event occurs, a new variable is added to the node, and binary constraints are checked as in a Constraint Satisfaction Problem (CSP). By solving the arc-consistency problem, it is possible to correctly select only a subset of nodes and objects whose consistency must be verified to leave the rest of the system unchanged.

To explain the process, let us suppose that at time t the state of the system is consistent and that on the module of camera C^j the set $T^j(t)$ of $K^j(t)$ objects is known. Then, a new object is detected at time $t+1$. We will refer to the new object as τ^j_{new}, where $new = K^j(t)+1$. Instead of searching for all the possible matches across all the cameras' views, we analyze the graph node N^j and compute the neighborhood of N^j, Ξ^j, as the set of nodes linked to it by means of an arc constraint:

$$\Xi^j = \left\{ N^i | \exists \alpha_{i,j} \right\} \tag{1}$$

Referring to Fig. 4, if a new object τ^1_4 is detected in N^1, the set $\Xi^1 = \{N^3, N^2\}$ is created.

Let us define as lp^j_{new} the lower support point of the new object, computed as the middle point of the bottom of the bounding box. For each element N^i of the set Ξ^j we must evaluate if the support point lp^j_{new} lies inside the overlapping zone $Z^{i,j}$ between camera C^j and camera C^i. In case lp^j_{new} is not contained in any overlapping zone, the new object is not visible in any other camera, therefore a new label x^j_{new} is assigned and the system state switches to consistent. Otherwise, a local search space referring to the new object τ^j_{new} is built. The local search space on camera C^i is denoted $\Sigma^{j,i}(\tau^j_{new})$, and it is composed by the set of objects τ^i_m such that:

$$\begin{aligned} &(i)\ N^i \in \Xi^j \\ &(ii)\ lp^j_{new} \in Z^{i,j} \\ &(iii)\ lp^i_m \in Z^{j,i} \end{aligned} \tag{2}$$

In other words, for each camera C^i whose view is overlapped with C^j (condition (i)) and in which the new object is visible (condition (ii)), the set of objects $T^i(t+1)$ at the time $t+1$ is considered. For each object of this set, the visibility on the camera C^j is checked by means of condition (iii) and, if it is visible, the object is considered as a candidate for consistent labeling. It is straightforward that each local search space obtained with this procedure is minimal and results in computational saving, especially if the matching procedure is complex and time consuming.

Additionally, reducing the search space improves the matching accuracy by reducing the probability of false matches.

Considering the above example, let us suppose that the new object τ_4^1 detected in the camera C^1 has its lower support point inside the overlapping zone with camera C^3, then the local search space for that camera is defined as $\Sigma^{1,3}(\tau_4^1) = \{\tau_1^3, \tau_2^3, \tau_3^3\}$. For the new object a hypothesis space Γ must be created for each candidate camera, considering all the possible candidate hypotheses. We must consider all the possible combinations since the new object may be composed by more than a single object already present in the system (i.e., it can be a group of people already visible on another camera). In our example, we obtain $\Gamma^{1,3}(\tau_4^1) = \{\tau_1^3, \tau_2^3, \tau_3^3, \{\tau_1^3, \tau_2^3\}, \{\tau_1^3, \tau_3^3\}, \{\tau_2^3, \tau_3^3\}, \{\tau_1^3, \tau_2^3, \tau_3^3\}\}$. Note that the hypothesis space also contains objects already correlated with other objects of the node N^1 (e.g., object τ_2^3 is correlated with object τ_1^1). The reason is that the new object could also have been created by a segmentation error: in this case, it will not correspond to an actual new object but to part of an existing one (τ_1^1). If the MAP process finds the correspondence between τ_4^1 and τ_2^3 (as well as with τ_1^2 in camera C^2), the system infers that also τ_1^1 and τ_4^1 are parts of the same object and must be merged.

The improvement due to the adoption of the CTG, instead of an exhaustive search, can be quantified by considering a generic setup of c cameras' systems. Let us consider that the number of object detected at time t on each camera system is expressed by $K^j(t)$. If at time t the whole system is consistent in terms of detected objects and at time $t+1$ a new object appears on camera C^i, without the CTG this new object must be matched against all the possible combinations that can be made with the objects appearing on each camera. The number of comparisons can be expressed using the following equation:

$$n_i(t+1) = \sum_{j=1, j\neq i}^{c} (2^{K^j(t+1)} - 1) \tag{3}$$

Using the CTG leads to a twofold contribution. Firstly, the number of camera systems involved in the searching process is reduced by considering only the overlapped ones; secondly, the number of objects is reduced considering only the ones inside the chosen overlapping zones. If we define as $\widetilde{K}^{j,i}(t+1)$ the number of objects visible at time $t+1$ on the camera system C^j and inside the overlapping zone between cameras C^j and C^i, the number of comparison $n_i'(t+1)$ using the CTG becomes:

$$n_i'(t+1) = \sum_{j=1\, j\neq i}^{c'} (2^{\widetilde{K}^{j,i}(t+1)} - 1) * \omega^{j,i} \tag{4}$$

where $\omega^{j,i}$ takes the value 1 if the CTG arc linking nodes N^j and N^i exists, and 0 otherwise. It holds by definition that $n \propto c \cdot 2^{K^j(t+1)}$ and $n' \propto c' \cdot 2^{\widetilde{K}^{j,i}(t+1)}$. Since $c' \leq c$ and $\widetilde{K}^{j,i}(t+1) \leq K^j(t+1)$, the CTG speedup is straightforward.

4.2 Bayesian-competitive Framework

To ensure system consistency, we define a Bayesian-competitive framework and a *maximum-a-posteriori* (MAP) estimator to choose the most probable label configuration to be assigned to the new object. To solve object matching across camera views and impose consistency, two main steps are performed for the new object τ_{new}^{j} appearing in the FoV of camera C^j. Having selected the hypothesis space $\Gamma^{j,i}$ for each camera C^i satisfying arc consistency of node of camera C^j in the CTG:

(1) for each hypothesis $\gamma_k^{j,i} \in \Gamma^{j,i}$, prior and likelihood are computed; a MAP estimator finds the most probable hypothesis on each camera C^i (called *intra-camera MAP*),
(2) if more than one node is present in Ξ^j, a second MAP estimator (called *inter-camera MAP*) detects the most probable among these hypotheses.

For each hypothesis $\gamma_k^{j,i} \in \Gamma^{j,i}$, the intra-camera MAP with the camera of node N^i is given by:

$$\begin{aligned} \tau_{new}^{j} \rightarrow \tau_{M}^{j} \text{ with } M = \left\{ x_m^i(t) | \tau_m^i \in \gamma_h^{j,i} \right\} \text{ where} \\ h = \arg\max_k \left(P\left(\gamma_k^{j,i} \,|\, \tau_{new}^{j} \right) \right) = \arg\max_k \left(P\left(\tau_{new}^{j} \,|\, \gamma_k^{j,i} \right) \, P\left(\gamma_k^{j,i} \right) \right) \end{aligned} \tag{5}$$

The prior indicates the probability of the existence of the hypothesis. This can be modelled with some geometrical considerations. Intuitively, the hypothesis of a group of people (e.g., $\{\tau_2^3, \tau_3^3\}$) has a high probability if the objects are close to each other in the image plane where the detection event occurs. This value is independent from the detection event itself. The prior probability is given by the homographic mapping between overlapped cameras computed in Appendix 2.B. In detail, each lp_k of each object τ_k included in the local search space is warped to camera's image plane where the new object appeared. The distance of the warped point from lp_{new} is used in prior computation. Since this contribution is *a priori*, no assumptions or information about the new object are used. Clutterness or isolation of warped lp points are considered as a discriminant element. More specifically, a hypothesis consisting of a single object will gain higher prior if the warped lp is far enough from the other objects' support points. On the other hand, a hypothesis consisting of two or more objects (i.e., a possible group) will gain higher prior if the objects it consists of result close to each other after the warping, and, at the same time, the whole group is far from other objects.

To achieve this, two contributions are combined. For the hypothesis of a group of people, for each pair of objects $\{\tau_a^i, \tau_b^i\}$ of the same hypothesis $\gamma_k^{j,i}$, an inner-

hypothesis distance is computed as the distance between warped lp of two objects:

$$Id(\tau_a^i, \tau_b^i) = \begin{cases} \left\| \left(H^{i,j}\mathbf{lp_a^i} \right) \times \left(H^{i,j}\mathbf{lp_b^i} \right) \right\| & \text{if } \{\tau_a^i, \tau_b^i\} \in \gamma_k^{j,i} \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

where $H^{i,j}$ is the homography matrix that warps the ground plane of camera C^i on camera C^j.

Furthermore, an outer-hypothesis distance is defined as the distance between the warped position of the object (τ_a) in the hypothesis and an object (τ_b) which is in the local search space but disjoint with the given hypothesis:

$$Od(\tau_a^i, \tau_b^i) = \begin{cases} \left\| \left(H^{i,j}\mathbf{lp_a^i} \right) \times \left(H^{i,j}\mathbf{lp_b^i} \right) \right\| & \text{if } \{\tau_a^i, \tau_b^i\} \in \Sigma^{j,i}(\tau_{new}^j) \wedge \left(\tau_a^i \in \gamma_k^{j,i} \wedge \tau_b^i \notin \gamma_k^{j,i} \right) \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

Using inner and outer distances, a relative score is associated to each hypothesis:

$$\sigma\left(\gamma_k^{j,i}\right) = \min_{\tau_a^i, \tau_b^i \in \Sigma^{j,i}(\tau_{new}^j)\ a \neq b} Od\left(\tau_a^i, \tau_b^i\right) - \max_{\tau_c^i, \tau_d^i \in \Sigma^{j,i}(\tau_{new}^j)\ c \neq d} Id\left(\tau_c^i, \tau_d^i\right) \tag{8}$$

The prior is $P\left(\gamma_k^{i,j}\right) = P_0 + \Delta\sigma\left(\gamma_k^{i,j}\right)$ with $P_0 = \frac{1}{|\Sigma^{i,j}(\tau_{new}^i)|}$ and with the partitioning constraints $\Delta\sigma\left(\gamma_k^{i,j}\right) < P_0$ and $\sum_{k=1}^{n} \Delta\sigma\left(\gamma_k^{i,j}\right) = 0$, since the sum of the probabilities must remain equal to 1 (i.e., the increments $\Delta\sigma$ must vanish each other).

