Bayesian-Competitive Consistent Labeling for People Surveillance

Simone Calderara, *Student Member*, *IEEE*,
Rita Cucchiara, *Member*, *IEEE*, and
Andrea Prati, *Member*, *IEEE*

Abstract—This paper presents a novel and robust approach to *consistent labeling* for people surveillance in multicamera systems. A general framework scalable to any number of cameras with overlapped views is devised. An offline training process automatically computes ground-plane homography and recovers epipolar geometry. When a new object is detected in any one camera, hypotheses for potential matching objects in the other cameras are established. Each of the hypotheses is evaluated using a prior and likelihood value. The prior accounts for the positions of the potential matching objects, while the likelihood is computed by warping the vertical axis of the new object on the field of view of the other cameras and measuring the amount of match. In the likelihood, two contributions (*forward* and *backward*) are considered so as to correctly handle the case of groups of people merged into single objects. Eventually, a maximum-a-posteriori approach estimates the best label assignment for the new object. Comparisons with other methods based on homography and extensive outdoor experiments demonstrate that the proposed approach is accurate and robust in coping with segmentation errors and in disambiguating groups.

Index Terms—Consistent labeling, multicamera video surveillance, epipolar geometry.

1 Introduction

DISTRIBUTED surveillance systems exploit multiple video streams to enhance observation capabilities in detecting objects and events in the scene. In most cases, the attention of the surveillance system is focused on people moving in the scene, aiming to track their movements, reconstruct their trajectories, and extract all the possible visual information about such targets. Multiple cameras provide a wider coverage of the scene and redundant data that help solve occlusions and improve accuracy. To be completely effective, the use of multiple cameras requires keeping consistent association among the different views of single persons by assigning the same label to different instances of the same person acquired from different cameras. This task, known as *consistent labeling* [15], can be very challenging when cameras are uncalibrated because groups cannot be easily disambiguated from single individuals.

This paper presents a novel approach to consistent labeling for distributed surveillance applications. The system has been designed to deal with people occluding each other and groups of people and is focused on cameras with overlapped fields of view (FOV), hereinafter called "*overlapped cameras*." We call this approach `HECOL` (`Homography and Epipolar-based COnsistent Labeling`).

The key contribution of this work is the exploitation of Bayesian statistics to solve the consistent labeling problem for both single individuals and groups in overlapped cameras. The devised observation model is based on the vertical axis of the detected object, suitably warped on the overlapped views. Warping is performed using only the geometrical relationship between image planes and avoiding the need to perform a full camera calibration.

First, an offline training process computes the camera relationship, the ground-plane homography, and the epipolar geometry (Section 3). Subsequently, in the online process, when a new object is detected ("*detection event*"), the consistent labeling is solved (Section 4). Specifically, upon each detection of a new object, hypotheses for potential matching objects in the other cameras are established. Each of the hypotheses is evaluated using a prior and likelihood value. The prior accounts for the positions of the potential matching objects in their respective camera views, while the likelihood is computed by warping the vertical axis of the new object on the FOV of the other cameras and computing the amount of match therein. The likelihood considers two contributions (*forward* and *backward*) for correctly handling both the cases of single individuals and groups. Eventually, a maximum-a-posteriori (MAP) approach estimates the best label assignment for the new object.

The proposed approach leads to an accuracy of 100 percent in the case of a single, correctly segmented person entering the scene and to very good results (96-100 percent accuracy) also in the presence of groups and segmentation errors (Section 5).

2 Related Works

Tracking multiple people in complex scenarios by using distributed cameras is a challenging task. In particular, disambiguation of multiple individuals in a group is one of the most critical points for correct tracking. In fact, no effective solutions have been proposed yet for the case of multiple people standing or walking very close to one another. Most single-view approaches rely on the tracking history prior to the group formation for detecting people grouping in the scene and, subsequently, correctly reacquiring separate targets after a possible splitting event. This task can be accomplished by either observing target trajectories as in [20] or focusing the attention on target appearance [13]. However, such methods cannot deal with people entering the scene already as a group. In order to solve this problem, projection histograms [8] 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 limited tilt angles.

However, as mentioned above, multiple consistent views may be exploited to disambiguate groups of people and also correct certain segmentation errors, but this approach becomes effective only if data from different views are correlated and made consistent. Consistent labeling from multiple cameras can be generally classified into three main categories: appearance-based, geometry-based, and mixed approach. *Appearance-based* methods base the matching essentially on the colors of the objects by exploiting a number of metrics computed from people's color histogram (e.g., [22]). The use of color information alone can lead to errors when camera sensors acquire images under different light conditions. To reinforce color-based matching, other visual features are used. For instance, in [21] and [26], matching reliability is improved by measuring face similarity. These methods work only if the camera's FOV assures a frontal view of the people's faces with a sufficient resolution. *Geometry-based* approaches exploit purely geometrical constraints between different views. With use of calibration, the relationships between overlapped cameras can be modeled in the 3D space and warping techniques can be applied with high accuracy, as proposed, e.g., in [19], [27], [28]. Conversely, a completely uncalibrated approach based on the image projections of overlapped cameras' field of view lines was first proposed by Khan and Shah in [15]. In that work, neither the problem of the disambiguation of groups nor that of simultaneous detections of new objects are addressed. Stauffer and Tieu in [24] proposed a method for building a graph representing the topology of a network of overlapped cameras directly from the tracking data. *Mixed* approaches combine both geometry and visual appearance. Different techniques are adopted to fuse information based on probabilistic

- *S. Calderara and R. Cucchiara are with the Dipartimento di Ingegneria dell'Informazione, University of Modena and Reggio Emilia, Via Vignolese, 905, 41100 Modena-Italy. E-mail: {simone.calderara, rita.cucchiara}@unimore.it.*
- *A. Prati is with the Dipartimento di Scienze e Metodi dell'Ingegneria, University of Modena and Reggio Emilia, Via Amendola, 2-Pad. Morselli, 42100 Reggio Emilia-Italy. E-mail: andrea.prati@unimore.it.*

Manuscript received 24 Apr. 2007; revised 31 Aug. 2007; accepted 1 Oct. 2007; published online 13 Dec. 2007.
Recommended for acceptance by H. Sawhney.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org and reference IEEECS Log Number TPAMI-2007-04-0234.

