IPSJ Transactions on Computer Vision and Applications Vol. 1 1–14 (Jan. 2009)

Regular Paper

Ground-Target Tracking in Multiple Cameras Using Collaborative Particle Filters and Principal Axis-Based Integration

Wei Du,[†1] Jean-Bearnard Hayet,[†2] Jacques Verly[†1] and Justus Piater[†1]

This paper presents a novel approach to tracking ground targets 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. Second, 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. The general framework of our approach allows us to track individual targets distributively and independently, which is of potential use in case that we are only interested in the trajectories of a few key targets and that we cannot track all the targets in the scene simultaneously.

1. Introduction

Tracking targets on the ground using multiple cameras is a basic task in many applications such as video surveillance and sports analysis. A commonly-used fusion strategy is to detect targets 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 [12),14),16),17),20)]. However, the automatic detection of targets entering and leaving the scene is itself a difficult problem. Moreover, 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 video-surveillance scenario. Top-down approaches are preferable in such situations.

In this paper, we present a novel top-down approach to ground-target 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 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 distinctive features. First, it doesn't require all targets to be tracked simultaneously. Instead of explicitly modeling the interactions between targets, we compute the consensus between cameras by having trackers in different cameras communicate. Second, it has a fully distributed architecture. All the computations are performed locally and only the target 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, which is much more precise than the direct fusion of targets' feet positions.

The rest of the paper is organized as follows. Section 2 introduces related work and highlights our contributions. Section 3 formulates the multi-camera tracking problem. Section 4 introduces the collaborative particle filters, including the principal axis-based integration and the boosted proposal functions. Experiments on both surveillance and soccer game scenarios are shown in Section 5.

†1 University of Liège, INTELSIG Laboratory, Department of Electrical Engineering and Computer Science, Belgium
†2 Centro de Investigación en Matemáticas (CIMAT), Mexico

This paper is an extension to a conference version which was published at ACCV'07 [11].

2. Related Work

Due to the plentiful advantages over single-camera tracking, multi-camera tracking is receiving growing attention in the field of computer vision. A popular class of approaches considers tracking with multiple cameras as a correspondence problem between tracks of targets seen from different viewpoints. When cameras are calibrated and a model of the site is available, it is possible to map targets in different cameras into a common world coordinate system, and the correspondence problem is to establish equivalence between targets at the same location. Collins et al. [7] developed a system for surveillance in the context of battlefield awareness using model-based geolocation. A simpler situation is when the 3D world degenerates into a known 2D ground plane. In this case, ground homographies have been widely used to do the image-to-scene mapping [2)–4),27)]. In case of simultaneously tracking of multiple targets, the geometric information is also used to reason about the occlusion situations between targets [12),14),16),17),20)].

Besides geometric locations, low-level features also measure the similarity of targets in different cameras. For instance, Orwell et al. [23] detected targets in each camera by background subtraction and matched them using color histograms. Cai et al. [5] established correspondences by matching a set of feature points with a Bayesian classification scheme. Krumm et al. [19] combined information in multiple stereo cameras; they performed background subtraction to detect human-shaped blobs in 3D, and used color histograms to identify targets. Mittal et al. [20] extended this work to multiple wide-baseline cameras; target segmentation and tracking were done by clustering points into 3D blobs using region-based stereo matching and volume intersection.

As no single feature is reliable enough for tracking in all cameras, the fusion of multiple features in the framework of Bayesian Networks was introduced by Chang et al. [6] and Dockstader et al. [8]. The first publication used Bayesian Networks to group features such as color, landmarks, location, and apparent height into targets, while the second tracked 2D semantic features in each camera and fused them by computing the confidence level of a camera using a Bayesian Network. Both publications handle occlusions by Bayesian Networks, which largely improves the robustness of the algorithms. However, when the disparity between cameras is large, both in location and in orientation, the reliability of feature matches is limited.

Due to their tremendous success in visual tracking, particle filters have also been adopted in multi-camera tracking [9),15)]. Most reported work performed particle filtering in 3D so that precise camera calibration is required to project particles into the image plane of each camera [18),21)]. The multi-camera information is often integrated by either the product of the likelihoods in all cameras [18)] or a selection of the best cameras that contain the most distinctive information [21)]. Both methods demand the collection and processing of the multi-camera observations at a central location, forming a centralized fusion framework.

Our approach is different from previous work in two main aspects. First, it is a top-down approach and allows individual targets to be tracked independently. Second, our approach has a decentralized architecture which is suitable for parallel implementation. The key is to let the cameras "talk" to each other so that each local tracker is able to take advantage of additional information in other cameras. In particular, a tree-structured graphical model is built to describe the relationships between target states in 3D and in different cameras. This model consists of a set of leaf nodes and a central node, and message passing between them enables the collaboration of the local trackers in all cameras. To get rid of the dependence on precise foot positions, we exploit the principal axes of the targets in the fusion of the multi-camera information, the intersections of which give better ground positions. At the same time, the fusion results in the ground plane are incorporated as proposal functions into each camera. This not only improves the tracking precision in each camera but also helps maintain consistencies between the local trackers. Experiments on sequences of video surveillance and soccer games show that acceptable performances can be achieved without explicitly modeling the interactions between targets.

3. Problem Formulation

Suppose L cameras are used and each camera collects one observation for a target at each time instant.

Denote the target state on the ground plane by $x_{t,0}$ and its states in different cameras by $x_{t,j}$, $j = 1, \ldots, L$. Here, we model the target in a camera as a rectangular region and the corresponding target in 3D as a cylinder. Therefore, $x_{t,0} = [u_{t,0}, v_{t,0}, h_{t,0}, w_{t,0}]$, where $[u_{t,0}, v_{t,0}]$ is the position of the target on the ground plane and $[h_{t,0}, w_{t,0}]$ is its 3D size. For simplicity, $[h_{t,0}, w_{t,0}]$ is often assumed fixed. Likewise, $x_{t,j} = [u_{t,j}, v_{t,j}, h_{t,j}, w_{t,j}]$, where $[u_{t,j}, v_{t,j}]$ is the position of the target in the image plane and $[h_{t,j}, w_{t,j}]$ its 2D size. Let $z_{t,j}$ denote 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.

