Entry Edge of Field of View for multi-camera tracking in distributed video surveillance*

Simone Calderara, Roberto Vezzani, Andrea Prati, Rita Cucchiara
University of Modena and Reggio Emilia, Italy

Abstract

Efficient solution to people tracking in distributed video surveillance is requested to monitor crowded and large environments. This paper proposes a novel use of the Entry Edges of Field of View (E^2oFoV) to solve the consistent labeling problem between partially overlapped views. An automatic and reliable procedure allows to obtain the homographic transformation between two overlapped views, without any manual calibration of the cameras. Through the homography, the consistent labeling is established each time a new track is detected in one of the cameras. A Camera Transition Graph (CTG) is defined to speed up the establishment process by reducing the search space. Experimental results prove the effectiveness of the proposed solution also in challenging conditions.

1. Introduction and Related Work

Automatically monitoring people in crowded environments such as metro stations, city markets, or public parks, has nowadays become feasible for many reasons. First, from the accuracy's point of view, human operators are likely to fail in monitoring crowded and cluttered environments through tens of cameras. Automatic techniques have reached a degree of maturity to be employed at least as a first automatic step to alert human operators, reducing their effort and the sources of distraction. Second, from the economical's point of view, the cost of mounting cameras and developing automatic solutions has declined to the point that is now dwarfed by the cost of hiring human operators to watch them. The collapse of the prices of cameras has contributed to increase the deployment of multiple cameras to 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 is still a challenging task, 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 subproblems: the 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. This second problem is often known as *consistent labeling*.

Approaches to consistent labeling 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 [9, 8]. Finally, *hybrid* approaches mix information about the geometry and the calibration with those provided by the visual appearance, and they are based on probabilistic information fusion [6] or on Bayesian Belief Networks (BBN) [1][5].

Geometry-based approaches can be further subdivided into calibrated and uncalibrated approaches. Among calibrated approaches, two particularly interesting papers are [12], in which homography is exploited to solve occlusions, and [11], that uses the epipolar lines. A very relevant example of the uncalibrated approaches is the work of Khan and Shah [7], 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. Through a learning phase 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. During the training phase, the correspondences between points belonging to two overlapped cameras 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. In such a situation, the head in the first camera is put in correspondence with the feet in the other one. However, this matching is reliable enough if the goal is only the consistent labeling at the camera handoff instant (as in [7]) and if the people have the same height. Instead, if an exact correspondence is required, for example to compute an homography transfor-

*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.

mation, we must verify that the matching points belong to the same real point (e.g., the feet).

In this paper we proposed an uncalibrated geometrical approach based on the Edge of Field of View, similar to [7]. To solve the above mentioned drawbacks, we introduce the concept of *Entry Edge of Fields of View* (E^2oFoV) to assure the consistency between the extracted lines and to compute a precise homography, used to establish the consistent labeling.

Besides being able to correctly track people, a multicamera system should also be very fast, since efficient reaction is a key point in video-surveillance. The consistent labeling can, indeed, result to be quite slow in crowded environments, where several cameras are present and many people are moving. For this reason, this paper also reports an algorithm to reduce the computational cost of the consistent labeling establishment. In particular, a graph-based model, named *camera transition graph* (CTG), generalizable for a set of N overlapped cameras, is employed to efficiently search for the best match between objects in two overlapped cameras, similarly to the Vision graph described in [4] or the topology graph presented in [10]. The paper reports 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 partially overlapped cameras monitoring an outdoor environment.

2 Detecting Overlapping Areas

The proposed approach belongs to the class of uncalibrated geometry-based techniques. 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 overlapped with at least another camera C^j. Let us call *3DFOV lines* $L^{i,s}$ the projection of the limits of the field of view (FOV) of a camera C^i on the ground plane ($z = 0$), corresponding to the intersection between the ground plane and the rectangular pyramid with its vertex at the camera optical center (the camera view frustum); s indicates the equation of the line in the image plane. In particular, four of them, $L^{i,s_h}, h = 1..4$ could be computed, with s_h corresponding to the image borders $x = 0$, $x = x_{max}$, $y = 0$, and $y = y_{max}$. They could be visible also by another camera; in such a situation we call *Edge of Field of View* $L^{i,s}_j$ the 3DFoV line corresponding to s of the Camera C^i seen by the camera C^j. The $L^{i,s}_j$ cannot be always computed, because sometimes is totally hidden by a large object (e.g. a column). For our purposes, partial visibility is sufficient.