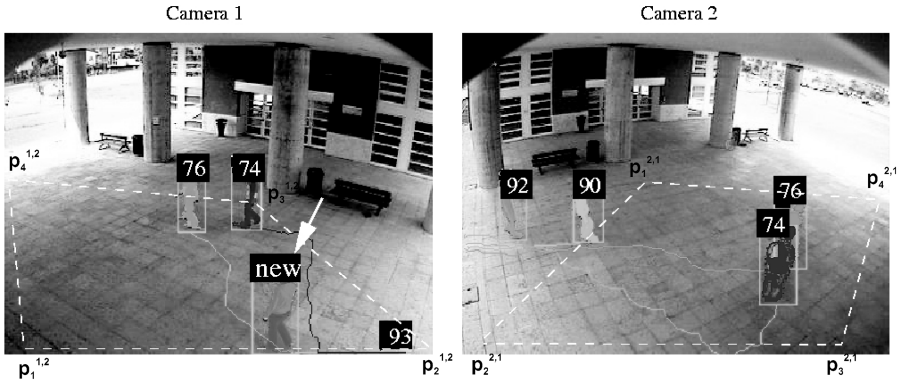


Fig. 1. An example of two overlapped cameras. The entry edges of FOV lines and the overlapping zone boundaries are shown. The new object τ is detected in C^1 (see the arrow). In camera C^2, three objects with labels 74, 76, and 90 are present in the overlapping zone. The hypothesis space Γ is obtained using the CTG constraints and contains $\{\{\tau_{74}\},\{\tau_{76}\},\{\tau_{90}\},\{\tau_{74},\tau_{76}\},\{\tau_{74},\tau_{90}\},\{\tau_{76},\tau_{90}\},\{\tau_{74},\tau_{76},\tau_{90}\}\}$, i.e., hypotheses of both single object and groups.

information fusion [12] or on Bayesian Belief Networks (BBN) [3], [6]. Dockstader and Teklap, for instance, in [6] use a Bayesian network to simultaneously perform spatial and temporal data fusion from multiple cameras by exploiting two Kalman-based predictors, one operating on image plane data and the other working on observation in a 3D Cartesian space. In [3], 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 inadequate in the case of segmentation errors and does not address the problem of groups of people. Color histograms are used by Krumm et al. in [16] together with stereo matching techniques. To overcome color delocalization problems, mixed color-spatial representations were proposed such as the correlograms used by Jiang et al. in [11] or the polar representation proposed by Kang et al. in [12]. Although these representations can be very suitable for the shape of a single person, they cannot be adopted in the case of groups of people.

The Bayesian framework has been widely adopted to statistically combine data acquired by multiple sensors [23]. In a network of disjointed cameras, Bayes rule has been applied to provide a globally optimal data association based on the objects' appearances and tracks in both traffic [9] and people surveillance [10]. In [23], Pasula et al. stress the fact that the data association has to be performed globally, by exhaustively considering all the hypotheses of association between any two objects in any camera pairs. However, the global hypothesis space is a high-dimensional space and finding the best hypotheses may be computationally unfeasible for real-time applications. In order to reduce the complexity of the search, Kettnaker and Zabih [14] propose the use of a linear programming technique. In the approach presented in this paper, instead, the high-dimensionality of the hypotheses space is dramatically reduced by enforcing local consistency by data warping between the overlapped cameras.

3 Offline Geometry Recovery

The HECOL approach is based on two separate processes. The first is an offline training process computing homography and epipolar geometry. The second is the real-time consistent labeling and will be detailed in the next section.

Let us consider a system composed of a set $\mathbf{C} = \{C^1, C^2, \ldots, C^n\}$ of n uncalibrated cameras with each camera C^i overlapped with at least another camera C^j. We aim to detect overlapping zones between cameras to compute a reliable ground-plane homography without the need of a complete camera calibration. Manually defining the overlapping zone could be possible, but is a tedious task which may lead to imprecision, thus an automatic approach is preferable. This could be done by using people trajectories [18] even when the camera streams are not synchronized [25].

In our system, both the overlapping zones and the epipoles are automatically extracted with a training procedure iterated for each pair of overlapped cameras with a single person moving around the scene. In such a simple scenario, the lower support point $\mathbf{lp}$ and the upper support point $\mathbf{up}$ of the detected moving person can be precisely extracted and their correspondences between two views can be easily established. These points are computed as the middle point of the bottom of the bounding box and the highest point of the shape, respectively. Moreover, the point $\mathbf{lp}$ approximates the contact with the ground plane.

In order to smooth the possible effects of the approximation in the numerical precision, the pairs of lower support points on the two cameras are not directly used to compute the homography. Ideally, the physical points of correspondence should be as distant as possible from each other in the ground plane and noncollinear. If the whole overlapping zone between two overlapped cameras could be known, the corners of this area would satisfy both these criteria. For this reason, we exploit a large number of pairs of $\mathbf{lp}$ points collected from each of the image borders to identify the lines delimiting the overlapped zones with a Least Square Optimization (LSQ) that minimizes the mean square error (MSE)

$$MSE = \sum_{k=1}^{n} (y_k - bx_k - a)^2, \tag{1}$$

where (x_k, y_k) are the coordinates of $\mathbf{lp}$ computed when the person enters in the image scene. This procedure is repeated for both C^i and C^j.

