Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2007, Article ID 38373, 15 pages
doi:10.1155/2007/38373

Research Article

Distributed Bayesian Multiple-Target Tracking in Crowded Environments Using Multiple Collaborative Cameras

Wei Qu,[1] Dan Schonfeld,[1] and Magdi Mohamed[2]

[1] *Multimedia Communications Laboratory, Department of Electrical and Computer Engineering, University of Illinois at Chicago, IL 60607-7053, USA*
[2] *Visual Communication and Display Technologies Lab, Physical Realization Research COE, Motorola Labs, Schaumburg, IL 60196, USA*

Received 28 September 2005; Revised 13 March 2006; Accepted 15 March 2006

Recommended by Justus Piater

Multiple-target tracking has received tremendous attention due to its wide practical applicability in video processing and analysis applications. Most existing techniques, however, suffer from the well-known "*multitarget occlusion*" problem and/or immense computational cost due to its use of high-dimensional joint-state representations. In this paper, we present a distributed Bayesian framework using multiple collaborative cameras for robust and efficient multiple-target tracking in crowded environments with significant and persistent occlusion. When the targets are in close proximity or present multitarget occlusions in a particular camera view, camera collaboration between different views is activated in order to handle the multitarget occlusion problem in an innovative way. Specifically, we propose to model the camera collaboration likelihood density by using epipolar geometry with sequential Monte Carlo implementation. Experimental results have been demonstrated for both synthetic and real-world video data.

1. INTRODUCTION AND RELATED WORK

Visual multiple-target tracking (MTT) has received tremendous attention in the video processing community due to its numerous potential applications in important tasks such as video surveillance, human activity analysis, traffic monitoring, and so forth. MTT for targets whose appearance is distinctive is much easier since it can be solved reasonably well by using multiple independent single-target trackers. In this situation, when tracking a specific target, all the other targets can be viewed as background due to their distinct appearance. However, MTT for targets whose appearance is similar or identical such as pedestrians in crowded scenes is a much more difficult task. In addition to all of the challenging problems inherent in single-target tracking, MTT must deal with multitarget occlusion, namely, the tracker must separate the targets and assign them correct labels.

Most early efforts for MTT use monocular video. A widely accepted approach that addresses many problems in this difficult task is based on a joint state-space representation and infers the joint data association [1, 2]. MacCormick and Blake [3] used a binary variable to identify foreground objects and proposed a probabilistic exclusion principle to penalize the hypothesis where two objects occlude. In [4], the likelihood is calculated by enumerating all possible association hypotheses. Isard and MacCormick [5] combined a multiblob likelihood function with the condensation filter and used a 3D object model providing depth ordering to solve the multitarget occlusion problem. Zhao and Nevatia [6, 7] used a different 3D shape model and joint likelihood for multiple human segmentation and tracking. Tao et al. [8] proposed a sampling-based multiple-target tracking method using background subtraction. Khan et al. [9] proposed an MCMC-based particle filter which uses a Markov random field to model motion interaction. Smith et al. [10] presented a different MCMC-based particle filter to estimate the multiobject configuration. McKenna et al. [11] presented a color-based system for tracking groups of people. Adaptive color models are used to provide qualitative estimates of depth ordering during occlusion. Although the above solutions, which are based on a centralized process, can handle the problem of multitarget occlusion in principle, they require a tremendous computational cost due to the complexity introduced by the high dimensionality of the joint-state representation which grows exponentially in terms of the number of objects tracked. Several researchers

proposed decentralized solutions for multitarget tracking. Yu and Wu [12] and Wu et al. [13] used multiple collaborative trackers for MTT modeled by a Markov random network. This approach demonstrates the efficiency of the decentralized method. However, it relies on the objects' joint prior and does not deal with the "*false labeling*" problem. The decentralized approach was carried further by Qu et al. [14] who proposed an interactively distributed multiobject tracking (IDMOT) framework using a magnetic-inertia potential model.

Monocular video has intrinsic limitations for MTT, especially in solving the multitarget occlusion problem, due to the camera's limited field of view and loss of the targets' depth information by camera projection. These limitations have recently inspired researchers to exploit multiocular videos, where expanded coverage of the environment is provided and occluded targets in one camera view may not be occluded in others. However, using multiple cameras raises many additional challenges. The most critical difficulties presented by multicamera tracking are to establish a consistent-label correspondence of the same target among the different views and to integrate the information from different camera views for tracking that is robust to significant and persistent occlusion. Many existing approaches address the label correspondence problem by using different techniques such as feature matching [15, 16], camera calibration and/or 3D environment model [17–19], and motion-trajectory alignment [20]. Khan and Shah [21] proposed to solve the consistent-labeling problem by finding the limits of the field of view of each camera as visible in the other cameras. Methods for establishing temporal instead of spatial label correspondences between nonoverlapping fields of view are discussed in [22–24]. Most examples of MTT presented in the literature are limited to a small number of targets and do not attempt to solve the multitarget occlusion problem which occurs frequently in crowded scenes. Integration of information from multiple cameras to solve the multitarget occlusion problem has been approached by several researchers. Static and active cameras are used together in [25]. Chang and Gong [26] used Bayesian networks to combine multiple modalities for matching subjects. Iwase and Saito [27] integrated the tracking data of soccer players from multiple cameras by using homography and a virtual ground image. Mittal and Davis [28] proposed to detect and track multiple objects by matching regions along epipolar lines in camera pairs. A particle filter-based approach is presented by Gatica-Perez et al. [29] for tracking multiple interacting people in meeting rooms. Nummiaro et al. [30] proposed a color-based object tracking approach with a particle filter implementation in multicamera environments. Recently, Du and Piater [31] presented a very efficient algorithm using sequential belief propagation to integrate multiview information for a single object in order to solve the problem of occlusion with clutter. Several researchers addressed the problem of 3D tracking of multiple objects using multiple camera views [32, 33]. Dockstader and Tekalp [34] used a Bayesian belief network in a central processor to fuse independent observations from multiple cameras for 3D position tracking. A different central process is used to integrate data for football player tracking in [35].

In this paper, we present a distributed Bayesian framework for multiple-target tracking using multiple collaborative cameras. We refer to this approach as Bayesian multiple-camera tracking (BMCT). Its objective is to provide a superior solution to the multitarget occlusion problem by exploiting the cooperation of multiocular videos. The distributed Bayesian framework avoids the computational complexity inherent in centralized methods that rely on joint-state representation and joint data association. Moreover, we present a paradigm for a multiple-camera collaboration model using epipolar geometry to estimate the *camera collaboration function* efficiently without recovering the targets' 3D coordinates.

The paper is organized as follows: Section 2 presents the proposed BMCT framework. Its implementation using the density estimation models of sequential Monte Carlo is discussed in Section 3. In Section 4, we provide experimental results for synthetic and real-world video sequences. Finally, in Section 5, we present a brief summary.

2. DISTRIBUTED BAYESIAN MULTIPLE-TARGET TRACKING USING MULTIPLE COLLABORATIVE CAMERAS

We use multiple trackers, one tracker per target in each camera view for MTT in multiocular videos. Although we illustrate our framework by using only two cameras for simplicity, the method can be easily generalized to cases using more cameras. The state of a target in camera A is denoted by $x_t^{A,i}$, where $i = 1,\ldots,M$ is the index of targets, and t is the time index. We denote the image observation of $x_t^{A,i}$ by $z_t^{A,i}$, the set of all states up to time t by $x_{0:t}^{A,i}$, where $x_0^{A,i}$ is the initialization prior, and the set of all observations up to time t by $z_{1:t}^{A,i}$. Similarly, we can denote the corresponding notions for targets in camera B. For instance, the "*counterpart*" of $x_t^{A,i}$ is $x_t^{B,i}$. We further use z_t^{A,J_t} to denote the neighboring observations of $z_t^{A,i}$, which "*interact*" with $z_t^{A,i}$ at time t, where $J_t = \{j_{l_1}, j_{l_2},\ldots\}$. We define a target to have "*interaction*" when it touches or even occludes with other targets in a camera view. The elements $j_{l_1}, j_{l_2},\ldots \in \{1,\ldots,M\}$, $j_{l_1}, j_{l_2},\ldots \neq i$, are the indexes of targets whose observations interact with $z_t^{A,i}$. When there is no interaction of $z_t^{A,i}$ with other observations at time t, $J_t = \varnothing$. Since the interaction structure among observations is changing, J may vary in time. In addition, $z_{1:t}^{A,J_{1:t}}$ represents the collection of neighboring observation sets up to time t.