An example is reported in Fig. 5, showing two images of the cameras associated to nodes N^1 and N^2. A detection event occurs generating τ_{new}^1. The hypothesis space $\Gamma^{1,2}\left(\tau_{new}^1\right)$ consists of single objects (τ_{74}^2, τ_{76}^2, and τ_{90}^2), of groups of two objects ($\{\tau_{74}^2, \tau_{76}^2\}$, $\{\tau_{76}^2, \tau_{90}^2\}$, $\{\tau_{74}^2, \tau_{90}^2\}$), and of the group composed by all the objects ($\{\tau_{74}^2, \tau_{76}^2, \tau_{90}^2\}$). Using the previously depicted prior model, the values reported in Table 2 are obtained. Consequently, the most probable hypothesis accounts for τ_{90}^2 as a single object, the next one accounts for the group $\{\tau_{74}^2, \tau_{76}^2\}$, and so on until the least probable hypothesis which considers all the objects as belonging to the same group.

It is important to underline that this prior model may seem complex but is very fast as far as run time computation is concerned. As a matter of fact, the warped position of lp to be used in equations 6 and 7 is directly available for objects that have a consistent label, and the computation of values through equations 8 is not computationally onerous.

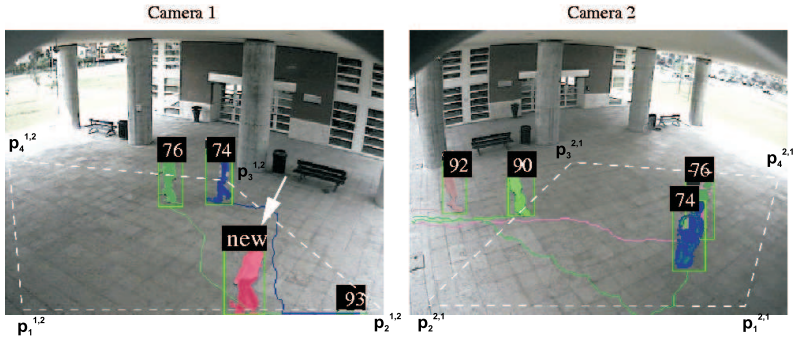


Fig. 5. Example of a detection event occurring on Camera C^1. The overlapping zone boundaries are shown with the dashed line and the new object is the one pointed by the arrow. Three people with labels 74, 76, and 90 are present and detected in both camera modules.

Hypothesis	σ value$[pxl]$	Prior value	Hypothesis	σ value$[pxl]$	Prior value
$\{\ \tau^2_{76}\ \}$	$+52$	$0,167$	$\{\ \tau^2_{76},\ \tau^2_{74}\ \}$	$+61$	$0,186$
$\{\ \tau^2_{74}\ \}$	$+52$	$0,167$	$\{\ \tau^2_{74},\ \tau^2_{90}\ \}$	-84	$0,100$
$\{\ \tau^2_{90}\ \}$	$+113$	$0,200$	$\{\ \tau^2_{76}, \tau^2_{90}\ \}$	-61	$0,110$
			$\{\ \tau^2_{76},\ \tau^2_{74},\ \tau^2_{90}\ \}$	-136	$0,070$

Table 2
Example of σ values and computed priors.

4.3 Likelihood Computation

Likelihood is computed by testing the fitness of each hypothesis against current evidence. As for prior probabilities, each hypothesis is evaluated separately. The main goal of this process is to distinguish between single hypotheses, group hypotheses and possible segmentation errors.

In this work, we only address geometrical properties and propose to exploit the vertical axis of the object as an invariant feature. In fact, the appearance of the person may change under different points of view, while in the ideal case, the axis of the object, taken on the image plane and suitably warped in the desired camera's image plane, does not change.

The axis of an object τ^j_k detected in the image plane I^j of camera C^j can be warped by exploiting the homography matrix and the knowledge of epipolar constraints among cameras. In absence of tilt angle, when camera's retinal plane is orthogonal to the ground plane, the inertial axis of a vertical object appears as a segment of a straight line parallel to the lateral border of the image. Even if the axis is vertical in its image plane, the warped axis in another camera's image plane is generally not perfectly vertical (Fig. 6). To compute the correct warping the camera's vertical vanishing point vp has been exploited. Considering a person in standing position while walking, his 3D principal inertial axis is orthogonal to the ground plane.

Since all the axes have the same 3D direction, their projection on each respective camera image plane will intersect at a vanishing point. The vanishing points are computed with the procedure described in [43].

The vertical vanishing point is then used to obtain warped axis inclination (Fig. 6). The inertial axis of the person is bounded by the lower and upper support point lp^i and up^i computed in I^i. With the homography, lp^i is projected into the point a_1^j of the image plane I^j:

$$a_1^j = H^{i,j}\mathbf{lp}^i \tag{9}$$

The warped axis will lie on a straight line passing through vp^j and a_1^j (Fig. 6.d). The ending point of the warped axis is computed by using up^i: since this point

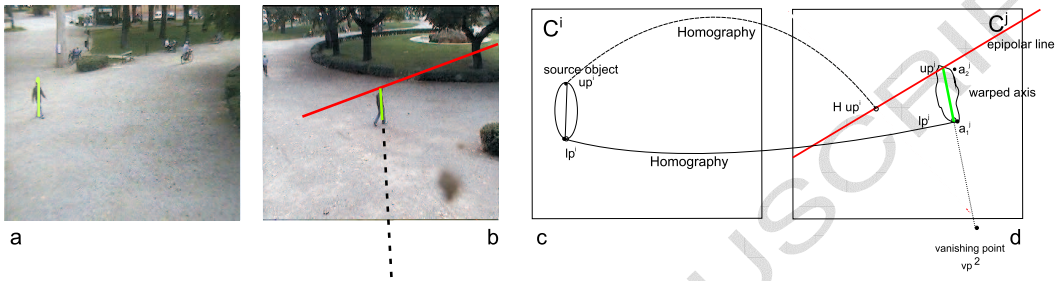


Fig. 6. Example of exploiting vanishing point and epipolar geometry to warp the axis of the object τ^i to the image plane I^j.

does not lie on the ground plane (on which the homography has been computed), its projection on the image I^j does not correspond to the up on C^j; however, the projected point lies on the epipolar line. Consequently, the ending point a_2^j is obtained as the intersection between the epipolar line and the straight line of the axis computed above:

$$a_2^j = \begin{cases} \langle \mathbf{e^i}, H^{l,i}\mathbf{up}_a^l \rangle \\ \langle \mathbf{vp^i}, H^{l,i}\mathbf{lp}_a^l \rangle \end{cases} \tag{10}$$

With the same process the axis of an object τ^j on C^j is warped on the segment $\langle a_1^i, a_2^i \rangle$ in I^i. In non-ideal situations, a measure of the axis correspondence must be defined by matching the pixels of the warped axis into the foreground blob ($FG(\tau)$) of the target object τ. The fitness measure $\xi_{\tau^i \to \tau^j}$ from the object τ^i to τ^j is defined as follows:

$$\xi_{\tau^i \to \tau^j} = \frac{\sum\limits_{\mathbf{y}_k^j \in I^j} \rho\left(\mathbf{y}_k^j, a_1^j, a_2^j, \tau^j\right)}{\|\langle a_1^j, a_2^j \rangle\|} \tag{11}$$

where $\mathbf{y}_k^j$ is the k-th point in the image I^j, and:

$$\rho\left(\mathbf{y}_k^j, a_1^j, a_2^j, \tau^j\right) = \begin{cases} 1 \text{ if } \left(\mathbf{y}_k^j \in FG(\tau^j)\right) \wedge \left(\mathbf{y}_k^j \in \langle a_1^j, a_2^j \rangle\right) \\ 0 \text{ otherwise} \end{cases} \tag{12}$$

Similarly, the axis of the object τ^j in the image plane I^j can be warped into I^i by computing the axis $\langle a_1^i, a_2^i \rangle$ with the same equations (9) and (10) as used previously, and $\xi_{\tau^j \to \tau^i}$ can be computed with equations (11) and (12) by swapping the i and j indexes. In the ideal case of correspondence between τ^i and τ^j, and for a person in a precise standing position (with their legs held together, so that lp and up correspond to feet and head, respectively), it holds that $\xi_{\tau^i \to \tau^j} = \xi_{\tau^j \to \tau^i} = 1$. Hence, in real situations, fitness should be close to 1. However, the computation of the axis and its warping is not error-free. Errors may be generated by some approximations in homography and epipole computation; non-negligible errors also come from the simplifications embodied in the computation of axis limits (lp and up). Moreover, in the case of groups of people, the lp and up points of the foreground blob of the object do not correspond to the limits of any real axis. In order to cope with these non-idealities and with groups of people, the formulation of the likelihood takes into account the matching of the object axis in both the cameras' image planes.

In the computation of the likelihood, we refer to *forward* contribution when the warping of the axis is made from the image plane in which the new object appears (I^j) to the image plane of the considered hypothesis (I^i). Thus, forward axis correspondence can be evaluated by computing the fitness of the new object τ_{new}^j with all the objects composing the given hypothesis $\gamma_k^{j,i}$ for camera C^i:

$$fp_{forward}\left(\tau_{new}^j | \gamma_k^{j,i}\right) = \frac{\sum\limits_{\tau_k^i \in \gamma_k^{j,i}} \xi_{\tau_{new}^j \to \tau_k^i}}{card\left(\Sigma^{j,i}\left(\tau_{new}^j\right)\right)}(1 - S) \tag{13}$$

where S measures the maximum range of variability of the fitness measure of the objects inside the given hypothesis:

$$S = \max_{\tau_k^i \in \gamma_k^{j,i}} (\xi_{\tau_{new}^j \to \tau_k^i}) - \min_{\tau_k^i \in \gamma_k^{j,i}} (\xi_{\tau_{new}^j \to \tau_k^i}) \tag{14}$$

The use of the cardinality of the local search space as a normalization coefficient weights each hypothesis according to the presence or absence of objects in the whole scene.

