DECENTRALIZED MULTIPLE CAMERA MULTIPLE OBJECT TRACKING

Wei Qu, Dan Schonfeld

ECE Department, University of Illinois
Chicago, IL, USA, 60607
{wqu, ds}@ece.uic.edu

Magdi Mohamed

Visual Communication & Display Technologies
Motorola Labs, Schaumburg, IL, USA, 60196
Magdi.Mohamed@motorola.com

ABSTRACT

In this paper, we present a novel decentralized Bayesian framework using multiple collaborative cameras for robust and efficient multiple object tracking with significant and persistent occlusion. This approach avoids the common practice of using a complex joint state representation and a centralized processor for multiple camera tracking. When the objects are in close proximity or present multi-object occlusions in a particular camera view, camera collaboration between different views is activated in order to handle the multi-object occlusion problem. Specifically, we propose to model the camera collaboration likelihood density by using epipolar geometry with particle filter implementation. The performance of our approach has been demonstrated on both synthetic and real-world video data.

1. INTRODUCTION AND RELATED WORK

Multiple Object Tracking (MOT) has received tremendous attention due to its numerous potential applications such as smart video surveillance and human computer interfaces. In addition to all of the challenging problems inherent in single object tracking, MOT has to deal with multi-object occlusion, namely, the tracker must separate the objects and assign them correct labels.

Most early efforts for MOT use monocular video. Many existing approaches that address the difficulties of this challenging task are based on a centralized process and using a joint state space representation [1]. Although these solutions based on a centralized process can handle the problem of multi-object occlusion in principle, the joint state representation introduces much high complexity and requires exponentially increased computational cost with the number of tracking objects [2]. Several researchers proposed decentralized solutions to multi-object tracking. Yu et al. [3] use multiple collaborative trackers for MOT modeled by a Markov random network. This approach demonstrates the efficiency of the decentralized method. The decentralized concept was carried further by Qu et al. [2] who proposed an Interactively Distributed Multi-Object Tracking (IDMOT) framework using a magnetic-inertia potential model for MOT.

Monocular video has intrinsic limitations for MOT, especially in solving multi-object occlusion, due to the camera's limited field of view and the loss of the objects' depth information by camera projection. These limitations have recently inspired researchers to exploit multi-ocular videos, where expanded coverage of the environment is provided and occluded objects in one camera view may not be occluded in others. However, using multiple cameras raises many additional challenges. The most critical difficulties presented by multi-camera tracking are to establish a consistent label correspondence of the same object among the different views and to integrate the information from different camera views for tracking that is robust to significant and persistent occlusion. Many approaches address the label correspondence problem [4]. Establishing temporal instead of spatial label correspondences between non-overlapping fields of view is discussed in [5]. Integration of information from multiple cameras to solve the multi-object occlusion problem has been investigated in [6], [7]. However, either joint state representation or different central processors have been used to integrate observations from multiple cameras.

In this paper, we extend the concept of decentralized tracking using one camera presented in [2], [3] into a more complicated context of multiple overlapping cameras. The objective is to provide an efficient solution to the multi-object occlusion problem by exploiting the cooperation of multi-ocular videos. This approach avoids the computational complexity inherent in centralized methods that rely on joint state representation and/or joint data association in the earlier multiple camera tracking approaches.

2. DECENTRALIZED BAYESIAN FORMULATION

We use multiple trackers, one tracker per object in each camera view for MOT in multi-ocular videos. Without loss of generality, we illustrate our framework by using two cameras. Similar to the notations in [2], we denote the state of an object in camera A by $x_t^{A,i}$, where $i = 1, \ldots, M$ is the index of objects, 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 initialization prior, the set of all observations

up to time t by $z_{1:t}^{A,i}$. Similarly, we can denote the notations for objects in camera *B*, for instance, the *"counterpart"* of $x_t^{A,i}$ is $x_t^{B,i}$. We further denote the interactive observations of $z_t^{A,i}$ at time t by z_t^{A,J_t} where $J_t = \{j_{l_1}, j_{l_2}, \ldots\}$. We define an object to have *"interaction"* when it touches or even occludes with other objects 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 objects whose observations interact with $z_t^{A,i}$. In addition, $z_{1:t}^{A,J_{1:t}}$ represents the collection of interactive observation sets up to time t.

