Trajectory Association across Multiple Airborne Cameras

Yaser Ajmal Sheikh, *Member*, *IEEE*, and
Mubarak Shah, *Fellow*, *IEEE*

Abstract—A camera mounted on an aerial vehicle provides an excellent means to monitor large areas of a scene. Utilizing several such cameras on different aerial vehicles allows further flexibility in terms of increased visual scope and in the pursuit of multiple targets. In this paper, we address the problem of associating trajectories across multiple moving airborne cameras. We exploit geometric constraints on the relationship between the motion of each object across cameras without assuming any prior calibration information. Since multiple cameras exist, ensuring coherency in association is an essential requirement, e.g., that transitive closure is maintained between more than two cameras. To ensure such coherency, we pose the problem of maximizing the likelihood function as a k-dimensional matching and use an approximation to find the optimal assignment of association. Using the proposed error function, canonical trajectories of each object and optimal estimates of intercamera transformations (in a maximum likelihood sense) are computed. Finally, we show that, as a result of associating trajectories across the cameras, under special conditions, trajectories interrupted due to occlusion or missing detections can be repaired. Results are shown on a number of real and controlled scenarios with multiple objects observed by multiple cameras, validating our qualitative models, and, through simulation, quantitative performance is also reported.

Index Terms—Applications, scene analysis, motion, sensor fusion, registration.

1 Introduction

The concept of a cooperative multicamera ensemble has recently received increasing attention from the research community. The idea is of great practical relevance since cameras typically have limited fields of view, but are now available at low costs. Thus, instead of having a single high-resolution camera with a wide field of view that surveys a large area, far greater flexibility and scalability can be achieved by observing a scene "through many eyes," using a multitude of lower-resolution COTS (commercial off-the-shelf) cameras. Several approaches with varying constraints have been proposed, highlighting the wide applicability of cooperative sensing in practice. For instance, the problem of associating objects across multiple *stationary* cameras with overlapping fields of view has been addressed in a number of papers, e.g., [25], [3], [4], [10], [24], [20], [9], [1], [22], and [19]. Extending the problem to associating across cameras with nonoverlapping fields of view, geometric, and appearance-based approaches has also been proposed recently, e.g., [14], [18], [6], [15], [31], [29], and [30]. Camera motion has also been studied, where correspondence is estimated across pan-tilt-zoom cameras [23], [7], [17]. In general, when using sensors in such a decentralized but cooperative fashion, knowledge of intercamera relationships becomes of paramount importance in understanding what happens in the environment. Without such information, it is difficult to tell, for instance, whether an object viewed in each of two cameras is the same object or a new object. Two cues available to infer this are the appearance and the motion of the object. For the interested reader, some notable papers describing the use of appearance information for association include [16], [29], and [30].

In this paper, we consider the problem of inferring the correct association based on the motion of the object, contained wholly in its trajectory, and describe an approach to recover the most likely association given our scene model. Furthermore, we address the problem of recovering the optimal estimates of the scene parameters from the given observations. We make two fundamental assumptions about the data: 1) That the altitude of the aerial vehicle upon which the camera is mounted is significantly high with respect to the ground (and, so, a planar assumption is viable) and 2) that at least one object is seen simultaneously between every pair of cameras for at least five frames. Given these assumptions and taking as input the timestamped trajectories of objects observed in each camera, we estimate the intercamera transformations, the association of each object across the views, and "canonical" trajectories, which are the best estimate (in a maximum likelihood sense) of the original object trajectories up to a 2D projective transformation. To that end, we describe an extension to the reprojection error for multiple views, providing a geometrically and statistically sound means of evaluating the likelihood of a candidate correspondence set. We formulate the problem of maximizing this joint likelihood function as a k-dimensional matching problem and use an approximation that maintains transitive closure. The estimated solution is verified using a strong global constraint for the complete set of correspondences across all cameras. We evaluated the proposed approach with both simulated and real data. During evaluation, the object association problem within each sequence (single camera tracking) is considered to have already been solved and the solution of this module in each camera is taken as input. The rest of the paper is organized as follows: Section 2 describes the estimation of intercamera relationships and the problem of association is posed as a maximum likelihood assignment. The problem of association is posed and solved in a graph-theoretic framework. Results on controlled and real sequences are shown in Section 3, with conclusions in Section 4.