2.1. Dynamic graphical modeling and conditional independence properties

The *graphical model* [36] is an intuitive and convenient tool to model and analyze complex dynamic systems. We illustrate the dynamic graphical model of two consecutive frames for multiple targets in two collaborative cameras in Figure 1. Each camera view has two layers: the hidden layer has circle

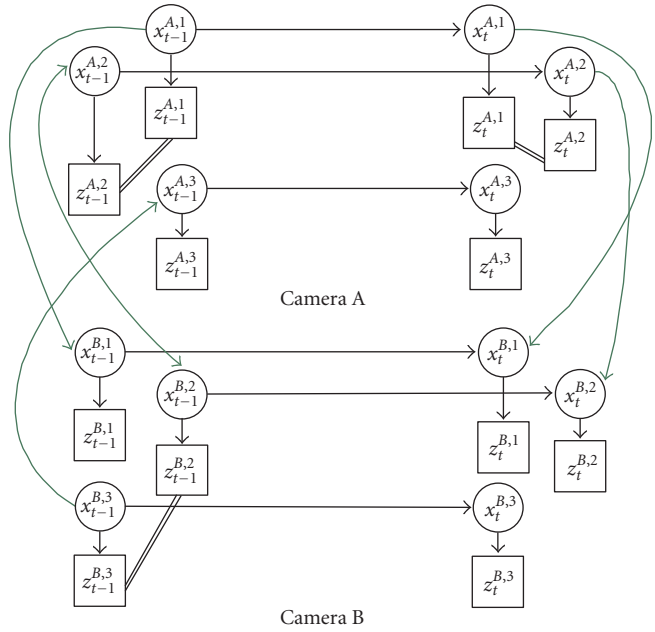


Figure 1: The dynamic graphical model for multiple-target tracking using multiple collaborative cameras. The directed curve link shows the "camera collaboration" between the counterpart states in different cameras.

nodes representing the targets' states; the observable layer has square nodes representing the observations associated with the hidden states. The directed link between consecutive states of the same target in each camera represents the state dynamics. The directed link from a target's state to its observation characterizes the local observation likelihood. The undirected link in each camera between neighboring observation nodes represents the "*interaction*." As it is mentioned, we activate the interaction only when the targets' observations are in close proximity or occlusion. This can be approximately determined by the spatial relation between the targets' trackers since the exact locations of observations are unknown. The directed curve link between the counterpart states of the same target in two cameras represents the "*camera collaboration*." This collaboration is activated between any possible collection of cameras only for targets which need help to improve their tracking robustness. For instance, when the targets are close to occlusion or possibly completely occluded by other targets in a camera view. The direction of the link shows "which target resorts to which other targets for help." This "*need-driven*"-based scheme avoids performing camera collaboration at all times and for all targets; thus, a tremendous amount of computation is saved. For example, in Figure 1, all targets in camera B at time t do not need to activate the camera collaboration because their observations do not interact with the other targets' observations at all. In this case, each target can be robustly tracked using independent trackers. On the other hand, targets 1 and 2 in camera A at time t activate camera collaboration since their observations interact and may undergo multitarget occlusion. Therefore, external information from other cameras may be helpful to make the tracking of these two targets more stable.

A graphical model like Figure 1 is suitable for centralized analysis using joint-state representations. However, in order to minimize the computational cost, we choose a completely distributed process where multiple collaborative trackers, one tracker per target in each camera, are used for MTT simultaneously. Consequently, we further decompose the graphical model for every target in each camera by performing four steps: (1) each submodel aims at one target in one camera; (2) for analysis of the observations of a specific camera, only neighboring observations which have direct links to the analyzed target's observation are kept. All the nodes of both nonneighboring observations and other targets' states are removed; (3) each undirected "*interaction*" link is decomposed into two different directed links for the different targets. The direction of the link is from the other target's observation to the analyzed target's observation; (4) since the "*camera collaboration*" link from a target's state in the analyzed camera view to its counterpart state in another view and the link from this counterpart state to its associated observation have the same direction, this causality can be simplified by a direct link from the grandparent node to its grandson as illustrated in Figure 2 [36]. Figure 3(a) illustrates the decomposition result of target 1 in camera A. Although we neglect some indirectly related nodes and links and thus simplify the distributed graphical model when analyzing a certain target, the neglected information is not lost but has been taken into account in the other targets' models. Therefore, when all the trackers are implemented simultaneously, the decomposed subgraphs together capture the original graphical model.

According to *graphical model theory* [36], we can analyze the *Markov properties*, that is, conditional independence

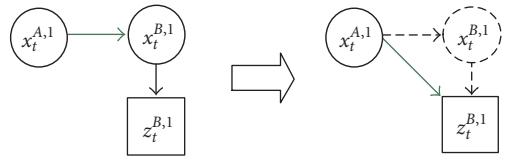


Figure 2: Equivalent simplification of camera collaboration link. The link causality from grandparent to parent then to grandson node is replaced by a direct link from grandparent to grandson node.

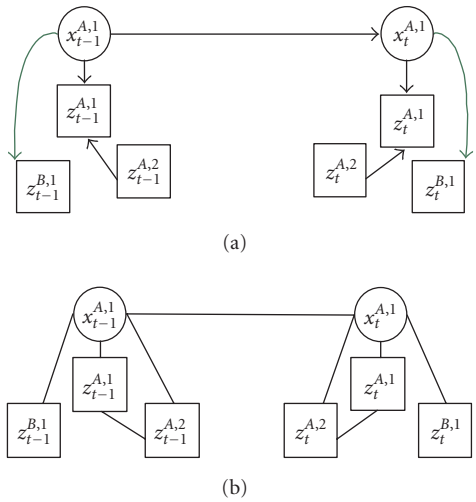


Figure 3: (a) Decomposition result for the target 1 in view A from Figure 1; (b) the moral graph of the graphical model in (a) for Markov property analysis.

properties [36, pages 69–70] for every decomposed graph on its corresponding *moral graphs* as illustrated in Figure 3(b). Then, by applying the *separation theorem* [36, page 67], the following *Markov* properties can be easily substantiated:

(i) $p(x_t^{A,i}, z_t^{A,J_t}, z_t^{B,i} \mid x_{0:t-1}^{A,i}, z_{1:t-1}^{A,i}, z_{1:t-1}^{A,J_{1:t-1}}, z_{1:t-1}^{B,i}) = p(x_t^{A,i}, z_t^{A,J_t}, z_t^{B,i} \mid x_{0:t-1}^{A,i})$,

(ii) $p(z_t^{A,J_t}, z_t^{B,i} \mid x_t^{A,i}, x_{0:t-1}^{A,i}) = p(z_t^{A,J_t}, z_t^{B,i} \mid x_t^{A,i})$,

(iii) $p(z_t^{A,i} \mid x_{0:t}^{A,i}, z_{1:t-1}^{A,i}, z_{1:t}^{A,J_{1:t}}, z_{1:t}^{B,i}) = p(z_t^{A,i} \mid x_t^{A,i}, z_t^{A,J_t}, z_t^{B,i})$,

(iv) $p(z_t^{B,i} \mid x_t^{A,i}, z_t^{A,i}) = p(z_t^{B,i} \mid x_t^{A,i})$,

(v) $p(z_t^{A,J_t}, z_t^{B,i} \mid x_t^{A,i}, z_t^{A,i}) = p(z_t^{A,J_t} \mid x_t^{A,i}, z_t^{A,i}) p(z_t^{B,i} \mid x_t^{A,i}, z_t^{A,i})$.

These properties have been used in the appendix to facilitate the derivation.

2.2. Distributed Bayesian tracking for each tracker

In this section, we present a Bayesian conditional density propagation framework for each decomposed graphical model as illustrated in Figure 3. The objective is to provide a generic statistical framework to model the interaction among cameras for multicamera tracking. Since we use multiple collaborative trackers, one tracker per target in each camera view, for multicamera multitarget tracking, we will dynamically estimate the posterior based on observations from both the target and its neighbors in the current camera view as well as the target in other camera views, that is, $p(x_{0:t}^{A,i} \mid z_{1:t}^{A,i}, z_{1:t}^{A,J_{1:t}}, z_{1:t}^{B,i})$ for each tracker and for each camera view. By applying Bayes's rule and the *Markov* properties derived in the previous section, a recursive conditional density updating rule can be obtained:

$$\begin{aligned} p\left(x_{0:t}^{A,i} \mid z_{1:t}^{A,i}, z_{1:t}^{A,J_{1:t}}, z_{1:t}^{B,i}\right) &= k_t p\left(z_t^{A,i} \mid x_t^{A,i}\right) p\left(x_t^{A,i} \mid x_{0:t-1}^{A,i}\right) p\left(z_t^{A,J_t} \mid x_t^{A,i}, z_t^{A,i}\right) \\ &\quad \times p\left(z_t^{B,i} \mid x_t^{A,i}\right) p\left(x_{0:t-1}^{A,i} \mid z_{1:t-1}^{A,i}, z_{1:t-1}^{A,J_{1:t-1}}, z_{1:t-1}^{B,i}\right), \end{aligned} \tag{1}$$

where

$$k_t = \frac{1}{p\left(z_t^{A,i}, z_t^{A,J_t}, z_t^{B,i} \mid z_{1:t-1}^{A,i}, z_{1:t-1}^{A,J_{1:t-1}}, z_{1:t-1}^{B,i}\right)}. \tag{2}$$

The derivation of (1) and (2) is presented in the appendix. Notice that the normalization constant k_t does not depend on the states $x_{0:t}^{A,i}$. In (1), $p(z_t^{A,i} \mid x_t^{A,i})$ is the local observation likelihood for target i in the analyzed camera view A, $p(x_t^{A,i} \mid x_{0:t-1}^{A,i})$ represents the state dynamics, which are similar to traditional Bayesian tracking methods. $p(z_t^{A,J_t} \mid x_t^{A,i}, z_t^{A,i})$ is the "*target interaction function*" within each camera which can be estimated by using the "*magnetic repulsion model*" presented in [14]. A novel likelihood density $p(z_t^{B,i} \mid x_t^{A,i})$ is introduced to characterize the collaboration between the same target's counterparts in different camera views. We call it a "*camera collaboration function.*"

When not activating the camera collaboration for a target and regarding its projections in different views as independent, the proposed BMCT framework can be identical to the IDMOT approach [14], where $p(z_t^{B,i} \mid x_t^{A,i})$ is uniformly distributed. When deactivating the interaction among the targets' observations, our formulation will further simplify to traditional Bayesian tracking [37, 38], where $p(z_t^{A,J_t} \mid x_t^{A,i}, z_t^{A,i})$ is also uniformly distributed.

