# Consistent Labeling for Multi-camera Object Tracking*

Simone Calderara[1], Andrea Prati[2], Roberto Vezzani[1], and Rita Cucchiara[1]

[1] Dipartimento di Ingegneria dell'Informazione - University of Modena and Reggio Emilia - Via Vignolese, 905 - 41100 Modena, Italy
[2] Dipartimento di Scienze e Metodi dell'Ingegneria - University of Modena and Reggio Emilia - Via Allegri, 13 - 42100 Reggio Emilia, Italy

**Abstract.** In this paper, we present a new approach to multi-camera object tracking based on the consistent labeling. An automatic and reliable procedure allows to obtain the homographic transformation between two overlapped views, without any manual calibration of the cameras. Object's positions are matched by using the homography when the object is firstly detected in one of the two views. The approach has been tested also in the case of simultaneous transitions and in the case in which people are detected as a group during the transition. Promising results are reported over a real setup of overlapped cameras.

## 1 Introduction

Crowed environments such as metro stations, city markets, or public parks, are very difficult to monitor, also for human operators. Moreover, the recent development in video acquisition hardware and the consequent relaxation of its price have made possible a broad deployment of hundreds of cameras. The availability of multiple points of view and the redundancy of information allow a more precise (though not unfailing) monitoring of complex scenes.

Despite of the complexity increase, multiple camera systems exhibit the undoubt advantages of covering wide areas and enhancing the managing of occlusions (by exploiting the different viewpoints). However, the automatic merge of the knowledge extracted from single cameras are still challenging tasks, especially in application of *distributed people tracking*. The goal is to track multiple people moving in an environment observed by multiple cameras, tightly connected, synchronized and with partially overlapped views.

The solution to this problem must deal with two sub-problems: a reliable tracking in each camera system and the preservation of the identity of the people moving from a camera's view to the one of another camera. For this second problem, often know as *consistent labeling*, many solutions have been addressed. Among these, geometrical approaches require camera calibration. In outdoor environments with many cameras, placed in high positions over poles at unknown

* This work was supported by the project L.A.I.C.A. (Laboratorio di Ambient Intelligence per una Città Amica), funded by the Regione Emilia-Romagna, Italy.

F. Roli and S. Vitulano (Eds.): ICIAP 2005, LNCS 3617, pp. 1206–1214, 2005.

distance, manual calibration could be difficult and time consuming to achieve. Thus, automatic camera calibration techniques have been proposed.

This paper focus on finding a reliable solution to the consistent labeling problem, also in challenging conditions such as the simultaneous presence of more people passing from a camera's field of view to the one of another camera. We exploit and improve a proposal of Khan-Shah [1] of edge of field of view (EOFOV) computation. The EOFOV lines, i.e. the boundaries between the field of views of partially overlapped cameras, are automatically created with a learning phase using a training video with a person walking along the limits of the field of view.

The paper presents a general model with a set of $N$ overlapped cameras, and an improved and generalized technique for EOFOV learning. Then, the consistency between the extracted lines is exploited to compute a precise homography, used to establish the consistent labeling. The paper reports extensive experimental work in which very complex situations of multiple people crossing simultaneously the border of the FOV are considered. Experiments have been provided in a real setup with pairs of partially overlapped cameras monitoring an outdoor environment.

## 2 Related Works

Approaches to multicamera tracking can be generally classified into three main categories: geometry-based, color-based, and hybrid approaches. The former exploits geometrical relations and constraints between the different views to perform the consistent labeling process. Instead, *color-based* approaches base the matching essentially on the color of the tracks. For example, in [2] a color space invariant to illumination changes is proposed and histogram-based information at low (texture) and mid (regions and blobs) level are adopted. Conversely, the work in [3] uses stereoscopic vision to match tracks, but when this matching is not sufficiently reliable, color histograms are used to solve ambiguities. Finally, *Hybrid* approaches, belonging to the third class, mix information about the geometry and the calibration with those provided by the visual appearance. These last methods are based on probabilistic information fusion [4] or on Bayesian Belief Networks (BBN) [5][6], and sometimes a learning phase is required [7].