According to the Bayes' rule, the recursive inference of the marginal posterior $p(x_{t,0}|Z^t)$ is formulated as

$$p(x_{t,0}|Z^t) \propto p(Z_t|x_{t,0}) \int p(x_{t,0}|x_{t-1,0}) p(x_{t-1,0}|Z^{t-1}) \mathrm{d}x_{t-1,0}, \tag{1}$$

where $p(Z_t|x_{t,0})$ is the multi-camera image likelihood function and $p(x_{t,0}|x_{t-1,0})$ is the prior state transition model.

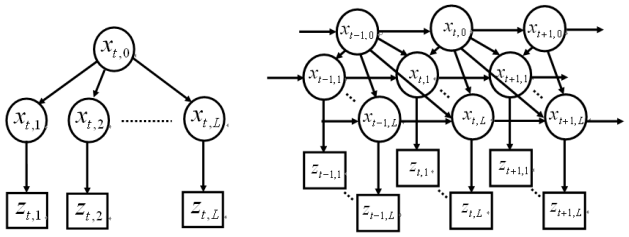
We assume that the observations in different cameras, $z_{t,j}$, $j = 1, \ldots, L$, are conditionally independent given $x_{t,0}$, i.e.

$$p(Z_t|x_{t,0}) = \prod_{j=1}^{L} p(z_{t,j}|x_{t,0}), \tag{2}$$

where $p(z_{t,j}|x_{t,0})$ is the image likelihood function in camera j. Substituting Eq. (2) into Eq. (1), we get the update equation for $p(x_{t,0}|Z^t)$,

$$p(x_{t,0}|Z^t) \propto \prod_{j=1}^{L} p(z_{t,j}|x_{t,0}) \int p(x_{t,0}|x_{t-1,0}) p(x_{t-1,0}|Z^{t-1}) \mathrm{d}x_{t-1,0}. \tag{3}$$

There are two reasons why direct inference using the above equations is intractable. First, a closed-form expression for Eq. (3) is only available in situations of linear state evolution models and Gaussian likelihood. Second, the multi-camera image observations are not produced by the target alone, but jointly by the target and scene clutter. Thus, $p(Z_t|x_{t,0})$ is easily affected by an inaccurate observation in one particular camera. For instance, if the target is occluded in one camera, the observation in this camera is either incomplete if the occlusion is introduced by clutter in the background or spurious if produced by similar objects. Consequently, $p(Z_t|x_{t,0})$ is biased by this inaccurate observation, even if accurate observations are available in other cameras. Therefore, we reformulate the problem of multi-camera tracking by explicitly modeling the dependencies between different cameras using graphical models.



(a) 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.

3.1 Model Description

Figure 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 not associated with any observation. One advantage of this model is that it is acyclic so that most inference algorithms such as belief propagation (BP) can produce the exact inference [31]. 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.

3.2 Collaborative Multi-Camera Tracking

Based on the dynamic Markov model in Fig. 1 (b), multi-camera tracking amounts to inferring each target state $x_{t,i}$, $i = 0, \ldots, L$, based on all observations Z^t collaboratively. A message passing scheme, the same as is used in BP, is adopted to pass messages from each camera to the ground plane. The local message from camera j is defined as

$$m_{0j}(x_{t,0}) \leftarrow \int p_j(z_{t,j}|x_{t,j})\psi^t_{0,j}(x_{t,0}, x_{t,j}) \int p(x_{t,j}|x_{t-1,j})p(x_{t-1,j}|Z^{t-1})\mathrm{d}x_{t-1,j}\mathrm{d}x_{t,j}, \tag{4}$$

where $\psi^t_{0,j}(x_{t,0}, x_{t,j})$ is the potential function that models the stochastic relation between $x_{t,0}$ and $x_{t,j}$. The belief in the ground plane $p(x_{t,0}|Z^t)$ is computed recursively by the message product and the propagation of the previous posterior,

$$p(x_{t,0}|Z^t) \propto \prod_{j=1,\ldots,L} m_{0j}(x_{t,0}) \int p(x_{t,0}|x_{t-1,0})p(x_{t-1,0}|Z^{t-1})\mathrm{d}x_{t-1,0}. \tag{5}$$

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 in camera j $p(x_{t,j}|Z^t)$ is computed as

$$p(x_{t,j}|Z^t) \propto p_j(z_{t,j}|x_{t,j}) \int\int p(x_{t,j}|x_{t-1,j})p(x_{t-1,j}|Z^{t-1}) \underline{p(x_{t,j}|x_{t-1,0})p(x_{t-1,0}|Z^{t-1})}\mathrm{d}x_{t-1,0}\mathrm{d}x_{t-1,j}. \tag{6}$$

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.

The above formulation shows that our algorithm involves both particle filters for propagating marginal posteriors over time, and a message passing scheme for having the particle filters collaborate. The inference based on this formulation is very efficient due to the simplicity of the graphical models used in Fig. 1.

4. 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. Here, we consider $x_{t,0}$ to be a 2D position in the ground plane, whereas $x_{t,j}$ is a vector corresponding to the bounding box of the tracked target in camera j.

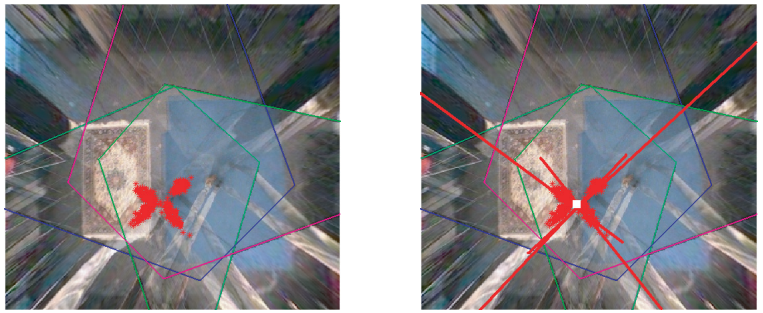
Fig. 2 The particle distributions in four cameras at a time instant. It can be seen that the foot positions are not precise although all the particles are placed at the right locations.