3. SEQUENTIAL MONTE CARLO IMPLEMENTATION

Since the posterior of each target is generally non-Gaussian, we describe in this section a nonparametric implementation of the derived Bayesian formulation using the sequential Monte Carlo algorithm [38–40], in which a particle set is employed to represent the posterior

$$p\left(x_{0:t}^{A,i} \mid z_{1:t}^{A,i}, z_{1:t}^{A,J_{1:t}}, z_{1:t}^{B,i}\right) \sim \left\{x_{0:t}^{A,i,n}, w_t^{A,i,n}\right\}_{n=1}^{N_p}, \tag{3}$$

where $\{x_{0:t}^{A,i,n},\ n = 1, \ldots, N_p\}$ are the samples, $\{w_t^{A,i,n},\ n = 1, \ldots, N_p\}$ are the associated weights, and N_p is the number of samples.

Considering the derived sequential iteration in (1), if the particles $x_{0:t}^{A,i,n}$ are sampled from the importance density $q(x_t^{A,i} \mid x_{0:t-1}^{A,i,n}, z_{1:t}^{A,i}, z_{1:t}^{A,J_{1:t}}, z_{1:t}^{B,i}) = p(x_t^{A,i} \mid x_{0:t-1}^{A,i,n})$, the

corresponding weights are given by

$$w_t^{i,n} \propto w_{t-1}^{i,n} p\left(z_t^{A,i} \mid x_t^{A,i,n}\right) p\left(z_t^{A,J_t} \mid x_t^{A,i,n}, z_t^{A,i}\right) p\left(z_t^{B,i} \mid x_t^{A,i,n}\right). \tag{4}$$

It has been widely accepted that better importance density functions can make particles more efficient [39, 40]. We choose a relatively simple function $p(x_t^{A,i} \mid x_{t-1}^{A,i})$ as in [37] to highlight the efficiency of using camera collaboration. Other importance densities such as reported in [41–44] can be used to provide better performance.

Modeling the densities in (4) is not trivial and usually has great influence on the performance of practical implementations. In the following subsections, we first discuss the target model, then present the proposed *camera collaboration likelihood model*, and further summarize other models for density estimation.

3.1. Target representation

A proper model plays an important role in estimating the densities. Different target models, such as the 2D ellipse model [45], 3D object model [34], snake or dynamic contour model [37], and so forth, are reported in the literature. In this paper, we use a five-dimensional parametric ellipse model which is quite simple, saves a lot of computational costs, and is sufficient to represent the tracking results for MTT. For example, the state $x_t^{A,i}$ is given by $(cx_t^{A,i}, cy_t^{A,i}, a_t^{A,i}, b_t^{A,i}, \rho_t^{A,i})$, where $i = 1, \ldots, M$ is the index of targets, t is the time index, (cx, cy) is the center of the ellipse, a is the major axis, b is the minor axis, and ρ is the orientation in radians.

3.2. Camera collaboration likelihood model

The proposed Bayesian conditional density propagation framework has no specific requirements of the cameras (e.g., fixed or moving, calibrated or not) and the collaboration model (e.g., 3D/2D) as long as the model can provide a good estimation of the density $p(z_t^{B,i} \mid x_t^{A,i})$. Epipolar geometry [46] has been used to model the relation across multiple camera views in different ways. In [28], an epipolar line is used to facilitate color-based region matching and 3D coordinate projection. In [26], match scores are calculated using epipolar geometry for segmented blobs in different views. Nummiaro et al. used an epipolar line to specify the distribution of samples in [30]. Although they are very useful in different applications as reported in the prior literature, these models are not suitable for our framework. Since generally hundreds or even thousands of particles are needed in sequential Monte Carlo implementation for multiple-target tracking in crowded scenes, the computation required to perform feature matching for each particle is not feasible. Moreover, using the epipolar line to facilitate importance sampling is problematic and is not suitable for tracking in crowded environments [30]. Such a camera collaboration model may introduce additional errors as discussed and shown in Section 4.2. On the other hand, we present a paradigm of camera collaboration likelihood model using sequential Monte Carlo implementation which does not require feature matching and recovery of the target's 3D coordinates, but only assumes that the cameras' epipolar geometry is known.

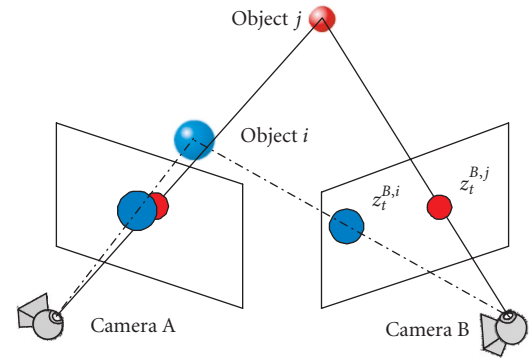


Figure 4: The model setting in 3D space for camera collaboration likelihood estimation.

Figure 4 illustrates the model setting in 3D space. Two targets i and j are projected onto two camera views. In view A, the projections of targets i and j are very close (occluding) while in view B, they are not. In such situations, we only activate the camera collaboration for trackers of targets i and j in view A but not in view B. We have considered two methods to calculate the likelihood $p(z_t^{B,i} \mid x_t^{A,i})$ without recovering the target's 3D coordinates: (1) mapping $x_t^{A,i}$ to view B and then calculating the likelihood there; (2) mapping the observation $z_t^{B,i}$ to camera view A and calculating the density there. The first way looks more impressive but is actually infeasible. Since usually hundreds or thousands of particles have to be used for MTT in crowded scenes, mapping each particle into another view and computing the likelihood requires an enormous computational effort. We have therefore decided to to choose the second approach. The observations $z_t^{B,i}$ and $z_t^{B,j}$ are initially found by tracking in view B. Then they are mapped to view A, producing $\hbar(z_t^{B,i})$ and $\hbar(z_t^{B,j})$, where $\hbar(\cdot)$ is a function of $z_t^{B,i}$ or $z_t^{B,j}$ characterizing the epipolar geometry transformation. After that, the collaboration likelihood can be calculated based on $\hbar(z_t^{B,i})$ and $\hbar(z_t^{B,j})$. Sometimes, a more complicated case occurs, for example, target i is occluding with others in both cameras. In this situation, the above scheme is initialized by randomly selecting one view, say, view B, and using IDMOT to find the observations. These initial estimates may be not very accurate; therefore, in this case, we iterate several times (usually twice is enough) between different views to get more stable estimates.

According to *epipolar geometry theory* [46, pages 237–259], a point in one camera view can find an epipolar line in another view. Therefore, $z_t^{B,i}$ which is represented by a circle model corresponds to an epipolar "*band*" in view A, which is $\hbar(z_t^{B,i})$. A more accurate location along this band for $\hbar(z_t^{B,i})$ can be obtained by feature matching. We find that two practical issues prevent us from doing so. Firstly, the

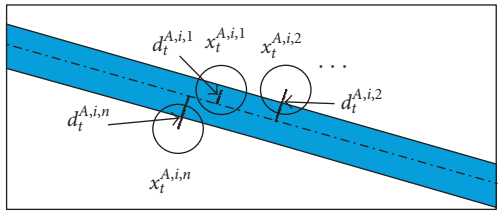


Figure 5: Calculating the camera collaboration weights for target i in view A. Circles instead of ellipses are used to represent the particles for simplicity.

wide-baseline cameras usually make the target's features vary significantly in different views. Moreover, the occluded target's features are interrupted or even completely lost. Secondly, the crowded scene means that there may be several similar candidate targets along this band. It usually happens that the optimal match may be a completely wrong location, and thus falsely guides the tracker away. Our experiments show that using the band $\hbar(z_t^{B,i})$ itself cannot only avoid the above errors but also provides useful spatial information for target location. Furthermore, the local image observation has already been considered in the local likelihood $p(z_t^{A,i} \mid x_t^{A,i})$ which provides information for estimating both the target's location and size.

Figure 5 shows the procedure used to calculate the collaboration weight for each particle based on $\hbar(z_t^{B,i})$. The particles $\{x_t^{A,i,1}, x_t^{A,i,2}, \ldots, x_t^{A,i,n}\}$ are represented by the circles instead of ellipse models for simplicity. Given the Euclidean distance $d_t^{A,i,n} = \|x_t^{A,i,n} - \hbar(z_t^{B,i})\|$ between the particle $x_t^{A,i,n}$ and the band $\hbar(z_t^{B,i})$, the collaboration weight for particle $x_t^{A,i,n}$ can be computed as

$$\phi_t^{A,i,n} = \frac{1}{\sqrt{2\pi}\Sigma_\phi} \exp\left\{ -\frac{\left(d_t^{A,i,n}\right)^2}{2\Sigma_\phi^2} \right\}, \tag{5}$$

where Σ_ϕ^2 is the variance which can be chosen as the bandwidth. In Figure 5, we simplify $d_t^{A,i,n}$ by using a *point-line distance* between the center of the particle and the middle line of the band. Furthermore, the camera collaboration likelihood can be approximated as follows:

$$p\left(z_t^{B,i} \mid x_t^{A,i}\right) \approx \sum_{n=1}^{N_p} \frac{\phi_t^{A,i,n}}{\sum_{n'=1}^{N_p} \phi_t^{A,i,n'}} \delta\left(x_t^{A,i} - x_t^{A,i,n}\right). \tag{6}$$

3.3. Interaction and local observation likelihood models

We have proposed a "*magnetic repulsion model*" to estimate the interaction likelihood in [14]. It can be used here similarly:

$$p\left(z_t^{A,J_t} \mid x_t^{A,i}, z_t^{A,i}\right) \approx \sum_{n=1}^{N_p} \frac{\varphi_t^{A,i,n}}{\sum_{n'=1}^{N_p} \varphi_t^{A,i,n'}} \delta\left(x_t^{A,i} - x_t^{A,i,n}\right), \tag{7}$$

where $\varphi_t^{A,i,n}$ is the interaction weight of particle $x_t^{A,i,n}$. It can be iteratively calculated by

$$\varphi_t^{A,i,n} = 1 - \frac{1}{\alpha} \exp\left\{ -\frac{\left(l_t^{A,i,n}\right)^2}{\Sigma_\varphi^2} \right\}, \tag{8}$$

where α and Σ_φ are constants. $l_t^{A,i,n}$ is the distance between the current particle's observation and the neighboring observation.