*Geometry-based approaches* can be further subdivided into calibrated and uncalibrated approaches. In [8], each camera processes the scene and obtains a set of tracks. Then, regions along the epipolar lines in pairs of cameras are matched and the mid-points of the matched segments are back-projected in 3D and then, with an homography, onto the ground plane to identify possible positions of the person within a probability distribution map (filtered with a Gaussian kernel). The probability distribution maps are then combined using outlier-rejection techniques to yield a robust estimate of the 2D position of the objects, which is then used to track them. A particularly interesting paper is reported in [9] in which homography is exploited to solve occlusions. Single camera processing is based on particle filter and on probabilistic tracking based on appearance to detect occlusions. Once an occlusion is detected, homography

is used to estimate the track position in the occluded view, by using the last valid positions of the track in it and the current position of the track in the other view (properly warped in the occluded one by means of the transformation matrix). A very relevant example of the uncalibrated approaches is the work of Khan and Shah [1]. Their approach is based on the computation of the so-called *Edges of Field of View*, i.e. the lines delimiting the field of view of each camera and, thus, defining the overlapped regions. Through a learning procedure in which a single track moves from one view to another, an automatic procedure computes these edges that are then exploited to keep consistent labels on the objects when they pass from one camera to the adjacent.

Our approach is a suitable modification of this proposal to compute, starting from the EOFOV lines extraction, the homography relation between the two ground plane in an automatic way. This transformation is adopted for the consistent labeling problem.

## 3 The Edge of Field of View

The proposed approach belongs to the class of uncalibrated geometry-based techniques and it is a suitable modification of the proposal reported by Khan and Shah in [1]. The basic idea relies on learning the calibration parameters by means of the creation of the so-called *edges of field of view* (EOFOV).

Let us suppose that the system is composed of a set $\mathbf{C} = \{C^1, C^2, ..., C^n\}$ of $n$ cameras, with each camera $C^i$ being overlapped with at least another camera $C^j$. Projecting the limits of the field of view (FOV) of a camera $C^i$ on the ground plane ($z = 0$), the so-called *3D FOV lines* [1] can be obtained. In particular, they correspond to the intersection between the ground plane and the rectangular pyramid with its vertex at the camera optical center (the camera view frustum). A 3D FOV line is denoted by $L^{i,s}$, where $s$ indicates the equation of the line in the 2D coordinates of the camera $C^i$ that generates the 3D FOV line. In particular, the four 3D FOV lines $L^{i,s_h} \,|\, h = 1 \ldots 4$ (where $s_h$ corresponds to the image borders $x = 0$, $x = x_{max}$, $y = 0$, and $y = y_{max}$) can be computed. A projection of a 3D FOV line of camera $C^i$ may be visible in another camera $C^j$ partially overlapped with $C^i$. The FOV line (in 2D) of the line $s$ of camera $C^i$ seen by the camera $C^j$ will be then denoted with $L^{i,s}_j$ and represents one of the EOFOV lines for the camera $C^j$. Each line $L^{i,s}_j$ divides the image on camera $C^j$ into two half-planes, one overlapped with camera $C^i$ and the other disjoint. The intersection of the overlapped semi-planes defined by the EOFOV lines from camera $C^i$ to camera $C^j$ defines the overlapping area $Z^i_j$. In Fig. 1 the placement of four cameras in our campus is depicted. Fig. 2(a) shows a view of cameras $C^1$ and $C^2$ and the EOFOV $L^{1,s}_2$ is depicted with $s$ correspondent to the image border $x = 0$ of $C^1$.

The EOFOV lines are created with a training procedure; the process is iterated for each pair ($C^i$, $C^j$) of partially overlapped cameras. To this aim, we need the correspondences of a certain number of points on the ground plane in the two considered views. Thus, as proposed in [1], during the training phase a