Given C^1 and C^2 two overlapped cameras, let us call *Entry Edges of Field of View* the obtained lines. From these lines, the overlapping zones $Z^{1,2}$ and $Z^{2,1}$ can be easily computed. The four corners of each of these zones define a set of four points in the ground plane $z = 0$ which are sufficient to compute the homography matrix H from camera C^1 to camera C^2 (whereas, H^{-1} indicates the homography matrix from C^2 to C^1). These points are reported in Fig. 1. Using *Singular Value Decomposition* (SVD), the H matrix can be computed as $\mathbf{p}_k^{1,2} = H\,\mathbf{p}_k^{2,1}$ with $k = 1, \ldots, 4$. Differently from [15], the corresponding $\mathbf{lp}$ are matched not when the object is first detected but only after it has entered the scene completely. This creates smaller but more precise overlapping zones.

During the same training phase, epipolar geometry is also recovered. Denoting as M_k a point in world 3D reference frame, let us call $\mathbf{m}_k^1$ and $\mathbf{m}_k^2$ its projections in the FOV of the two cameras C^1 and C^2, respectively. Epipolar constraints ensure that given $\mathbf{m}_k^1$ then $\mathbf{m}_k^2$ must lie on a line obtained as the projection on C^2 of the 3D-line $\langle c^1, M_k \rangle$ passing through the optical center c_1 of C^1 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.

Points $\mathbf{up}$ do not lay on the ground plane and thus are exploited for epipole computation. To compute epipole location using a single plane, the parallax property of projective images is used: Given the 3D point M_k not lying on the ground plane and its projection $\mathbf{m}_k^1$ on C^1, it is possible to find two correspondences in the image plane of C^2. The former is the real projection of M_k on C^2, $\mathbf{m}_k^2$, while the latter is the point in C^2 computed through the homographic transformation H supposing that $\mathbf{m}_k^1$ lies on the ground plane on camera C^1. The line computed from these points must be an epipolar line since it passes through the correspondences of the same point on the image plane of C^1. Given at least two lines, the epipole can be computed by means of LSQ as the intersection of such lines. This method is mathematically correct but the use of LSQ with upper support points can be strongly unstable, as shown also in [17]. Therefore, we apply the RANSAC technique [7] which has a high computational cost, due to its iterative nature, but is performed offline and does not affect the runtime system performance at all.

4 Bayesian-Competitive Consistency Labeling

After this initial training phase, consistent labeling is established. The system works independently of the adopted technique for detection and tracking of moving objects, with the only requirement that the shape of the moving object is extracted almost entirely.

Let us define with τ a new object appeared on a given camera at detection event. The multicamera system must check whether τ corresponds to a completely new object or to one which is already present in the FOV of other cameras. Moreover, the system should deal with groups and identify the objects composing them. The exhaustive search may be very computationally-expensive if there are many cameras and many objects. Thus, a graph model (called *Camera Transition Graph* (CTG)) has been defined in order to reduce the search space of multicamera matching, similarly to the proposal in [24]. The CTG is an undirected graph where each node corresponds to a camera and each arc indicates the presence of an overlapping zone between the corresponding two cameras. The set of objects detected and tracked in a camera at that time is associated to the corresponding node. By parsing the graph, a new detected object τ is associated with the subset of K potential matching objects satisfying the camera topology constraints. These K objects are combined to form the hypothesis space Γ that contains all the $(2^K - 1)$ possible subsets, with both single object and groups.

For the sake of clarity, let us suppose we have only two overlapped cameras C^1 and C^2, as shown in the example reported in Fig. 1, where the corresponding Γ set is described in the figure caption. The extension to a number of N cameras overlapped pairwise is straightforward and requires few modifications that will be discussed later in this section.

After the creation of the hypothesis space Γ, a MAP estimator is adopted to find the most probable hypothesis γ_i being:

$$i = \arg\max_k(p(\gamma_k \mid \tau)) = \arg\max_k(p(\tau \mid \gamma_k)\, p(\gamma_k)). \tag{2}$$

To evaluate the maximum posteriori the prior of each hypothesis γ_k and the likelihood of the new object τ given the hypothesis must be computed. The next two sections will detail the computation of the prior and likelihood, respectively.

4.1 Prior Computation

The prior of a given hypothesis γ_k is not computed by means of a specific pdf, but is heuristically evaluated by assigning a value proportional to a score σ_k. Obviously, since this contribution is a priori of the new object, no information about τ must be used.

The score σ_k accounts for the distance between objects calculated after the homographic warping. Let us suppose the new object τ appears on camera C^1. The lower support point ($\mathbf{lp}$) of each of the K objects in C^2 is warped to the image plane of C^1. Cluttering or isolation of warped $\mathbf{lp}$ points are considered as a discriminating element as follows: A hypothesis consisting of a single object will gain higher prior if the warped $\mathbf{lp}$ 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 that compose the hypothesis are close to each other after the warping, and, at the same time, the whole group is far from other objects.

The score σ_k of the hypothesis γ_k is computed as the difference of two contributions, the first accounting for the distances among the objects within γ_k (*within-hypothesis distance*) and the second for the distances from the remaining objects (*between-hypotheses distance*). Specifically, the within-hypothesis distance is computed as follows:

$$Wd = \max_{\{\tau_a,\tau_b\}\in\gamma_k} \left\| \left(H^{-1}\mathbf{lp_a}\right) \times \left(H^{-1}\mathbf{lp_b}\right) \right\|, \tag{3}$$

where H is the ground-plane homography matrix from C^1 to C^2, and the between-hypotheses distance as:

$$Bd = \min_{\tau_a\in\gamma_k,\tau_b\in\Gamma-\{\gamma_k\},a\neq b} \left\| \left(H^{-1}\mathbf{lp_a}\right) \times \left(H^{-1}\mathbf{lp_b}\right) \right\|. \tag{4}$$

The score is computed as $\sigma_k = Bd - Wd$ and the prior assigned proportional to it. Since Γ is a partition of the complete hypotheses set, priors must be normalized in order to sum up to one.

