Multi-camera People Tracking by Collaborative Particle Filters and Principal Axis-Based Integration

Wei Du and Justus Piater

University of Liège, Department of Electrical Engineering and Computer Science
Montefiore Institute, B28, Sart Tilman Campus, B-4000 Liège, Belgium
weidu.montefiore.ulg.ac.be, justus.piater@ulg.ac.be

Abstract. This paper presents a novel approach to tracking people in multiple cameras. A target is tracked not only in each camera but also in the ground plane by individual particle filters. These particle filters collaborate in two different ways. First, the particle filters in each camera pass messages to those in the ground plane where the multi-camera information is integrated by intersecting the targets' principal axes. This largely relaxes the dependence on precise foot positions when mapping targets from images to the ground plane using homographies. Secondly, the fusion results in the ground plane are then incorporated by each camera as boosted proposal functions. A mixture proposal function is composed for each tracker in a camera by combining an independent transition kernel and the boosted proposal function. Experiments show that our approach achieves more reliable results using less computational resources than conventional methods.

1 Introduction

Tracking people in multiple cameras is a basic task in many applications such as video surveillance and sports analysis. A commonly-used fusion strategy is to detect people in each camera with bottom-up approaches such as background subtraction and color segmentation, and then to calculate the correspondences between cameras using the camera calibrations, or more often, the ground homographies. In order to reason about occlusions between targets, this fusion strategy usually requires all targets to be correctly detected and tracked [8,4,2,6,5]. However, sometimes we may be interested in the trajectories of only a few key targets, for instance, the star players in a soccer game or a few suspects in a surveillance scenario. Top-down approaches are preferable in such situations.

In this paper, we present a novel top-down approach to people tracking by multiple cameras. The approach is based on collaborative particle filters, i.e., we track a target not only in each camera but also in the ground plane by individual particle filters. These particle filters collaborate in two different ways. First, the particle filters in each camera pass messages to those in the ground plane where the multi-camera information is integrated using the homographies

Y. Yagi et al. (Eds.): ACCV 2007, Part I, LNCS 4843, pp. 365–374, 2007.

of each camera. Such a fusion framework usually relies on precise foot positions of the targets, which are often not provided by the particle filters in the cameras. To overcome the imprecise foot positions as well as the uncertainties of the camera calibrations, we exploit the principal axes of the targets during integration, which greatly improves the precision of the fusion results. These fusion results are then incorporated by the trackers in each camera as boosted proposal functions. A mixture proposal function is composed for each tracker in a camera by combining an independent transition kernel and the boosted proposal function, from which new particles are generated for the next time instant.

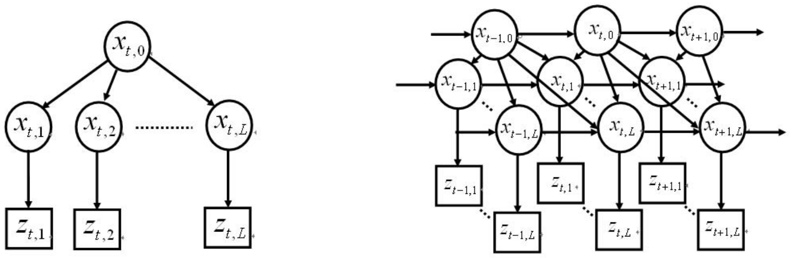
Our approach has several distinct features. First, it doesn't require all targets to be tracked simultaneously. Instead of having different target trackers interact, we compute the consensus between cameras by having different camera trackers communicate. Second, it has a fully distributed architecture. All the computations are performed locally and only the filter estimates are exchanged between the cameras and the fusion module. Third, the fusion of the multi-camera information is done by intersecting the targets' principal axes. Experiments on both surveillance and soccer scenarios show that our approach achieves more reliable results using less computational resources than conventional methods.