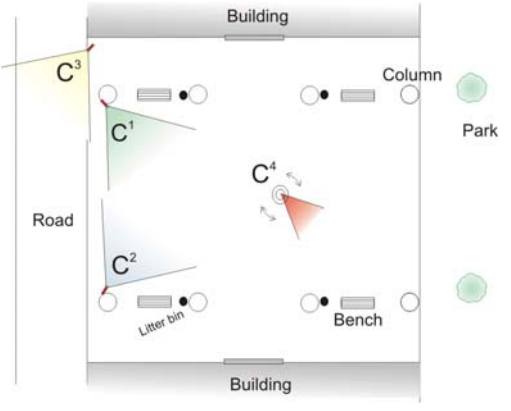



**Fig. 1.** Map of our real setup

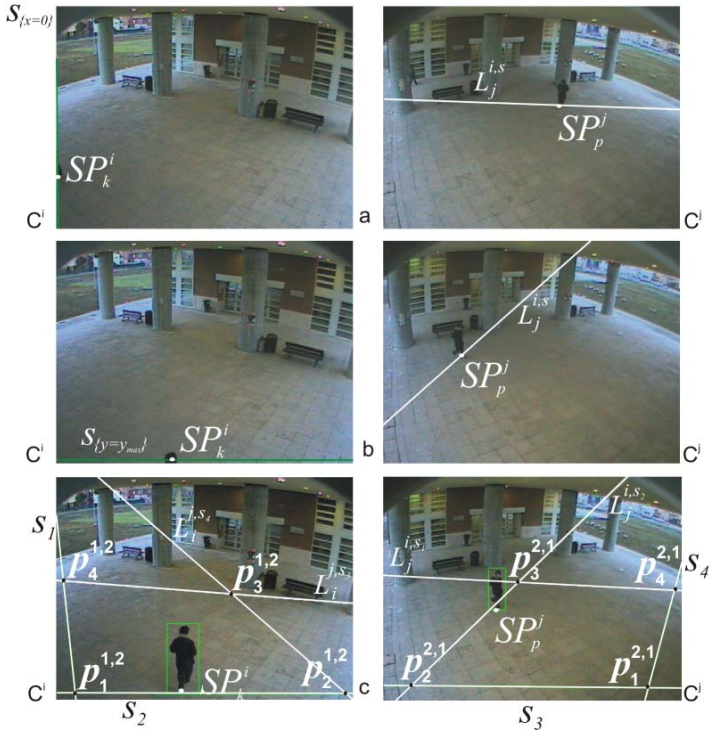



**Fig. 2.** Examples of EOFOV creation: (a) correct creation, (b) incorrect creation (using Khan-Shah approach), and (c) the proposed solution.

single person moves freely in the scene, with the minimum requirements to pass through at least two points of each limit of the FOV of two overlapped cameras.

Let us call $O_k^i$ the object segmented and tracked with label $k$ in the camera $C^i$ and $SP_k^i$ the point of contact with the ground plane (*support point*, hereinafter).

The support point can be easily computed as the middle point of the bottom of the bounding box of the blob.

Given the constraint to have a single moving person in the training video, problems of consistent labeling do not occur. Thus, when the object is detected in the $C^j$ camera too, and tracked with the $p$ label, it is directly associated to $O_k^i$. Therefore, in this moment (known as the moment of "camera handoff") the support point $SP_k^i$ can be directly associated to $SP_p^j$ (if it is visible). In this case the point $SP_k^i$ lies on the EOFOV line $L_i^{j,s}$ for the camera $C^i$. The equation of each line $L_i^{j,s}$ is computed by collecting a set of coordinates of the support point $SP_k^i$ detected at the camera handoff and exploiting a Least Square optimization (Fig. 2(a) shows $SP_k^i$ and $SP_p^j$ at the camera handoff instant).

In the method proposed in [1], the points $SP$ are extracted at the camera handoff moment. This can bring to false correspondences, as in the case of a person entering from the bottom of the image (Fig. 2(b)). In such a situation, the head in $C^i$ is in correspondence with the feet in $C^j$. However, this matching is reliable enough if the goal is only the consistent labeling at the camera handoff instant (as in [1]) and if the persons have the same height. Instead, if an exact correspondence is required, for example to compute an homography transformation, we must verify that the matching points belong to the same real point (e.g., the feet).

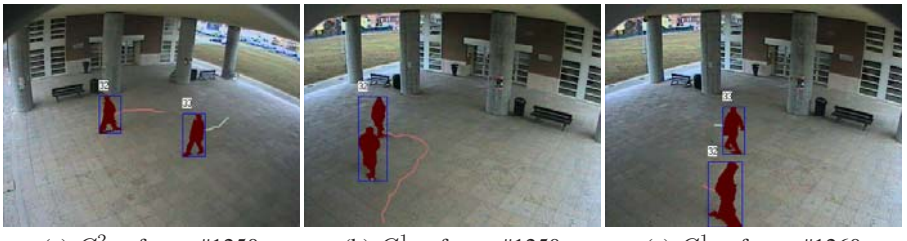
To this aim, we modified the approach by delaying the computation of the EOFOV lines to the moment in which the object is completely entered the scene of the new camera (see Fig. 2(c)). This can bring to a displacement of the line with respect to the actual limit of the image, but it assures the correct match of the feet's position in the two views. The amount of the displacement depends on the position of the feet w.r.t. the image limit. As a consequence, the actual FOV lines $s$ are neither coincident nor parallel to the image border. In Fig. 2(c) the lines $(L_j^{i,s_1}, L_j^{i,s_2})$, $(L_i^{j,s_3}, L_i^{j,s_4})$ correspondent to $(s_1, s_2)$ in $C^i$, $(s_3, s_4)$ in $C^j$ respectively are depicted.