Different cues have been proposed to estimate the local observation likelihood [37, 47, 48]. We fuse the target's color histogram [47] with a PCA-based model [48] together, namely, $p(z_t^{A,i} \mid x_t^{A,i}) = p_c \times p_p$, where p_c and p_p are the likelihood estimates obtained from the color histogram and PCA models, respectively.

3.4. Implementation issues

3.4.1. New target initialization

For simplicity, we manually initialize all the targets in the experiments. Many automatic initialization algorithms such as reported in [6, 21] are available and can be used instead.

3.4.2. Triangle transition algorithm for each target

To minimize computational cost, we do not want to activate the camera collaboration when targets are far away from each other since a single-target tracker can achieve reasonable performance. Moreover, some targets cannot utilize the camera collaboration even when they are occluding with others if these targets have no projections in other views. Therefore, a tracker activates the camera collaboration and thus implements the proposed BMCT only when its associated target "*needs*" and "*could*" do so. In other situations, the tracker will degrade to implement IDMOT or a traditional Bayesian tracker such as multiple independent regular particle filters (MIPFs) [37, 38].

Figure 6 shows the triangle algorithm transition scheme. We use *counterpart epipolar consistence loop checking* to check if the projections of the same target in different views are on each other's epipolar line (band).

Every target in each camera view is in one of the following three situations.

(i) *Having good counterpart*. The target and its counterpart in other views satisfy the epipolar consistence loop checking. Only such targets are used to activate the camera collaboration.

(ii) *Having bad counterpart*. The target and its counterpart do not satisfy the epipolar consistence loop checking which means that at least one of their trackers made a mistake. Such targets will not activate the camera collaboration to avoid additional error.

(iii) *Having no counterpart*. Occurs when the target has no projection in other views at all.

The targets "*having bad counterpart*" or "*having no counterpart*" will implement a degraded BMCT, namely, IDMOT.

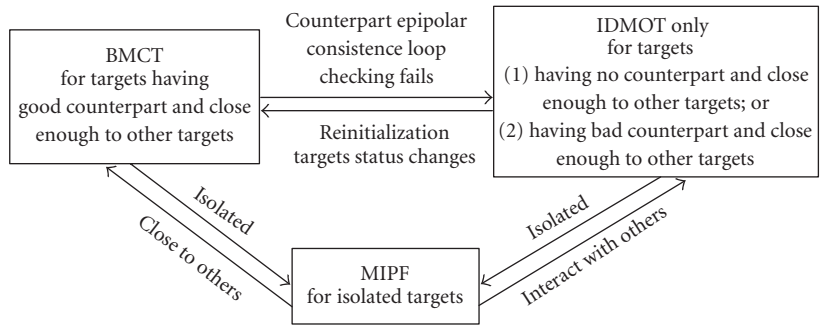


Figure 6: Triangle transition algorithm; BMCT: the proposed distributed Bayesian multiple collaborative cameras multiple-target tracking approach; IDMOT: interactively distributed multiple-object tracking [14]; MIPF: multiple independent regular particle filters [38].

The only chance for these trackers to be upgraded back to BMCT is after reinitializtion, where the status may change to "*having good counterpart.*"

Within a camera view, if the analyzed tracker is isolated from other targets, it will only implement MIPF to reduce the computational costs. When it becomes closer or interacts with other trackers, it will activate either BMCT or IDMOT according to the associated target's status. This triangle transition algorithm guarantees that the proposed BMCT using multiocular videos can work better and is never inferior to monocular video implementations of IDMOT or MIPF.

3.4.3. Target tracker killing

The tracker has the capability to decide that the associated target disappeared and should be killed in two cases: (1) the target moves out of the image; or (2) the tracker loses the target and tracks clutter. In both situations, the epipolar consistence loop checking fails and the local observation weights of the tracker's particles become very small since there is no target information any more. On the other hand, in the case where the tracker misses its associated target and follows a false target, we do not kill the tracker and leave it for further comparison.

3.4.4. Pseudocode

Algorithm 1 illustrates the pseudocode of the proposed BMCT using sequential Monte Carlo implementation for target i in camera A at time t.

// Regular Bayesian tracking such as MIPF.
Draw particles $x_t^{A,i,n} \sim q(x_t^{A,i} \mid x_{0:t-1}^{A,i,n}, z_{1:t}^{A,i}, z_{1:t}^{A,J_{1:t}}, z_{1:t}^{B,i})$.
Local observation weighting: $w_t^{A,i,n} = p(z_t^{A,i} \mid x_t^{A,i,n})$.
Normalize $(w_t^{A,i,n})$.
Temporary estimate $\hat{z}_t^{A,i} \sim \hat{x}_t^{A,i} = \sum_{n=1}^{N_p} w_t^{A,i,n} x_t^{A,i,n}$.
// Camera collaboration qualification checking.
IF (epipolar consistence loop checking is OK) // BMCT.
(i) IF ($\hat{z}_t^{A,i}$ is close to others) // Activate camera collaboration.
  (1) Collaboration weighting: $\phi_t^{A,i,n} = p(z_t^{B,i} \mid x_t^{A,i,n})$.
  (2) IF ($\hat{z}_t^{A,i}$ is touching others) // Activate target interaction.
    (a) FOR $k = 1 \sim K$ // Interaction iteration:
      Interaction weighting: $\varphi_t^{A,i,n,k}$ // (8).
      ...
    (b) Reweighting $w_t^{A,i,n} = w_t^{A,i,n} \times \varphi_t^{A,i,n,K}$.
  (3) Reweighting $w_t^{A,i,n} = w_t^{A,i,n} \times \phi_t^{A,i,n}$.
ELSE // IDMOT only.
(ii) IF ($\hat{z}_t^{A,i}$ is touching others).
    (1) FOR $k = 1 \sim K$ // Interaction iteration:
      Interaction weighting: $\varphi_t^{A,i,n,k}$ // (8).
      ...
    (2) Reweighting $w_t^{A,i,n} = w_t^{A,i,n} \times \varphi_t^{A,i,n,K}$.
Normalize $(w_t^{A,i,n})$.
Estimate $\hat{x}_t^{A,i} = \sum_{n=1}^{N_p} w_t^{A,i,n} x_t^{A,i,n}$.
Resample $\{x_t^{A,i,n}, w_t^{A,i,n}\}$.

Algorithm 1: Sequential Monte Carlo implementation of BMCT algorithm.

4. EXPERIMENTAL RESULTS

To demonstrate the effectiveness and efficiency of the proposed approach, we performed experiments on both synthetic and real video data with different conditions. We have used 60 particles per tracker in all the simulations for our approach. Different colors and numbers are used to label the targets. To compare the proposed BMCT against the state of the art, we also implemented multiple independent particle filters (MIPF) [37], interactively distributed multiple-object tracking (IDMOT) [14], and color-based multicamera tracking (CMCT) [30]. For all real-world videos, the fundamental matrix of epipolar geometry is estimated by using the algorithm proposed by Hartley and Zisserman [46, pages 79–308].

4.1. Synthetic videos

We generate synthetic videos by assuming that two cameras are widely set at a right angle and at the same height above the ground. This setup makes it very easy to compute the epipolar line. Six soccer balls move independently within the overlapped scene of the two views. The difference between

Figure 7: Tracking results of the synthetic videos: (a) multiple independent particle filters [37]; (b) interactively distributed multiple-object tracking [14]; (c) the proposed BMCT.

the target's projections in the different views is neglected for simplicity since only the epipolar geometry information and not the target's features are used to estimate the camera collaboration likelihood. The change in the target's size is also neglected since the most important concern of tracking is the target's location. Various multitarget occlusions are frequently encountered when the targets are projected onto each view. A two-view sequence, where each view has 400 frames with a resolution of 320×240 pixels, is used to demonstrate the ability of the proposed BMCT for solving multitarget occlusions. For simplicity, only the color histogram model [47] is used to estimate the local observation likelihood.

Figure 7 illustrates the tracking results of (a) MIPF, (b) IDMOT, and (c) BMCT. MIPF suffers from severe multitarget occlusions. Many trackers (circles) are "hijacked" by targets with strong local observation, and thus lose their associated targets after occlusion. Equipped with magnetic repulsion and inertia models to handle target interaction, IDMOT has improved the performance by separating the occluding targets and labeling them correctly for many targets. The white link between the centers of the occluding targets shows the interaction. However, due to the intrinsic limitations of monocular video and the relatively simple inertia model, this approach has two failure cases. In camera 1, when targets 0 and 2 move along the same direction and persist in a severe occlusion, due to the absence of distinct inertia directions, the blind magnetic repulsion separates them with random directions. Coincidentally, a "strong" clutter nearby attracts one tracker away. In camera 2, when targets 2 and 5 move along the same line and occlude, their labels are falsely exchanged due to their similar inertia. By using biocular videos simultaneously and exploiting camera collaboration, BMCT rectifies these problems and tracks all of the targets robustly. The epipolar line through the center of a particular target is mapped from its counterpart in another view and reveals when the camera collaboration is activated. The color of the epipolar line is an indicator of the counterpart's label.