2 Trajectory Association across Cameras

In this section, an unsupervised approach is presented to estimating the intercamera relationships in terms of the interframe homography. We describe how the likelihood that trajectories, observed in different cameras, originating from the same world object, is estimated. The use of this, in turn, for multiple object assignment across multiple cameras is then described next. The scene is modeled as a plane in 3-space, $\mathbf{\Pi}$, with K moving objects, observed by N cameras. The kth object moves along a trajectory on $\mathbf{\Pi}$, represented by a time-ordered set of points. A particular object k, present in the field of view of camera n, is denoted as O_k^n and the imaged location of O_k^n at time t is $\mathcal{X}_k^n(t) = (x_{k,t}^n, y_{k,t}^n, \lambda_{k,t}^n)^{\mathrm{T}} \in \mathbb{P}^2$, the homogeneous coordinates of the point in sequence n. The imaged trajectory of O_k^n is the sequence of points $\mathcal{X}_k^n = \{\mathcal{X}_k^n(i), \mathcal{X}_k^n(i+1), \ldots \mathcal{X}_k^n(j)\}$. When referring to inhomogeneous coordinates, we will refer to a point as $\mathbf{x}_k^n(t) = (x_{k,t}^n/\lambda_{k,t}^n, y_{k,t}^n/\lambda_{k,t}^n)^{\mathrm{T}} \in \mathbb{R}^2$. For two cameras, an association or correspondence $c_{k,l}^{n,m}$ is an ordered pair (O_k^n, O_l^m) that represents the hypothesis that O_k^n and O_l^m are images of the same object. Formally, it defines the event, $c_{k,l}^{n,m} \doteq \{O_k^n \text{ and } O_l^m \text{ arise from the same object in the world}\}$, $l = 1, \cdots, \mathbf{z}(m)$, and $c_{k,0}^{n,m} \doteq \{O_k^n \text{ was not viewed in camera } m\}$, where $\mathbf{z}(m)$ is the number objects observed in camera m. Since these events are mutually exclusive and exhaustive, $\sum_{l=0}^{\mathbf{z}(\mathbf{m})} p(c_{k,l}^{n,m} | \mathcal{X}_k^n, \mathcal{X}_l^m) = 1$. Similarly, for more than two cameras, a correspondence $c_{i,j,\ldots l}^{m,n,\ldots p}$ is a hypothesis defined by the tuple $(O_i^m, O_j^n, \ldots O_l^p)$. Note that O_1^1 does not necessarily correspond to O_1^2; the numbering of objects in each sequence is in the order of detection. Thus, the problem is to find the set of associations C such that $c_{i,j,\ldots l}^{m,n,\ldots p} \in C$ if and only if $O_i^m, O_j^n, \ldots O_l^p$ are images of the same object in the world. Graphical illustration allows us to more clearly represent these different relationships (Fig. 1). We abstract the problem of tracking objects across cameras as follows: Each

- *Y.A. Sheikh is with the Robotics Institute, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213. E-mail: yaser@cs.cmu.edu.*
- *M. Shah is with the School of Electrical Engineering and Computer Science, University of Central Florida, Orlando, FL 32816. E-mail: shah@cs.ucf.edu.*

Manuscript received 5 Feb. 2006; revised 20 Sept. 2006; accepted 11 June. 2007; published online 25 July 2007.
Recommended for acceptance by H. Sawhney.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-0121-0206.
Digital Object Identifier no. 10.1109/TPAMI.2007.70750.

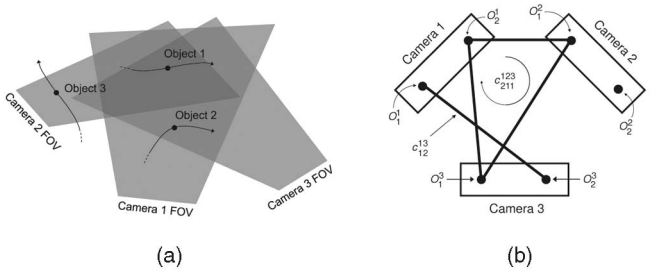

Fig. 1. (a) Three trajectories observed in three cameras. (b) The graph associated with the scenario in (a). In this instance, Object 1 is visible in all cameras, and the association across the cameras is represented by c_{211}^{123}. Object 2 is visible only in Camera 1 and Camera 3 and therefore an edge exists only between Camera 1 and 3. Object 3 is visible only in the field of view of Camera 2; therefore, there is a unconnected node in the partition corresponding to Camera 2.

observed trajectory is modeled as a node and the graph is partitioned into N partitions, one for each of the N cameras. A hypothesized association, c, between two observed objects (nodes) is represented as an edge between the two nodes. This N-partite representation is illustrated in Fig. 1. At a certain instant of time, we have $\mathbf{z}(n)$ trajectories for the nth camera corresponding to the objects visible in that camera. The measured image positions of objects $\mathbf{x}_k^n = \{\mathbf{x}_k^n(i), \mathbf{x}_k^n(i+1), \ldots \mathbf{x}_k^n(j)\}$ are described in terms of the true image positions $\bar{\mathbf{x}}_k^n = \{\bar{\mathbf{x}}_k^n(i), \bar{\mathbf{x}}_k^n(i+1), \ldots \bar{\mathbf{x}}_k^n(j)\}$. It is assumed in this work that the trajectories are compensated for by global egomotion of the camera, through the estimation of frame-to-frame homographies, and are therefore in a single coordinate system for each camera. The sample is assumed to be corrupted by independent normally distributed measurement noise, $\mu = 0$ and covariance matrix $\mathbf{R}^n(i)$,[1] that is, $\mathbf{x}_k^n(i) = \bar{\mathbf{x}}_k^n(i) + \epsilon$, $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{R}^n(i))$.