We propose to estimate the posterior of an object in a camera based on not only the interactive observations from the same camera but also the counterpart observations from all the other related cameras, i.e., $p(x_{0:t}^{A,i}|z_{1:t}^{A,i}, z_{1:t}^{A,J_{1:t}}, z_{1:t}^{B,i})$. Three assumptions are made in derivation: (i) similar to [8], [9], we assume observations in different time are independent, both mutually and with respect to the dynamical process; (ii) given an object's state and observation in a particular camera, its counterpart observations in other cameras are conditionally independent with other objects' observations in this camera view; (iii) given an object's state, the associated observations in different cameras are conditionally independent.

$$
\begin{aligned}
& p(x_{0:t}^{A,i}|z_{1:t}^{A,i}, z_{1:t}^{A,J_{1:t}}, z_{1:t}^{B,i}) \\
&= \frac{p(z_t^{A,i}|x_t^{A,i}, z_t^{A,J_t}, z_t^{B,i}) p(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}|z_{1:t-1}^{A,i}, z_{1:t}^{A,J_{1:t}}, z_{1:t}^{B,i})} \quad (1) \\
&= \frac{p(z_t^{A,i}|x_t^{A,i}) p(z_t^{A,J_t}, z_t^{B,i}|x_t^{A,i}, z_t^{A,i})}{p(z_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,J_t}, z_t^{B,i}|x_t^{A,i})} \\
&\quad \cdot \frac{p(x_t^{A,i}, z_t^{A,J_t}, z_t^{B,i}|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(z_t^{A,J_t}, z_t^{B,i}|z_{1:t-1}^{A,i}, z_{1:t-1}^{A,J_{1:t-1}}, z_{1:t-1}^{B,i})} \\
&\quad \cdot p(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}) \quad (2) \\
&= \frac{1}{k_t} p(z_t^{A,i}|x_t^{A,i}) p(z_t^{A,J_t}, z_t^{B,i}|x_t^{A,i}, z_t^{A,i}) \\
&\quad \cdot p(x_t^{A,i}, z_t^{A,J_t}, z_t^{B,i}|x_{0:t-1}^{A,i}) \frac{1}{p(z_t^{A,J_t}, z_t^{B,i}|x_t^{A,i})} \\
&\quad \cdot p(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}) \quad (3) \\
&= \frac{1}{k_t} p(z_t^{A,i}|x_t^{A,i}) p(z_t^{A,J_t}, z_t^{B,i}|x_t^{A,i}, z_t^{A,i}) p(x_t^{A,i}|x_{0:t-1}^{A,i}) \\
&\quad \cdot p(z_t^{A,J_t}, z_t^{B,i}|x_t^{A,i}, x_{0:t-1}^{A,i}) \frac{1}{p(z_t^{A,J_t}, z_t^{B,i}|x_t^{A,i})} \\
&\quad \cdot p(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}) \quad (4) \\
&= \frac{1}{k_t} p(z_t^{A,i}|x_t^{A,i}) p(z_t^{A,J_t}, z_t^{B,i}|x_t^{A,i}, z_t^{A,i}) p(x_t^{A,i}|x_{0:t-1}^{A,i}) \\
&\quad \cdot p(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}) \quad (5) \\
&= \frac{1}{k_t} p(z_t^{A,i}|x_t^{A,i}) p(z_t^{B,i}|x_t^{A,i}, z_t^{A,i}) p(z_t^{A,J_t}|x_t^{A,i}, z_t^{A,i}) \\
&\quad \cdot p(x_t^{A,i}|x_{0:t-1}^{A,i}) p(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}) \quad (6) \\
&= \frac{1}{k_t} p(z_t^{A,i}|x_t^{A,i}) p(z_t^{B,i}|x_t^{A,i}) p(z_t^{A,J_t}|x_t^{A,i}, z_t^{A,i}) \\
&\quad \cdot p(x_t^{A,i}|x_{0:t-1}^{A,i}) p(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}). \quad (7)
\end{aligned}
$$