Backward contribution is computed similarly to the previous contribution, except that warping source and destination camera are swapped. Each object in the hypothesis $\gamma_k^{j,i}$ is warped to the image plane of the source camera C^j, with the same axis warping technique used for forward contribution, and matched against object τ_{new}^j as reported in the following equation:

$$fp_{backward}\left(\tau_{new}^j | \gamma_k^{j,i}\right) = \frac{\sum\limits_{\tau_k^i \in \gamma_k^{j,i}} \xi_{\tau_k^i \to \tau_{new}^j}}{card\left(\Sigma^{j,i}\left(\tau_{new}^j\right)\right)}(1 - S) \tag{15}$$

where S now becomes

$$S = \max_{\tau_k^i \in \gamma_k^{j,i}} (\xi_{\tau_k^i \to \tau_{new}^j}) - \min_{\tau_k^i \in \gamma_k^{j,i}} (\xi_{\tau_k^i \to \tau_{new}^j}) \tag{16}$$

The S parameter modulates the contribution according to the range of variability of the fitness within the hypothesis γ. Hence, hypotheses having, for the objects composing it, high values close to each other will be preferred to hypotheses having wide ranges of values.

At the end, the maximum of the two contributions is computed and taken as likelihood value:

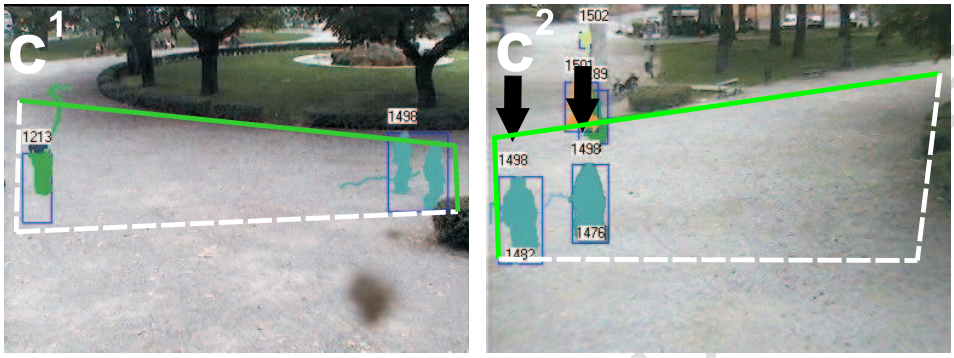
$$P\left(\tau_{new}^j \,|\, \gamma_k^{j,i}\right) = \max(fp_{backward}, fp_{forward}) \tag{17}$$

The choice of the maximum value can be explained by the fact that both the contributions represent the degree of compatibility of the hypothesis, but account for two different points of view. In case when a simple one-to-one assignment must be made, both contributions should be high. In some cases, one of the contributions can be quite unreliable, i.e. can tend to zero: this happens when the points used for the warping of the axis are not correctly extracted due to the position of the person in the scene. The use of the maximum value ensures the contribution where the extraction of support points is generally more accurate and suitable for the matching will be used.

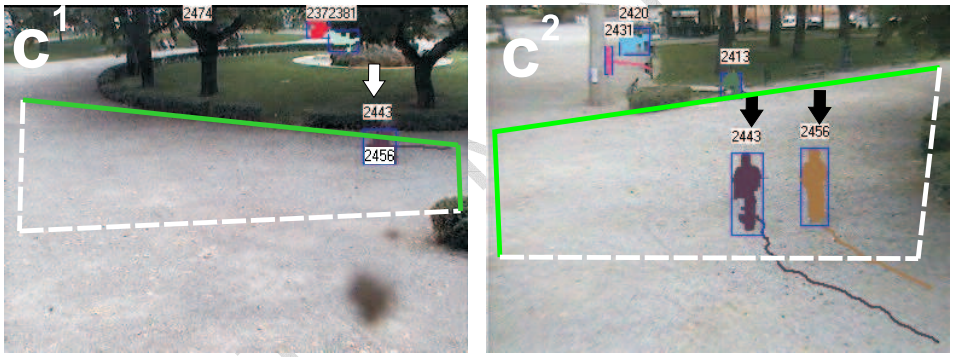
The effectiveness of the double backward/forward contribution is evident in the full characterization of groups of people. Forward contribution, in fact, is useful to distinguish people in a group which had been previously detected as a single object. This is the case in which a group is already inside the scene and the group's components appear one at a time in the FoV of another camera. This situation will be referred as *"group inside"* hereinafter. In particular, after warping, each axis of the new objects will intersect the same group's foreground blob on the chosen camera. This information can be exploited to characterize the new objects as belonging to the existing group. An example is reported in Fig. 7(a), where the group labeled with 1498 in the image of camera C^1 is detected as two separate people (id #1476 and #1482) in the image of camera C^2. The two objects are evaluated separately and for each of them the warped axis falls in the blob of the object #1498 on camera C^1, therefore they have a forward contribution very close to 1. The backward contribution, on the other hand, can be low since the axis of the object labeled 1498 in C^1 does not precisely correspond to the axis of any person composing the group and its warping on C^2 will be imprecise.

Merged transitions (i.e., cases in which people appear in a new camera as merged in a group; this case will be referred as *"group enter"*) can be solved by exploiting the fact that in the other camera's module the two objects are detected as separated (Fig. 7(b)). Fig. 8 shows a specific example for explaining the case of group enter.

In Fig. 8(a) (magnified in Fig. 8(b)) a new object appears and it is detected as a single object in camera C^2. In the camera C^1, instead, the two objects are detected as separated (Fig. 8(c) and 8(d)). The forward likelihood of the new object is warped on the image plane of camera C^1, obtaining the dotted line in Fig. 8(d). This line intersects only the object τ^1_{111}, therefore the forward likelihood is zero for object τ^1_{110}. The backward likelihood, on the other hand, is obtained by warping the solid axes of Fig. 8(d) back on the camera C^2 and obtaining the dotted lines in Fig. 8(b). As a consequence the resulting maximum likelihood is assigned to the hypothesis containing both the objects τ^1_{110} and τ^1_{111}.

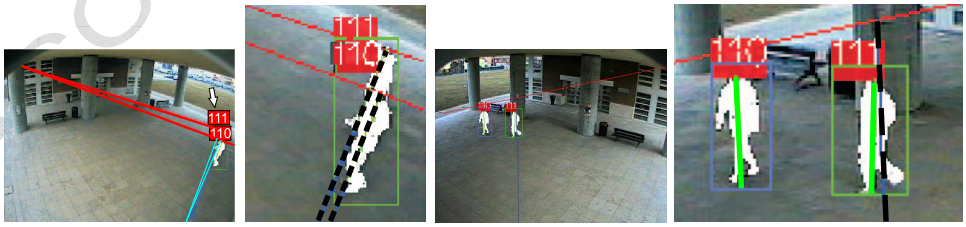


(a) Group inside



(b) Group enter

Fig. 7. Examples of full group characterization



(a) C^2 (b) Zoom of Fig. (a) (c) C^1 (d) Zoom of Fig. (c)

Fig. 8. Example of group detection and full group characterization based on disjoint objects in an overlapped view.

4.4 Inter-camera Bayesian assignment

The previous Bayesian-competitive framework, which takes into account the priors and the likelihood computed with backward and forward contributions, selects the most probable hypothesis of mapping the new object with a single object or a group of objects. This *intra-camera* MAP is computed for each couple of overlapped cameras according to the CTG graph. Global system consistency is assured by transitivity properties of subsequent assignments on different cameras. For instance, if at time t camera C^1 is consistent with camera C^2, and, at time $t+1$, an object τ on camera C^3 enters in both overlapping zones, selecting τ representation on either C^1 or C^2 does not affect consistency. Nevertheless, in complex scenes, more hypotheses could have similar a-posteriori probability. In order to reinforce the assignment, the same intra-camera MAP is computed for all pairs of overlapped cameras. Then, an *inter-camera* MAP selects the hypothesis that has the maximum a-posteriori probability among them:

$$P\left(C^i \mid \tau^j_{new}\right) \propto P\left(\tau^j_{new} \mid C^i\right) = \max_{\gamma^{j,i}_k \in \Gamma^{j,i}(\tau^j_{new})} P\left(\tau^j_{new} \mid \gamma^{j,i}_k\right) \tag{18}$$

Priors have all the same value considering all the overlapped camera views equally probable. Eventually, the label is assigned to the new object according to the winning hypothesis. If the chosen hypothesis identifies a group, all the labels of objects composing the group are assigned as identifiers, as in Fig. 7(a). Using equation (18), label assignment is expressed by the following formula:

$$\begin{aligned} &\tau^j_{new} \rightarrow \tau^j_M \text{ with } M = \left\{x^i_m(t) | \tau^i_m \in \gamma^{j,i}_h\right\} \text{ where} \\ &h = \arg\max_k \left(P\left(\gamma^{j,i}_k \mid \tau^j_{new}\right)\right) \text{ with } \; i = \arg\max_n \left(P\left(C^n \mid \tau^j_{new}\right)\right) \end{aligned} \tag{19}$$

5 Experimental Results

The presented system has been tested on different setups: in the Campus of University (see, for instance, Fig. 2) and in a public park in Reggio Emilia (see Fig. 11), Italy. It is the final demonstrator of the project LAICA (in Italian "Laboratorio di Ambient Intelligence per una Città Amica", in English "Laboratory of Ambient Intelligence for a Friendly City"), a large project funded by Regione Emilia Romagna, Italy.