4.1 Principal Axis-based Integration

The ground-plane particle filter integrates the multi-camera information according to Eqs. (4) and (5). 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, the sizes of the targets cannot usually 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. Here, it is simply the vertical line in the middle of a rectangle associated to a particle. 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 [14),17)], illustrated in **Figs. 2** and **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, a prior motion model $p(x_{t,0}|x_{t-1,0})$ 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. (4). The principal axes are incorporated in the potential function $\psi^t_{0,j}(x_{t,0}, x_{t,j})$ in Eq. (4). 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



(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).

distances of the ground particles to these projected principal axes and converts them to probability densities, given by

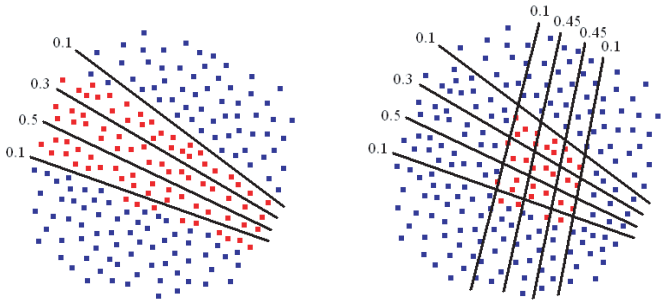
$$\psi^t_{0,j}(x^n_{t,0}, x^m_{t,j}) \propto \exp(-\mathrm{dist}^2(x^n_{t,0}, \mathrm{project}(H_j, x^m_{t,j}))), \tag{7}$$

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

$$w^{j,n}_{t,0} \propto \sum_{m=0}^{N} \pi^m_{t,j} \psi^t_{0,j}(x^n_{t,0}, x^m_{t,j}), \qquad \pi^n_{t,0} \propto \prod_{j=1}^{L} w^{j,n}_{t,0}, \tag{8}$$

where $w^{j,n}_{t,0}$ is the message weight of $x^n_{t,0}$ from camera j, and $\pi^n_{t,0}$ and $\pi^m_{t,j}$ are the belief weights of $x^n_{t,0}$ and $x^m_{t,j}$. Intuitively, the closer a ground-plane particle is to all the principal axes, the larger its weight is, as illustrated in **Fig. 4**.

Note that a target is tracked in the ground plane in the same way as in each camera, although there are no image observations. The ground-plane particles are evaluated by the incoming messages from the cameras. There are two reasons why we don't directly use the intersections of the principal axes as the fusion results. First, computing these intersections is computationally intensive, with a complexity of $O(N^L)$, where N is the number of particles in each camera and L is the number of cameras. Our approach is much more efficient with a complexity of only $O(LN^2)$. Second, by using an individual tracker in the ground plane, we can incorporate a prior motion model to enforce smooth motion.



(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.

4.2 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 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 [10]. 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,

$$p(x_{t,j}|x_{t-1,j}, x_{t-1,0}) \propto \alpha p(x_{t,j}|x_{t-1,j}) + (1-\alpha)p(x_{t,j}|x_{t-1,0}). \tag{9}$$

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. The sizes of the targets can be estimated when the full camera calibrations are available or assumed to be a Gaussian distribution with a mean of the sizes at the previous time instant. For the latter case, the particles from $p(x_{t,j}|x_{t-1,0})$ are moved up by a distance d to avoid the problem of the imprecise foot positions, where d is learned from the previous time instant.

In a sense, the fusion results at $t-1$ are used as boosted proposal functions by each camera [22]. 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.

4.3 Observation Model

The tracking algorithm requires an observation model in each camera for computing the likelihood $p_j(z_{t,j}|x_{t,j})$, $j = 1, \ldots, L$. Following Pérez et al. [25], a classical observation model based on Hue-Saturation-Value (HSV) color histograms is adopted due to the ease of implementation and the advantage of being insensitive to illumination effects.

Thus, in camera j at time t, the color model of the target of interest $\mathbf{q}_{t,j}(s_{t,j}^{(n)}) = \{q_{t,j}(b; s_{t,j}^{(n)})\}_{b=1}^{B}$ associated with a sampled particle $s_{t,j}^{(n)}$ is obtained by a histogramming technique, which assigns a probability to each of the B color bins. This model is compared to a previously-learned reference color model $\mathbf{q}_j^* = \{q_j^*(b)\}_{b=1}^{B}$, and the Bhattacharyya distance is computed to measure the similarity,

$$D[\mathbf{q}_j^*, \mathbf{q}_{t,j}(s_{t,j}^{(n)})] = \left[1 - \sum_{b=1}^{B} \sqrt{q_j^*(b), q_{t,j}(b; s_{t,j})}\right]^{\frac{1}{2}}. \tag{10}$$

Once the distance between the color histograms is computed, we use the image likelihood function

$$p_j(z_{t,j}|x_{t,j} = s_{t,j}^{(n)}) \propto \exp\left(-\lambda D^2[\mathbf{q}_j^*, \mathbf{q}_{t,j}(s_{t,j}^{(n)})]\right), \tag{11}$$

where λ is fixed to 20 in all our experiments. As for the bin numbers, the default setting is 10 for all the Hue, Saturation, and Value channels.

To model the spatial layout of the color distribution, a multi-part color model is obtained by splitting the tracked region into subregions, each with an individual color model. In our work, a two-part, upper body and lower body, color model is adopted for modeling pedestrians in video-surveillance scenarios and players in soccer-broadcasting scenarios. The effectiveness of this method has repeatedly been demonstrated [22),25)] and is confirmed by this work.

To speed up tracking, the above procedure was implemented using integral histograms [26]. The construction of the integral histograms in an arbitrary rectangular region demands the construction of an integral image for each bin, at a cost proportional to the size of the region times the number of bins. After the construction of the integral histograms, the evaluation of a particle is nearly as cheap as the comparison of two histograms, which is very efficient [26].

4.4 Discussion

In multi-camera tracking, a big problem is how to deal with inconsistent observations due to e.g. occlusions in some of the cameras. To overcome this problem, many approaches assess the tracking performance to obtain a confidence weight for each camera [1),2),4),13),27)]. As the confidence weights are computed according to the quality of the estimates, the risk of failure is high if the targets being tracked are approached by other, similar objects. Our approach is different from this strategy in that each camera is treated equally with no explicit preference. However, as pointed out by Sun, et al. [29], the asymmetric message passing mechanism guarantees that the information is propagated mainly from high-confidence cameras to low-confidence cameras due to the smaller entropy of the messages in this direction. An extreme example is when a target is completely occluded in a camera. Then, this camera still "contributes" by propagating mostly uniformly distributed beliefs. Although this camera is not informative, it will not affect the tracking results at other cameras. Consequently, the propagation of incorrect information is avoided.