Particle filters are conventional in multi-camera tracking. Most previous work performed particle filtering in 3D so that precise camera calibration is required to project particles into the image plane of each camera [9,7]. The multi-camera information is often integrated by either the product of the likelihoods in all cameras [7] or a selection of the best cameras that contain the most distinctive information [9]. In our previous work, we proposed a different approach to fusion that combined particle filters and belief propagation, where particle filters collaborated with each other via a message passing procedure [1]. To match ground-plane target positions using homographies, the foot positions of the tracked people have to be detected. This, however, is a difficult and error-prone task if done separately for each camera. In this paper, we address the precision and computational issues. We relax the dependence on precise foot positions by exploiting the principal axes of the targets, the intersections of which give better ground positions. At the same time, we improve the speed over our previous system [1] by incorporating the fusion results from the ground plane as proposal functions into each camera.

The rest of the paper is organized as follows. Section 2 formulates the multi-camera tracking problem. Section 3 introduces the collaborative particle filters, including the principal axis-based integration and the boosted proposal functions. Experiments on sequences of video surveillance and soccer games are shown in Section 4.

2 Problem Formulation

Suppose L cameras are used and each camera collects one observation for each target at each time instant. Denote the target state on the ground plane by $x_{t,0}$ and its states in different cameras by $x_{t,j}$, $j = 1, \ldots, L$. Let $z_{t,j}$ denote



(a) tree-structured graphical model (b) dynamic Markov model

Fig. 1. Graphical models for modeling the dependencies at time t and for modeling the evolution of the system in time

the observation in camera j at time t, $Z_t = \{z_{t,1}, \ldots, z_{t,L}\}$ the multi-camera observation at time t, and $Z^t = \{Z_1, \ldots, Z_t\}$ the multi-camera observations up to time t.

Fig. 1(a) shows the graphical model that models the dependencies between target states in the ground plane and at different cameras at time t. We assume that the $x_{t,j}$, $j = 1, \ldots, L$, are independent given $x_{t,0}$ so that a tree-structured model is formed. Note that $x_{t,0}$ is associated with no observation. Connecting the graphical models at different times results in a dynamic Markov model, shown in Fig. 1(b), that describes the evolution of the system over time. As all the $x_{t,j}$ depend on $x_{t,0}$, we add temporal links from $x_{t-1,0}$ to $x_{t,j}$. The addition of these temporal links is beneficial to the design of the proposal functions, shown in the next section.

In both models in Fig. 1, each directed link from $x_{t,0}$ to $x_{t,j}$, $j = 1, \ldots, L$, represents a message passing process and is associated with a potential function $\psi^t_{0,j}(x_{t,0}, x_{t,j})$. The directed link from $x_{t,j}$ to $z_{t,j}$, $j = 1, \ldots, L$, represents the observation process and is associated with a likelihood function $p_j(z_{t,j}|x_{t,j})$. In Fig. 1(b), the directed links from $x_{t-1,i}$ to $x_{t,i}$, $i = 0, \ldots, L$, and from $x_{t-1,0}$ to $x_{t,j}$, $j = 1, \ldots, L$ represent the state transition processes and are associated with motion models $p(x_{t,i}|x_{t-1,i})$ and $p(x_{t,j}|x_{t-1,0})$ respectively.

Thus, we infer each $x_{t,i}$, $i = 0, \ldots, L$, based on all Z^t. A message passing scheme, the same as is used in belief propagation, is adopted to pass messages from each camera to the ground plane. The message from camera j is defined as

$$m_{0j}(x_{t,0}) \leftarrow \int p_j(z_{t,j}|x_{t,j})\psi^t_{0,j}(x_{t,0}, x_{t,j}) \int p(x_{t,j}|x_{t-1,j})p(x_{t-1,j}|Z^{t-1})\mathrm{d}x_{t-1,j}\mathrm{d}x_{t,j}. \tag{1}$$

The belief $p(x_{t,0}|Z^t)$ is computed recursively by the message product and the propagation of the previous posterior,

$$p(x_{t,0}|Z^t) \propto \prod_{j=1,\ldots,\, L} m_{0j}(x_{t,0}) \times \int p(x_{t,0}|x_{t-1,0})p(x_{t-1,0}|Z^{t-1})\mathrm{d}x_{t-1,0}. \tag{2}$$

Note that the same message and belief update equations are used in our previous work [1].