As an example, referring to Fig. 1, the group composed by objects #74 and #76 will gain higher prior than the one composed by the objects #74 and #90.

4.2 Likelihood Computation

Likelihood is computed by testing the fitness of each hypothesis against current evidence. The main goal is to distinguish between single hypothesis, group hypotheses, and possible segmentation errors.

We propose to exploit only geometrical properties in order to avoid the uncertainties due to color variation and to adopt the vertical axis of the object as an invariant feature. In the ideal case, the axis of the object does not change if is taken on the image plane and warped in the desired camera's image plane.

The axis of the object τ can be warped correctly with only the homography matrix and the knowledge of epipolar constraints among cameras. In the absence of tilt angle, i.e., when camera's retinal plane is orthogonal to the ground plane, the inertial axis of a physically vertical object appears as a segment of a straight line parallel to the lateral border of the image. However, in a more general situation, the projection of the axis in the image plane is generally not perfectly vertical (see, for instance, Fig. 2). Conversely, considering people in a standing position, all the 3D principal inertial axes are orthogonal to the ground plane and their projection on each respective camera image plane will intersect at the camera's vanishing point $\mathbf{vp}$.

The vertical vanishing point (computed by a robust technique as described in [1]) is then used to obtain warped axis inclination (Fig. 2). The lower support point $\mathbf{lp}$ of τ is projected on camera C^2 by using the homography matrix. The corresponding point on the image plane of camera C^2 is denoted as $\mathbf{a_1} = H\mathbf{lp}$, where H is the previously defined homography matrix from C^1 to C^2. The warped axis will lie on a straight line passing through $\mathbf{vp^2}$ and $\mathbf{a_1}$ (Fig. 2d). The end point of the warped axis is computed by using the upper support point $\mathbf{up}$. Since this point does not lie on the ground plane, its projection on the image of camera C^2 does not correspond to the actual upper support point; however, the projected point lies on the epipolar line. Consequently, the axis' ending point $\mathbf{a_2}$ is obtained as the intersection between the epipolar line $\langle \mathbf{e^2}, H\mathbf{up}\rangle$ and line $\langle \mathbf{vp}^2, H\mathbf{lp}\rangle$ passing through the axis.

Based on geometrical constraints, the warped axis $\langle \mathbf{a_1}, \mathbf{a_2}\rangle$ of τ in the image plane of C^2 is univocally identified but its computation is not error free. Errors may be generated by approximations in homography and epipole computation. Nonnegligible errors may also derive from the simplifications inherent to the computation of axis limits $\mathbf{lp}$ and $\mathbf{up}$ (e.g., the presence of undetected shadows could result in an incorrect localization of $\mathbf{lp}$). Moreover, in the case

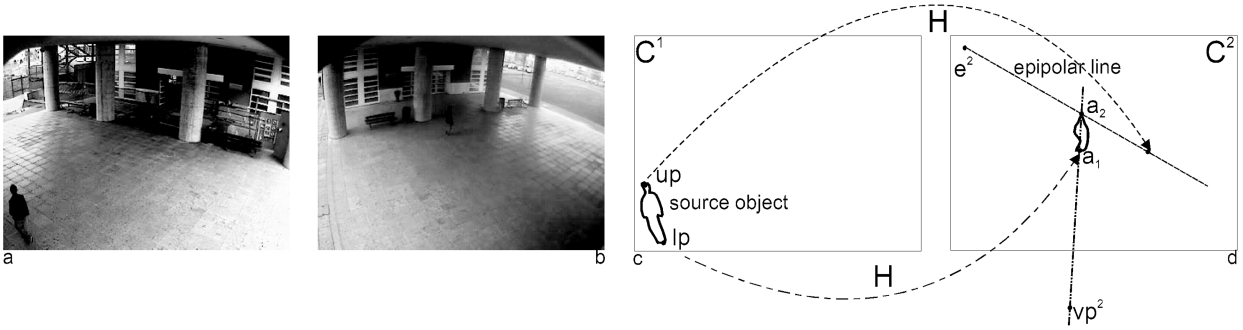


Fig. 2. Example of exploiting vanishing point and epipolar geometry to warp the axis of the object τ to the image plane of camera C^1.

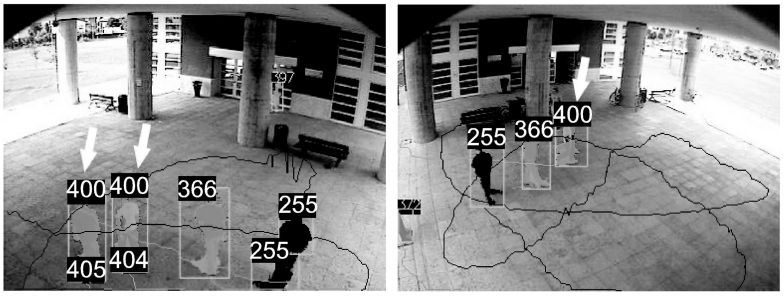


Fig. 3. Examples of "group fully within FOV" (τ_{400}) and segmentation error (τ_{255}). (a) C^1 at frame #459 of video GV1. (b) C^2 at frame #459 of video GV2.

of groups of people, the **lp** and **up** points of a foreground blob including more people do not correspond to the limits of any physically vertical axis.