In the denominator of (2), densities $p(z_t^{A,i}|z_{1:t-1}^{A,i}, z_{1:t}^{A,J_{1:t}}, z_{1:t}^{B,i})$ and $p(z_t^{A,J_t}, z_t^{B,i}|z_{1:t-1}^{A,i}, z_{1:t-1}^{A,J_{1:t-1}}, z_{1:t-1}^{B,i})$ are all unrelated to $x^{A,i}$, thus their product can be regarded as a constant k_t. According to assumption (i), in (1) we make a simplification $p(z_t^{A,i}|x_{0:t}^{A,i}, z_{1:t-1}^{A,i}, z_{1:t}^{A,J_{1:t}}, z_{1:t}^{B,i}) = p(z_t^{A,i}|x_t^{A,i}, z_t^{A,J_t}, z_t^{B,i})$. Similarly, we make simplifications in (3), (5) respectively by assumption (i). In (6), we use the assumption (ii). From (6) to (7), we have exploited the assumption (iii). In (7), $p(z_t^{A,i}|x_t^{A,i})$ is the local observation likelihood, $p(x_t^{A,i}|x_{0:t-1}^{A,i})$ is the state dynamics. $p(z_t^{A,J_t}|x_t^{A,i}, z_t^{A,i})$ is the *"interactive likelihood"* between the tracked object's observation and its interactive observations in the same camera similar to [2]. The main novelty of this paper is that we introduce an additional likelihood density $p(z_t^{B,i}|x_t^{A,i})$ called a *"camera collaboration likelihood"* to characterize the collaboration between the same object's counterparts in different views. When not activating the camera collaboration for an object and regarding its projections in different views as independent, the proposed framework can degrade to the IDMOT approach [2] by switching $p(z_t^{B,i}|x_t^{A,i})$ to a uniform distribution.

3. DENSITY ESTIMATION

We describe a particle filtering implementation [8] of the derived Bayesian formulation in this section. A particle set $\{x_{0:t}^{A,i,n}, w_t^{A,i,n}\}_{n=1}^{N_p}$ is employed to represent the posterior $p(x_{0:t}^{A,i}|z_{1:t}^{A,i}, z_{1:t}^{A,J_{1:t}}, z_{1:t}^{B,i})$, where $\{x_{0:t}^{A,i,n}, n = 1, \ldots, N_p\}$ are the particles, $\{w_t^{A,i,n}, n = 1, \ldots, N_p\}$ are associated weights and N_p is the number of particles. According to the *sequential importance sampling theory* [8], considering the derived sequential iteration equation (7), if the particles $x_{0:t}^{A,i,n}$ are sampled from the density $p(x_t^{A,i}|x_{0:t-1}^{A,i,n})$ which is modeled as a Gaussian random walk, the corresponding weights are given by

$$
w_t^{i,n} \propto w_{t-1}^{i,n} p(z_t^{A,i}|x_t^{A,i,n}) p(z_t^{A,J_t}|x_t^{A,i,n}, z_t^{A,i}) p(z_t^{B,i}|x_t^{A,i,n}) \quad (8)
$$

In (8), the local likelihood $p(z_t^{A,i}|x_t^{A,i,n})$ can be calculated by fusing object's color histogram with a PCA-based model similar to [3], [2]; The interactive likelihood $p(z_t^{A,J_t}|x_t^{A,i}, z_t^{A,i})$ can be estimated similarly by the *"magnetic repulsion model"* presented in [2]; The camera collaboration likelihood can be estimated by the model discussed as follows.

3.1. Camera Collaboration Likelihood Model

The proposed framework has no specific requirement of the camera collaboration model as long as it can give a relatively

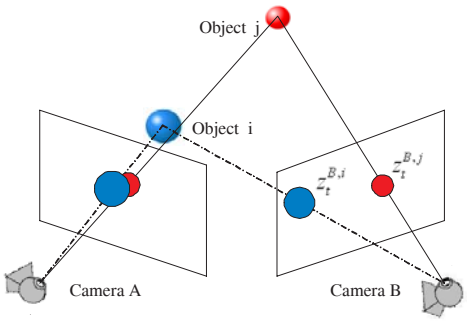

Fig. 1. The model setting in 3D space for camera collaboration likelihood estimation.