The inference of $x_{t,j}$, $j = 1, \ldots, L$, is done by nearly standard particle filters, except that the fusion results at $t-1$ are taken into consideration. The belief $p(x_{t,j}|Z^t)$ is computed as

$$\begin{aligned} p(x_{t,j}|Z^t) \;&\propto\; p(x_j|z_j) \times \qquad (3) \\ &\int\!\!\int p(x_{t,j}|x_{t-1,j})p(x_{t-1,j}|Z^{t-1})\underline{p(x_{t,j}|x_{t-1,0})p(x_{t-1,0}|Z^{t-1})}\mathrm{d}x_{t-1,0}\mathrm{d}x_{t-1,j}. \end{aligned}$$

The underlined terms incorporate the fusion results as a boosted proposal function. In other words, the fusion module is used by each camera as a coupled process.

3 Collaborative Particle Filters

All the inference processes formulated above, in the ground plane and for each camera, are performed by individual but collaborative particle filters. Details are given below.

3.1 Principal Axis-Based Integration

The ground-plane particle filter integrates the multi-camera information according to Eqs. 1 and 2. For tracking ground targets, homographies are often used to map the foot positions from each camera to the ground plane. However, a large number of particles are required to estimate precise foot positions, which significantly slows down the tracking system. With a small number of particles, usually the sizes of the targets cannot be estimated precisely. We overcome this problem by exploiting the principal axes of the targets.

The principal axis of a target is defined as the vertical line from the head of the target to the feet. It has been shown that the principal axes of a target in different cameras intersect in the ground plane, and computing the intersection point yields very robust fusion results [4,6], illustrated in Fig. 3. We exploit this effect in our multi-camera integration.

The idea is to sample particles in the ground plane by importance sampling, and to evaluate these particles by passing messages from each camera. Here, $p(x_{t,0}|x_{t-1,0})$ is used as the proposal function from which new particles for $x_{t,0}$ are sampled. Each of these ground-plane particles receives messages from each camera, and a message weight is computed using Eq. 1. The principal axes are incorporated in the potential function $\psi^t_{0,j}(x_{t,0}, x_{t,j})$ in Eq. 1. In general, the principal axes of the particles in a camera are projected to the ground plane using the homographies. The potential function measures the distances of the ground particles to these projected principal axes and converts them to probability densities, given by

$$\psi^t_{0,j}(x^n_{t,0}, x^m_{t,j}) \;\propto\; \exp(-\mathrm{dist}^2(x^n_{t,0}, \mathrm{project}(H_j, x^m_{t,j}))), \qquad (4)$$

where $x^n_{t,0}$ and $x^m_{t,j}$ are the nth ground-plane particle and mth particle in camera j, H_j is the homography from camera j to the ground plane, dist() computes

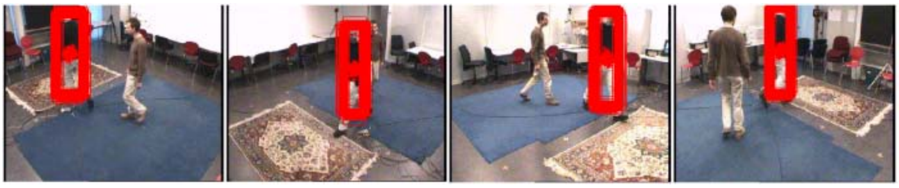
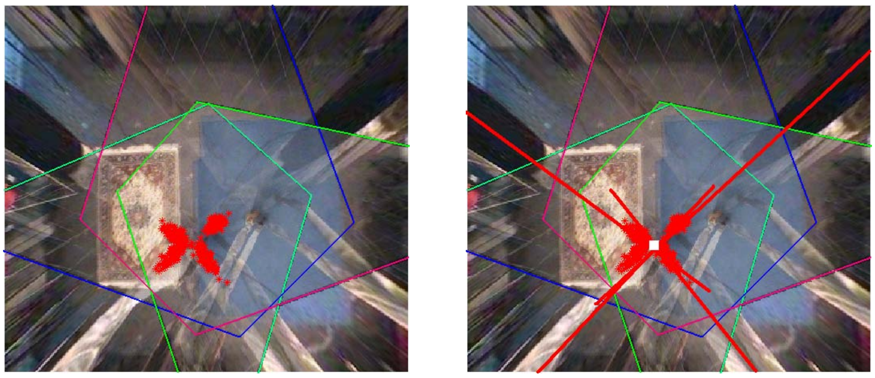


Fig. 2. The particle distributions in four cameras at a time instant. It can be seen that the foot positions are not precise although all the particles are placed at the right location.