## 4 Consistent Labeling Approach

The approach proposed in [1] establishes the consistent labeling only in the exact moment of the camera handoff from $C^i$ to $C^j$. This technique assign the object $O_p^j$ in the camera $C^j$ to the object in the camera $C^i$ with the minimum distance from the EOFOV $L_i^{j,s}$ corresponding to the side $s$ from which the object enters. This approach has two main limits: if two or more objects cross simultaneously (Fig. 3), then an incorrect labeling could be established; if two or more objects are merged from the view of $C^j$ at the moment of the camera handoff, but then they separate, the consistent labeling with the corresponding labels of $C^i$ can not be recovered.

This last limit is evident in Fig. 4: camera $C^2$ sees the two objects ($O_{32}^2$ and $O_{33}^2$) as separated (Fig. 4(a)), but they are merged in a single object when they appear in camera $C^1$ and only the label 32 is assigned to it(Fig. 4(b)).

**Fig. 3.** Example of simultaneous crossing of two objects



(a) $C^2$ at frame #1250 (b) $C^1$ at frame #1250 (c) $C^1$ at frame #1260

**Fig. 4.** Example of simultaneous crossing of two objects

We propose to overcome these problems by means of homography, thus extending the matching search to the whole zone of overlap of field of view. For two overlapped cameras $C^i$ and $C^j$, the training procedure computes the overlapping areas $Z_j^i$ and $Z_i^j$. The four corners of each of these areas defines a set of four points, $P_j^i = \{p_1^{i,j}, p_2^{i,j}, p_3^{i,j}, p_4^{i,j}\}$ and $P_i^j = \{p_1^{j,i}, p_2^{j,i}, p_3^{j,i}, p_4^{j,i}\}$, where the subscripts indicate corresponding points in the two cameras (see Fig. 2(c)). These four associations between points of the camera $C^i$ and points of the camera $C^j$ on the same plane $z = 0$ are sufficient to compute the homography matrix $H_j^i$ from camera $C^i$ to camera $C^j$. Obviously, the matrix $H_i^j$ can be easily obtained with the equation $H_i^j = (H_j^i)^{-1}$.

Each time a new object $O_k^i$ is detected in the camera $C^i$ in the overlapping area (not only at the moment of the camera handoff), its support point $SP_k^i$ is projected in $C^j$ by means of the homographic transformation. Calling $(x_{SP_k^i}, y_{SP_k^i})$ the coordinates of the support point $SP_k^i$, we can write the projected point in homogeneous coordinates:

$$[a, b, c]^T = H_j^i \left[ x_{SP_k^i}, y_{SP_k^i}, 1 \right] \tag{1}$$

The projected point $\widetilde{SP_k^j}$ corresponds on the image plane of $C^j$ to the projective coordinates $\widetilde{x^j} = \frac{a}{c}$ and $\widetilde{y^j} = \frac{b}{c}$. These coordinates could not correspond to the support point of an actual object. For the match with $O_k^i$ we select the object in

$C^j$ whose support point is at the minimum distance in the 2D plane from these coordinates:

$$O_k^i \longleftrightarrow O_p^j \mid p = \arg\min_q \; D(\widetilde{SP_k^j}, SP_q^j) \;\; \forall q \in \mathbf{O^j} \tag{2}$$

where $D(\cdot)$ denotes the Euclidean distance and $\mathbf{O^j}$ is the set of objects detected in $C^j$. The results achieved with this approach in the two cases above reported are shown in Fig. 3 and Fig. 4, respectively, where the correct label assignment is achieved.

## 5 Experimental Results

To test the system, we have installed four partially overlapped cameras in our department (see Fig. 2 for two snapshots and Fig. 1 for a map). The tests were carried out using a single camera probabilistic and appearance based tracking module [10]. EOFOV lines of the two cameras have been computed over a training video of 8000 frames. As an evidence of the goodness of the automatically obtained homography we report in Fig. 5 the mosaic image of two frames obtained merging a frame of a camera with a homographically distorted frame of the other camera.

To test the consistent labeling algorithm, instead, we have tested the system not only in the simple conditions of the training phase, but also in presence of simultaneous transitions of more than one person at a time (Sync. Trans.) and in presence of transitions in which two people are merged (Merged Trans.) in a single track during the camera handoff and split far from the EOFOV.