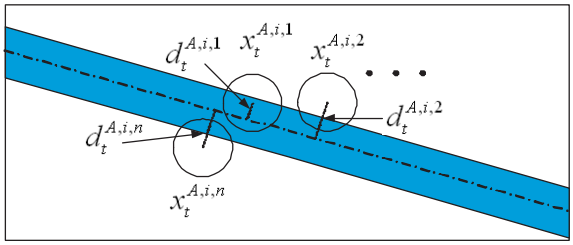

Fig. 2. Calculating the camera collaboration weights for object i in view A.

good estimation of density $p(z_t^{B,i}|x_t^{A,i})$. Here we present a paradigm using particle filter implementation without recovering object's 3D coordinates but only assuming cameras' epipolar geometry is known.

Fig. 1 illustrates the model setting in 3D space. Two objects i and j are projected to two camera views. In view A, the projections of object i and j are occluding while in view B they are not. $z_t^{B,i}$ and $z_t^{B,j}$ are roughly estimated by doing tracking in view B firstly. 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, object i is occluded 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 roughly estimate the observations. We do admit these initial estimates may be not very accurate, therefore, in this case, we do iteration several times (usually twice is enough) between different views to get more stable estimates. According to *Epipolar Geometry Theory* [10, p.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 corresponds to an epipolar *"band"* in view A, i.e., $\hbar(z_t^{B,i})$.

Fig. 2 shows the procedure 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. 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,

$$\phi_t^{A,i,n} = \frac{1}{\sqrt{2\pi}\Sigma_\phi} exp\{-\frac{(d_t^{A,i,n})^2}{2\Sigma_\phi^2}\} \quad (9)$$

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

$$p(z_t^{B,i}|x_t^{A,i}) \approx \sum_{n=1}^{N_p} \frac{\phi_t^{A,i,n}}{\sum_{n'=1}^{N_p} \phi_t^{A,i,n'}} \delta(x_t^{A,i} - x_t^{A,i,n}) \quad (10)$$

where $\delta(\cdot)$ is the Dirac delta function.

4. EXPERIMENTAL RESULTS

The performance of our approach has been demonstrated on both synthetic and real-world data. We use a five dimension parametric ellipse model including the center, size and orientation parameters to represent an object's state. Different colors and numbers are used to label the objects. Without code optimization, the C++ implementation runs stably at $8 \sim 12$ frames per second for the testing sequences on a 3.2GHz Pentium IV PC.

We generate the synthetic videos by assuming two cameras are widely set at right angle and at the same height above the ground. Six soccer balls moves differently within the overlapped scene of the two views. Various multi-object occlusions are frequent when the objects are projected onto each view. In Fig. 3, we compare the tracking results of (a) Multiple Independent Particle Filter (MIPF) [9], (b) IDMOT and (c) the proposed approach. MIPF severely suffers from multi-object occlusion problem. A lot of trackers are "hijacked" by the objects with strong local observation and thus lose their associated objects after occlusion. Equipped with magnetic repulsion and inertia models to handle the object interaction, IDMOT has improved the performance separating the occluding objects and labeling them correctly for many objects. However, due to the intrinsic limitations of monocular video, it still has failure cases. By using bi-ocular videos and exploiting camera collaboration, the proposed approach tracks all the objects robustly. The epipolar line through the center of a object is mapped from its counterpart in another view and reveals when the camera collaboration is activated.

The sequence `UnionStation` is captured at a railway station using two widely separated digital cameras with a resolution of 320×240 pixels and a frame rate of 25 frames per second. The crowded scene has various persistent and significant multi-object occlusions when pedestrians passing by each other. Each view sequence consists of 697 frames.