In the park, tens of cameras were installed for standard CCTV-based video surveillance. Fig. 9 shows a map of the park with snapshots acquired from two example

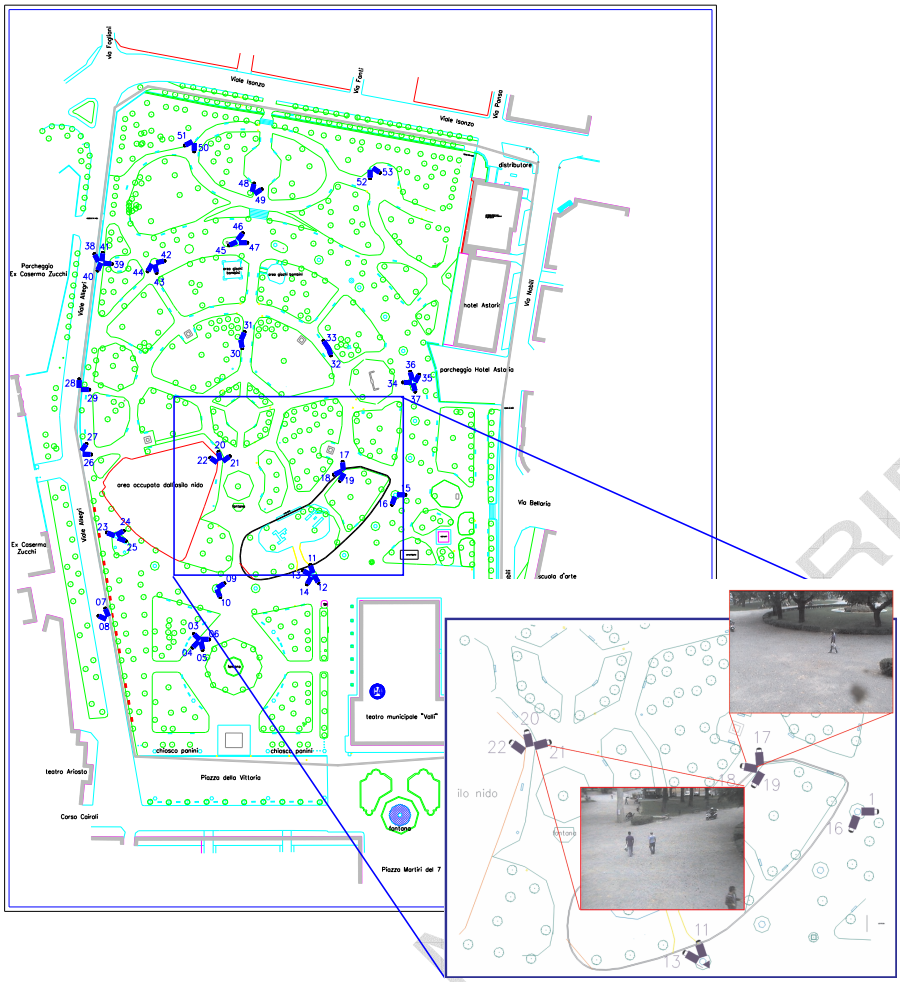


Fig. 9. Map of the public park in Reggio Emilia, Italy, used as real scenario.

cameras. Since the acquisition system was previously installed without camera calibration, the training phase for geometry recovery described in Appendix 2 has been adopted.

The testing area is particularly challenging since shadows and reflexes, illumination variations, fixed objects (e.g., trees and benches), and waving trees make segmentation and tracking processes particularly complex. As a consequence, often segmentation errors arise and new small objects are detected in the image. Moveover, several people enter from different directions and simultaneous camera-handoff is frequent. Finally, the monitored area presents a non-uniform illumination so that the local threshold adaptation described in Appendix 1 is crucial.

Tests have been performed considering several video sequences acquired during ordinary work days in different luminance conditions for two or three partially overlapped cameras. In acquiring the videos, no constraints have been imposed on people's trajectories or behaviors in order to test the system in actual conditions. The system works in real time on a server that samples frames from each camera

source and multiple threads are issued to provide tracking from each single camera module. When a detection event occurs on a camera module, the corresponding thread sends a synchronization event to the other threads and the consistent labeling is established.

	Single (tot. 114)	Group enter (tot. 25)	Group inside (tot. 21)	Overall (tot. 160)
Homography only	96	17	0	113
Forward only	103	**23**	0	126
Backward only	98	0	**19**	117
HECOL	**109**	**23**	**19**	**151**

Table 3
Experimental results obtained observing an actual test set of 97 minutes video. In bold the best performance for each type of detection event.

Among the different videos analyzed, more than 1 hours of videos have been manually annotated to create ground truths. The achieved results are reported in Table 3. This table highlights the performance of the system with respect to the different situations mentioned above (i.e., single person, group inside and group enter). Moreover, the last column reports the overall accuracy of the system. The first two rows of the table show the results with homography only (no epipole computation) and with a Bayesian classifier with forward contribution only (i.e., it does exploit neither priors nor backward contribution). This method corresponds to the approach used in [44], which essentially bases the object matching process on the support points' distances measured on the inter-camera ground plane homographic mosaic. The rest of the table reports the improvements in system performance by adding backward contribution and Bayesian-competitive reasoning.

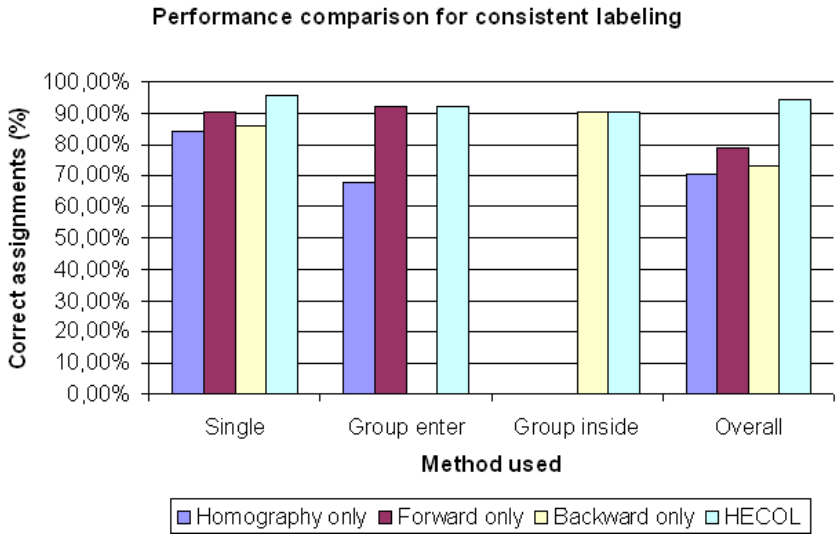
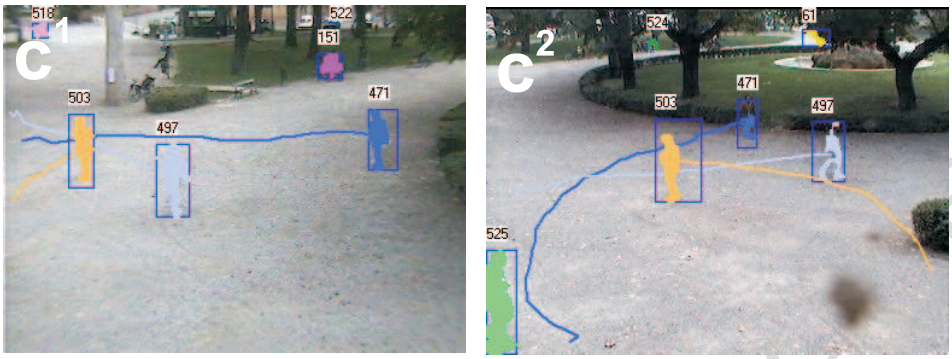


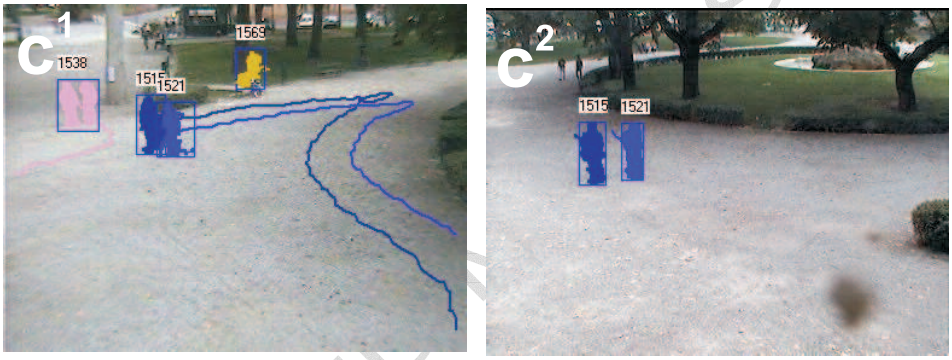
Fig. 10. Performance comparison.

The graph in Fig. 10 summarizes the comparison, for the different type of detection event, between the different implemented methods. In counting the correct assignment in the case of segmentation errors each piece in which the person is segmented has been evaluated independently, and considered correct if labeled consistently.

Results on group detection demonstrate that each of the two contributions in the MAP estimation concurs to solve different situations. In particular, the backward contribution is more suitable in assigning correct labels in the case of “group enter”, but cannot detect new objects in case of a group of people entered together and split in a second time (“group inside”). Conversely, in this case, the forward contribution is able to achieve a percentage of correct classification of 92%. From these results, the advantage of merging forward and backward contributions is evident.



(a)



(b)

Fig. 11. Examples of system working on real cases.

Snapshots in Figs. 11(a) and 11(b) show the output of the system in real situations: in the upper row, five moving individuals are tracked, and two of them are visible from both cameras; in the lower, two people are detected as a group from one FoV and as two disjoint objects in another, but consistent labeling is guaranteed.

A second test session has been performed to evaluate the overall performance in terms of computational time required. Tests have been carried out using two cameras on a desktop personal computer P4 at 3GHz with 1GB RAM running Linux Fedora Core 3 kernel 2.6 brand, with a frame rate of about seven fps (frames per second) for each module (one for each camera). Communication between modules on different cameras is achieved by using IPC (inter-process-communication) to communicate data and events from/to consistent labeling process.