The EoFoV $L^{i,s}_j$ divides the image on camera C^j into two half-planes, one overlapped with camera C^i and the other one disjoint. The intersection of the overlapped semi-planes defined by the EoFoV lines from camera C^i to camera C^j generates the overlapping area $AoFoV^i_j$.

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 [7], during the training phase a single person moves freely in the scene, with the minimum requirement to pass through at least two points of each 3DFoV.

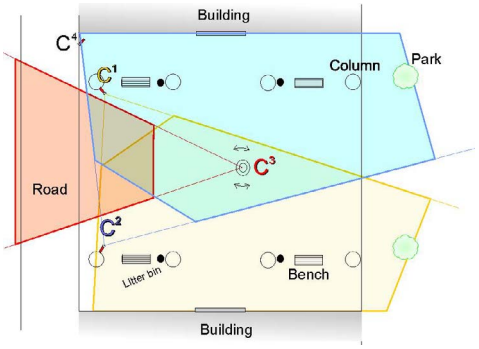


Figure 1: Map of our real setup.

Therefore, even if the EoFoV is not completely visible, it can be computed if we are able to detect a sufficient number of reliable corresponding points belonging to the ground plane $z = 0$. In Fig. 1 a Map of the setup at the Modena's Campus is depicted where four cameras partially overlapped (one PTZ and three fixed cameras) have been installed. In Fig. 2 the synchronized views of C^1 and C^2 are shown.

In Fig. 2.a the EoFoV $L^{1,s}_2$ and $L^{2,s}_1$ are indicated. Let us suppose to have in Fig. 2.a only the person P (e.g. the one labelled as 3). When he enters in the FoV (camera handoff) as in Fig. 2.a of C^1, his support point is computed and matched with the support point of the correspondent shape K detected in C^2. The support point SP is defined as the middle point of the bottom of the bounding box of the blob, with the assumption that the training person is walking in a standing position. Therefore, collecting several matching pairs (SP^i_P, SP^j_K), the EoFoV can be computed with a Least Square Optimization. In the example in Fig. 2.b the $L^{2,s}_1$ corresponds approximately to the border of the image, being computed at the camera handoff moment.

However, there are cases in which, at the moment of the camera handoff, the detected parts of the person do not lie on the ground plane, as in Fig. 2.c, where the head is detected. Thus, matching a head's point in this camera with the SP in the other camera is incorrect and causes an erroneous EoFoV computation.

To avoid this problem, we define *Entry EoFoV (E^2oFoV)* as the EoFoV that is computed with the matching of