The principal assumption upon which the similarity between two trajectories is evaluated is that, due to the altitude of the aerial camera, the scene can be well approximated by a plane in 3-space and, as a result, a homography exists between any two frames of any sequence ([12]). From this assumption of planarity, it follows that a homography $\mathbf{H}_{k,l}^{n,m}$ must exist between any two trajectories that correspond, i.e., for any association hypothesis $c_{k,l}^{n,m}$. This constraint can be exploited to compute the likelihood that 2D trajectories observed by two different cameras originate from the same 3D trajectory in the world—in other words, to estimate $p(c_{k,l}^{n,m}|\mathcal{X}_k^n, \mathcal{X}_l^m)$ (which we describe presently). Furthermore, we show how this can be extended to multiple views to evaluate $p(c_{i,j,\ldots k}^{n,m,\ldots l}|\{\mathcal{X}_i^n, \mathcal{X}_j^m, \ldots \mathcal{X}_k^l\})$. By assuming conditional independence between each association c, the probability of a candidate solution C given the trajectories in multiple cameras is

$$p(C|\{\mathcal{X}\}) = \prod_{c_{i,j,\ldots k}^{n,m,\ldots l} \in C} p\Big(c_{i,j,\ldots k}^{n,m,\ldots l} | \Big\{\mathcal{X}_i^n, \mathcal{X}_j^m, \ldots \mathcal{X}_k^l\Big\}\Big). \tag{1}$$

We are interested in the Maximum Likelihood solution,

$$C^* = \arg\max_{C \in \mathcal{C}} p(C|\{\mathcal{X}\}), \tag{2}$$

where $\mathcal{C}$ is the space of solutions.

2.1 Evaluating the Likelihood of Associations

Consider first the straightforward case of several objects observed by *two* airborne cameras. This can be modeled by constructing a complete bipartite graph $G = (U, V, E)$ in which the vertices $U = \{u(\mathcal{X}_1^n), u(\mathcal{X}_2^n) \ldots u(\mathcal{X}_{\mathbf{z}(n)}^n)\}$ represent the trajectories in Sequence n and $V = \{v(\mathcal{X}_1^m), v(\mathcal{X}_2^m) \ldots v(\mathcal{X}_{\mathbf{z}(m)}^m)\}$ represent the trajectories in Sequence m and E represents the set of edges between any pair of trajectories from U and V. The bipartite graph is complete because any two trajectories may match hypothetically. The weight of each edge is the probability of correspondence of Trajectory $\mathcal{X}_l^m$ and Trajectory $\mathcal{X}_k^n$, as defined in (8). By finding the maximum matching of G, we find a unique set of correspondence C', according to the *maximum likelihood* solution,

$$C' = \arg\max_{C \in \mathcal{C}} \sum_{c_{k,l}^{n,m} \in C} \log p\Big(c_{k,l}^{n,m} | \mathcal{X}_k^n, \mathcal{X}_l^m\Big), \tag{3}$$

where $\mathcal{C}$ is the solution space. Several algorithms exist for the efficient maximum matching of a bipartite graph, for instance [21] or [13], which are $O(n^3)$ and $O(n^{2.5})$ respectively. To evaluate the likelihood of association between trajectories in two cameras, we need to evaluate $p(c_{k,l}^{n,m}|\mathcal{X}_k^n, \mathcal{X}_l^m)$. The evaluation of this likelihood is complicated by the imaging process, so, despite the fact that trajectories in correspondence can be viewed as "samples" from a single trajectory on the plane $\mathbf{\Pi}$, the coordinates of the samples are not registered. We can compute $p(c_{k,l}^{n,m}|\mathcal{X}_k^n, \mathcal{X}_l^m)$ by computing the maximum likelihood estimate of the homography, $\mathbf{H}_{k,l}^{n,m}$, and two new trajectories, $\bar{\mathcal{X}}_k^n$ and $\bar{\mathcal{X}}_l^m$, related *exactly* by $\mathbf{H}_{k,l}^{n,m}$, as described in [12], by minimizing the reprojection error. The reprojection error is a cost function that explicitly minimizes the *transfer* error between the trajectories and was first proposed by Sturm in [32], continued in further work with Chum et al. in [5]. Using this estimate of the homography and the "true" trajectories,

$$\begin{aligned} p\Big(c_{k,l}^{n,m}|\mathcal{X}_k^n, \mathcal{X}_l^m\Big) &\propto L\Big(\mathcal{X}_k^n, \mathcal{X}_l^m | c_{k,l}^{n,m}; \bar{\mathcal{X}}_k^n, \mathbf{H}_{k,l}^{n,m}\Big) \\ &= L\Big(\mathcal{X}_k^n | c_{k,l}^{n,m}; \bar{\mathcal{X}}_k^n, \mathbf{H}_{k,l}^{n,m}\Big) L\Big(\mathcal{X}_l^m | c_{k,l}^{n,m}; \bar{\mathcal{X}}_k^n, \mathbf{H}_{k,l}^{n,m}\Big). \end{aligned} \tag{4}$$