One advantage of our approach is that it has a fully distributed architecture. Image observations in different cameras are processed locally and only messages are exchanged between the ground plane and each camera. Moreover, as a byproduct, it also allows different observation models to be used in different

cameras to better characterize varying properties of the targets across views. In a sense, each camera is used as a black box and only exchanges messages with the ground plane. It doesn't have to know what kinds of observation models are being used in other cameras and how they are implemented. This is beneficial from an implementation point.

5. Results

We tested our method on both video-surveillance and soccer-broadcasting sequences. In all experiments, we manually initialized the targets of interest in the first frame of each sequence and learned the reference color models. The reference models were updated gradually with exponentially forgetting the past models, or were kept unchanged when dramatic changes to the models indicated occlusions [30]. 100 particles were sampled for each particle filter.

In the first experiment, we tested our approach on challenging video-surveillance sequences of multiple people walking indoors. The sequences were taken by four cameras separated by wide baselines and contain heavy occlusions. The cameras are only partially calibrated with ground and head homographies available. The ground homography refers to the common homography from the image plane to the ground plane, while the head homography refers to the homography from the image plane to the head plane, i.e., the plane parallel to the ground plane but 1.75 m higher [⋆1]. For the first two cameras, both ground and head homographies are available. For the other two, the head planes happen to be orthogonal to the image plane and there are thus no head homographies given. However, the height where the head plane is located in the image view is provided for them. All the calibration information was exploited to determine not only the foot positions but also the head positions of the targets.

Figure 5 shows the results of selectively tracking three people in the crowd. **Figure 6** shows the tracking results of one person on the ground plane. In this experiment, static cameras with narrow viewing angles are used, making the targets quite large in the images. Thus, a more complicated multi-cue observation model is adopted. We first construct a hierarchical color model that further splits the upper body and lower body into subregions and continues to split until the subregions become small enough. A cascade comparison with a reference hierarchical model is then performed which halts when the similarity score is smaller than a threshold at a level of the hierarchy. After a particle is evaluated by the color likelihood function, background subtraction is then used to attract the particle to the regions with moving objects. In a way, the color and motion information is used in turn to evaluate the particles [24], shown in **Fig. 7**.

Experimental comparisons are difficult to perform because our method tracks targets independently, whereas most other approaches have to track all targets simultaneously even if we are only interested in one particular target. Two possible alternatives are a traditional particle filter that computes the products of the likelihoods in all cameras [18] and a fusion module that integrates results by individual particle filters in the cameras. Since only partial camera calibrations are available, the former method requires very precise foot positions to be detected. For the latter, the lack of the feedback from the fusion module may cause inconsistency between the local trackers in the cameras. Thus, we only compared with these two methods which are referred to as *likelihood product* and *centralized fusion* respectively, shown in **Fig. 8**. *Likelihood product* needs camera calibrations to project particles from 3D to each camera. Here, we used both the ground and head homographies to reason about the height of the target in each image. As expected, the method is sensitive to the precisions of the foot and head positions of the targets and fails quickly during the first big occlusion. For *centralized fusion*, we ran standard particle filters in each camera and computed the intersections of the principal axes of their estimates using both the ground and head homographies. Due to the lack of the collaboration between cameras, the local trackers work separately and fail one by one when occlusions occur.

In the second experiment, we tracked several soccer players in three Pan-Tilt-Zoom cameras, shown in **Fig. 9**. The ground homography for each camera was obtained online either by using known features such as border lines on the field where enough of them are visible, e.g., near the penalty area, or by accumulating small estimates of motion between consecutive frames where no or few features are visible [13]. Although we don't explicitly model the interactions between play-

⋆1 It is assumed that all people have a fixed height of 1.75 m. This fixed-height assumption significantly improved the robustness of the tracking algorithm.