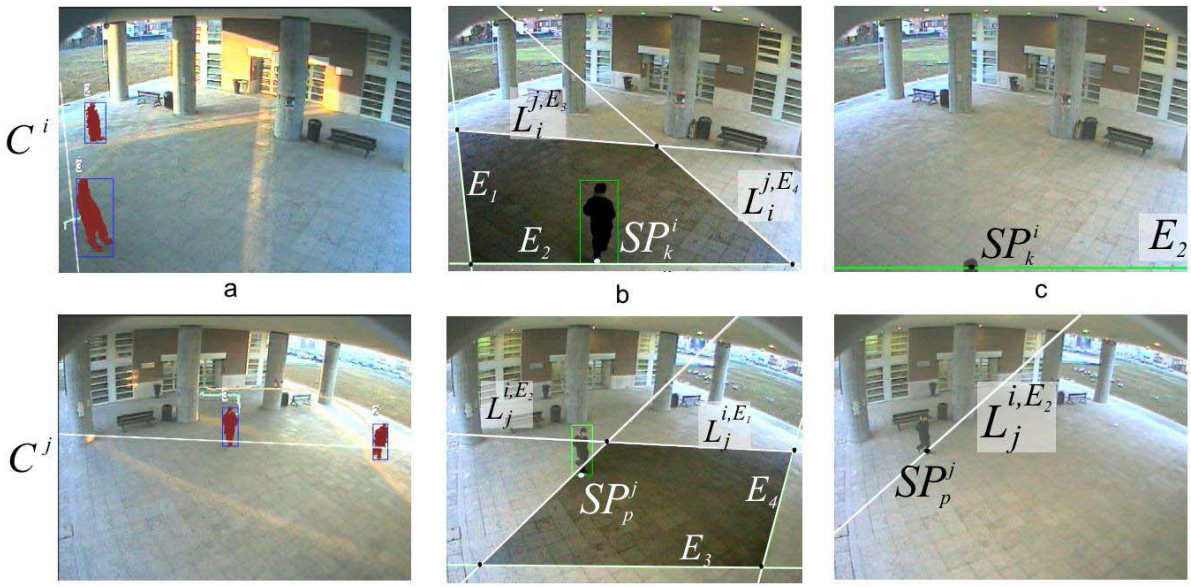


Figure 2: a. Simultaneous camera handoff of two tracks; b. E^2oFoV creation; b. $EoFoV$ creation (using Khan-Shah approach)

(SP_P^i, SP_K^j) extracted by the bounding boxes of *totally* visible people, i.e. after the camera handoff when the blob does not "touch" the image border anymore. This approach 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. As a consequence, the actual FOV lines E_h are neither coincident nor parallel to the image borders. In Fig. 2.b the E^{2}oFoV lines $(L_j^{i,E_1}, L_j^{i,E_2})$, $(L_i^{j,E_3}, L_i^{j,E_4})$, correspondent to (E_1, E_2) in C^i, (E_3, E_4) in C^j respectively, are depicted. Fig. 2.b shows also the overlapping FoV, named Area of FoV $AoFoV$, delimited with the E^{2}oFoV and E_h lines.

3 Consistent Labeling with Homography

In order to propose a general approach we define the solution of consistent labeling not only at the camera handoff but whenever it is necessary to exploit homography.

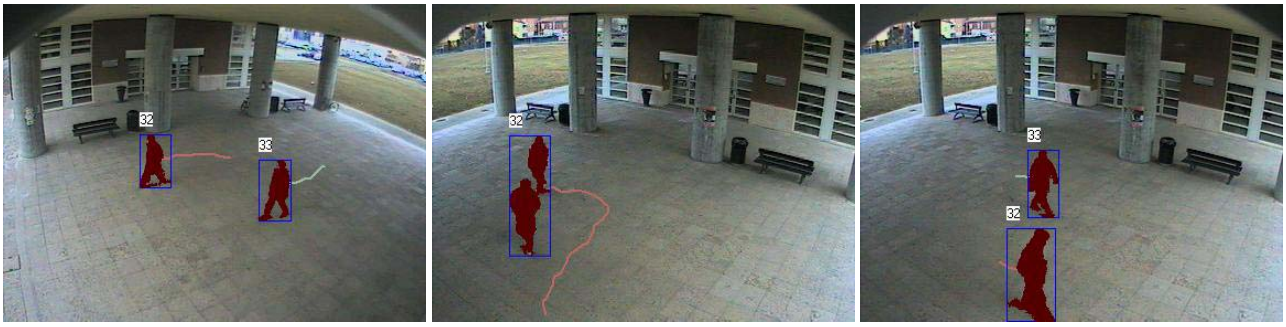
For two overlapped cameras C^i and C^j, the training procedure computes the $AoFoV_j^i$ and $AoFoV_i^j$ areas. The four corners of each of these area define a set of four points, $P_j^i = \{p_1^{i,j}, p_2^{i,j}, p_3^{i,j}, p_4^{i,j}\}$ and $P_i^j = \{p_1^{j,i}, p_2^{j,i}, p_3^{j,i}, p_4^{j,i}\}$, where 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 is detected in the camera C^i in the overlapping area (not only at the moment of the camera handoff), its support point is projected in C^j by means of the homographic transformation. The coordinates of the projected point could not correspond to the support point of an actual object. For the match we select the object in C^j whose support point is at the minimum Euclidean distance in the 2D plane from these coordinates.

This approach is an efficient tradeoff between classical approaches that verify correspondences at the camera handoff only as in [7], and complex methods of 3D reconstruction that find correspondences at each frame preventing any real time implementation[11]. Instead the matching is verified whenever a new track is detected in an image and also a 1-to-n or n-to-m match could be computed for coping with the cases where more people passe through the EoFoV at the same time as in Fig. 2.a. Moreover, as in Fig. 3, some people could be initially detected as a group and thus matched with a single track. For instance, in Fig. 3 a group is matched to a single object (label 32). Nevertheless, whenever the people can be detected separately (after 10 frames) the correct consistent labeling is recovered.