The proportionality follows from Bayes theorem, assuming a uniform prior on all associations and ignoring the constant evidence term. Since the errors at each point are assumed independent, the conditional probability of the association given the trajectories in the pair of sequences can be estimated,

$$\begin{aligned} &L\Big(\mathcal{X}_k^n, \mathcal{X}_l^m |\, c_{k,l}^{n,m}; \mathbf{H}_{k,l}^{n,m}, \bar{\mathcal{X}}_k^n\Big) = \\ &\prod_i \frac{1}{2\pi \|\mathbf{R}^n(i)\|^{\frac{1}{2}} \|\mathbf{R}^m(i)\|^{\frac{1}{2}}} e^{-\frac{1}{2}\left(d(\mathcal{X}_k^n(i), \bar{\mathcal{X}}_k^n(i))_{\mathbf{R}^n(i)} + d(\mathcal{X}_l^m(i), \bar{\mathcal{X}}_l^m(i))_{\mathbf{R}^m(i)}\right)}, \end{aligned} \tag{5}$$

where $d(\cdot)_{\mathbf{R}}$ is the Mahalanobis distance and $\mathbf{R}^n(i)$ is the error covariance matrix,

$$\begin{aligned} &d\big(\mathcal{X}_k^n(i), \bar{\mathcal{X}}_k^n(i)\big)_{\mathbf{R}^n(i)} + d\big(\mathcal{X}_l^m(i), \bar{\mathcal{X}}_l^m(i)\big)_{\mathbf{R}^m(i)} = \\ &\quad \big(\mathbf{x}_k^n(i) - \bar{\mathbf{x}}_k^n(i)\big)^T \mathbf{R}^{n\,-1}(i) \big(\mathbf{x}_k^n(i) - \bar{\mathbf{x}}_k^n(i)\big) \\ &\quad + \big(\mathbf{x}_l^m(i) - \bar{\mathbf{x}}_l^m(i)\big)^T \mathbf{R}^{m\,-1}(i) \big(\mathbf{x}_l^m(i) - \bar{\mathbf{x}}_l^m(i)\big). \end{aligned} \tag{6}$$

Thus, to estimate the data likelihood, we compute the optimal estimates of the homography and the true trajectory (upto a homography) and use them to evaluate (5).

This formulation generalizes to *multiple* airborne cameras by considering k-partite hypergraphs instead of the bipartite graphs considered previously, shown in Fig. 2. Once again, we wish to find the set of associations C',

$$C' = \arg\max_{C \in \mathcal{C}} \sum_{c_{k,l\ldots m}^{p,q\ldots r} \in C} \log p\Big(c_{k,l\ldots m}^{p,q\ldots r} | \mathcal{X}_k^p, \mathcal{X}_l^q \ldots, \mathcal{X}_m^r\Big). \tag{7}$$

To evaluate the inner probability, for instance, $p(c_{1,1,\ldots 1}^{1,2,\ldots,N}|\mathcal{X}_1^1, \mathcal{X}_1^2, \ldots, \mathcal{X}_1^N)$, we proceed by computing the maximum likelihood estimate of a set of $N-1$ homographies and one "canonical" trajectory related to each view by the set of homographies (see

1. The covariance matrix is time indexed since it has been transformed by a homography while compensating for frame-to-frame motion. This also captures the inherent error in the estimated frame-to-frame homography that causes drift. Details on transforming the covariance matrix from the frame coordinate to the reference coordinate are available in [8] or [12]. This paper also discusses the use of first order analysis, implicitly justifying the use of our error model.

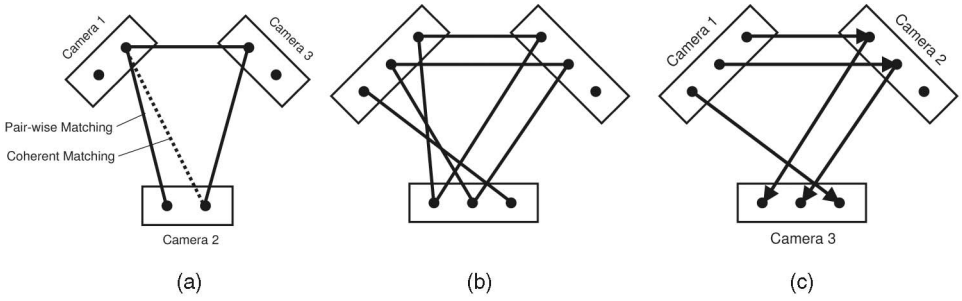

Fig. 2. Tracking across three moving cameras. (a) An impossible matching. Transitive closure in matching is an issue for matching in three of more cameras. (b) Missing observations. This matching shows the case of missing observations, with three objects in the scene, each visible in two cameras at a time. (c) The digraph associated with (b).