The velocity of the proposed method is heavily dependent on the cardinality of the local search space. The exploitation of CTG allows a drastic reduction of search

times. The accuracy results previously reported in Table 3 have been obtained with an unlimited search range, that is using the constraints described in equation 2 in all the overlapping zones. In order to increase the frame rate, the hypothesis space Γ, described in Section 4.1, is suitably filtered based on the Euclidean distance between the lower support point of the candidate object, τ^j_{new}, and the warped lower support point distance of candidate objects. Before adding an object to Γ, its lower support point is warped to candidate object's image plane using homography. Then, if the Euclidean distance between these two points is above a fixed threshold, the object is discarded from the hypotheses and is not taken into account in the consistent labeling process.

This filtering method can significantly increase the performance of the system at the expenses of a lower accuracy. Discarding objects from the hypothesis space reduces the number of comparisons for obtaining object matches, but fails in all cases where lower support points of either hypothesis objects or new objects are imprecisely extracted.

Frame rate gain and accuracy loss, as consequences of different distance thresholds, are shown in the graph of Fig. 12. The frame rate, after the complete segmentation, tracking, consistent labeling and people's shape storing, varies from 12 to 7 fps.

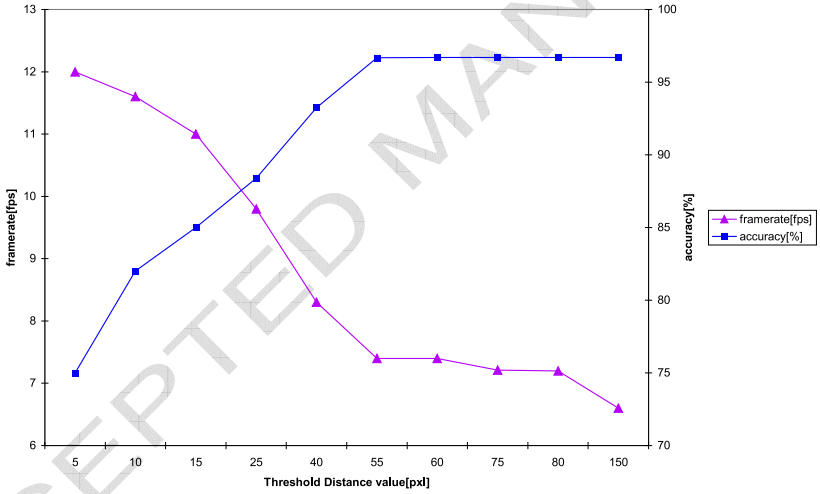


Fig. 12. Accuracy and computational load as a function of the distance filtering threshold. Accuracy is evaluated as $\frac{\#correct\ label\ assignment}{\#label\ assignment}$.

6 Conclusions

In this paper, a complete system for multi-camera video surveillance is described. The system is described in its overall architecture and main components, but most of the attention is devoted to a novel approach to consistent labeling in multiple

overlapped cameras. The consistent labeling method has been initially devised for camera handoff events, i.e. when an object passes in the FoV of a new camera. In simple situations, where single, well-segmented and well-tracked objects pass through the edges of the FoV, homographic mosaic alone can be sufficient to establish consistent labels. We propose to build the homographic mapping off-line without using calibration parameters: a training video with a single person walking is used to automatically compute Entry Edges of Field of View that define the overlapping zone whose vertices are used in the computation of the homography.

Unfortunately, this simple approach is not sufficient in many complex real cases: many new objects may appear in the middle of the overlapping zone due to segmentation errors; groups of people can enter as merged in the scene but, due to perspective, they can be perceived as separated by another camera; many people may pass in the FoV of another camera simultaneously. To cope with all these cases, a more complex approach must be conceived.

To find the best matching with a new object, we warp the object's axis on overlapped cameras and compute a fitness measure for each possible match. Since the axis does not lie on the ground plane (on which the homography has been computed), homography is not sufficient and epipolar geometry must be used. To do this, epipolar lines are computed by means of RANSAC optimization, directly in the same training step used to compute the homography. The warped axis provides, in a Bayesian framework, a measure of the goodness of a given object matching. MAP framework is exploited to choose the best matching and, consequently, the best label assignment. Eventually, the problem is generalized in the case of N overlapped cameras by exploiting a Camera Transition Graph (CTG) used to reduce the search space by checking arc consistency on the graph.

Experiments have been carried out in real park scenario in the city of Reggio Emilia, Italy. The resulting accuracy proves that the joint use of forward and backward contributions guarantees the best percentage of correct assignment.

Acknowledgments

This work is partially supported by the project BESAFE (Behavior lEarning in Surveilled Areas with Feature Extraction) funded by NATO Science for Peace programme (2006-2008) and by the project FREE SURF funded by Italian MIUR Ministry (2007-2008).

Appendix 1 - Reliable Background Suppression

This appendix describes the main steps of the algorithm used for background suppression to improve the SAKBOT approach [10].

Let us denote with BG_t the background model available at time t and with I_t the current input frame. We will also refer to $BG_t(i,j)$ to denote the vector (R,G,B) of pixel (i,j) in the background model. The same applies for $I_t(i,j)$.

A Background bootstrapping

A crucial task in every background suppression approach is the background initialization or *bootstrapping*, which needs to be both fast and accurate. Unfortunately, it is often impossible to have a clear background for many frames in order to compute statistics for building the model. Therefore, it is important to implement a method that can initialize the background model as quickly as possible even starting from "dirty" frames.

Our approach basically partitions the image into blocks (of 16×16 pixels) and selectively updates the background model with a block whenever a sufficiently high number of pixels within the block are not in motion. Motion is evaluated with a thresholded single difference between two consecutive frames. If more than 95% of the pixels in the block are detected as not in motion, the model is updated in that area with the pixels not in motion. If this happens for more than 10 times (even non consecutive), the whole block is considered "stable" and no longer evaluated. Once all the blocks have been set to "stable" the background model is ready. To avoid deadlocks and to speed up the bootstrapping, if no new blocks change state to "stable" for two consecutive frames, the threshold for the single difference is increased.

B Background Updating and Foreground Extraction

After the bootstrapping stage, the background model is updated using a temporal median. A fixed k-sized circular buffer is used to collect values of each pixel over time. In addition to the k values, the current background model $BG_t(i,j)$ is sampled and added to the buffer to account for the last reliable background information available. These $n = k + 1$ values are then ordered according to their grey-level intensity, and the median value is used as an estimate for the current background model.

Once the background model has been created and updated, the foreground is extracted frame by frame using the background differencing technique. The difference between the current image I_t and the background model BG_t is computed:

$$M_t(i,j) = \frac{(I_t(i,j) - BG_t(i,j)) \cdot \mathbf{i^T}}{3} \tag{B.1}$$

where $\mathbf{i^T}$ is the 1×3 identity vector. $M_t(i,j)$ is the foreground mask containing the grey-level information of the difference. The mask is then binarized using two different thresholds: a low threshold T_{low} to filter out the noisy pixel extracted due to small intensity variations; a high threshold T_{high} to identify the pixels where a large intensity variation occurs. Both these thresholds are local, i.e. they have different values for each pixel of the image. Let $b_p(i,j)$ be the value at position p inside the ordered circular buffer b of pixel (i,j) and, consequently, $b_{\frac{k+1}{2}+1}$ the median. The thresholds are computed as follows:

$$T_{low}(i,j) = \lambda \left(b_{\frac{k+1}{2}+l} - b_{\frac{k+1}{2}-l} \right) \tag{B.2}$$

$$T_{high}(i,j) = \lambda \left(b_{\frac{k+1}{2}+h} - b_{\frac{k+1}{2}-h} \right) \tag{B.3}$$

where λ is a fixed multiplier, while l and h are fixed scalar values. We experimentally set $\lambda = 7$, $l = 2$ and $h = 4$, for a buffer of $n = 9$ values. It is straightforward to see that, being the vector b ordered, T_{high} is always higher or equal than T_{low}. Our experiments demonstrated that these settings perform well in most common surveillance scenarios. The reason for the adoption of dynamic per-pixel thresholds is trivially explained by the fact that using fixed, per-frame thresholds makes the system less reactive to local illumination changes.

The final binarized motion mask B_t is obtained as composition of the two binarized motion masks computed respectively using the low and the high thresholds: a pixel is marked as foreground in B_t if it is presented in the low-thresholded binarized mask AND it is spatially connected to at least one pixel present in the high-thresholded binarized mask. Finally, the list MVO_t of moving objects at time t is extracted from B_t using a two-pass labeling algorithm. Each j^{th} element MVO_t^j of the list is a candidate foreground object. A further refinement of the list is performed discarding all the small objects.

Foreground points resulting from the background subtraction could be used for the selective background update, i.e. avoiding to include them in the background updating process; nevertheless, in this case, all the errors made during background subtraction will consequently affect the selective background update. A particularly critical situation arises in the case of "ghosts", preventing the area to be updated in the background image forever, causing a *deadlock*. Our approach substantially overcomes this problem since it performs selectivity not by reasoning on single moving points, but on detected and recognized moving objects. This object-level

reasoning proved much more reliable and less sensitive to noise than point-based selectivity [10].

C Object Validation

An object-level validation step is performed in order to remove all the moving objects generated by small motion of the background, for example by waving trees. This validation is performed accounting for joint contribution coming from color information and gradient of the objects.