In order to improve the robustness to these errors, we account also for the dual process that can be performed for each of the K potential matching objects: The axis of the object in C^2 is warped on the segment $\langle \mathbf{a_1}, \mathbf{a_2} \rangle$ on camera C^1. The measure of axis correspondence is not merely the distance between axes $\langle \mathbf{a_1}, \mathbf{a_2} \rangle$ and $\langle \mathbf{lp}, \mathbf{up} \rangle$; it is defined as the number of matching pixels between the warped axis and the foreground blob of the target object. This makes it easier to define a normalized value for quantifying the matching. Accordingly, the fitness measure $\varphi_{\tau_a \to \tau_b}$ from the object τ_a in a generic camera C^i to τ_b in a generic camera C^j is defined as the number of pixels resulting from the intersection between the warped axis and the foreground blob of τ_b normalized by the length (in pixels) of the warped axis itself. The reversed fitness measure $\varphi_{\tau_b \to \tau_a}$ is computed similarly by reverting the warping order. In the ideal case of correspondence between τ_a and τ_b, $\varphi_{\tau_a \to \tau_b} = \varphi_{\tau_b \to \tau_a} = 1$. However, in the case of errors in the **lp** and **up** computation, the warped axis could fall partially outside of the foreground blob, lowering the fitness measure.

In the likelihood definition, we refer to *forward* contribution when the fitness is calculated from the image plane in which the new object appears (camera C^1) to the image plane of the considered hypothesis (camera C^2). Thus, generalizing for hypotheses containing more than one object (group hypotheses), forward axis correspondence can be evaluated by computing the fitness of the new object τ with all the objects composing the given hypothesis γ_k for camera C^2

$$fp_{forward}(\tau|\gamma_k) = \frac{\sum_{\tau_m \in \gamma_k} \varphi_{\tau \to \tau_m}}{K \cdot S_f}. \tag{5}$$

S_f measures the maximum range of variability of the forward fitness measure of the objects inside the given hypothesis

$$S_f = \max_{\tau_m \in \gamma_k}(\varphi_{\tau \to \tau_m}) - \min_{\tau_n \in \gamma_k}(\varphi_{\tau \to \tau_n}). \tag{6}$$

The use of the normalizing factor K (i.e., the number of potential matching objects on C^2) weighs each hypothesis according to the presence or absence of objects in the whole scene.

Backward contribution is computed similarly from the hypotheses space to the observed object

$$fp_{backward}(\tau|\gamma_k) = \frac{\sum_{\tau_m \in \gamma_k} \varphi_{\tau_m \to \tau}}{K \cdot S_b}, \tag{7}$$

where S_b is defined as

$$S_b = \max_{\tau_m \in \gamma_k}(\varphi_{\tau_m \to \tau}) - \min_{\tau_n \in \gamma_k}(\varphi_{\tau_n \to \tau}). \tag{8}$$

At the end, the likelihood is defined as $p(\tau \,|\, \gamma_k) = \max(fp_{backward}, fp_{forward})$. The use of the maximum value ensures that 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. In fact, it can handle three different cases: the "*group fully within FOV*" case, the "*group entering FOV*" case, and the case of *segmentation errors*.

The first is the case in which a group is already inside the scene while the group's components appear one at a time in the FOV of another camera. In this particular situation, forward contribution is useful to distinguish people in a group which had been previously detected as single objects. After the warping, each axis of the new objects will intersect the same group's foreground blob on the chosen camera. Therefore, the new objects are labeled as belonging to the existing group. An example is reported in Figs. 3a and 3b, where the group labeled 400 in the image of camera C^2 is detected as two separate people (labels #404 and #405) in the image of camera C^1. The two objects are evaluated separately and for each of them the warped axis falls within the blob of the object #400 on camera C^2, with a forward contribution very close to one. The backward contribution, on the other hand, can be low since the axis of the object labeled 400 in C^2 does not precisely correspond to the axis of any person composing the group and its warping on C^1 will be imprecise.

The second case is when, as in Fig. 4a (magnified in Fig. 4b), two people appear in a new camera detected as a single blob. The group disambiguation can be solved by exploiting the fact that in

Fig. 4. Example of "group entering FOV." (a) C^2 at frame #432 of video GV1. (b) Zoom of (a). (c) C^1 at frame #432 of video GV1. (d) Zoom of (c).

the other camera the two objects are detected as separated (Figs. 4c and 4b). The forward likelihood of the new object is warped on the image plane of camera C^1, obtaining the dotted line in Fig. 4d. This line intersects only object τ_{111}, therefore the forward likelihood is zero for object τ_{110}. The backward likelihood, on the other hand, is obtained by warping the solid axes of both people of Fig. 4d back on the camera C^2 and obtaining the dotted lines in Fig. 4b. As a consequence the resulting maximum likelihood is assigned to the hypothesis $\{\tau_{110}, \tau_{111}\}$ that is higher than those assigned to the separate hypotheses $\{\tau_{110}\}$ and $\{\tau_{111}\}$.

Backward contribution is also useful to solve the case of *segmentation errors*, in which a person has been erroneously extracted by the object detection system as two separate objects, but a full view of the person exists from the past in an overlapped camera. This is the case of object #255 in Fig. 3a. This object enters the FOV of camera C^1 and is detected as split into two separate objects. As shown in Fig. 3b, a full view of the person exists on camera C^2. In this particular case, both the forward and backward contributions are suitable to solve the consistent labeling problem for the lower part of the object on C^1, while in the case of the upper part of the object, forward contribution will be extremely low since the **lp** of the object does not lie on the ground plane. The backward contribution, on the other hand, will be high and very close to one. Both contributions cooperate in correctly assigning label #255 to both extracted parts of the object on C^1.

4.3 Extension to N Cameras

The previous sections have focused on the case of a system composed of only two cameras, C^1 and C^2. The same approach still holds when N pairwise overlapped cameras are present. In fact, global system consistency is ensured by the transitive property of subsequent assignments on different cameras.