Fig. 3). Each homography relates the one camera coordinate frame with the canonical reference frame (less one because the homography from the canonical reference to itself is the identity matrix). The canonical trajectory is estimated that best describes all the observations in each camera simultaneously. Using these estimates of the $N-1$ homographies and the canonical trajectory, we have

$$p\left(c_{1,1,\ldots 1}^{1,2,\ldots,N}|\mathcal{X}_1^1,\mathcal{X}_1^2,\ldots,\mathcal{X}_1^N\right) \propto L\left(\{\mathcal{X}_1^1,\mathcal{X}_1^2,\ldots,\mathcal{X}_1^N\}|\left\{\mathbf{H}_{1,1}^{1,2},\ldots,\mathbf{H}_{1,1}^{N-1,N}\right\},\bar{\mathcal{X}}_1^1\right), \tag{8}$$

where the *pdf* of $L(\{\mathcal{X}_1^1,\mathcal{X}_1^2,\ldots,\mathcal{X}_1^N\}|\{\mathbf{H}_{1,1}^{1,2},\ldots,\mathbf{H}_{1,1}^{1,N}\},\bar{\mathcal{X}}_1^1)$ is[2]

$$L\left(\{\mathcal{X}_1^1,\mathcal{X}_1^2,\ldots,\mathcal{X}_1^N\}|\left\{\mathbf{H}_{1,1}^{1,2},\ldots,\mathbf{H}_{1,1}^{N-1,N}\right\},\bar{\mathcal{X}}_i\right)=\prod_i \frac{1}{\left(2\pi\|\mathbf{R}\|\right)^{\frac{N}{2}}}e^{-d_r/2}, \tag{9}$$

where

$$d_r = d\left(\mathcal{X}_1^1(i),\bar{\mathcal{X}}_1(i)\right)_{\mathbf{R}}+\sum_{j=2}^{N} d\left(\mathcal{X}_1^j(i),\mathbf{H}_{1,1}^{1,j}\bar{\mathcal{X}}_1(i)\right)_{\mathbf{R}}. \tag{10}$$

The Direct Linear Transform algorithm or RANSAC can be used as an initial estimate, followed by a Levenberg-Marquardt minimization over $9(N-1)+2\Delta t$ variables: $9(N-1)$ unknowns for the set of homographies and $2\Delta t$ unknowns for the canonical Δt 2D points. Equation (9) is used to compute the maximum likelihood estimates of the homography and the canonical trajectory and then used to evaluate the probability of the association hypothesis. Two important properties of the reprojection error for two cameras are also inherited by this multicamera reprojection error: 1) invariance to the choice of canonical reference (since the estimated trajectories are exactly related by the estimated intercamera homographies) and 2) invariance to rigid transformations. The maximum likelihood estimate of the canonical trajectory is also the maximum likelihood estimate of the true world trajectory up to a projective transformation.

However, it is known that the k-dimensional matching problem is NP-Hard for $k \geq 3$ ([26]). A possible approximation that is sometimes used is pairwise, bipartite matching; however, such an approximation is unacceptable in the current context since it is vital that transitive closure is maintained while tracking. The requirements of consistency in the tracking of objects across cameras is illustrated in Fig. 2. Instead, to address the complexity involved while accounting for consistent association we use the method proposed in [28]. A weighted digraph $D=(V,E)$ is constructed such that $\{V_1,V_2,\ldots V_k\}$ partitions V, where each partition corresponds to a moving camera. Direction is obtained by assigning an arbitrary order to the cameras (for instance, by enumerating them) and directed edges exist between every node in partition V_i and every node in partition V_j, where $i>j$ (due to the ordering). By forbidding the existence of edges against the ordering of the cameras, D is constructed as an acyclic digraph. This can be expressed as $E=\{v(\mathcal{X}_k^p)v(\mathcal{X}_l^q)|v(\mathcal{X}_k^p)\in V_p, v(\mathcal{X}_l^q)\in V_q\}$, where $e=v(\mathcal{X}_k^p)v(\mathcal{X}_l^q)$ represents an edge and $q>p$. The solution to the original association problem is then equivalent to finding the edges of maximum matching of the split G^* of the digraph D. It should be noted that, with this approach, we need only define pairwise edge-weights. Fig. 2 shows a possible solution and its corresponding digraph. It should be noted that, during the construction of the graph, we need to ensure that "left-over" objects are not assigned association. In order to avoid this, we prune all edges whose edge weights are below a certain likelihood. This is equivalent to ignoring measurements outside a "validation" region, as described in [2], ensuring that association hypotheses with low likelihoods are ignored.