4 Camera Transition Graph

When a new track is detected in camera C^i the system must check whether it is a completely new track or it is already present in other cameras. This check can be very complex and computationally expensive if many cameras with many



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

Figure 3: Example of simultaneous crossing of two merged objects (frame #1250, and split after entered at frame #1260

tracks are present. To this aim, a graph model has been defined to exploit camera relationships in reducing the search space of the multi-camera matching process.

In the proposed model a graph is built using information acquired during the training phase. The graph is called *Camera Transition Graph* (CTG), because it incorporates information about camera position and it models possible tracks handoff among overlapped views.

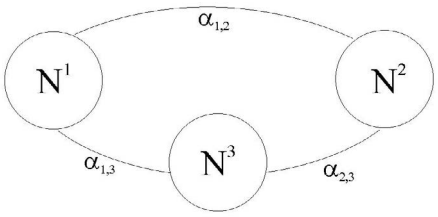


Figure 4: Example of Camera Transition Graph (CTG).

The problem can be viewed as a Constraint Satisfaction Problem (CSP). The CTG is a symmetric graph where each node N^i is the set of objects visible and tracked in an instant in a camera C^i, and each arc $\alpha_{i,j} = \alpha_{j,i}$ indicates the presence of a common $AoFoV$ between C^i and C^j and needs that the constraint of consistency must be verified if a track is visible inside the $AoFoV$.

In the CTG for each node N^j we denote $T^j(t) = \left\{\tau_i^j(t) | \, i = 1, \ldots, k^j(t)\right\}$ the set of variable at the time t, that are the tracks detected and with $x_i^j(t)$ their correspondent assigned labels (the instanced values for the variables).

The unary constraint at each node is that two distinct tracks must have different labels and they must be conserved during the time. This is the typical tracking problem from a single camera and the unary constraints must be checked at each frame. Instead, the binary constraint of "consistent labeling" over the $AoFoV$ is verified only when needed.

The state of the whole system at any time can be either *consistent* or *inconsistent*. We refer to a consistent state whenever the constraints are satisfied and all the projections of the same person on multiple cameras are marked with the same label, while different people have different labels.

If a new track is detected at time $t+1$, the system is switched to inconsistency, and the system must check if it appears also in other cameras or not. Exploiting graph theory, specifically solving an arc-consistency problem on the CTG, it is possible to correctly select only cameras and tracks generating inconsistencies and leaving the rest of the system unchanged.

Suppose that at time t the state of the system is *consistent*. Suppose also that on camera C^l of the node N^l, at time $t+1$ is detected a new track $\tau^l_{k^l(t)+1}$ labelled $x^l_{k^l(t)+1}(t+1)$. Instead of searching for possible matches across cameras' views, we analyze the graph node corresponding to camera C^l, N^l, and we compute the set Ξ^l of nodes linked to it by means of a consistency constraint arc:

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

Let us call $SP^l_{k^l(t)+1}$ the support point of the new track, computed as reported in Section 2. For each element N^i of the set Ξ^l we must evaluate if the support point $SP^l_{k^l(t)+1}$ lies inside the $AoFoV^i_l$ between camera C^l and camera C^i on the image plane of C^l. To this aim, we used a boolean function ϕ defined as follows:

$$\phi\left(N^l, N^i, SP\right) = \begin{cases} 1 & if \; SP \in Z^i_l \\ 0 & otherwise \end{cases} \tag{2}$$

In the case that ϕ returns zero for each node N^i in the set Ξ^l the new track is not visible in any other overlapped camera, thus, a new label is assigned and the system is consistent.