The gradient is computed with respect to both spatial and temporal coordinates:

$$\begin{aligned}\frac{\partial I_t(i,j)}{\partial(x,t)} &= I_{t-\Delta t}(i-1,j) - I_t(i+1,j)\\ \frac{\partial I_t(i,j)}{\partial(y,t)} &= I_{t-\Delta t}(i,j-1) - I_t(i,j+1)\end{aligned} \tag{C.1}$$

For stationary points, we can approximate the past sample $I_{t-\Delta t}$ with the background model BG_t:

$$\begin{aligned}\frac{\partial I_t(i,j)}{\partial(x,t)} &= BG_t(i-1,j) - I_t(i+1,j)\\ \frac{\partial I_t(i,j)}{\partial(y,t)} &= BG_t(i,j-1) - I_t(i,j+1)\end{aligned} \tag{C.2}$$

The gradient module is obtained by using the following equation:

$$G_t = \left\{ g(i,j) \mid g(i,j) = \sqrt{\left\| \frac{\partial I_t(i,j)}{\partial(x,t)} \right\|^2 + \left\| \frac{\partial I_t(i,j)}{\partial(y,t)} \right\|^2} \right\} \tag{C.3}$$

From our tests, it is evident that this gradient module is quite robust against small motions in the background, mainly thanks to the use of temporal partial derivative. Moreover, the joint spatio-temporal derivative makes the object gradient computation more accurate, since it also detects gradient in the inner parts of the object.

Given the list of moving objects MVO_t, the gradient G_t is compared, for each pixel (i,j) of each moving object MVO_t^h, with the gradient (in the spatial domain) of the background GBG_t in order to evaluate the coherence with it. This *gradient*

coherence GC_t is evaluated over a $k \times k$ neighborhood as follows:

$$GC_t(i,j) = \min_{\substack{i-k \leq x \leq i+k \\ j-k \leq y \leq j+k}} |G_t(i,j) - GBG_t(x,y)| \tag{C.4}$$

where $|.|$ represents the absolute value of ., since G_t and GBG_t are grey-level images.

Unfortunately, when the gradient module (either G_t or GBG_t) is close to zero, data are not reliable. To overcome to this problem, we combine the gradient coherence with the *color coherence* CC_t:

$$CC_t(i,j) = \min_{\substack{i-k \leq x \leq i+k \\ j-k \leq y \leq j+k}} \|I_t(i,j) - BG_t(x,y)\| \tag{C.5}$$

where $\|.\|$ represents the norm in the RGB color space.

The overall validation score VS_t^h for the considered MVO_t^h is the normalized sum of the per-pixel validation score, obtained by multiplying the two coherence measures reported above:

$$VS_t^h = \frac{\sum_{(i,j) \in MVO_t^h} GC_t(i,j) * CC_t(i,j)}{N_t^h} \tag{C.6}$$

where N_t^h is the area of MVO_t^h. This value is then thresholded and, if below the threshold, MVO_t^h is discarded and its pixels are marked as belonging to background.

D Fast Ghost Suppression

As mentioned above, one of the problem of selective background updating is the possible creation of ghosts. Therefore, it is necessary to implement a method to detect ghosts and force them into the background model. The approach used is similar to that used for background bootstrapping (see Appendix 1.A), but at region level instead of pixel level.

All the validated objects are used to build an image called A_t that accounts for the number of times when a pixel is detected as still by the single difference:

$$A_t(i,j) = \begin{cases} A_{t-1}(i,j) + 1 & \text{if } SD(i,j) < T_{SD} \\ A_{t-1}(i,j)/2 & \text{otherwise} \end{cases} \tag{D.1}$$

where $SD(i,j)$ is the single difference between two consecutive frames and T_{SD} a suitable threshold. A valid object MVO_t^h is classified as ghost if:

$$\frac{\sum_{(i,j)\in MVO_t^h} A_t(i,j)}{N_t^h} > T_{ghost} \tag{D.2}$$

where T_{ghost} is the threshold on the percentage of points of the MVO_t^h stopped for sufficient time.

Practically speaking, in the case of pixels belonging to a ghost, the single difference will be lower than the threshold and will start increasing the value in A_t. When the sum of the accumulator values for the MVO exceeds the threshold T_{ghost}, the MVO is forced into the background.

Our experiments demonstrate that this approach yields to less total false results than state-of-the-art techniques, such as the mixture of Gaussian proposed by Stauffer and Grimson in [40]. Experiments have been carried out on the dataset provided at the web address `http://mmc36.informatik.uni-augsburg.de/VSSN06_OSAC/`. Fig. D.1 shows the results of the proposed method compared with those achieved with a mixture of Gaussians (MoG), in terms of false negatives and false positives at pixel level. It is clear that our method achieves lower overall false results, even though it is outperformed by MoG in terms of false negatives.

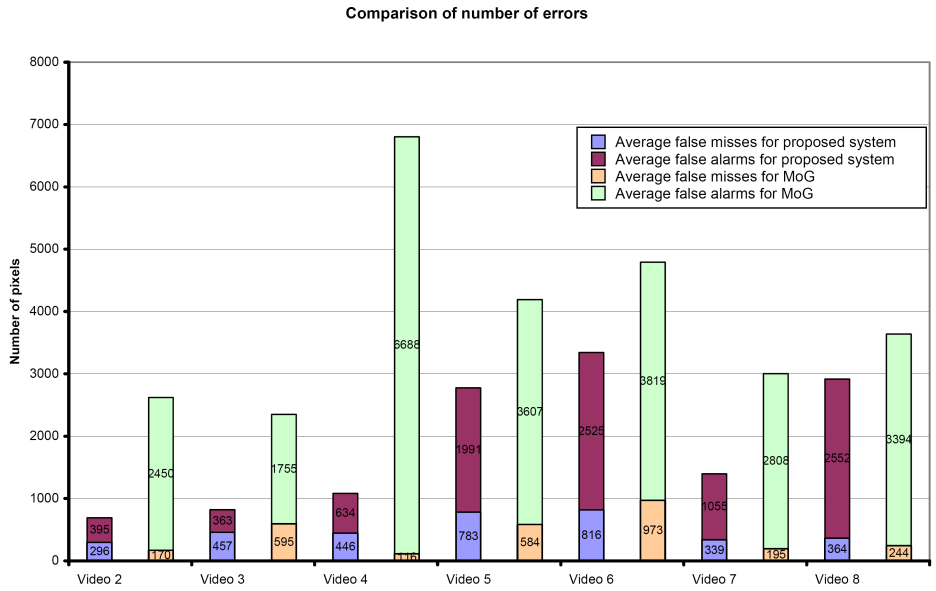


Fig. D.1. Comparative results with respect to mixture of Gaussians (MoG).

Appendix 2 - Geometry Recovery

Let us consider of a system composed of a set $\mathbf{C} = \{C^1, C^2, ..., C^n\}$ of n uncalibrated cameras, with each camera C^i overlapped with at least another camera C^j.

We aim at detecting overlapping zones between cameras to compute a reliable homography without the need of a complete camera calibration. An off-line training process is used. In the same process, additional data are extracted to recover epipolar geometry. Manually defining the overlapping zone could be possible, but is a tedious task which may lead to imprecision. The ground-plane homography can be computed automatically using people trajectories [45] and even when the camera streams are not synchronized [46]. In our system, the overlapping zones and the epipoles are automatically extracted with a training procedure iterated for each pair of partially overlapped cameras with a single person moving around the scene.

A Detecting Overlapping Zones

In order to compute the homography between two planes, corresponding points lying on the two planes must be extracted. Since our target planes are the 2D projection of the ground plane ($z = 0$) on the images of the two overlapped cameras, we need to extract the lower support points (lp) that approximate the contact with the ground plane. Thus, at each frame, the lp point and the upper support point up of the detected moving people can be extracted from each FoV. Lower support point is computed as reported in section 4.1, while the upper support point is the highest point of the shape. It is important that the persons shape and, consequently, the lp points are extracted only when it has entered the scene completely.

However, it is worth noting that the choice of the matching points directly affects the numeric precision and stability of the inter-view mapping itself. In particular, Vincent and Laganibre in [47] remark that using redundant match configurations (i.e., more than four matches) can lead to instability due to the introduction of degenerative elements. For this reason, directly using a set of lower support points extracted from a trajectory of the walking person as corresponding points for the homography computation can lead to imprecise values. Ideally, the corresponding points should be as distant as possible and non-collinear. If the whole overlapping zone between two overlapped cameras could be known, the corners of this area would satisfy both of these criteria: in fact, they are the farthest points visible in both the cameras and they are non-collinear. For this reason, instead of directly using the lower support points lp to compute the homography, we exploit the lp to identify the overlapped zone and then use the corners of the zone to compute the homography.

If we imagine that the camera view frustum (i.e., the rectangular pyramid with its vertex at the camera optical center) intersects the ground plane ($z = 0$), we obtain four lines. These lines correspond to the projection of the limits of the field of view of a camera on the ground plane; let us denote these as $L^{i,s}$, where i indicates the camera C^i being considered and s the corresponding line in the image plane.

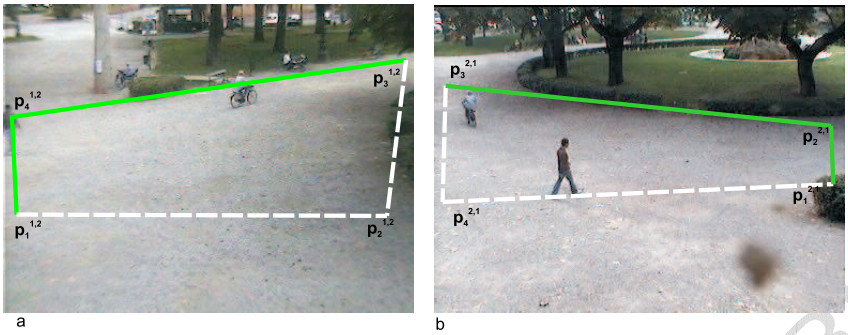


Fig. A.1. Examples of the process for detecting the overlapping zones

When the $L^{i,s}$ lines of camera C^i are also visible by another camera C^j, they are denoted as $L^{i,s}_j$; as in [16], we shall call these Edges of Field of View, or EoFoV in short. Each EoFoV $L^{i,s}_j$ divides the image on camera C^j into two half-planes, one overlapped with the FoV of camera C^i and the other disjoined. The intersection of the overlapped semi-planes defined by the EoFoV lines from camera C^i to camera C^j generates the overlapping zone $Z^{i,j}$. In actual cases, the generic overlapping zone $Z^{i,j}$ is delimited by a combination of EoFoV lines $L^{i,s}_j$ (when visible) and image-projected lines $L^{j,s}$ (see Fig. A.1, where solid lines indicate EoFoV, while dotted ones are image-projected $L^{i,s}$ lines).