(a) Mapping particles to the ground.

(b) Mapping principal axes to the ground.

Fig. 3. Comparison between homography-based integration and principal axis-based integration. In (a), the projections of the particles (the red stars) from the images in Fig 2 to the ground have a large variance, making the integration imprecise. In contrast, in (b), the intersection of the principal axes (the red lines) of four selected particles yields a more precise foot position (the white square).

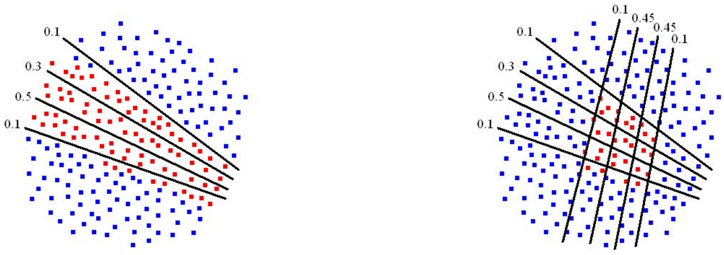
the distance between a point and a line, and project() maps the principal axis to the ground. The message and belief weights are then computed by

$$w_{t,0}^{j,n} \propto \sum_{m=0}^{N} \pi_{t,j}^{m} \psi_{0,j}^{t}(x_{t,0}^{n}, x_{t,j}^{m}), \ \ \pi_{t,0}^{n} \propto \prod_{j=1}^{L} w_{t,0}^{j,n}, \qquad (5)$$

where $w_{t,0}^{j,n}$ is the message weight of $x_{t,0}^{n}$ from camera j, and $\pi_{t,0}^{n}$ and $\pi_{t,j}^{m}$ are the belief weights of $x_{t,0}^{n}$ and $x_{t,j}^{m}$. Intuitively, the closer a ground-plane particle is to all the principal axes, the larger its weight is, as illustrated in Fig. 4.

3.2 Boosted Proposal Functions

A target is tracked in each camera by a particle filter. Due to the occlusions or other image noise, feedback from the fusion module is expected to improve



(a) Camera 1 passes messages

(b) Camera 2 passes messages

Fig. 4. An illustration of evaluating ground-plane particles using two cameras. The ground-plane particles are evaluated according to the distances to the projected principal axes. (a) After the first camera passes messages to the ground plane, all the particles along the principal axes (red dots) have larger weights than those further away (blue dots). The weights of the camera particles are shown at one end of the corresponding principal axes. (b) After the second camera passes messages, only those ground-plane particles that are close to the intersections have large weights.

the tracking performance in a camera. A similar message passing procedure was adopted in our previous work to pass messages from the ground plane to each camera, which proved computationally expensive. We propose here a different method to incorporate this feedback.

Note that in the dynamic Markov model in Fig. 1(b), for each $x_{t,j}$, $j = 1, \ldots, L$, there is an extra temporal link from $x_{t-1,0}$ besides that from $x_{t-1,j}$. This enables us to design a mixture proposal function for importance sampling,

$$p(x_{t,j}|x_{t-1,j}, x_{t-1,0}) \propto \alpha p(x_{t,j}|x_{t-1,j}) + (1-\alpha)p(x_{t,j}|x_{t-1,0}). \quad (6)$$

Thus, we sample particles from both $p(x_{t,j}|x_{t-1,j})$ and $p(x_{t,j}|x_{t-1,0})$, i.e., αN particles are sampled from $p(x_{t,j}|x_{t-1,j})$ and the other $(1-\alpha)N$ from $p(x_{t,j}|x_{t-1,0})$. Parameter α specifies a trade-off between two proposal functions and is set to 0.5 in our experiments. To sample from $p(x_{t,j}|x_{t-1,0})$, we fit a Gaussian distribution to $x_{t-1,0}$ and propagate it to each camera using the homographies.