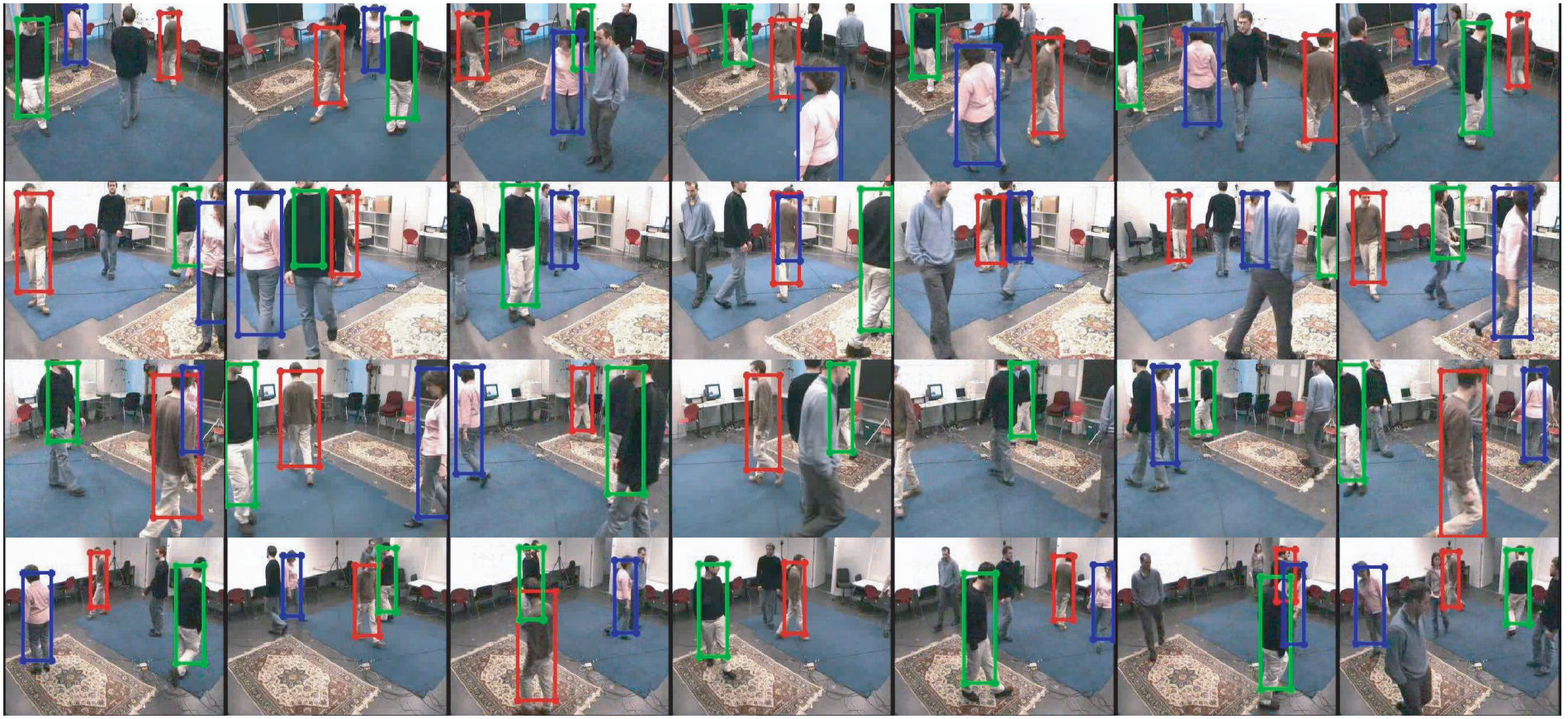


Fig. 5 Results of tracking three people in an indoor environment with four cameras. Each column shows four simultaneous views.

ers, the feedback from the ground tracker to each camera compensates for the occlusions in some of the cameras. However, at a point, due to a heavy occlusion that occurs in all cameras, a tracker jumps from one target to another. Such failures are expected especially when the total occlusions come from similar objects such as teammate players. Nevertheless, the feedback from the ground tracker enforces the consistency between the local trackers, even if they collectively follow the wrong target. **Figure 10** shows the particle distributions at the time instant when the jump begins. This problem can be partially solved by tracking multiple targets simultaneously.

6. Conclusions 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

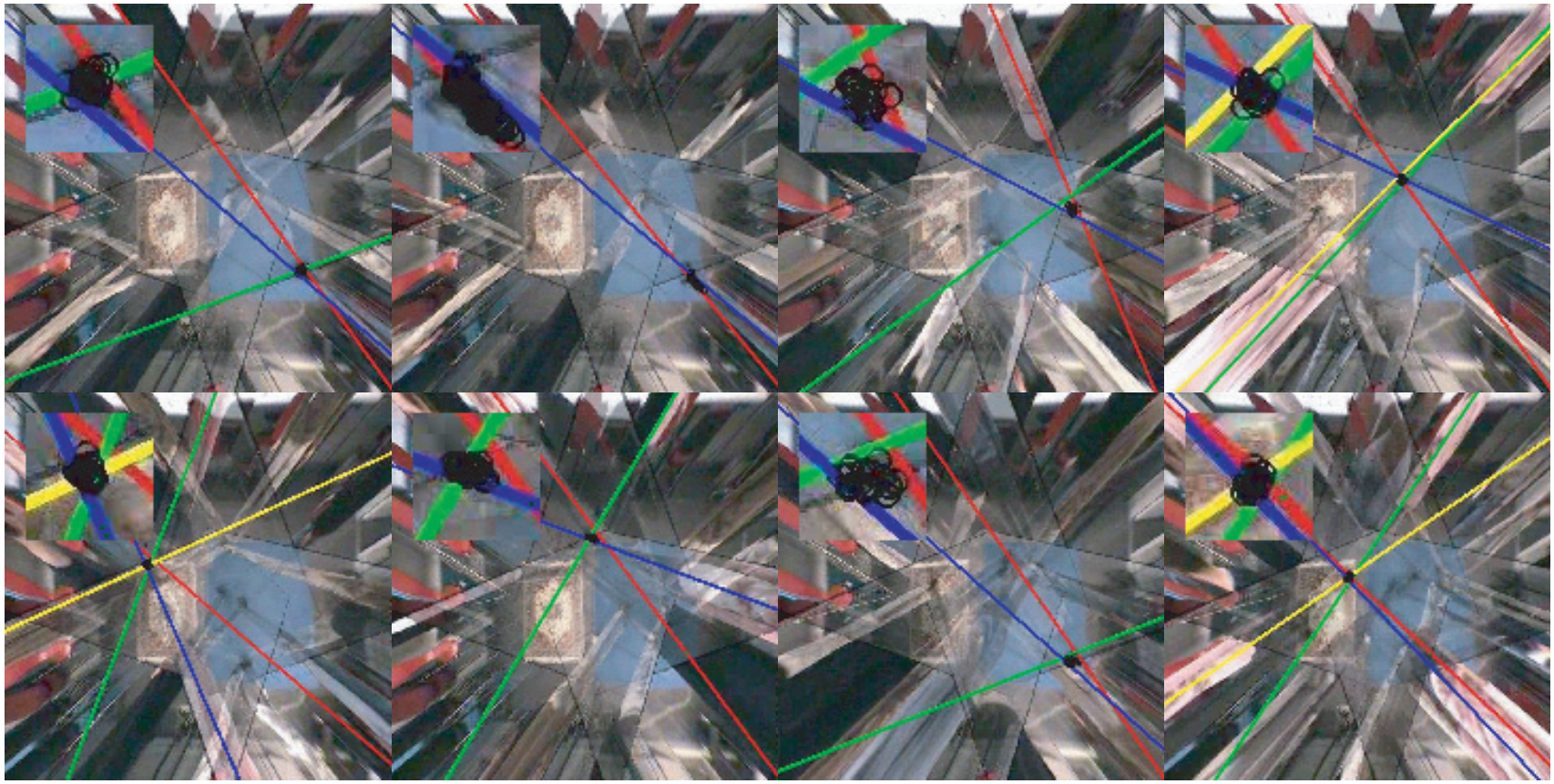


Fig. 6 Results of tracking one person on the ground plane. The colored lines are the principal axes of the estimates in the cameras. The black dots are the ground particles.