When more than two cameras overlap simultaneously it is possible to take into account more information than in the pairwise case. To account for this situation, the proposed approach is suitably modified by adding an additional step that selects the best assignment from all the possible hypotheses coming from each camera.

In detail, when a detection event occurs on C^1, for each camera C^j overlapped with C^1 (i.e., nodes connected to the node of C^1 in the CTG) the best local assignment hypothesis is chosen using the maximum-a-posteriori framework. A second MAP estimator detects the most probable among these hypotheses. In complex scenes more hypotheses could have similar a posteriori probability but it may exist a particular view where the hypothesis assignment is easier. The second MAP stage has the purpose to choose this view. This can be easily done using the previously computed posteriors and Bayes rule

$$p(C^j \mid \tau) \propto p(\tau \mid C^j) = \max_{\gamma_k \in \Gamma} p(\gamma_k \mid \tau). \quad (9)$$

The camera posterior is evaluated for each camera C^j that overlaps with C^1 assuming that all the views of overlapped cameras are equally probable. Eventually, the label is assigned to the new object according to the winning hypothesis on the winning camera. If the chosen hypothesis identifies a group, all the labels of objects composing the group are assigned as identifiers, as shown in Figs. 3a and 3b.

5 Experimental Results

The system being discussed has been tested on a test bed created at our campus and consists of four cameras (three fixed CCD cameras and a PTZ camera). The testing area is particularly challenging since it is partially covered by a porch. Hence shadows, illumination variations, and fixed objects (e.g., benches and columns) make segmentation and tracking processes nontrivial.

Tests have been performed considering several video sequences acquired during ordinary working days in different environmental conditions for two or three partially overlapped cameras. A total of about 90 minutes of video footage has been evaluated with ground truths. In acquiring the videos, no constraints have been imposed on people's trajectories in order to stress the system. 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. Object detection is based on background subtraction and shadow removal as proposed in the Sakbot system described in [4]. Object tracking from each camera is based on probabilistic, appearance-based tracking as described in [5]. Then, when a detection event occurs on a camera, the corresponding thread sends a synchronization event to the other threads and the consistent labeling is established.

For a quantitative measure, six videos of adequate duration have been collected. The first video (labeled SHV—Single Handoff Video—12,600 frames) contains only camera handoffs of one person at a time, but from several people of different appearance and different moving behavior. The video identified by SEV (Segmentation Error Video—16,200 frames), instead, contains several segmentation errors due to the poor illumination conditions and the appearance of the people in the scene not contrasting with that of the background. Such segmentation errors result in people being not completely detected or split in more parts, as shown in Fig. 3a. Videos GV1 (Group Video—9,900 frames) and GV2 (14,400 frames) include several situations in which groups of people are formed. We distinguished between two possible situations: "*group entering FOV*" and "*group fully within FOV*," as defined in Section 4.2. Video GV1 contains both these situations, while video GV2 only contains the first one. Additionally, video GV2 also contains the so-called *simultaneous handoffs*, i.e., the case in which more people enter simultaneously the overlapping zone from the same entry line (Figs. 4a and 4c). Finally, videos MV1 (Mixed Video—22,500 frames) and MV2 (9,000 frames) report a mixture of all these situations and are denoted by a high degree of complexity.

Table 1 reports the achieved results for the various aforementioned cases. The last column reports the overall accuracy of the system. Proceeding from top to bottom, the table first shows the results with homography only (no epipole computation) and then with a Bayesian classifier with only forward contribution. This method is similar to the approach used in [2], which essentially bases the object matching process on the support points' distances measured on the intercamera ground plane homographic mosaic.

TABLE 1
This Table Reports the Achieved Results for the Various Experiments

	# Single HO		# Segm.err.		# Grp enter.		# Grp within		# Simul. HO		Overall
	Total	Correct	Total	Correct	Total	Correct	Total	Correct	Total	Correct	Accuracy
Only Homography											
SHV	137	128	0	0	0	0	0	0	0	0	93.43%
SEV	28	28	76	30	0	0	7	7	0	0	58.56%
GV1	30	29	0	0	15	5	40	39	0	0	85.88%
GV2	78	73	0	0	35	16	0	0	19	18	81.06%
MV1	39	34	21	10	30	15	18	14	14	13	70.49%
MV2	52	50	38	16	30	15	26	24	18	18	75.00%
Avg.	342/364 (93.96%)		56/135 (41.48%)		51/110 (46.36%)		84/91 (92.31%)		49/51 (96.08%)		77.40%
Forward Contribution Only											
SHV	137	135	0	0	0	0	0	0	0	0	98.54%
SEV	28	28	76	34	0	0	7	7	0	0	62.16%
GV1	30	30	0	0	15	6	40	40	0	0	89.41%
GV2	78	77	0	0	35	16	0	0	19	18	84.09%
MV1	39	35	21	10	30	15	18	15	14	13	72.13%
MV2	52	51	38	19	30	15	26	24	18	18	77.43%
Avg.	356/364 (97.80%)		63/135 (46.67%)		52/110 (47.27%)		86/91 (94.51%)		49/51 (96.08%)		80.63%
Backward Contribution Only											
SHV	137	131	0	0	0	0	0	0	0	0	95.62%
SEV	28	28	76	65	0	0	7	3	0	0	86.49%
GV1	30	30	0	0	15	15	40	18	0	0	74.12%
GV2	78	76	0	0	35	35	0	0	19	18	97.73%
MV1	39	36	21	16	30	30	18	7	14	13	83.61%
MV2	52	52	38	35	30	30	26	12	18	17	89.02%
Avg.	353/364 (96.98%)		116/135 (85.93%)		110/110 (100.00%)		40/91 (43.96%)		48/51 (94.12%)		87.76%
HECOL											
SHV	137	137	0	0	0	0	0	0	0	0	**100.00%**
SEV	28	28	76	75	0	0	7	7	0	0	**99.10%**
GV1	30	30	0	0	15	15	40	40	0	0	**100.00%**
GV2	78	78	0	0	35	35	0	0	19	18	**99.24%**
MV1	39	39	21	19	30	30	18	16	14	14	**96.72%**
MV2	52	52	38	36	30	30	26	24	18	18	**97.56%**
Avg.	364/364 (**100.00%**)		130/135 (**96.30%**)		110/110 (**100.00%**)		87/91 (**95.60%**)		50/51 (**98.04%**)		**98.77%**