In a sense, the fusion results at $t-1$ are used as boosted proposal functions by each camera. This is beneficial not only in maintaining consistency between the particle filters at different nodes but also in speeding up the tracking algorithm. The sampled particles are evaluated using the image likelihood as is done in standard particle filters.

4 Results

We tested our method on both video surveillance and soccer game sequences. We manually initialized the targets of interest in the first frames of the sequences and sampled 100 particles for each filter.

Figure 5 shows the results of tracking a pedestrian in PETS sequences and a comparison with a reference method [7], which tracks a target in 3D by a particle

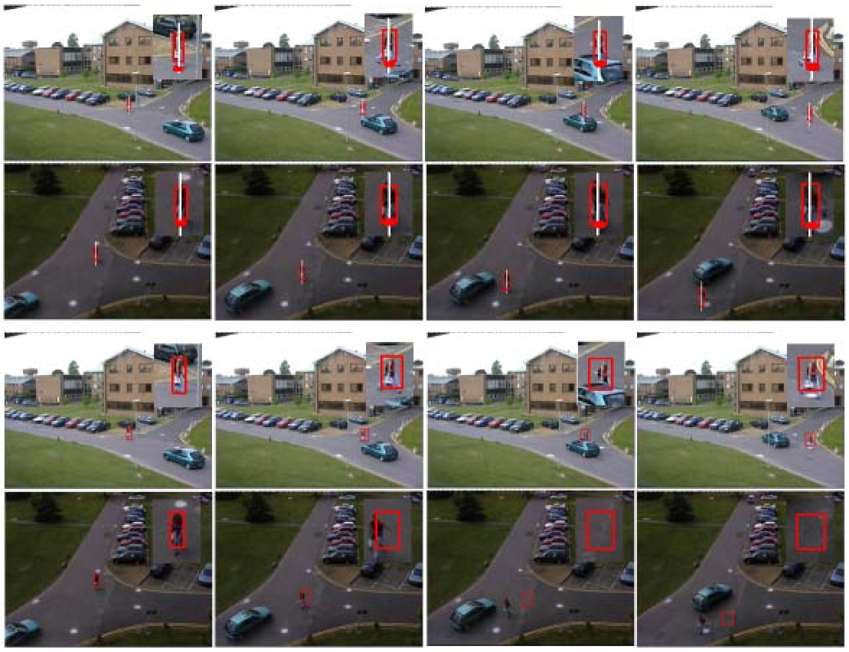


Fig. 5. The results of tracking a pedestrian in PETS sequences with our approach (top rows) and with a reference method [7] (bottom rows). In the latter method [7], we initialize a tracker in one camera and project the particles to another camera using the homographies. Here, due to imprecise foot positions, the estimates are projected to wrong positions.

filter and evaluates the particles by the product of the likelihoods in all cameras. In this experiment, we adopted a classic color observation model and evaluate each particle by matching the color histogram to a reference model [11]. The figure shows that particle filters do not estimate precise foot positions; thus, mapping the particles between cameras or between cameras and the ground plane using homographies is imprecise. As a result, using this method [7], most particles in one camera are projected to wrong positions in another camera so that only one camera contributes to the tracking. On the other hand, due to the use of the principal axes, our method integrates information from both cameras and achieves more reliable results.

Figure 6 shows the results of tracking two selected people in an indoor environment with four cameras. In this experiment, we adopted a hierarchical multi-cue observation model and evaluated each particle first by a color likelihood function and then by a background-subtraction likelihood function [10]. We also assumed that the sizes of the people were fixed and could be inferred from their ground positions [2]. Thus, the only parameters of interest were the positions in the images and in the ground plane. A comparison with our previous work [1] shows that the new approach achieved similar results but was approximately twice faster.