4.2. Indoor videos

The *Hall* sequences are captured by two low-cost surveillance cameras in the front hall of a building. Each sequence

Figure 8: Tracking results of the proposed BMCT on the indoor gray videos *Hall*.

has 776 frames with a resolution of 640 × 480 pixels and a frame rate of 29.97 frames per second (fps). Two people loiter around generating frequent occlusions. Due to the grayness of the images, many color-based tracking approaches are not suitable. We use this video to demonstrate the robustness of our BMCT algorithm for tracking without color information. Background subtraction [49] is used to decrease the clutter and enhance the performance. An intensity histogram, instead of color histogram, is combined with a PCA-based model [48] to calculate the local observation likelihood. Benefiting from not using color or other feature-based matching but only exploiting the spatial information provided by epipolar geometry, our camera collaboration likelihood model still works well for gray-image sequence. As expected, BMCT produced very robust results in each camera view as shown for sample frames in Figure 8.

The *UnionStation* videos are captured at Chicago Union Station using two widely separated digital cameras at different heights above the ground. The videos have a resolution of 320 × 240 pixels and a frame rate of 25 fps. Each view sequence consists of 697 frames. The crowded scene has various persistent and significant multitarget occlusions when pedestrians pass by each other. Figure 9 shows the tracking results of (a) IDMOT, (b) CMCT, and (c) BMCT. We used the ellipses' bottom points to find the epipolar lines. Although IDMOT was able to resolve many multitarget occlusions by handling the targets within each camera view, IDMOT still made mistakes during severe multitarget occlusions because of the intrinsic limitations of monocular videos. For example, in view B, tracker 2 falsely attaches to a wrong person when the right person is completely occluded by target 6 for a long time. CMCT [30] is a multicamera target tracking approach which also uses epipolar geometry and a particle filter implementation. The original CMCT needs to prestore the target's multiple color histograms for different views. For practical tracking, however, especially in crowded environments, this prior knowledge is usually not available. Therefore, in our implementation, we simplified this approach by using the initial color histograms of each target from multiple cameras. Then, these histograms are updated using the adaptive model proposed by Nummiaro et al. in [47]. The original CMCT is also based on the best front-view selection scheme and only outputs one estimate each time for a target. For a better comparison, we keep all the estimates from the different cameras. Figure 9(b) shows the tracking results using the CMCT algorithm. As indicated in [30], using an epipolar line to direct the sample distribution and (re)initialize the targets is problematic. When there are several candidate targets in the vicinity of the epipolar lines, the trackers may attach to wrong targets, and thus lose the correct target. It can be seen from Figure 9(b) that CMCT did not produce satisfactory results, especially when multitarget occlusion occurs. Compared with the above approaches, BMCT shows very robust results separating targets apart and assigning them with correct labels even after persistent and severe multitarget occlusions as a result of using both target interaction within each view and camera collaboration between different views. The only failure case of BMCT occurs in camera 2, where target 6 is occluded by target 2. Since there is no counterpart of target 6 appearing in camera 1, the camera collaboration is not activated and only IDMOT instead of BMCT is implemented. By using more cameras, such failure cases could be avoided. The quantitative performance and speed comparisons of all these methods will be discussed later.

The *LivingRoom* sequences are captured with a resolution of 320 × 240 pixels and a frame rate of 15 fps. We use them to demonstrate the performance of BMCT for three collaborative cameras. Each sequence has 616 frames and contains four people moving around with many severe multiple-target occlusions. Figure 10 illustrates the tracking results. By modeling both the camera collaboration and target interaction within each camera simultaneously, BMCT solves the multitarget occlusion problem and achieves a very robust performance.

4.3. Outdoor videos

The test videos *Campus* have two much longer sequences, each of which has 1626 frames. The resolution is 320 × 240 and the frame rate is 25 fps. They are captured by two cameras set outdoors on campus. Three pedestrians walk around with various multitarget occlusions. The proposed BMCT achieves stable tracking results on these videos as can be seen in the sample frames in Figure 11.

4.4. Quantitative analysis and comparisons

4.4.1. Computational cost analysis

There are three different likelihood densities which must be estimated in our BMCT framework: (1) local observation

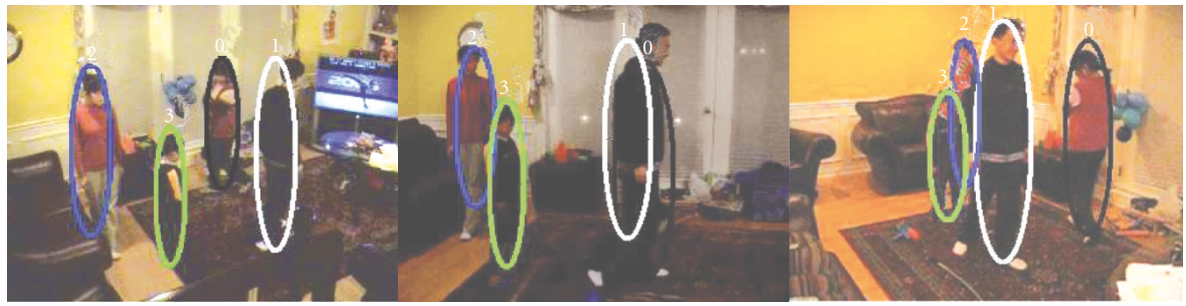
Figure 9: Tracking results of the videos *UnionStation*: (a) interactively distributed multiple-object tracking (IDMOT) [14]; (b) color-based multicamera tracking approach (CMCT) [30]; (c) the proposed BMCT.

likelihood $p(z_t^{A,i} \mid x_t^{A,i})$; (2) target interaction likelihood $p(z_t^{A,J_t} \mid x_t^{A,i}, z_t^{A,i})$ within each camera; and (3) camera collaboration likelihood $p(z_t^{B,i} \mid x_t^{A,i})$. The weighting complexity of these likelihoods are the main factors which impact the entire system's computational cost. In Table 1, we compare the average computation time of the different likelihood weightings in processing one frame of the synthetic sequences using BMCT. As we can see, compared with the most time-consuming component which is the local observation likelihood weighting of traditional particle filters, the computational cost required for camera collaboration is negligible. This is because of two reasons: firstly, a tracker activates the camera collaboration only when it encounters potential multitarget occlusions; and secondly, our epipolar geometry-based camera collaboration likelihood model avoids feature matching and is very efficient.

The computational complexity of the centralized approaches used for multitarget tracking in [9, 29, 33] increases exponentially in terms of the number of targets and cameras since the centralized methods rely on joint-state representations. The computational complexity of the proposed distributed framework, on the other hand, increases linearly with the number of targets and cameras. In Table 2, we compare the complexity of these two modes in terms of the number of targets by running the proposed BMCT and a joint-state representation-based MCMC particle filter (MCMC-PF) [9]. The data is obtained by varying the number of targets on the synthetic videos. It can be seen that under the condition of achieving reasonable robust tracking performance, both the required number of particles and the speed of the proposed BMCT vary linearly.

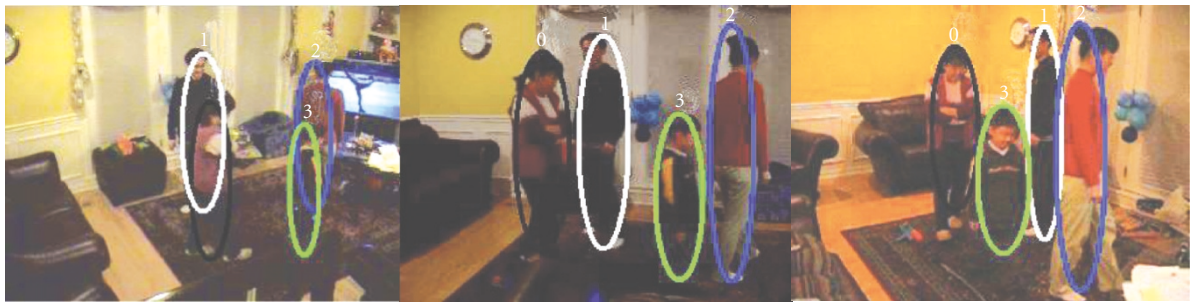
4.4.2. Quantitative performance comparisons

Quantitative performance evaluation for multiple-target tracking is still an open problem [50, 51]. Unlike single-target tracking and target detection systems, where standard metrics are available, the varying number of targets and the frequent multitarget occlusion make it challenging to provide an exact performance evaluation for multitarget tracking approaches [50]. When using multiple cameras, the target label correspondence across different cameras further increases the difficulty of the problem. Since the main concern of tracking is the correctness of the tracker's location and



CAM1#042 CAM2#042 CAM3#042

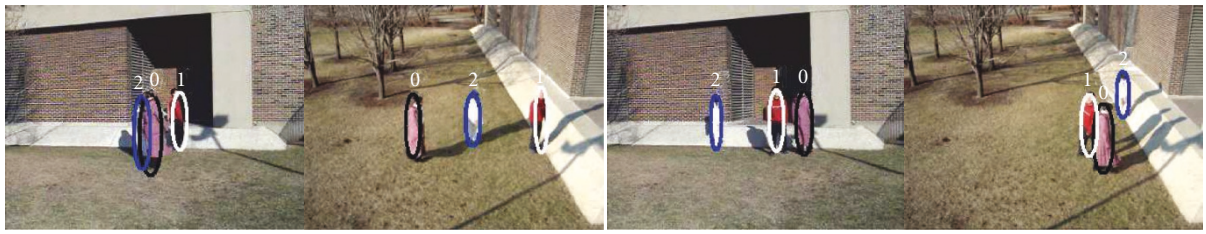
(a)



CAM1#128 CAM2#128 CAM3#128

(b)

Figure 10: Tracking results of the proposed BMCT on the trinocular videos *LivingRoom*.