Once a global association solution has been obtained, using this approximation we evaluate $p(C|\mathcal{X})$ as follows: We observe that all homographies mapping pairs of corresponding trajectories in sequences p and q are equal (up to a scale factor) and are, in turn, the same homography that maps the reference coordinate of sequence p to that of sequence q. Since all of the objects lie on the same plane, the homography relating the image of the trajectory of any object $\mathbf{H}_{k,l}^{p,q}$ in Sequence p to the image of the trajectory of that object in Sequence q is the same as the homography $\mathbf{H}_{i,j}^{p,q}$ relating any other object's trajectories in the two sequences (i.e., $i\neq p$ and $j\neq q$). Since these trajectories lie on the scene plane, these homographies are equal to $\mathbf{H}^{p,q}$, the homography that related the images of sequence p to the images of sequence q. This allows us to express $p(C|\{\mathcal{X}\})$ as

$$p(C|\{\mathcal{X}\}) = \prod_i \frac{1}{\left(2\pi\|\mathbf{R}\|\right)^{\frac{N}{2}}}e^{-d_r/2}, \tag{11}$$

where

$$d_r = \sum_k \left(d(\mathcal{X}_k^1(i),\bar{\mathcal{X}}_k^1(i))_{\mathbf{R}}+\sum_{j=2}^{N} d(\mathcal{X}_k^j(i),\mathbf{H}^{1,j}\bar{\mathcal{X}}_k(i))_{\mathbf{R}}\right). \tag{12}$$

This equation differs from (10) in the subscript of $\mathbf{H}$ since the homography and all canonical trajectories are simultaneously estimated. By using all trajectories between cameras, the spatial

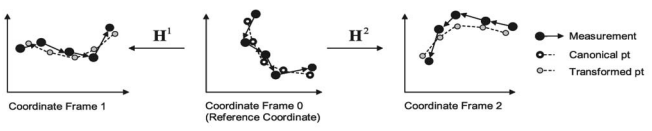

Fig. 3. A canonical trajectory and a set of homographies are estimated that minimize the multicamera reprojection error.

2. For notational convenience, we assume the covariance matrices are all equal.

Objective

Given object trajectories from all cameras for $\Delta t \geq 5$, estimate the inter-camera spatial transformations.

Algorithm

1) **Number cameras arbitrarily**
2) **For all pairwise** c**, compute** $p(c_{k,l}^{n,m}|\mathcal{X}_k^n, \mathcal{X}_l^m)$
3) **Compute Associations**: Find the maximum matching of the association graph as described in [28].
4) **Evaluate Global Likelihood Function**: Using the estimated maximum matching, compute the canonical trajectories and the maximum likelihood estimate of the inter-frame homographies.
5) **Repair Trajectories** For each unassociated trajectory, evaluate association likelihood with respect to all canonical trajectories, and re-associate broken trajectories.
6) **Concurrent Visualization**: Use the inter-frame homographies to construct a concurrent mosaic using all videos simultaneously.

Fig. 4. Algorithm for object association across moving cameras.

separation of different trajectories enforces a strong noncollinear constraint on association despite the near collinear motion of individual objects. In this way, even with relatively small durations of observation, the correct correspondence of objects can be discerned. Once again the optimal value of the set of homographies and the canonical trajectories are estimated using Levenberg-Marquardt minimization and the "goodness of fit" is measured. A degenerate case exists where all objects in the scene move in a straight line, but this case can be detected and estimation can be delayed until at least one object in the scene moves in a noncollinear motion. The algorithm is summarized in Fig. 4. The computational complexity of the proposed algorithm depends on the number of cameras, N, the average number of objects per camera, $\hat{K}$, and the average length of overlap time each object is observed in each pair of cameras, $\hat{t}$. The step with the highest computational cost is the construction of the association graph in Step 2. The number of pairwise associations is $\frac{N(N-1)\hat{K}^2}{2}$ and a naive implementation of the algorithm to evaluate each association requires a nonlinear minimization algorithm, for instance, the Levenberg-Marquardt algorithm. Its complexity is therefore $O(\frac{N(N-1)\hat{K}^5\hat{t}^3}{2})$; however, this can be improved to $O(\frac{N(N-1)\hat{K}^3\hat{t}}{2})$ by exploiting the sparse structure of the Jacobian during minimization. It should also be noted that, in general, both N and K are small numbers ($N = 3$ and $K = 6$ were the maximum values encountered during our experimentation).

2.2 Repairing Trajectories

During single camera tracking, object trajectories can sometimes be interrupted because of missing detections, noise, specularities, or feature similarity to the background. Trajectory interruption can also occur due to scene events like occlusion of the object by some other object, such as clouds, bridges, or tree cover, or due to the exiting and reentering of an object from the field of view. This causes the object's motion to be recorded by two different trajectories. Fig. 5 shows trajectories in two cameras, plotted in space and time. In the second camera, the second trajectory is interrupted as the object exited and reentered the scene. Several methods have been proposed to account for this problem at the single camera level using predictive methods. However, we show that the canonical tracks and the estimated intercamera homographies can be used to repair broken trajectories in a straightforward way. Since matching ensures a one-to-one correspondence, all such broken trajectories should be unassociated after matching. For each free trajectory $\mathcal{X}_i^n$, we evaluate with respect to each canonical trajectory $\bar{\mathcal{X}}_j$,

$$j^* = \arg\max_{j \in 1 \ldots N} p\left(\mathcal{X}_i^n | \bar{\mathcal{X}}_j; \mathbf{H}^n\right). \tag{13}$$