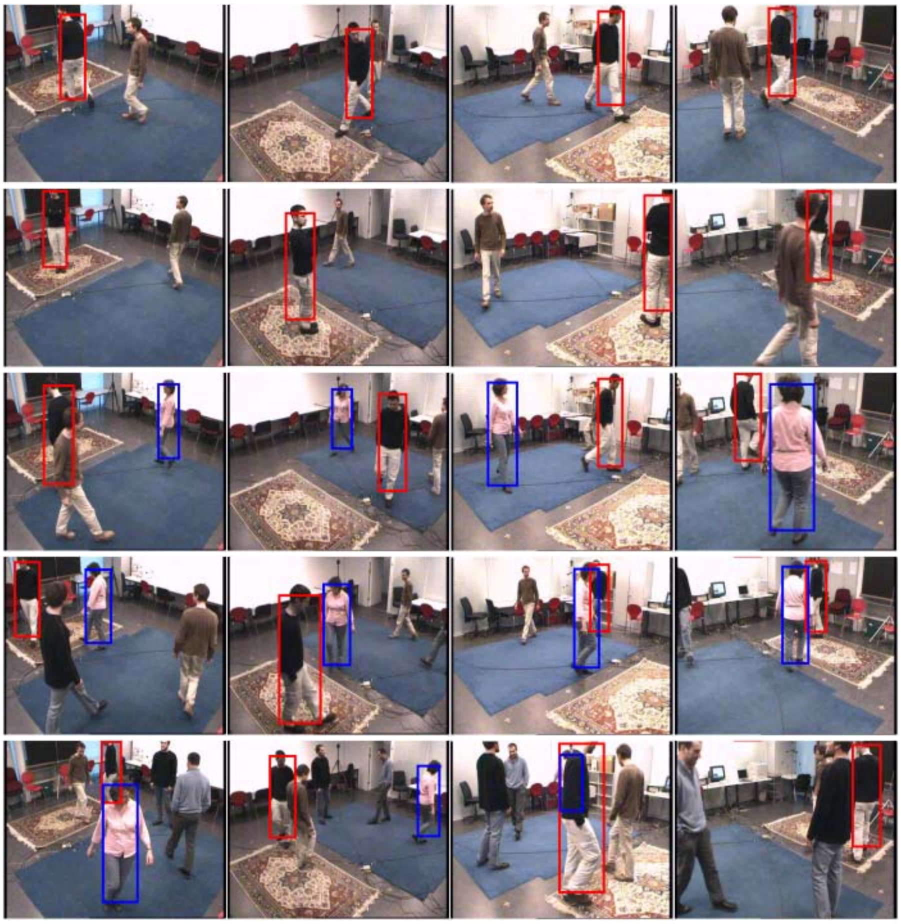


Fig. 6. The results of tracking two people in an indoor environment with four cameras. Each row shows four simultaneous views. In this experiment, both the head and the ground homographies of each camera are available. The fixed-size assumption significantly improved the robustness of the algorithm.

Figure 7 shows the results of tracking several soccer players in three cameras. Due to the interactions between the players, the feedback from the fusion module to each camera becomes critical, without which the trackers in different cameras fail one by one. In this experiment, the same observation model as in the PETS experiment was used and the homographies of each camera were obtained on-line by using a field model and by accumulating motion estimates between consecutive frames [3]. Note in the figure that the estimated foot positions do not coincide with the bottom of the bounding boxes, but are more precise than these thanks to the multiple-camera fusion using principal axes. At one point, due to a heavy occlusion that occurs in all cameras, a tracker jumps from one

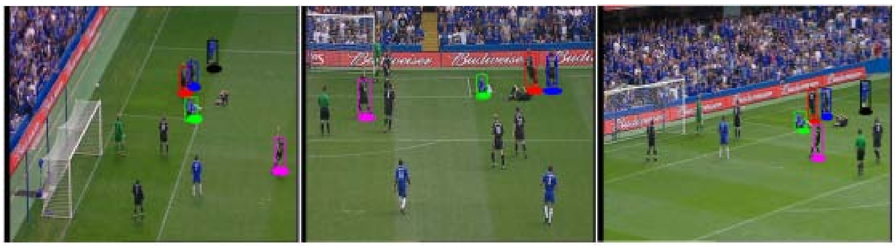


Fig. 7. The results of tracking several soccer players in the last frames of the three sequences. The ellipses under the rectangles are the fusion results in the ground plane.