When the object P and K are detected by the modules of camera C^i and C^j, respectively, and the objects' shapes do not intersect the image borders anymore, the pair of points (lp^i_P, lp^j_K) is collected. This is iterated several times. Collecting several matching pairs (lp^i_P, lp^j_K), the lines delimiting overlapping zones can be computed with a Least Square Optimization (LSQ). In general, given a set of samples $x_i, y_i | i = 1 \dots n$, the LSQ linear regression minimizes the mean square error (MSE):

$$MSE = \sum_{i=1}^{n} (y_i - bx_i - a)^2 \tag{A.1}$$

being (x_k, y_k) the coordinates of lp for both C^i and C^j.

B Homography computation using overlapping zones

The four corners of each of the overlapping zones $Z^{i,j}$ and $Z^{j,i}$ define a set of four points in the ground plane, $p^{i,j} = \{p^{i,j}_1, p^{i,j}_2, p^{i,j}_3, p^{i,j}_4\}$ and $p^{j,i} = \{p^{j,i}_1, p^{j,i}_2, p^{j,i}_3, p^{j,i}_4\}$, where the subscripts indicate corresponding points in the two cameras (see Fig.

A.1). These four associations between points of the two cameras C^i and C^j on the same plane $z = 0$ are sufficient to compute the homography matrix $H^{i,j}$ from camera C^i to camera C^j:

$$\mathbf{p_k^{i,j}} = H^{i,j}\ \mathbf{p_k^{j,i}}, \quad k = 1, \ldots, 4 \tag{B.1}$$

where bold notation is used, hereinafter, in order to indicate the representation of an image point in projective coordinates, which can be expressed in the explicit form as:

$$\begin{bmatrix} x_k^{i,j} \\ y_k^{i,j} \\ 1 \end{bmatrix} = \begin{bmatrix} h_{1,1}^{i,j} & h_{1,2}^{i,j} & h_{1,3}^{i,j} \\ h_{2,1}^{i,j} & h_{2,2}^{i,j} & h_{2,3}^{i,j} \\ h_{3,1}^{i,j} & h_{3,2}^{i,j} & 1 \end{bmatrix} \begin{bmatrix} x_k^{j,i} \\ y_k^{j,i} \\ 1 \end{bmatrix} \tag{B.2}$$

Since eight parameters must be estimated in order to compute the homography matrix, at least four pairs of points are needed to solve the equation system using a direct method. When more than four matching points are provided the homography matrix can be estimated, for example, in a least square framework using *Singular Value Decomposition* [48].

C Epipolar geometry recovery for coupled overlapped cameras

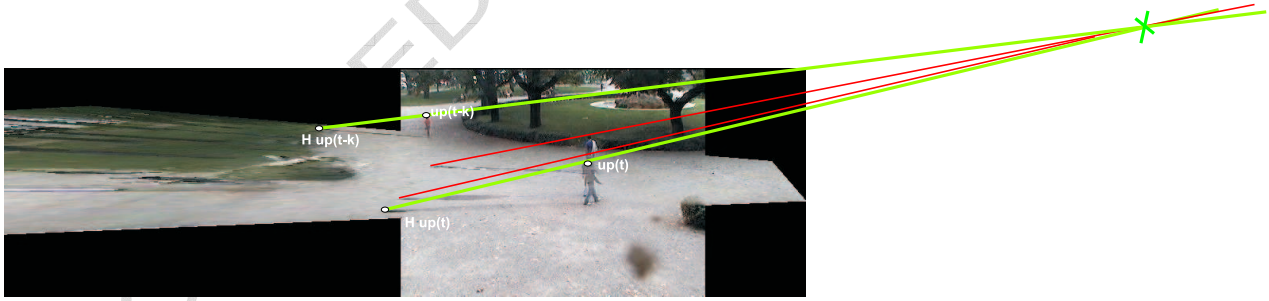


Fig. C.1. Example of exploiting parallax to compute epipole location. Green lines are kept after RANSAC optimization leading to estimated epipole indicated by the cross.

Denoting as M_k a point in world 3D reference frame, let us call $\mathbf{m_k^i}$ and $\mathbf{m_k^j}$ its projections in the FoV of the two cameras C^i and C^j, respectively. Epipolar constraints ensure that, given $\mathbf{m_k^i}$ on C^i, $\mathbf{m_k^j}$ on C^j must lie on a line $l^j_{\mathbf{m_k^i}}$ obtained as the projection on C^j of the 3D-line $\langle c^i, M_k \rangle$ passing through the optical center c_i of C^i and the 3D point M_k. Given the matrix form of this relation, it is possible to introduce the fundamental matrix F as the singular matrix that incorporates all the necessary information for the computation of line projections directly using points

on the image plane:

$$l^j_{\mathbf{m^i_k}} = F\mathbf{m^i_k} \tag{C.1}$$

As a consequence, the epipolar constraint is expressed by the following equation:

$$\mathbf{m^j_k}^T F\mathbf{m^i_k} = 0 \tag{C.2}$$

Given a 3D plane Π and its projections π^i and π^j on the image planes of camera C^i and C^j, the relation between them is expressed by the homographic matrix H. To compute epipole location using a single plane, the parallax property of projective images is exploited. In particular, given a 3D point M_k not lying on the plane Π, and its projection $\mathbf{m^i_k}$ on C^i, it is possible to find two correspondences in the image plane of C^j. The former is the real projection of M_k on C^j, $\mathbf{m^j_k}$, while the latter is the point in C^j computed through the homographic transformation H given the hypothesis that $\mathbf{m^i_k}$ lies on the plane π^i. In Fig. C.1, examples are reported where the upper support points up of a person (at the frame t and $t-k$) are projected on the homographic mosaic image. The two resulting lines intersect in the epipole.

The line computed from these points must be an epipolar line since it passes through the images of the same point of image plane I^i. Given at least two lines, the epipole can be easily located (by means of LSQ) as the intersection of these lines:

$$e^j = \min_{p^j}\left(\sum_k d^2\left(p^j, l^j_{\mathbf{m^i_k}}\right)\right) \tag{C.3}$$

where $l^j_{\mathbf{m^i_k}} = \langle \mathbf{m^j_k}, H\mathbf{m^i_k}\rangle$ [1] and $d(.)$ is an L_2 distance, e.g. the Euclidean distance. After epipole computation, given a generic image point m^i_k, the epipolar line can be obtained using the following equation:

$$l^j_{\mathbf{m^i_k}} = F\mathbf{m^i_k} = e^j \times \left(H\mathbf{m^i_k}\right) \tag{C.4}$$

where F is the fundamental matrix.

This method is mathematically correct but the use of LSQ with upper support points can be strongly unstable, as shown also in [34]. Therefore, we apply the RANSAC technique [35] instead to evaluate epipole location. The proposed method can be summarized in three main steps:

(1) during the previous E^{2}oFoV training phase, correspondences between object's head projections in both cameras involved are sampled at each frame and de-

[1] The notation $\langle a, b\rangle$ indicates the line passing through the points a and b.

	LSQ		RANSAC	
Sample number	50	100	50	100
Epipole C^1 (ground truth=(412,35))	$(450, 10)$	$(419, 25)$	$(413, 36)$	$(413, 36)$
Epipole C^2 (ground truth=(32,28))	$(-400, -330)$	$(-347, -228)$	$(30, 26)$	$(33, 30)$
MSE value C^1	8.156	4.630	3.116	3.495
MSE value C^2	10.918	6.417	2.989	2.361
Iteration number C^1	n/a	n/a	924	535
Iteration number C^2	n/a	n/a	621	1237

Table C.1
Comparison of the results achieved in an actual case with RANSAC (with $Th = 0.9$ and $Th_1 = 5$) and LSQ.

noted as $up_k^i(t)$ and $up_k^j(t)$, where the superscript indicates the camera image plane, the subscript the 3D object's identifier and t the time of sampling;

(2) after choosing a camera, C^j, we randomly compute two epipolar lines taken from the sample set of N frames using equation (C.3) and locate candidate epipole e^j;

(3) for each sampled match, except the ones used for computation, we evaluate accuracy of estimation building epipolar lines through e^j and head's points on C^i as shown in the following equation:

$$\phi(e^j) = \frac{\sum\limits_{up_k^i(t)\in\mathbf{S}} \varphi\left(e^j, up_k^i(t)\right)}{N} \tag{C.5}$$

with

$$\varphi(e^j, up_k^i(t)) = \begin{cases} 1 \text{ if } d^2\left(e^j, l^j_{\mathbf{up_k^i(t)}}\right) < Th_1 \\ 0 \text{ otherwise} \end{cases} \tag{C.6}$$

and $\mathbf{S} = \{(up_k^i(t), up_k^j(t)),\ t = 1, ..., N\ \}$.

If the *accuracy* parameter ϕ is greater than a fixed threshold Th the process is arrested; otherwise, the process is iterated by randomly choosing another couple of samples. To speed up the process, after a fixed number of iterations, threshold Th is lowered by a fixed amount and the whole process is restarted.

The results obtained using this method for epipole location are shown in Table C.1 and compared with LSQ relaxation technique. Strong difference in accuracy respect LSQ computation is motivated by RANSAC outlier rejection. In fact, with the proposed approach, outliers in point match are directly rejected during evaluation phase, so that location estimation is not affected by false matches. The solution obtained by the RANSAC technique is the best intersection of two epipolar lines computed from the extracted samples. Instead, using the LSQ relaxation, a new solution is generated leading to the possible case of having an estimated epipole near, in terms of mean square error, to the epipolar lines considered but not lying on any of these. As shown in Table C.1, the number of iterations of the RANSAC technique can vary significantly according to the order used in choosing samples. This cannot be predictable because lines are selected randomly and the ending condition is strictly influenced by the chosen accuracy of the estimated solution (thresholds

Th and Th_1). However, the high computational cost of the iterative method adopted can be easily sustained since this stage is performed off-line and does not affect the overall system performances at all.