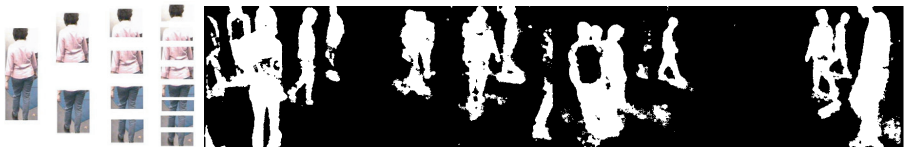


Fig. 7 The left figure shows the hierarchical color model and the right figure shows the result of background subtraction based on a Gaussian-mixture model [28].

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,

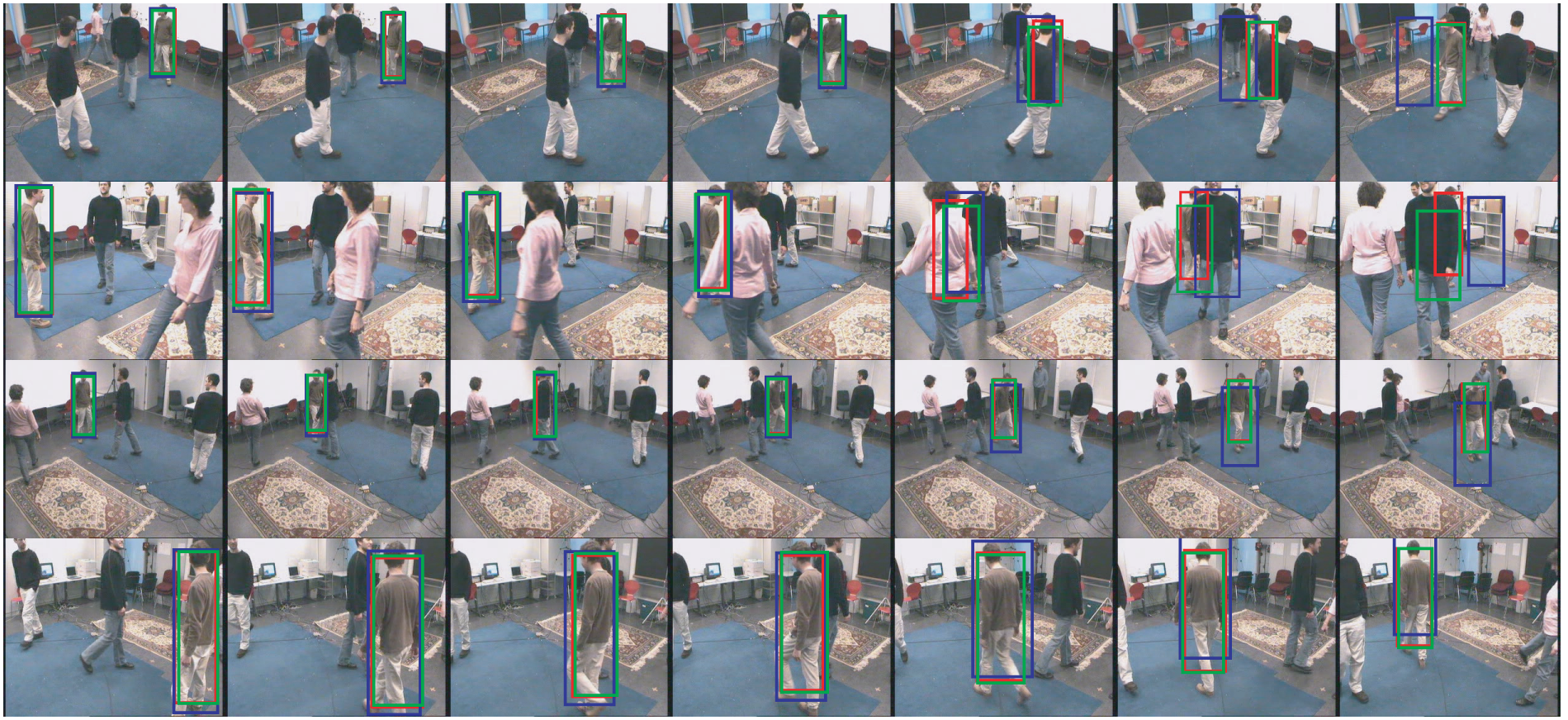


Fig. 8 Comparisons of our approach with *likelihood product* and *centralized fusion*. The red rectangles are results by our approach, the blue ones by *likelihood product*, and the green ones by *centralized fusion*. All methods are able to handle short occlusions in one camera, shown in the third column. However, as *centralized fusion* consists of individual particle filters that work separately, it cannot handle dramatic occlusions such as the one at the end of the second row. As a result, individual trackers in different cameras produce inconsistent results. On the other hand, *likelihood product* depends too much on the precision of the foot and head positions, and therefore is also sensitive to dramatic occlusions. In the fifth column, the target is occluded in two cameras, introducing large uncertainties to the particle filter in the ground plane. Our approach integrates the advantages of both *likelihood product* and *centralized fusion* in that individual trackers are deployed in different cameras as well as in the ground plane while likelihoods are combined everywhere.