The last column reports the overall accuracy of the system. Proceeding from top to bottom, the table first shows the results with homography only (no epipole computation), then with a Bayesian classifier with only forward contribution, then with the backward contribution only and, finally, with the complete Bayesian-competitive labeling.

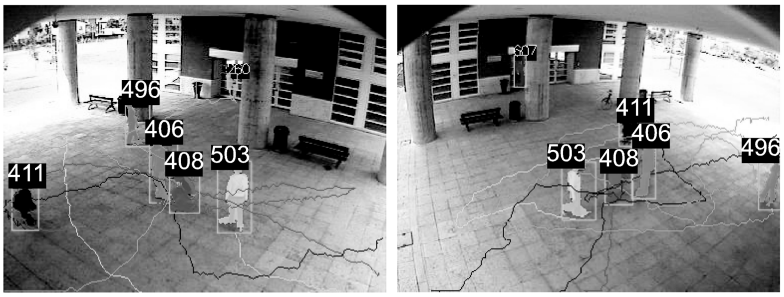
The rest of the table reports the accuracy with the backward contribution only and the complete Bayesian-competitive labeling. A visual summary of these results organized by category is shown in Fig. 6.

When counting correct label assignments in the case of segmentation errors, each separate object obtained from the over-segmentation of a person has been evaluated independently, and considered correct if labeled consistently. The low average accuracy (41.48 percent) of the pure homography-based method is due to the incorrect warping of $\mathbf{lp}$ in the case of segmentation errors. The basic method based only on the homography proves good results only for simple, single camera handoffs (about 94 percent of accuracy).

The use of forward and backward likelihood contributions can overcome the effect of many segmentation errors. In Fig. 3a (frame #459 of video GV1), the person labeled with 255 is split into two different objects in C^1. The proposed method recovers from this segmentation issue assigning the same label to both parts. The average accuracy increases to 46.67 percent, 85.93 percent, and 96.30 percent by using, respectively, only backward contribution, only forward contribution, and both contributions.

In this paper, group disambiguation has been achieved by using a warping technique across overlapped views. Thanks to the opportunistic placement of cameras, the proposed method is statistically successful. Results on group disambiguation (columns 6-9 in Table 1) 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 "group entering FOV" case, doubling the average accuracy from less than 50 percent to a 100 percent of accuracy. Given the implicit duality of the problem, in the "group fully within FOV" case the most important contribution is the forward one, capable of increasing from 43.96 percent of average accuracy (using only backward contribution) to 95.60 percent using full MAP estimation. Snapshots in Figs. 5a and 5b show the output of the system in complex situations, where many people are moving at the same time and significantly occluding each other in one of the views.

The overall performance has also been evaluated 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 (interprocess-communication) to communicate data and events from/to consistent labeling process. A working prototype is currently in operation in a public park in Reggio Emilia, Italy, for real-time surveillance of people in crowded environments.



(a) (b)

Fig. 5. Examples of system working on actual cases with two cameras. (a) C^1 at frame #4294 of video MV1. (b) C^2 at frame #4294 of video MV1.

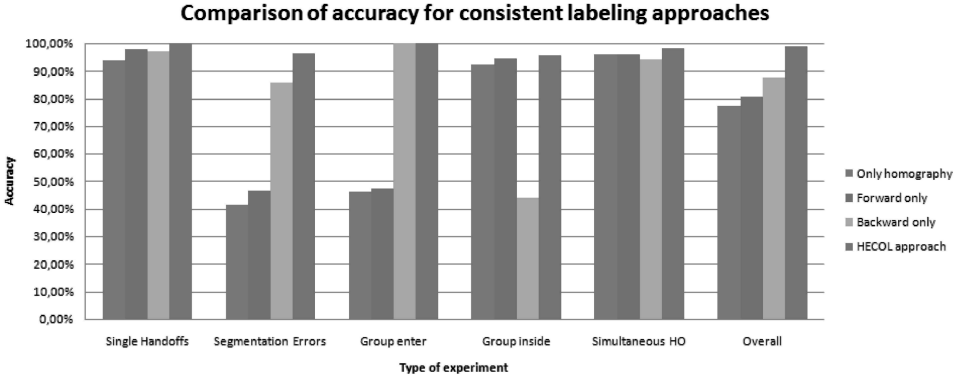


Fig. 6. Summary of the experimental results.

ACKNOWLEDGMENTS

The authors would like to thank Professor Massimo Piccardi for the support in revising the paper. 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).