References

[1] C. Wren, A. Azarbayejani, T. Darrell, A. Pentland, Pfinder: real-time tracking of the human body, IEEE Trans. on PAMI 19 (7) (1997) 780–785.

[2] R. Collins, A. Lipton, T. Kanade, H. Fujiyoshi, D. Duggins, Y. Tsin, D. Tolliver, N. Enomoto, O. Hasegawa, P. Burt, L. Wixson, A System for Video Surveillance and Monitoring: VSAM Final Report, Technical report CMU-RI-TR-00-12, Robotics Institute, Carnegie Mellon University.

[3] I. Haritaoglu, D. Harwood, L. Davis, W4: real-time surveillance of people and their activities, IEEE Trans. on PAMI 22 (8) (2000) 809–830.

[4] M. Isard, J. MacCormick, Bramble: A bayesian multiple-blob tracker, in: Proc. Int. Conf. Computer Vision, Vol. 2, 2001, pp. 34–41.

[5] O. Lanz, Approximate bayesian multibody tracking., IEEE Trans. on PAMI 28 (9) (2006) 1436–1449.

[6] R. Cucchiara, A. Prati, R. Vezzani, A system for automatic face obscuration for privacy purposes, Pattern Recognition Letters 27 (15) (2006) 1809–1815.

[7] E. Newton, L. Sweeney, B. Malin, Preserving privacy by de-identifying face images, IEEE Transactions on Knowledge and Data Engineering 17 (2) (2005) 232 – 243.

[8] R. Cucchiara, C. Grana, G. Tardini, R. Vezzani, Probabilistic people tracking for occlusion handling, in: Proc. of International Conference on Pattern Recognition (ICPR 2004), Vol. 1, 2004, pp. 132–135.

[9] S. Zhou, R. Chellappa, B. Moghaddam, Visual tracking and recognition using appearance-adaptive models in particle filters, IEEE Trans. on Image Processing 13 (11) (2004) 1491–1506.

[10] R. Cucchiara, C. Grana, M. Piccardi, A. Prati, Detecting moving objects, ghosts and shadows in video streams, IEEE Trans. on PAMI 25 (10) (2003) 1337–1342.

[11] J. Kang, I. Cohen, G. Medioni, Continuous tracking within and across camera streams, in: Proc. of IEEE Int'l Conference on Computer Vision and Pattern Recognition, Vol. 1, 2003, pp. I–267 – I–272.

[12] C. Stauffer, K. Tieu, Automated multi-camera planar tracking correspondence modeling, in: Proc. of IEEE Int'l Conference on Computer Vision and Pattern Recognition, Vol. 1, 2003, pp. 259–266.

[13] J. Orwell, P. Remagnino, G. Jones, Multi-camera colour tracking, in: Proc. of Second IEEE Workshop on Visual Surveillance, (VS'99), 1999, pp. 14–21.

[14] S. Chang, T.-H. Gong, Tracking multiple people with a multi-camera system, in: Proc. of IEEE Workshop on Multi-Object Tracking, 2001, pp. 19–26.

[15] Z. Yue, S. Zhou, R. Chellappa, Robust two-camera tracking using homography, in: Proc. of IEEE Intl Conf. on Acoustics, Speech, and Signal Processing, Vol. 3, 2004, pp. 1–4.

[16] S. Khan, M. Shah, Consistent labeling of tracked objects in multiple cameras with overlapping fields of view, IEEE Trans. on PAMI 25 (10) (2003) 1355–1360.

[17] S. Dockstader, A. Tekalp, Multiple camera tracking of interacting and occluded human motion, Proc. of the IEEE 89 (10) (2001) 1441–1455.

[18] H. Tsutsui, J. Miura, Y. Shirai, Optical flow-based person tracking by multiple cameras, in: Proc. 2001 Int. Conf. on Multisensor Fusion and Integration in Intelligent Systems, 2001, pp. 91–96.

[19] J. Krumm, S. Harris, B. Meyers, B. Brumitt, M. Hale, S. Shafer, Multi-camera multi-person tracking for easyliving, in: Proc. of IEEE Intl Workshop on Visual Surveillance, 2000, pp. 3–10.

[20] Q. Zhou, J. Aggarwal, Object tracking in an outdoor environment using fusion of features and cameras, Image and Vision Computing 24 (11) (2006) 1244–1255.

[21] J. Black, T. Ellis, Multi camera image tracking, Image and Vision Computing 24 (11) (2006) 1256–1267.

[22] K. Nummiaro, E. Koller-Meier, T. Svoboda, D. Roth, L. Van Gool, Color-based object tracking in multi-camera environments, in: DAGM03, 2003, pp. 591–599.

[23] M. H. Tan, S. Ranganath, Multi-camera people tracking using bayesian networks, in: Information, Communications and Signal Processing, 2003 and the Fourth Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint Conference of the Fourth International Conference on, Vol. 3, 2003, pp. 1335 –1340.

[24] A. Mittal, L. Davis, Unified multi-camera detection and tracking using region-matching, in: Proc. of IEEE Workshop on Multi-Object Tracking, 2001, pp. 3–10.

[25] A. Mittal, L. Davis, M2tracker: A multi-view approach to segmenting and tracking people in a cluttered scene, IJCV 51 (3) (2003) 189–203.

[26] L. Jiang, C. S. Chua, Y. K. Ho, Color based multiple people tracking, in: IEEE (Ed.), Control, Automation, Robotics and Vision, 2002. ICARCV 2002. 7th International Conference on , Volume: 1 , 2-5 Dec. 2002, 2002, pp. Pages:309 – 314 vol.1.

[27] W. Niu, J. Long, D. Han, J. Wang, Human activity detection and recognition for video surveillance, in: Proc. of IEEE Intl Conference on Multimedia and Expo, Vol. 1, 2004, pp. 719–722.

[28] M. Capellades, D. Doermann, D. DeMenthon, R. Chellappa, An appearance based approach for human and object tracking, in: Proc. of IEEE Int'l Conference on Image Processing, Vol. 2, 2003, pp. II – 85–8 vol.3.

[29] J. Kang, I. Cohen, G. Medioni, Object reacquisition using invariant appearance model, in: ICPR, Vol. 4, 2004, pp. 759–762.

[30] F. Cupillard, F. Bremond, M. Thonnat, Group behavior recognition with multiple cameras, in: Proc. of IEEE Workshop on Applications of Computer Vision, 2002, pp. 177–183.

[31] Q. Luong, O. Faugeras, The fundamental matrix: Theory,algorithms and stability analysis, International Journal of Computer Vision 17 (1996) 43–75.

[32] H. Hartley, In defence of the 8-point algorithm, in: Proc. 5th International Conference on Computer Vision, Cambridge(MA), 1995, pp. 1064–1070.

[33] A. Bartoli, P. Sturm, Nonlinear estimation of the fundamental matrix with minimal parameters, Pattern Analysis and Machine Intelligence, IEEE Transactions on 26 (3) (2004) 426 – 432.

[34] Q. T. Luong, R. Deriche, O. Faugeras, T. Papadopoulo, On determining the fundamental matrix: analysis of different methods and experimental result, Tech. rep., RR, INRIA (1993).

[35] M. A. Fischler, R. C. Bolles, Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography, Comm. of the ACM 24 (1981) 381–395.

[36] H. Pasula, S. J. Russell, M. Ostland, Y. Ritov, Tracking many objects with many sensors, in: Proc of International Joint Conferences on Artificial Intelligence, 1999, pp. 1160–1171.

[37] T. Huang, S. J. Russell, Object identification in a bayesian context, in: Proc of International Joint Conferences on Artificial Intelligence, 1997, pp. 1276–1283.

[38] O. Javed, Z. Rasheed, K. Shafique, M. Shah, Tracking across multiple cameras with disjoint views, in: Proc. of IEEE Intl Conference on Computer Vision, Vol. 2, 2003, pp. 952–957.

[39] V. Kettnaker, R. Zabih, Bayesian multi-camera surveillance, in: Proc. of IEEE Int'l Conference on Computer Vision and Pattern Recognition, Vol. 2, 1999, pp. 253–259.

[40] C. Stauffer, W. Grimson, Learning patterns of activity using real-time tracking, IEEE Trans. on PAMI 22 (8) (2000) 747–757.

[41] A. Senior, Tracking people with probabilistic appearance models, in: Proc. of Int'l Workshop on Performance Evaluation of Tracking and Surveillance (PETS) systems, 2002, pp. 48–55.

[42] R. Cucchiara, C. Grana, A. Prati, R. Vezzani, Probabilistic posture classification for indoor surveillance, IEEE Trans. on Systems, Man, and Cybernetics - Part A 35 (1) (2005) 42–54.

[43] C. Brauer-Burchardt, K. Voss, Robust vanishing point determination in noisy images, in: Proc. of Int'l Conference on Pattern Recognition, Vol. 1, 2000, pp. 559–562.

[44] S. Calderara, R. Vezzani, A. Prati, R. Cucchiara, Entry edge of field of view for multi-camera tracking in distributed video surveillance, in: Proc. of IEEE International Conference on Advanced Video and Signal-Based Surveillance, 2005, pp. 93–98.

[45] F. Lv, T. Zhao, R. Nevatia, Self-calibration of a camera from video of a walking human, in: Proc. of Int'l Conference on Pattern Recognition, Vol. 1, 2002, pp. 562–567.

[46] G. Stein, Tracking from multiple view points: Self-calibration of space and time, in: Proc. of IEEE Int'l Conference on Computer Vision and Pattern Recognition, Vol. 1, 1999, pp. 521–527.

[47] E. Vincent, R. Laganiere, Detecting planar homographies in an image pair, in: Proc. on Intl Symposium on Image and Signal Processing and Analysis (ISPA), 2001, pp. 182 – 187.

[48] J. C. Nash, The Singular-Value Decomposition and Its Use to Solve Least-Squares Problems., 2nd Edition, Compact Numerical Methods for Computers: Linear Algebra and Function Minimisation, Ed. Bristol, England: Adam Hilger, 1990, Ch. 3, pp. 30–48.