Otherwise, the search space $\Sigma^l_{k^l(t)+1}$ for the new track $\tau^l_{k^l(t)+1}$ is composed by the set of tracks τ^i_m such that:

$$\begin{aligned}&(i) N^i \in \Xi^l\\&(ii) \phi\left(N^l, N^i, SP^l_{k^l(t)+1}\right) = 1\\&(iii) \phi\left(N^i, N^l, SP^i_m\right) = 1\end{aligned} \quad (3)$$

In other words, for each camera C^i whose view is overlapped with C^l (condition (i) of equation 3) and in which the new track is visible (condition (ii) of equation 3), the set of tracks $T^i(t+1)$ at the time $t+1$ is considered. For each track of this set, the visibility on the camera C^l is checked by means of function ϕ (condition (iii)) and if it is visible, the track is considered as a candidate for the consistent labeling. It is evident that the search space $\Sigma^l_{k^l(t)+1}$ obtained with this procedure is minimal and that results in computational saving, especially if the matching procedure is complex and time consuming, as in the case that the track appearance is used.

5 Experimental Results

To test the system, we have installed four partially overlapped cameras in our department (see Fig. 1 for a map). The tests were carried out using a single camera probabilistic and appearance-based tracking module [3].

This approach permits to maintain the appearance of the track and thus to compute its Support Point also when it is partially occluded. It has been used for posture classification in [2]. E^{2}oFoV and $AoFoV$ of the three 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.a the mosaic image of three frames obtained merging a frame of a camera with the homographically distorted frames of the other two cameras. If the homography transformation between at least one camera and the ground plane is manually given, a sort of bird-eye view of the scene can be obtained (Fig. 5.b), in which the trajectory of a person is automatically drawn. In the figure some examples of the projections in the three views of the person are superimposed.

Some snapshots of the output of the system (in non-trivial conditions) after the consistent labeling assignment are reported in Fig. 6.

The track graphs in Fig. 5.c, and Fig. 7 report, for each person P_i, the slot of time (in frames) in which it is visible by the three cameras (C^1, C^2, and C^3) of our real setup. The color of the bars corresponds to the identifier assigned by the consistent labeling algorithm. In particular the graph in Fig. 5.c is related to the same sequence used to construct the trajectory of Fig. 5.b.

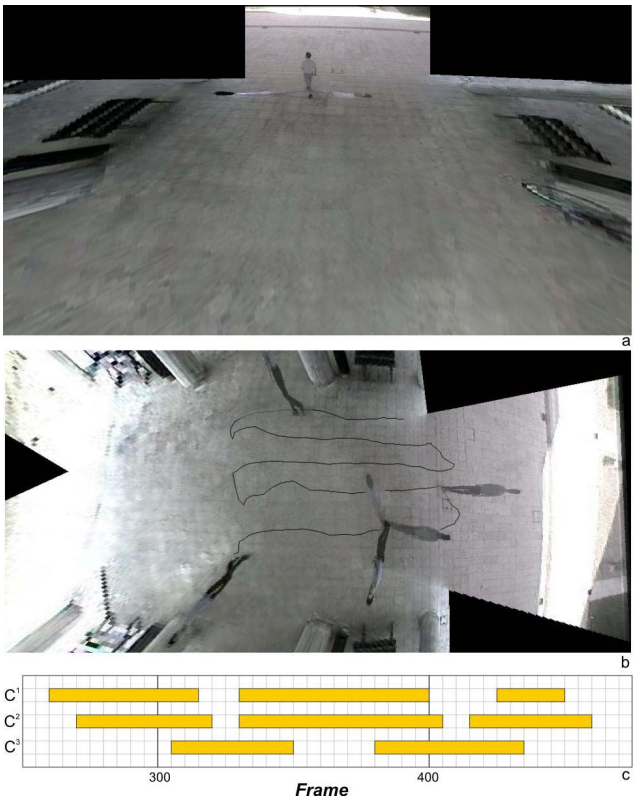


Figure 5: a. Automatically obtained mosaic image through homography; b. A bird-eye view of the scene with a superimposed trajectory; c. visibility over the three different cameras of the track used to generate the trajectory