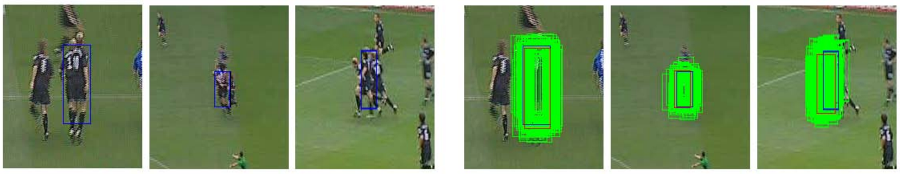


Fig. 8. The particle distributions at the time when the tracker is about to jump to a different player, which happens here because the players involved are very close both in space and in appearance in all three views. The green rectangles are the sampled particles, the blue are the estimates, and the red are the predictions of the fusion results at the previous time.

target to another. In such situations, multi-camera systems without feedback between cameras are susceptible to mismatched targets. In our system, thanks to the feedback from the ground-plane tracker, the trackers at each camera remain consistent, even if they collectively follow the wrong target. Figure 8 shows the particle distributions at the time instant when the jump begins. This problem can be partially solved by tracking multiple targets simultaneously.

5 Conclusion and Future Work

This paper presents a novel approach to ground-plane tracking of targets in multiple cameras. Different from previous work, our approach is not based on bottom-up detection or segmentation methods. Instead, we infer target states in each camera and in the ground plane by collaborative particle filters. Message passing and boosted proposal functions are incorporated in the collaboration between the trackers in each camera and the fusion module. Principal axes are exploited in the multi-camera integration, which enables us to handle the imprecise foot positions and some calibration uncertainties. In doing so, we achieve robust results using relatively little computational resources. We are currently adapting this approach to multi-target, multi-camera tracking, which involves the modeling of the target interactions and data association across cameras.

Acknowledgement

The authors wish to thank J. Berclaz and F. Fleuret for sharing their data.

References

1. Du, W., Piater, J.: Multi-view object tracking using sequential belief propagation. In: Asian Conference on Computer Vision, Hyderabad, India (2006)
2. Fleuret, F., Berclaz, J., Lengagne, R., Fua, P.: Multi-camera people tracking with a probabilistic occupancy map. IEEE Transactions on Pattern Analysis and Machine Intelligence (2007)
3. Hayet, J.-B., Piater, J., Verly, J.: Robust incremental rectification of sports video sequences. In: British Machine Vision Conference, Kingston, UK, pp. 687–696 (2004)
4. Hu, W.-M., Hu, M., Zhou, X., Tan, T.-N., Lou, J., Maybank, S.J.: Principal axis-based correspondence between multiple cameras for people tracking. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(4), 663–671 (2006)
5. Khan, S.M., Shah, M.: A multiview approach to tracking people in crowded scenes using a planar homography constraint. In: ECCV, pp. 98–109 (2006)
6. Kim, K., Davis, L.S.: Multi-camera tracking and segmentation of occluded people on ground plane using search-guided particle filtering. In: ECCV, pp. 98–109 (2006)
7. Kobayashi, Y., Sugimura, D., Sato, Y.: 3d head tracking using the particle filter with cascaded classifiers. In: BMVC (2006)
8. Mittal, A., Davis, L.S.: M2tracker: A multi-view approach to segmenting and tracking people in a cluttered scene. Internatial Journal of Computer Vision 51(3), 189–203 (2003)
9. Nummiaro, K., Koller-Meier, E., Svoboda, T., Roth, D., van Gool, L.: Color-based object tracking in multi-camera environment. In: 25th Pattern Recognition Symposium, DAGM (2003)
10. Pérez, P., Vermaak, J., Blake, A.: Data fusion for visual tracking with particles. Proceeding of the IEEE 92(3), 495–513 (2004)
11. Pérez, P., Hue, C., Vermaak, J., Gangnet, M.: Color-based probabilistic tracking. In: European Conference on Computer Vision, Copenhagen, Denmark, vol. 1, pp. 661–675 (2002)