$p(\mathcal{X}_i^n|\bar{\mathcal{X}}_{j^*}, \mathbf{H}^n)$ is evaluated asymmetrically,

$$p(\mathcal{X}_i^n|\bar{\mathcal{X}}_{j^*}, \mathbf{H}^n) \propto \prod_k \frac{1}{\sqrt{2\pi}\|\mathbf{R}^n(k)\|^{\frac{1}{2}}} e^{-\frac{1}{2}\left(d(\mathcal{X}_i^n(k), \bar{\mathcal{X}}_i(k))_{\mathbf{R}^n(k)}\right)}. \tag{14}$$

If this is greater than an empirical threshold $\gamma(k)$ and if there is no temporal overlap between $\mathcal{X}_i^n$ and $\mathcal{X}_j^n$ (the trajectory in Camera n currently associated with $\mathcal{X}_j$), then $\mathcal{X}_i^n$ and $\mathcal{X}_j^n$ are reconnected and both associated to $\mathcal{X}_j$—the trajectory is repaired. It is noteworthy here that, unlike single camera methods, the duration of occlusion is irrelevant as long as the object is continuously viewed in any other camera.

3 Results

In this section, we report the quantitative performance of the algorithm on simulations. We also report the experimental performance of our trajectory association approach qualitatively for data from airborne cameras and in a controlled setting. In each experiment, we demonstrate that the proposed approach is able to accurately associate trajectories across multiple moving cameras despite short durations of observations, nearly linear motion, and noisy detections. It should be noted that we do not model errors in object association *within* each sequence, i.e., we expect that the tracks within a sequence are correct. If errors in single camera tracking do exist, it is unlikely that the track will generate an association in another sequence. However, it should also be noted that, although erroneous single camera tracking might generate false negatives (no association where one exists), it is exceedingly unlikely that this will produce false positives (an incorrect association).

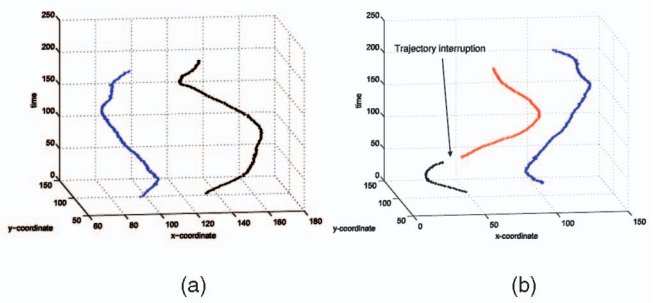

Fig. 5. Trajectory interruption. (a) Complete trajectories observed in Camera 1. (b) The second trajectory (black) is interrupted as the object exits and then reenters the field of view. The reentering trajectory is recorded as a new trajectory (red).

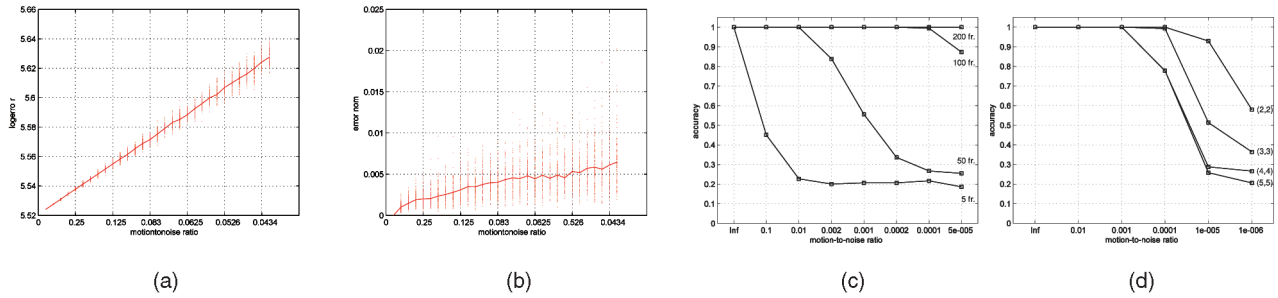

Fig. 6. Accuracy of the estimated parameters. (a) The log-likelihood of the canonical tracks as the motion-to-noise ratio was increased across three cameras observing three objects. (b) The error norm of the estimated to the true homography. One hundred iterations were run for each noise level, which are plotted (dots) along with the median value (line). Association accuracy with regard to number of cameras, number of objects, number of frames, and motion-to-noise ratio. The horizontal axis is not progressing linearly. (c) For 10 cameras with 10 objects, the percentage of correct associations to the total number of associations. (d) As the number of cameras and objects increase linearly, for a fixed 60 frames, the association accuracy decreases. The results are the average of 100 runs.