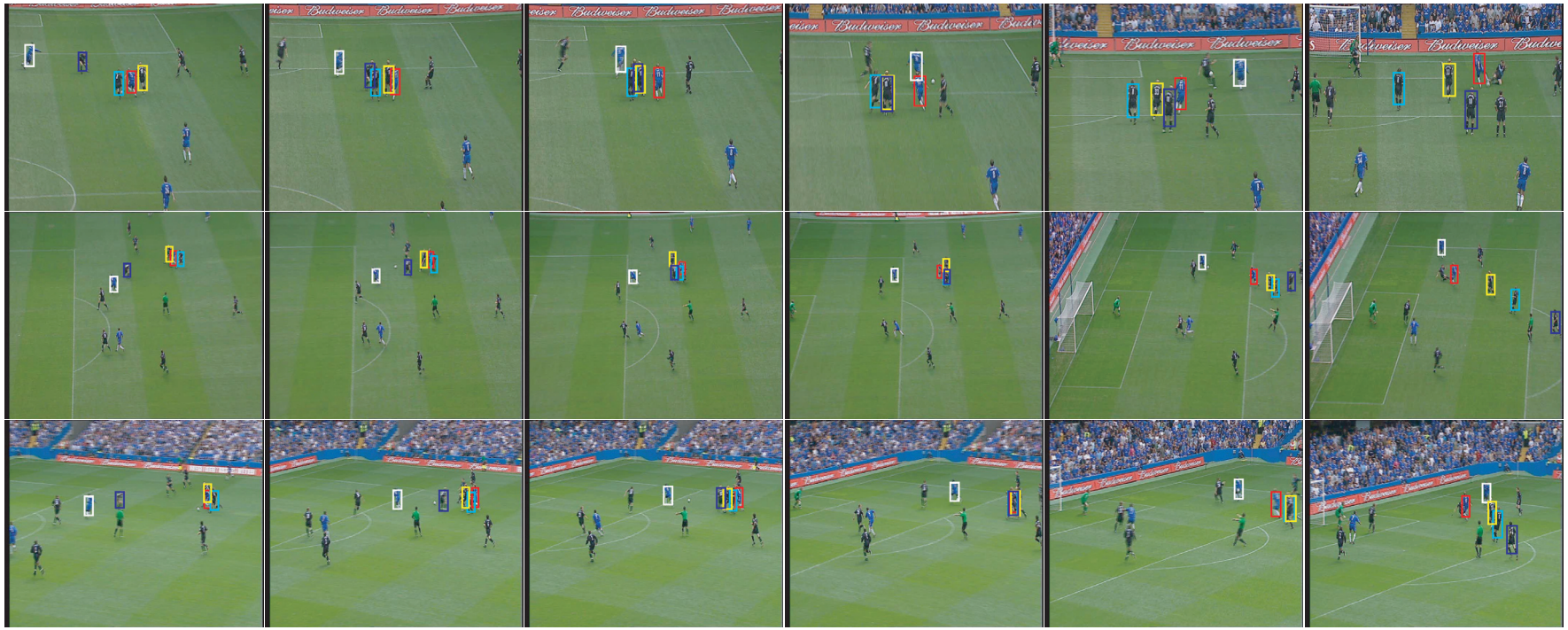


Fig. 9 Results of tracking several soccer players in the last frames of the three sequences. The switch between the blue and cyan trackers occurs at the fourth column.

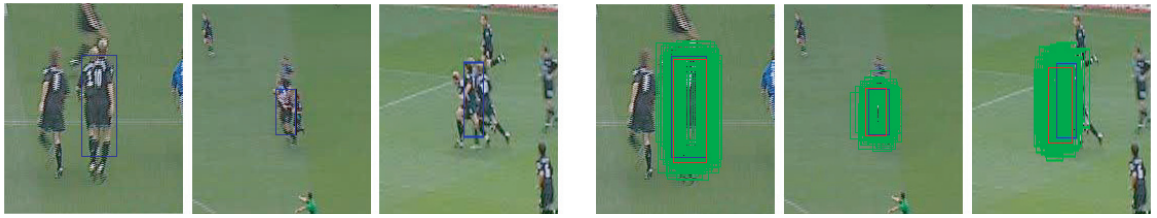


Fig. 10 The particle distributions at the time when the tracker is about to jump to a different player, which happens here because the players involved are very close both in space and in appearance in all three views. The green rectangles are the sampled particles, the blue ones are the estimates, and the red ones are the predictions of the fusion results at the previous time.

which involves the modeling of the target interactions and data association across cameras.

Acknowledgments The authors wish to thank Jérôme Berclaz and François Fleuret for providing the data in the first experiment.

References

1) Bar-Shalom, Y. and Li, X.R.: *Multitarget-Multisensor Tracking: Principles and Techniques*, YBS Publishing (1995).
2) Beugnon, C., Singh, T., Llinas, J. and Saha, R.K.: Adaptive track fusion in a multisensor environment, *the Third International Conference on Information Fusion*, pp.24–31, Paris, France (2000).
3) Black, J. and Ellis, T.: Multi camera image tracking, *Image and Vision Computing*, Vol.24, No.11, pp.1256–1267 (2006).
4) Borg, M., Brémond, F., Ferryman, J., Fusier, F., Thirde, D., Thonnat, M. and Valentin, V.: Video surveillance for aircraft activity monitoring, *International Conference on Advanced Video and Signal based Surveillance*, pp.16–21, Como, Italy (2005).
5) Cai, Q. and Aggarwal, J.K.: Automatic tracking of human motion in indoor scenes across multiple synchronized video streams, *International Conference on Computer Vision*, pp.356–262, Bombay, India (1998).
6) Chang, T-H., Gong, S. and Ong, E-J.: Tracking multiple people under occlusion with multiple cameras, *British Machine Vision Conference*, pp.566–575, Bristol, UK (2000).
7) Collins, R.T., Lipton, A.J., Fujiyoshi, H. and Kanade, T.: Algorithms for cooperative multisensor surveillance, *Proceeding of the IEEE*, Vol.89, No.10, pp.1456–1477 (2001).
8) Dockstader, S.L. and Tekalp, A.M.: Multiple camera fusion for multi-object tracking, *IEEE Workshop on Multi-Object Tracking*, pp.95–102, Vancouver, Canada (2001).
9) Doucet, A., deFreitas, N. and Gordon, N.: *Sequential Monte Carlo Methods in Practice*, Springer-Verlag, New York (2001).
10) Du, W. and Piater, J.: Multi-view object tracking using sequential belief propagation, *Asian Conference on Computer Vision*, pp.684–693, Hyderabad, India (2006).
11) Du, W. and Piater, J.: Multi-camera people tracking by collaborative particle filters and principal axis-based integration, *Asian Conference on Computer Vision*, pp.365–374, Tokyo, Japan (2007).
12) Fleuret, F., Berclaz, J., Lengagne, R. and Fua, P.: Multi-camera people tracking with a probabilistic occupancy map, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.30, No.2, pp.267–282 (2008).
13) Hayet, J.-B., Mathes, T., Czyk, J., Piater, J., Verly, J. and Macq, B.: A modular multi-camera framework for team sports tracking, *International Conference on Advanced Video and Signal based Surveillance*, pp.493–498, Como, Italy (2005).
14) Hu, W.-M., Hu, M., Zhou, X., Tan, T.-N., Lou, J. and Maybank, S.J.: Principal axis-based correspondence between multiple cameras for people tracking, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.27, No.4, pp.663–671 (2006).
15) Isard, M. and Blake, A.: Condensation – conditional density propagation for visual tracking, *International Journal of Computer Vision*, Vol.29, No.2, pp.5–28 (1998).
16) Khan, S.M. and Shah, M.: A multiview approach to tracking people in crowded scenes using a planar homography constraint, *European Conference on Computer Vision*, pp.98–109, Graz, Austria (2006).
17) Kim, K. and Davis, L.S.: Multi-camera tracking and segmentation of occluded people on ground plane using search-guided particle filtering, *European Conference on Computer Vision*, pp.98–109, Graz, Austria (2006).
18) Kobayashi, Y., Sugimura, D. and Sato, Y.: 3D head tracking using the particle filter with cascaded classifiers, *British Machine Vision Conference*, pp.37–46, Edinburgh, UK (2006).
19) Krumm, J., Harris, S., Meyersand, B., Brumitt, B., Hale, M. and Shafer, S.: Multi-camera multi-person tracking for Easyliving, *IEEE Workshop on Visual Surveillance*, pp.3–10, Dublin, Ireland (2000).
20) Mittal, A. and Davis, L.S.: M2tracker: A multi-view approach to segmenting and tracking people in a cluttered scene, *Internatial Journal of Computer Vision*, Vol.51, No.3, pp.189–203 (2003).
21) Nummiaro, K., Koller-Meier, E., Svoboda, T., Roth, D. and VanGool, L.: Color-based object tracking in multi-camera environment, *25th Pattern Recognition Symposium, DAGM*, pp.591–599, Magdeburg, Germany (2003).
22) Okuma, K., Taleghani, A., deFreitas, N., Little, J.J. and Lowe, D.G.: A boosted particle filter: multitarget detection and tracking, *European Conference on Computer Vision*, pp.28–39, Prague, Czech Republic (2004).
23) Orwell, J., Remagnino, P. and Jones, G.A.: Multi-camera colour tracking, *IEEE Workshop on Visual Surveillance*, pp.14–21, Fort Collins, CO (1999).
24) Pérez, P., Vermaak, J. and Blake, A.: Data fusion for visual tracking with particles, *Proceeding of the IEEE*, Vol.92, No.3, pp.495–513 (2004).
25) Pérez, P., Hue, C., Vermaak, J. and Gangnet, M.: Color-based probabilistic tracking, *European Conference on Computer Vision*, Vol.1, pp.661–675, Copenhagen, Denmark (2002).
26) Porkili, F.: Integral histogram: A fast way to extract histograms in cartesian spaces, *IEEE Conference on Compute Vision and Pattern Recognition*, pp.829–836, San Diego, CA (2005).
27) Snidaro, L., Foresti, G.L., Niu, F. and Varshney, P.K.: Sensor fusion for video surveillance, *the Seventh International Conference on Information Fusion*, pp.739–