REFERENCES

[1] C. Brauer-Burchardt and K. Voss, "Robust Vanishing Point Determination in Noisy Images," *Proc. Int'l Conf. Pattern Recognition,* vol. 1, pp. 559-562, 2000.
[2] S. Calderara, R. Vezzani, A. Prati, and R. Cucchiara, "Entry Edge of Field of View for Multi-Camera Tracking in Distributed Video Surveillance," *Proc. IEEE Int'l Conf. Advanced Video and Signal-Based Surveillance,* pp. 93-98, 2005.
[3] S. Chang and T.-H. Gong, "Tracking Multiple People with a Multi-Camera System," *Proc. IEEE Workshop Multi-Object Tracking,* pp. 19-26, 2001.
[4] R. Cucchiara, C. Grana, M. Piccardi, and A. Prati, "Detecting Moving Objects, Ghosts and Shadows in Video Streams," *IEEE Trans. Pattern Anaylsis and Machine Intelligence,* vol. 25, no. 10, pp. 1337-1342, Oct. 2003.
[5] R. Cucchiara, C. Grana, G. Tardini, and R. Vezzani, "Probabilistic People Tracking for Occlusion Handling," *Proc. Int'l Conf. Pattern Recognition,* vol. 1, pp. 132-135, Aug. 2004.
[6] S. Dockstader and A. Tekalp, "Multiple Camera Tracking of Interacting and Occluded Human Motion," *Proc. IEEE,* vol. 89, no. 10, pp. 1441-1455, Oct. 2001.
[7] M.A. Fischler and R.C. Bolles, "Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography," *Comm. ACM,* vol. 24, pp. 381-395, 1981.
[8] I. Haritaoglu, D. Harwood, and L. Davis, "W4: Real-Time Surveillance of People and Their Activities," *IEEE Trans. Pattern Anaylsis and Machine Intelligence,* vol. 22, no. 8, pp. 809-830, Aug. 2000.
[9] T. Huang and S.J. Russell, "Object Identification in a Bayesian Context," *Proc Int'l Joint Conf. Artificial Intelligence,* pp. 1276-1283, 1997.
[10] O. Javed, Z. Rasheed, K. Shafique, and M. Shah, "Tracking Across Multiple Cameras with Disjoint Views," *Proc. IEEE Int'l Conf. Computer Vision,* vol. 2, pp. 952-957, 2003.
[11] L. Jiang, C.S. Chua, and Y.K. Ho, "Color Based Multiple People Tracking," *Proc. IEEE Int'l Conf. Control, Automation, Robotics, and Vision,* vol. 1, pp. 309-314, 2002.
[12] J. Kang, I. Cohen, and G. Medioni, "Continuous Tracking Within and Across Camera Streams," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition,* vol. 1, pp. I-267-I-272, 2003.
[13] J. Kang, I. Cohen, and G. Medioni, "Object Reacquisition Using Invariant Appearance Model," *Proc. Int'l Conf. Pattern Recognition,* vol. 4, pp. 759-762, Aug. 2004.
[14] V. Kettnaker and R. Zabih, "Bayesian Multi-Camera Surveillance," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition,* vol. 2, pp. 253-259, June 1999.
[15] S. Khan and M. Shah, "Consistent Labeling of Tracked Objects in Multiple Cameras with Overlapping Fields of View," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 25, no. 10, pp. 1355-1360, Oct. 2003.
[16] J. Krumm, S. Harris, B. Meyers, B. Brumitt, M. Hale, and S. Shafer, "Multi-Camera Multi-Person Tracking for Easyliving," *Proc. IEEE Int'l Workshop Visual Surveillance,* pp. 3-10, 2000.
[17] Q.T. Luong, R. Deriche, O. Faugeras, and T. Papadopoulo, "On Determining the Fundamental Matrix: Analysis of Different Methods and Experimental Result," technical report, RR, INRIA, 1993.
[18] F. Lv, T. Zhao, and R. Nevatia, "Self-Calibration of a Camera from Video of a Walking Human," *Proc. Int'l Conf. Pattern Recognition,* vol. 1, pp. 562-567, Aug. 2002.
[19] A. Mittal and L. Davis, "M2tracker: A Multi-View Approach to Segmenting and Tracking People in a Cluttered Scene," *Int'l J. Computer Vision,* vol. 51, no. 3, pp. 189-203, Feb. 2003.
[20] W. Niu, J. Long, D. Han, and J. Wang, "Human Activity Detection and Recognition for Video Surveillance," *Proc. IEEE Int'l Conf. Multimedia and Expo,* vol. 1, pp. 719-722, June 2004.
[21] K. Nummiaro, E. Koller-Meier, T. Svoboda, D. Roth, and L. Van Gool, "Color-Based Object Tracking in Multi-Camera Environments," *Proc. 25th DAGM Symp. Pattern Recognition,* pp. 591-599, 2003.
[22] J. Orwell, P. Remagnino, and G. Jones, "Multi-Camera Colour Tracking," *Proc. Second IEEE Workshop Visual Surveillance,* pp. 14-21, June 1999.
[23] H. Pasula, S.J. Russell, M. Ostland, and Y. Ritov, "Tracking Many Objects with Many Sensors," *Proc. Int'l Joint Conf. Artificial Intelligence,* pp. 1160-1171, 1999.
[24] C. Stauffer and K. Tieu, "Automated Multi-Camera Planar Tracking Correspondence Modeling," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition,* vol. 1, pp. 259-266, June 2003.
[25] G. Stein, "Tracking from Multiple View Points: Self-Calibration of Space and Time," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition,* vol. 1, pp. 521-527, June 1999.
[26] M.H. Tan and S. Ranganath, "Multi-Camera People Tracking Using Bayesian Networks," *Proc. 2003 Joint Conf. Fourth Int'l Conf. Information, Communications, and Signal Processing,* vol. 3, pp. 1335-1340, 2003.
[27] Z. Yue, S. Zhou, and R. Chellappa, "Robust Two-Camera Tracking Using Homography," *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing,* vol. 3, pp. 1-4, 2004.
[28] Q. Zhou and J. Aggarwal, "Object Tracking in an Outdoor Environment Using Fusion of Features and Cameras," *Image and Vision Computing,* vol. 24, no. 11, pp. 1244-1255, Nov. 2006.