6 Conclusions

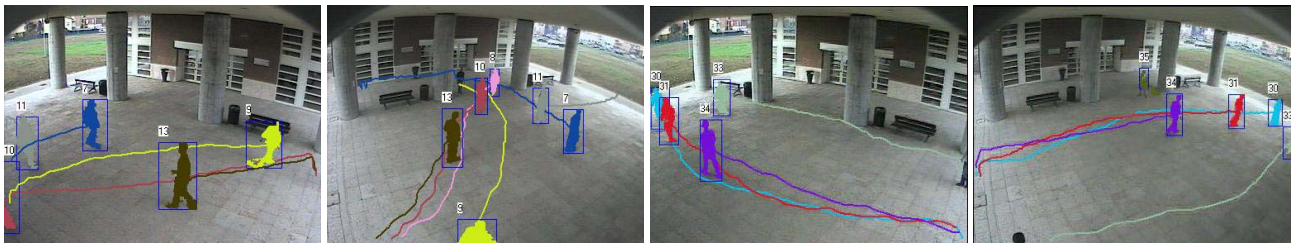
This paper presents a new method for establishing consistent labeling in a multi-camera system. Its main contributions can be summarized as follows:

1. the computation of reliable E^2oFoV lines to obtain without calibration the correct area of overlap $AoFoV$

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.

4. the automatic generation of a *Camera Transition Graph* that models the topology of the network of cameras reporting the field of view overlaps; this graph is used to reduce the search space during the consistent labeling phase.



(a) C^1 at frame #783 (b) C^2 at frame #783 (c) C^1 at frame #1080 (d) C^2 at frame #1080

Figure 6: Some snapshots of the output of the system after consistent labeling.

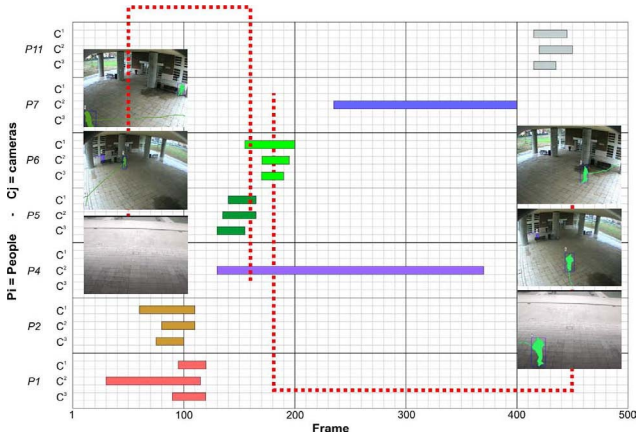


Figure 7: Visibility and labels (indicated with the color of the bars) of the tracks in a test sequence

The reported experiments demonstrate the accuracy of the proposed method, also in situations with many people overlapped and only partially visible.

References

[1] S. Chang and T.-H. Gong. Tracking multiple people with a multi-camera system. In *Proc. of IEEE Workshop on Multi-Object Tracking*, pages 19–26, 2001.

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

[3] R. Cucchiara, C. Grana, and G. Tardini. Track-based and object-based occlusion for people tracking refinement in indoor surveillance. In *Proc. of ACM 2nd International Workshop on Video Surveillance & Sensor Networks*, pages 81–87, 2004.

[4] D. Devarajan and R. Radke. Distributed metric calibration of large camera networks. In *First Workshop on Broadband Advanced Sensor Networks (BASENETS)*, 2004.

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

[6] Jinman Kang, I. Cohen, and G. Medioni. Continuous tracking within and across camera streams. In *Proc. of IEEE Int'l Conference on Computer Vision and Pattern Recognition*, volume 1, pages I–267 – I–272, 2003.

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

[8] J. Krumm, S. Harris, B. Meyers, B. Brumitt, M. Hale, and S. Shafer. Multi-camera multi-person tracking for easyliving. In *Proc. of IEEE Intl Workshop on Visual Surveillance*, pages 3–10, 2000.

[9] J. Li, C.S. Chua, and Y.K. Ho. Color based multiple people tracking. In *Proc. of IEEE Intl Conf. on Control, Automation, Robotics and Vision*, volume 1, pages 309–314, 2002.

[10] D. Makris, T. Ellis, and J. Black. Bridging the gaps between cameras. In *Proc. of IEEE Int'l Conference on Computer Vision and Pattern Recognition*, volume 2, pages 205–210, 2004.

[11] A. Mittal and L. Davis. Unified multi-camera detection and tracking using region-matching. In *Proc. of IEEE Workshop on Multi-Object Tracking*, pages 3–10, 2001.

[12] Z. Yue, S.K. Zhou, and R. Chellappa. Robust two-camera tracking using homography. In *Proc. of IEEE Intl Conf. on Acoustics, Speech, and Signal Processing*, volume 3, pages 1–4, 2004.