746, Stockholm, Sweden (2004).
28) Stauffer, C. and Grimson, W.E.L.: Adaptive background mixture models for real-time tracking, *IEEE Conference on Computer Vision and Pattern Recognition*, Vol.2, pp.246–252, Fort Collins, CO (1999).
29) Sun, J., Zheng, N. and Harry, S.: Stereo matching using belief propagation, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.25, No.7, pp.787–800 (2003).
30) Wang, H., Suter, D. and Schindler, K.: Effective appearance model and similarity measure for particle filtering and visual tracking, *European Conference on Computer Vision*, pp.606–618, Graz, Austria (2006).
31) Yedidia, J.S., Freeman, W.T. and Weiss, Y.: Generalized belief propagation, *Advances in Neural Information Processing Systems* (*NIPS*), Vol.13, pp.689–695 (2000).

(Received February 29, 2008)
(Accepted September 25, 2008)
(Released January 30, 2009)

(Communicated by *Yasushi Yagi*)

Wei Du received his Ph.D. from Institute of Computing Technology, Chinese Sciences of Academy, Beijing, China, in 2002. He's been working, as a postdoctoral researcher, on developing robust tracking systems with INRIA, Rocquencourt, France, University of Hamburg, Germany, and now, University of Liège, Belgium. His current research interests are the architecture of distributed tracking systems using multiple cameras and multiple cues.

Jean-Bernard Hayet graduated from Ecole Nationale Superieure de Techniques Avancees in Paris in 1999. He got his Master degree from Université Pierre et Marie Curie, Paris VI and his Ph.D. degree from Université Toulouse III following his doctoral work at LAAS-CNRS, in 2003. After a post-doctoral stay at Université de Liège in Belgium, he joined the Research Center in Mathematics (CIMAT) in Guanajuato, Mexico in 2007. His interest include mobile robot navigation and computer vision, and more precisely the use of local features and scene geometry, from landmark navigation to on-line planar scene rectification.

Jacques Verly received the Ingénieur Electronicien degree from the University of Liège, Belgium, in 1975. Through a sponsorship of the Belgian American Educational Foundation (BAEF), he attended Stanford University, Stanford, Calif., where he received the M.S. and Ph.D. degrees in electrical engineering in 1976 and 1980, respectively. From 1980 to 2000, he was at MIT Lincoln Laboratory, Lexington, Mass, where he carried out research in several different areas, including image processing and computer vision for a variety of imaging sensors, such as visible, laser radar, fully polarimetric SAR, and IR. Since 2000, he has been a Professor in the Department of Electrical Engineering and Computer Science (also known as the Institut Montefiore) of the University of Liège, Belgium. He is a Founder of the Signal and Image Exploitation Laboratory (INTELSIG). His current research interests are principally in medical imaging (image-guided surgery), radar signal processing (space-time adaptive processing), and object tracking in video streams (for video surveillance and sports analysis). He has about 200 publications and 2 US patents. He is a CRB Fellow of the Belgian American Educational Foundation.

Justus Piater graduated with highest honors from the University of Magdeburg, Germany. He was awarded a Fulbright scholarship and obtained his Ph.D. in computer science at the University of Massachusetts Amherst, USA, in 2001. A recipient of a European Marie-Curie Individual Fellowship, he was a postdoctoral researcher at INRIA Rhône-Alpes, France, from 2000 to 2002. He currently is a professor of computer science at the University of Liège, Belgium, where he directs the Computer Vision research group. His research interests include computer vision and machine learning, with a focus on visual learning, closed-loop interactive vision, and video analysis. He has published about 70 papers at international conferences and journals.