CAM1#133 CAM2#133 CAM1#210 CAM2#210

Figure 11: Tracking results of the proposed BMCT on the outdoor videos *Campus*.

label, we compare the tracking performance quantitatively by defining the *false position rate* FR_p and the *false label rate* FR_l as follows:

$$\mathrm{FR}_p = \frac{\text{the number of position failures}}{\text{the total number of targets in all cameras}},$$
$$\mathrm{FR}_l = \frac{\text{the number of label failures}}{\text{the total number of targets in all cameras}}, \quad (9)$$

where a *position failure* is defined as the absence of a tracker associated with one of the tracking targets in a camera and a *label failure* is defined as a tracker associated with a wrong target. Table 3 presents the statistical performance using different tracking algorithms on the *UnionStation* videos. It can be observed that BMCT outperforms the other approaches with much lower false position and label rates.

Table 1: Average computational time comparison of different likelihood weightings.

Local observation likelihood	Target interaction likelihood	Camera collaboration likelihood
0.057 s	0.0057 s	0.003 s

4.4.3. Speed comparisons

Speed comparisons of the different tracking approaches are presented in Table 3. We implemented all of the algorithms independently in C++ without code optimization on a 3.2 GHz Pentium IV PC. The number of particles used has an important impact on both the performance and speed for particle filter-based tracking approaches. In principle, using more particles will increase the stability of the tracking

Table 2: Complexity analysis in terms of the number of targets.

Total targets number		4	5	6
Total particles	MCMC-PF	500	1100	2800
	BMCT	400	500	600
Speed (fps)	MCMC-PF	8.5 ∼ 9	2.1 ∼ 3	0.3 ∼ 0.5
	BMCT	13.8 ∼ 15	11 ∼ 12	9 ∼ 10.5

Table 3: Quantitative performance and speed comparisons.

Method	MIPF	IDMOT	CMCT	BMCT
$FR_p^\dagger$	36.8%	11.6%	24.6%	2.3%
$FR_l^\dagger$	40.1%	15.3%	26.5%	1.6%
Particles per tracker	60	60	100	60
Speed (fps)	11 ∼ 13.5	10.5 ∼ 12	4 ∼ 5.3	9 ∼ 10

† FR_p is the false position rate and FR_l is the false label rate.

performance; yet it will also increase the running time. Determination of the optimal number of particles to achieve a proper tradeoff between stability and speed of particle filters is still an open problem. In our simulations, we compare the tracking results visually and select the minimal number of particles which does not degrade the performance significantly. It can be observed that the proposed BMCT outperforms the CMCT in speed and achieves very close efficiency compared with MIPF and IDMOT.

5. CONCLUSION

In this paper, we have proposed a Bayesian framework to solve the multitarget occlusion problem for multiple-target tracking using multiple collaborative cameras. Compared with the common practice of using a joint-state representation whose computational complexity increases exponentially with the number of targets and cameras, our approach solves the multicamera multitarget tracking problem in a distributed way whose complexity grows linearly with the number of targets and cameras. Moreover, the proposed approach presents a very convenient framework for tracker initialization of new targets and tracker elimination of vanished targets. The distributed framework also makes it very suitable for efficient parallelization in complex computer networking application. The proposed approach does not recover the targets' 3D locations. Instead, it generates multiple estimates, one per camera, for each target in the 2D image plane. For many practical tracking applications such as video surveillance, this is sufficient since the 3D target location is usually not necessary and 3D modeling will require a very expensive computational effort for precise camera calibration and nontrivial feature matching. The merits of our BMCT algorithm compared with 3D tracking approaches are that it is fast, easy to implement, and can achieve robust tracking results in crowded environments. In addition, with the necessary camera calibration information, the 2D estimates can also be projected back to recover the target's 3D location in the world coordinate system. Furthermore, we have presented an efficient camera collaboration model using epipolar geometry with sequential Monte Carlo implementation. It avoids the need for recovery of the targets' 3D coordinates and does not require feature matching, which is difficult to perform in widely separated cameras. The preliminary experimental results have demonstrated the superior performance of the proposed algorithm in both efficiency and robustness for multiple-target tracking in different conditions.

Several problems remain unresolved and will be the subject of future research. (1) The local likelihood model and the camera collaboration model are critical for practical implementation. Although very effective, the current versions of these models are based only on image (e.g., color, appearance) and geometry information. We plan to integrate multiple cues such as audio and other sensor information into our framework. (2) The simulation results provided are based on fixed camera locations. However, the proposed BMCT framework is not limited to fixed cameras. We plan to incorporate promising camera calibration methods or panoramic background modeling techniques to extend the application to moving cameras. (3) Three-dimensional tracking is essential in many applications such as stereo tracking and virtual reality applications. We plan to extend our framework for 3D tracking from multiple camera views.

APPENDIX

DERIVATION OF THE DENSITY UPDATING RULE

We will now derive the recursive density update rule (1) in Section 2.2 using the *Markov* properties presented in Section 2.1:

$$\begin{aligned} &p\left(x_{0:t}^{A,i} \mid z_{1:t}^{A,i}, z_{1:t}^{A,J_{1:t}}, z_{1:t}^{B,i}\right) \\ &= \frac{p\left(z_t^{A,i} \mid x_{0:t}^{A,i}, z_{1:t-1}^{A,i}, z_{1:t}^{A,J_{1:t}}, z_{1:t}^{B,i}\right) p\left(x_{0:t}^{A,i} \mid z_{1:t-1}^{A,i}, z_{1:t}^{A,J_{1:t}}, z_{1:t}^{B,i}\right)}{p\left(z_t^{A,i} \mid z_{1:t-1}^{A,i}, z_{1:t}^{A,J_{1:t}}, z_{1:t}^{B,i}\right)} \\ &= \frac{p\left(z_t^{A,i} \mid x_t^{A,i}, z_t^{A,J_t}, z_t^{B,i}\right) p\left(x_{0:t}^{A,i} \mid z_{1:t-1}^{A,i}, z_{1:t}^{A,J_{1:t}}, z_{1:t}^{B,i}\right)}{p\left(z_t^{A,i} \mid z_{1:t-1}^{A,i}, z_{1:t}^{A,J_{1:t}}, z_{1:t}^{B,i}\right)}. \end{aligned} \tag{A.1}$$

In (A.1), we use *Markov* property (iii) presented in Section 2.1, that is, $p(z_t^{A,i} \mid x_{0:t}^{A,i}, z_{1:t-1}^{A,i}, z_{1:t}^{A,J_{1:t}}, z_{1:t}^{B,i}) = p(z_t^{A,i} \mid x_t^{A,i}, z_t^{A,J_t}, z_t^{B,i})$. We further derive the densities $p(z_t^{A,i} \mid x_t^{A,i}, z_t^{A,J_t}, z_t^{B,i})$ and $p(x_{0:t}^{A,i} \mid z_{1:t-1}^{A,i}, z_{1:t}^{A,J_{1:t}}, z_{1:t}^{B,i})$, respectively, as follows.

For $p(z_t^{A,i} \mid x_t^{A,i}, z_t^{A,J_t}, z_t^{B,i})$, we have

$$\begin{aligned} &p\left(z_t^{A,i} \mid x_t^{A,i}, z_t^{A,J_t}, z_t^{B,i}\right) \\ &\quad = p\left(z_t^{A,i} \mid x_t^{A,i}\right) \frac{p\left(z_t^{A,J_t}, z_t^{B,i} \mid x_t^{A,i}, z_t^{A,i}\right)}{p\left(z_t^{A,J_t}, z_t^{B,i} \mid x_t^{A,i}\right)}. \end{aligned} \tag{A.2}$$