In Table 1 we have reported the obtained results; the number of camera transitions correctly identified (in which the consistent labeling is verified) and the number of wrong correspondences are reported in the last two columns of

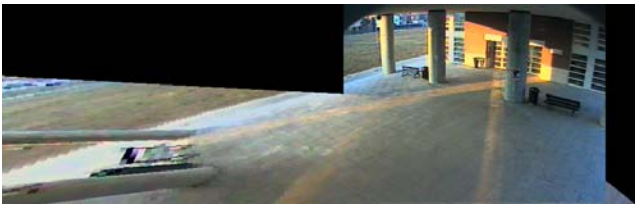
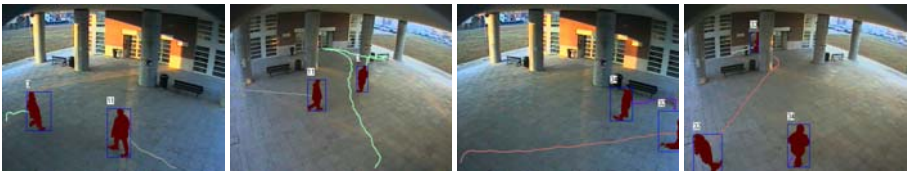


**Fig. 5.** Automatically obtained mosaic image through homography

**Table 1.** Experimental results

| | Sync. Trans. | Merged Trans. | N frames | N transitions | Correct | Incorrect |
|---|---|---|---|---|---|---|
| Video 1 | No | No | 8500 | 41 | 39 | 2 |
| Video 2 | No | No | 3000 | 5 | 5 | 0 |
| Video 3 | Yes | No | 1800 | 14 | 13 | 1 |
| Video 4 | Yes | Yes | 2000 | 7 | 6 | 1 |
| Video 5 | Yes | Yes | 500 | 2 | 2 | 0 |



(a) $C^1$ at frame #776 (b) $C^2$ at frame #776 (c) $C^1$ at frame #1490 (d) $C^2$ at frame #1490

**Fig. 6.** Some snapshots of the output of the system after consistent labeling

the table. It is evident that the system has a very high accuracy. The incorrect matches are mainly due to errors in the lower modules, i.e. in the segmentation and the tracking algorithms. Some snapshots of the output of the system after the consistent labeling assignment are reported in Fig. 6.

## 6 Conclusions

This paper presents a new method for establishing consistent labeling in a multi-camera system. Its main contributions can be summarized as follows:

1. the improvement of the automatic calibration procedure proposed in [1] to overcome to known problems; this procedure is based on the computation of the edges of field of view (EOFOV) lines;
2. the computation of the homography matrices between two overlapped views by using the EOFOV lines;
3. the exploitation of the homographic transformation to establish consistent labeling in the whole overlapping area, in order to recover the correct labels in the case of objects that enter as merged and then split.

The reported experiments demonstrate the accuracy of the proposed method, also in difficult situations.

## References

1. Khan, S., Shah, M.: Consistent labeling of tracked objects in multiple cameras with overlapping fields of view. IEEE Trans. on PAMI 25 (2003) 13551360
2. Li, J., Chua, C., Ho, Y.: Color based multiple people tracking. In: Proc. of IEEE Intl Conf. on Control, Automation, Robotics and Vision. Volume 1. (2002) 309314
3. Krumm, J., Harris, S., Meyers, B., Brumitt, B., Hale, M., Shafer, S.: Multi-camera multiperson tracking for easyliving. In: Proc. of IEEE Intl Workshop on Visual Surveillance. (2000) 310
4. Kang, J., Cohen, I., Medioni, G.: Continuous tracking within and across camera streams. In: Proc. of IEEE Intl Conference on Computer Vision and Pattern Recognition. Volume 1. (2003) I267 I272
5. Chang, S., Gong, T.H.: Tracking multiple people with a multi-camera system. In: Proc. of IEEE Workshop on Multi-Object Tracking. (2001) 1926
6. Dockstader, S., Tekalp, A.: Multiple camera tracking of interacting and occluded human motion. Proc. of the IEEE 89 (2001) 14411455

7. Chang, T., Gong, S., Ong, E.: Tracking multiple people under occlusion using multiple cameras. In: Proc. of British Machine Vision Conf. Volume 2. (2000) 566576
8. Mittal, A., Davis, L.: Unified multi-camera detection and tracking using region-matching. In: Proc. of IEEE Workshop on Multi-Object Tracking. (2001) 310
9. Yue, Z., Zhou, S., Chellappa, R.: Robust two-camera tracking using homography. In: Proc. of IEEE Intl Conf. on Acoustics, Speech, and Signal Processing. Volume 3. (2004) 14
10. Cucchiara, R., Grana, C., Tardini, G.: Track-based and object-based occlusion for people tracking refinement in indoor surveillance. In: Proc. of ACM 2nd InternationalWorkshop on Video Surveillance & Sensor Networks. (2004) 8187