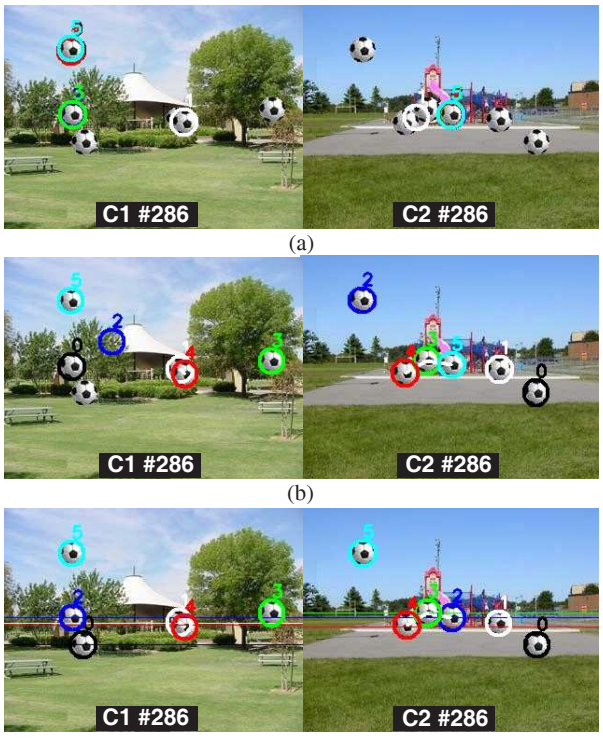

Fig. 3. Comparison of the tracking results on synthetic videos. (a) MIPF; (b) IDMOT; (c) The proposed approach.

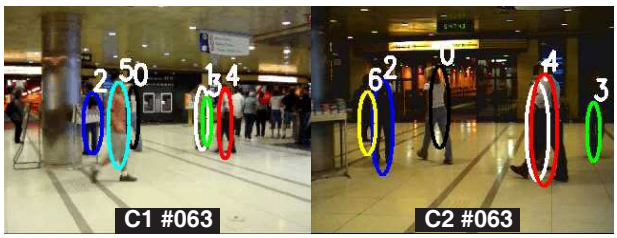

Fig. 4. Tracking results of the sequence `UnionStation`.

The fundamental matrix of epipolar geometry is estimated by using the algorithm proposed by Hartley and Zisserman [10, p.79-308]. Fig. 4 shows the tracking results using the proposed approach. It can be seen that objects are tracked robustly and assigned with correct labels even after persistent and severe multi-object occlusions due to using both interaction within each view and the camera collaboration between different views. The only failure case occurs when two objects present occlusion in one view and there is no counterparts of these objects appearing in another view. In this scenario, no camera collaboration is activated at all. By using more cameras to cover the scene, such kind or failure cases should be decreased.

Compared with the centralized approaches [6], whose computational complexity increases exponentially with the number of objects and cameras due to using a joint state representation, the computational complexity of the proposed approach increases linearly in terms of the number of objects and cameras. In table 1, we show the computation analysis in terms of the number of objects for our approach. As we can see, the requirement of the number of particles to achieve reasonable robust tracking performance is increased linearly.

Table 1. Computation analysis in terms of the number of objects for the proposed approach

Objects	5	6	7
Particles	600	720	840
Speed (fps)	$11 \sim 12$	$9 \sim 10$	$8.1 \sim 8.6$

5. REFERENCES

[1] M. Isard and J. MacCormick, "Bramble: A bayesian multiple-blob tracker," in *ICCV*, 2001.

[2] Wei Qu, Dan Schonfeld, and Magdi Mohamed, "Real-time interactively distributed multi-object tracking using a magnetic-inertia potential model," in *ICCV*, 2005.

[3] Ting Yu and Ying Wu, "Collaborative tracking of multiple targets," in *CVPR*, 2004.

[4] Sohaib Khan and Mubarak Shah, "Consistent labeling of tracking objects in multiple cameras with overlapping fields of view," *PAMI*, vol. 25, no. 10, pp. 1355–1360, 2003.

[5] Omar Javed, Zeeshan Rasheed, K. Shafique, and Mubarak Shah, "Tracking across multiple cameras with disjoint views," in *ICCV*, 2003.

[6] Tao Zhao, Manoj Aggarwal, Rakesh Kumar, and Harpreet Sawhney, "Real-time wide area multi-camera stereo tracking," in *CVPR*, 2005.

[7] Shiloh L. Dockstader and A. Murat Tekalp, "Multiple camera tracking of interacting and occluded human motion," in *Proc. IEEE*, 2001, vol. 89, pp. 1441–1455.

[8] M. Sanjeev Arulampalam, Simon Maskell, N. Gordon, and Tim Clapp, "A tutorial on particle filters for online nonlinear/non-gaussian baysian tracking," *IEEE Trans. Signal Processing*, vol. 50, no. 2, pp. 174–188, 2002.

[9] Michael Isard and Andrew Blake, "Condensation – conditional density propagation for visual tracking," *Int. J. Computer Vision*, vol. 29, no. 1, pp. 5–28, 1998.

[10] Richard Hartley and Andrew Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge University Press, United Kingdom, 2004.