For $p(x_{0:t}^{A,i} \mid z_{1:t-1}^{A,i}, z_{1:t}^{A,J_{1:t}}, z_{1:t}^{B,i})$, we have

$$\begin{aligned}
&p\left(x_{0:t}^{A,i} \mid z_{1:t-1}^{A,i}, z_{1:t}^{A,J_{1:t}}, z_{1:t}^{B,i}\right) \\
&\quad= \frac{p\left(x_t^{A,i}, z_t^{A,J_t}, z_t^{B,i} \mid x_{0:t-1}^{A,i}, z_{1:t-1}^{A,i}, z_{1:t-1}^{A,J_{1:t-1}}, z_{1:t-1}^{B,i}\right)}{p\left(z_t^{A,J_t}, z_t^{B,i} \mid z_{1:t-1}^{A,i}, z_{1:t-1}^{A,J_{1:t-1}}, z_{1:t-1}^{B,i}\right)} \\
&\qquad\times p\left(x_{0:t-1}^{A,i} \mid z_{1:t-1}^{A,i}, z_{1:t-1}^{A,J_{1:t-1}}, z_{1:t-1}^{B,i}\right) \\
&\quad= \frac{p\left(x_t^{A,i}, z_t^{A,J_t}, z_t^{B,i} \mid x_{0:t-1}^{A,i}\right)}{p\left(z_t^{A,J_t}, z_t^{B,i} \mid z_{1:t-1}^{A,i}, z_{1:t-1}^{A,J_{1:t-1}}, z_{1:t-1}^{B,i}\right)} \qquad \text{(A.3)} \\
&\qquad\times p\left(x_{0:t-1}^{A,i} \mid z_{1:t-1}^{A,i}, z_{1:t-1}^{A,J_{1:t-1}}, z_{1:t-1}^{B,i}\right) \\
&\quad= \frac{p\left(z_t^{A,J_t}, z_t^{B,i} \mid x_t^{A,i}, x_{0:t-1}^{A,i}\right)}{p\left(z_t^{A,J_t}, z_t^{B,i} \mid z_{1:t-1}^{A,i}, z_{1:t-1}^{A,J_{1:t-1}}, z_{1:t-1}^{B,i}\right)} \\
&\qquad\times p\left(x_t^{A,i} x_{0:t-1}^{A,i}\right) p\left(x_{0:t-1}^{A,i} \mid z_{1:t-1}^{A,i}, z_{1:t-1}^{A,J_{1:t-1}}, z_{1:t-1}^{B,i}\right) \\
&\quad= \frac{p\left(z_t^{A,J_t}, z_t^{B,i} \mid x_t^{A,i}\right)}{p\left(z_t^{A,J_t}, z_t^{B,i} \mid z_{1:t-1}^{A,i}, z_{1:t-1}^{A,J_{1:t-1}}, z_{1:t-1}^{B,i}\right)} \\
&\qquad\times p\left(x_t^{A,i} \mid x_{0:t-1}^{A,i}\right) p\left(x_{0:t-1}^{A,i} \mid z_{1:t-1}^{A,i}, z_{1:t-1}^{A,J_{1:t-1}}, z_{1:t-1}^{B,i}\right). \qquad \text{(A.4)}
\end{aligned}$$

In (A.3), *Markov* property (i) is used. In (A.4), property (ii) is applied.

By substituting (A.2) and (A.4) back into (A.1) and rearranging terms, we have

$$\begin{aligned}
&p\left(x_{0:t}^{A,i} \mid z_{1:t}^{A,i}, z_{1:t}^{A,J_{1:t}}, z_{1:t}^{B,i}\right) \\
&\quad= p\left(z_t^{A,i} \mid x_t^{A,i}\right) p\left(x_t^{A,i} \mid x_{0:t-1}^{A,i}\right) p\left(z_t^{A,J_t}, z_t^{B,i} \mid x_t^{A,i}, z_t^{A,i}\right) \\
&\qquad\times \frac{p\left(x_{0:t-1}^{A,i} \mid z_{1:t-1}^{A,i}, z_{1:t-1}^{A,J_{1:t-1}}, z_{1:t-1}^{B,i}\right)}{p\left(z_t^{A,i}, z_t^{A,J_t}, z_t^{B,i} \mid z_{1:t-1}^{A,i}, z_{1:t-1}^{A,J_{1:t-1}}, z_{1:t-1}^{B,i}\right)} \\
&\quad= p\left(x_t^{A,i} \mid x_{0:t-1}^{A,i}\right) p\left(z_t^{A,J_t} \mid x_t^{A,i}, z_t^{A,i}\right) p\left(z_t^{B,i} \mid x_t^{A,i}, z_t^{A,i}\right) \\
&\qquad\times \frac{p\left(z_t^{A,i} \mid x_t^{A,i}\right) p\left(x_{0:t-1}^{A,i} \mid z_{1:t-1}^{A,i}, z_{1:t-1}^{A,J_{1:t-1}}, z_{1:t-1}^{B,i}\right)}{p\left(z_t^{A,i}, z_t^{A,J_t}, z_t^{B,i} \mid z_{1:t-1}^{A,i}, z_{1:t-1}^{A,J_{1:t-1}}, z_{1:t-1}^{B,i}\right)} \qquad \text{(A.5)} \\
&\quad= p\left(z_t^{A,i} \mid x_t^{A,i}\right) p\left(x_t^{A,i} \mid x_{0:t-1}^{A,i}\right) p\left(z_t^{A,J_t} \mid x_t^{A,i}, z_t^{A,i}\right) \\
&\qquad\times \frac{p\left(z_t^{B,i} \mid x_t^{A,i}\right) p\left(x_{0:t-1}^{A,i} \mid z_{1:t-1}^{A,i}, z_{1:t-1}^{A,J_{1:t-1}}, z_{1:t-1}^{B,i}\right)}{p\left(z_t^{A,i}, z_t^{A,J_t}, z_t^{B,i} \mid z_{1:t-1}^{A,i}, z_{1:t-1}^{A,J_{1:t-1}}, z_{1:t-1}^{B,i}\right)}. \qquad \text{(A.6)}
\end{aligned}$$

In (A.5), *Markov* property (v) is used. In (A.6), property (iv) is applied.

ACKNOWLEDGMENTS

We would like to thank the anonymous reviewers for constructive comments and suggestions. We wish to express our gratitude to Lu Wang from the University of Southern California for his help with estimating the epipolar geometry of the real-world videos. We would also like to thank Yuanyuan Jia for her help.

REFERENCES

[1] Y. Bar-Shalom and A. G. Jaffer, *Tracking and Data Association*, Academic Press, San Diego, Calif, USA, 1998.

[2] C. Hue, J.-P. L. Cadre, and P. Pérez, "Sequential Monte Carlo methods for multiple target tracking and data fusion," *IEEE Transactions on Signal Processing*, vol. 50, no. 2, pp. 309–325, 2002.

[3] J. MacCormick and A. Blake, "A probabilistic exclusion principle for tracking multiple objects," *International Journal of Computer Vision*, vol. 39, no. 1, pp. 57–71, 2000.

[4] N. Gordon, "A hybrid bootstrap filter for target tracking in clutter," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 33, no. 1, pp. 353–358, 1997.

[5] M. Isard and J. MacCormick, "BraMBLe: a Bayesian multiple-blob tracker," in *Proceedings of 8th IEEE International Conference on Computer Vision (ICCV '01)*, vol. 2, pp. 34–41, Vancouver, BC, Canada, July 2001.

[6] T. Zhao and R. Nevatia, "Tracking multiple humans in crowded environment," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '04)*, vol. 2, pp. 406–413, Washington, DC, USA, June-July 2004.

[7] T. Zhao and R. Nevatia, "Tracking multiple humans in complex situations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 9, pp. 1208–1221, 2004.

[8] H. Tao, H. Sawhney, and R. Kumar, "A sampling algorithm for detection and tracking multiple objects," in *Proceedings of IEEE International Conference on Computer Vision (ICCV '99) Workshop on Vision Algorithm*, Corfu, Greece, September 1999.

[9] Z. Khan, T. Balch, and F. Dellaert, "An MCMC-based particle filter for tracking multiple interacting targets," in *Proceedings of 8th European Conference on Computer Vision (ECCV '04)*, vol. 4, pp. 279–290, Prague, Czech Republic, May 2004.

[10] K. Smith, D. Gatica-Perez, and J.-M. Odobez, "Using particles to track varying numbers of interacting people," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '05)*, vol. 1, pp. 962–969, San Diego, Calif, USA, June 2005.

[11] S. J. McKenna, S. Jabri, Z. Duric, A. Rosenfeld, and H. Wechsler, "Tracking groups of people," *Computer Vision and Image Understanding*, vol. 80, no. 1, pp. 42–56, 2000.

[12] T. Yu and Y. Wu, "Collaborative tracking of multiple targets," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '04)*, vol. 1, pp. 834–841, Washington, DC, USA, June-July 2004.

[13] Y. Wu, G. Hua, and T. Yu, "Tracking articulated body by dynamic Markov network," in *Proceedings of 9th IEEE International Conference on Computer Vision (ICCV '03)*, vol. 2, pp. 1094–1101, Nice, France, October 2003.

[14] W. Qu, D. Schonfeld, and M. Mohamed, "Real-time interactively distributed multi-object tracking using a magnetic-inertia potential model," in *Proceedings of 10th IEEE International Conference on Computer Vision (ICCV '05)*, vol. 1, pp. 535–540, Beijing, China, October 2005.

[15] Q. Cai and J. K. Aggarwal, "Tracking human motion in structured environments using a distributed-camera system,"

IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 21, no. 11, pp. 1241–1247, 1999.

[16] A. Utsumi and J. Ohya, "Multiple-camera-based human tracking using non-synchronous observations," in *Proceedings of 4th Asian Conference on Computer Vision (ACCV '00)*, pp. 1034–1039, Taipei, Taiwan, January 2000.

[17] P. H. Kelly, A. Katkere, D. Y. Kuramura, S. Moezzi, S. Chatterjee, and R. Jain, "An architecture for multiple perspective interactive video," in *Proceedings of 3rd ACM International Conference on Multimedia (ACM Multimedia '95)*, pp. 201–212, San Francisco, Calif, USA, November 1995.

[18] J. Black and T. Ellis, "Multiple camera image tracking," in *Proceedings of 2nd IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS '01)*, Kauai, Hawaii, USA, December 2001.

[19] C. Stauffer and K. Tieu, "Automated multi-camera planar tracking correspondence modeling," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '03)*, vol. 1, pp. 259–266, Madison, Wis, USA, June 2003.

[20] L. Lee, R. Romano, and G. Stein, "Monitoring activities from multiple video streams: establishing a common coordinate frame," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 758–767, 2000.

[21] S. Khan and M. Shah, "Consistent labeling of tracked objects in multiple cameras with overlapping fields of view," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 10, pp. 1355–1360, 2003.

[22] V. Kettnaker and R. Zabih, "Bayesian multi-camera surveillance," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '99)*, vol. 2, pp. 253–259, Fort Collins, Colo, USA, June 1999.

[23] O. Javed, Z. Rasheed, K. Shafique, and M. Shah, "Tracking across multiple cameras with disjoint views," in *Proceedings of the 9th IEEE International Conference on Computer Vision (ICCV '03)*, vol. 2, pp. 952–957, Nice, France, October 2003.

[24] A. Rahimi, B. Dunagan, and T. Darrell, "Simultaneous calibration and tracking with a network of non-overlapping sensors," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '04)*, vol. 1, pp. 187–194, Washington, DC, USA, June-July 2004.

[25] C. Micheloni, G. L. Foresti, and L. Snidaro, "A network of cooperative cameras for visual surveillance," *IEE Proceedings, Vision, Image & Signal Processing*, vol. 152, no. 2, pp. 205–212, 2005.

[26] T.-H. Chang and S. Gong, "Tracking multiple people with a multi-camera system," in *Proceedings of IEEE Workshop on Multi-Object Tracking*, pp. 19–26, Vancouver, BC, Canada, July 2001.

[27] S. Iwase and H. Saito, "Parallel tracking of all soccer players by integrating detected positions in multiple view images," in *Proceedings of the 17th International Conference on Pattern Recognition (ICPR'04)*, vol. 4, pp. 751–754, Cambridge, UK, August 2004.

[28] A. Mittal and L. Davis, "Unified multi-camera detection and tracking using region-matching," in *Proceedings of IEEE Workshop on Multi-Object Tracking*, pp. 3–10, Vancouver, BC, Canada, July 2001.

[29] D. Gatica-Perez, J.-M. Odobez, S. Ba, K. Smith, and G. Lathoud, "Tracking people in meetings with particles," in *Proceedings of International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS '05)*, Montreux, Switzerland, April 2005.

[30] K. Nummiaro, E. B. Koller-Meier, T. Svoboda, D. Roth, and L. V. Gool, "Color-based object tracking in multi-camera environments," in *Proceedings of 25th DAGM Symposium on Pattern Recognition*, pp. 591–599, Magdeburg, Germany, September 2003.

[31] W. Du and J. Piater, "Multi-view tracking using sequential belief propagation," in *Proceedings of Asian Conference on Computer Vision (ACCV '06)*, Hyderabad, India, January 2006.

[32] H. Tsutsui, J. Miura, and Y. Shirai, "Optical flow-based person tracking by multiple cameras," in *Proceedings of International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI '01)*, pp. 91–96, Baden-Baden, Germany, August 2001.

[33] T. Zhao, M. Aggarwal, R. Kumar, and H. Sawhney, "Real-time wide area multi-camera stereo tracking," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '05)*, vol. 1, pp. 976–983, San Diego, Calif, USA, June 2005.

[34] S. L. Dockstader and A. M. Tekalp, "Multiple camera tracking of interacting and occluded human motion," *Proceedings of the IEEE*, vol. 89, no. 10, pp. 1441–1455, 2001.

[35] M. Xu, J. Orwell, L. Lowey, and D. Thirde, "Architecture and algorithms for tracking football players with multiple cameras," *IEE Proceedings, Vision, Image & Signal Processing*, vol. 152, no. 2, pp. 232–241, 2005.

[36] J. Whittaker, *Graphical Models in Applied Mathematical Multivariate Statistics*, John Wiley & Sons, Chichester, UK, 1990.

[37] M. Isard and A. Blake, "CONDENSATION—conditional density propagation for visual tracking," *International Journal of Computer Vision*, vol. 29, no. 1, pp. 5–28, 1998.

[38] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking," *IEEE Transactions on Signal Processing*, vol. 50, no. 2, pp. 174–188, 2002.

[39] N. Bergman, "Recursive Bayesian estimation: navigation and tracking applications," Ph.D. dissertation, Linköping University, Linköping, Sweden, 1999.

[40] A. Doucet, "On sequential simulation-based methods for Bayesian filtering," Tech. Rep. CUED/F-INFENG/TR 310, Cambridge University, Department of Engineering, Cambridge, UK, 1998.

[41] W. Qu and D. Schonfeld, "Detection-based particle filtering for real-time multiple-head tracking applications," in *Image and Video Communications and Processing 2005*, vol. 5685 of *Proceedings of SPIE*, pp. 411–418, San Jose, Calif, USA, January 2005.

[42] K. Okuma, A. Taleghani, N. de Freitas, J. J. Little, and D. G. Lowe, "A boosted particle filter: multitarget detection and tracking," in *Proceedings of 8th European Conference on Computer Vision (ECCV '04)*, vol. 1, pp. 28–39, Prague, Czech Republic, May 2004.

[43] Y. Rui and Y. Chen, "Better proposal distributions: object tracking using unscented particle filter," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '01)*, vol. 2, pp. 786–793, Kauai, Hawaii, USA, December 2001.

[44] N. Bouaynaya, W. Qu, and D. Schonfeld, "An online motion-based particle filter for head tracking applications," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '05)*, Philadelphia, Pa, USA, March 2005.

[45] S. Birchfield, "Elliptical head tracking using intensity gradients and color histograms," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '98)*, pp. 232–237, Santa Barbara, Calif, USA, June 1998.

[46] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge University Press, Cambridge, UK, 2004.
[47] K. Nummiaro, E. B. Koller-Meier, and L. J. Van Gool, "Object tracking with an adaptive color-based particle filter," in *Proceedings of the 24th DAGM Symposium on Pattern Recognition*, pp. 353–360, Zurich, Switzerland, September 2002.
[48] B. Moghaddam and A. Pentland, "Probabilistic visual learning for object representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 696–710, 1997.
[49] O. Tuzel, F. Porikli, and P. Meer, "A Bayesian approach to background modeling," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (MVIV '05)*, vol. 3, pp. 58–58, San Diego, Calif, USA, June 2005.
[50] J. Black, T. Ellis, and P. Rosin, "A novel method for video tracking performance evaluation," in *Proceedings of Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, Nice, France, October 2003.
[51] http://vspets.visualsurveillance.org/.

Wei Qu was born in 1977. He received the B.E. degree in automatic control from Beijing Institute of Technology, Beijing, China, in 2000. From 2000 to 2002, he was a Research Assistant in Institute of Automation, Chinese Academy of Sciences, Beijing, China. In 2002, he became a Ph.D. Student in Electrical and Computer Engineering Department, University of Illinois at Chicago, where he received the M.S. degree in electrical and computer engineering in 2005. He is currently a Ph.D. candidate in Multimedia Communications Laboratory, Electrical and Computer Engineering Department, University of Illinois at Chicago. From 2004 to 2005, he was a Research Assistant with Visual Communications and Display Technologies Lab, Motorola Labs, Schaumburg, Ill. In the summer of 2005, he was a Research Intern with Mitsubishi Electric Research Laboratories (MERL), Cambridge, Mass. He has authored over 20 technical papers in various journals and conferences. He is a student member of the Institute of Electrical and Electronics Engineers (IEEE), International Neural Network Society (INNS), and International Society of Optical Engineering (SPIE). His current research interests are video tracking, retrieval, multimedia processing, pattern recognition, computer vision, and machine learning.

Dan Schonfeld was born in Westchester, Pennsylvania, on June 11, 1964. He received the B.S. degree in electrical engineering and computer science from the University of California at Berkeley, and the M.S. and Ph.D. degrees in electrical and computer engineering from the Johns Hopkins University, in 1986, 1988, and 1990, respectively. In 1990, he joined the University of Illinois at Chicago, where he is currently an Associate Professor in the Department of Electrical and Computer Engineering. He has authored over 100 technical papers in various journals and conferences. He was a co-author of a paper that won the Best Student Paper Award in Visual Communication and Image Processing 2006. He was also a co-author of a paper that was a finalist in the Best Student Paper Award in Image and Video Communication and Processing 2005. He has served as an Associate Editor of the IEEE Transactions on Image Processing on Nonlinear Filtering as well as an Associate Editor of the IEEE Transactions on Signal Processing on Multidimensional Signal Processing and Multimedia Signal Processing. His current research interests are in signal, image, and video processing; video communications; video retrieval; video networks; image analysis and computer vision; pattern recognition; and genomic signal processing.

Magdi Mohamed received the B.Sc. (electrical engineering) degree from the University of Khartoum, Sudan, and the M.S. (computer science) and Ph.D. (electrical and computer engineering) degrees from the University of Missouri-Columbia, in 1983, 1990, and 1995, respectively. During 1985–1988, he worked as a Teaching Assistant at the Computer Center, University of Khartoum. He also worked as a Computer Engineer at ComputerMan, Khartoum, Sudan, and as an Electrical and Communication Engineer at SNBC in Omdurman, Sudan. From 1991 to 1995 he was a Research and Teaching Assistant in the Department of Electrical and Computer Engineering, University of Missouri-Columbia. He worked as a Visiting Professor in the Computer Engineering and Computer Science Department at the University of Missouri-Columbia from 1995 to 1996. He has been with Motorola Inc. since 1996 where he is currently working as a Principal Staff Engineer at Motorola Labs, Physical Realization Research Center of Excellence, in Schaumburg, Ill, USA. His research interests include offline and online handwriting recognition, optical motion tracking, digital signal and image processing, computer vision, fuzzy sets and systems, neural networks, pattern recognition, parallel and distributed computing, nonlinear and adaptive modeling techniques, machine intelligence, fractals and chaos theory.