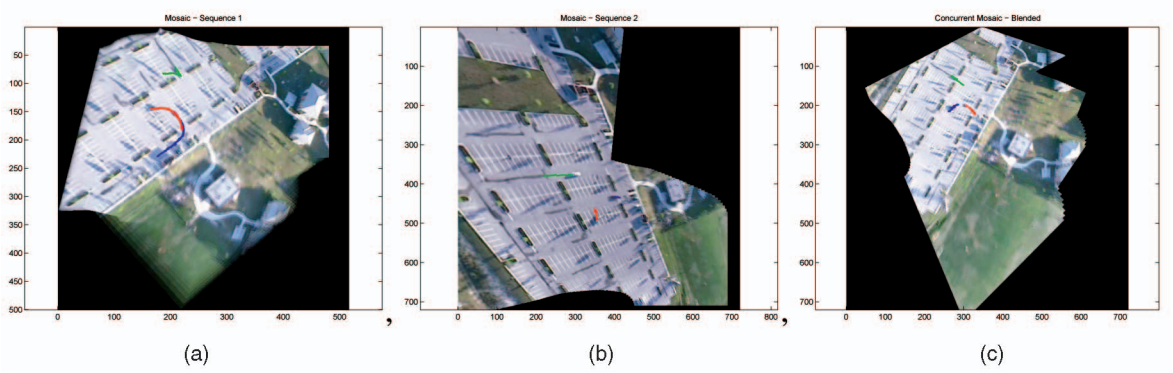

Fig. 7. Second UAV experiment—short temporal overlap. Despite a very short duration of overlap, correct correspondence was estimated. (a) Mosaic of Sequence 1. (b) Mosaic of Sequence 2. (c) Concurrent visualization of two sequences. The two mosaics were blended using a quadratic color transfer function. Information about the objects and their motion is compactly summarized in the concurrent mosaic.

In order to run simulations, a generator was designed to randomly synthesize data for quantitative experimentation. The camera parameters included the number of cameras and the number of frames of observation and the object parameters included the number of objects and the mean and variance of the object motion $(\hat{\rho}, \sigma_\rho)$. For each object, an initial position, $X(0)$ and $Y(0)$, was determined by sampling from a uniform distribution over a spatial support region, assuming the world plane $\mathbf{\Pi}$ was the plane $Z = 0$. To closely imitate the smooth motion of real-world objects, the object motion $(\rho, \Delta\theta)$ was sampled from the normal distributions $\mathcal{N}(\hat{\rho}, \sigma_\rho)$ and $\mathcal{N}(0, \sigma_\theta)$ and initial θ was a (single) sample from a uniform distribution over the interval $[-\pi, \pi]$. For each camera, a reference to frame homography $\mathtt{P}$ was randomly generated by sampling from a uniform distribution over the support of the camera extrinsic and intrinsic parameters and the imaged trajectories of each object in each camera are generated as $\mathcal{X}(t) = \mathtt{P}[X(t)\ \ Y(t)\ \ 1]^T + \epsilon$, where ϵ is the zero-mean measurement noise that is specified by a noise variance parameter σ_ϵ. The ratio ρ/σ_ϵ is referred to as the motion-to-noise ratio, measuring the expected strength of noise. In order to analyze the accuracy of the estimated intercamera homography as the ratio of mean motion to noise variance, we recorded the mean squared error of difference between the maximum likelihood estimate of the homography and the true homography over 100 runs. At each run, a new set of trajectories and homographies was generated. As expected, the estimation error decreased as the number of frames increase and the objects began to show more noncollinear motion, shown in Fig. 6. We then analyzed the quality of the estimate of the canonical tracks with respect to the ground truth by computing the average log-likelihood of the canonical frame given the ground truth. Here, too, the average of 100 runs was taken. We then analyzed the association accuracy with respect to larger increase in noise as the number of cameras and objects increased. Fig. 6c reports the association accuracy 10 trajectories viewed across of 10 cameras as the number of frames increase. The motion-to-noise ratio was varied from infinity (divide-by-zero) to 5×10^{-5}, while the number of frames was tested for 5, 50, 100, and 200 frames. Clearly, as the number of frames increased, the accuracy increased too. One hundred runs were executed (with randomly generated trajectories) per noise strength and the average accuracy was reported. The accuracy is shown in Fig. 6d as it varies with respect to the number of cameras/objects. As expected, as the number of cameras and objects decrease, the accuracy of the approach reduces too. The trajectory length was 60 frames (2 seconds at 30 fps). It can be noted that the motion-to-noise ratio in both experiments is *not* linearly increasing.

We conducted experiments on data collected by cameras mounted on unmanned aerial vehicles (UAVs). In these experiments, two UAVs mounted with cameras viewed real scenes with moving cars. The objects exited and entered the field of view and all three objects were only briefly visible together in the field of view. The individual trajectories of each sequence, on a single registered coordinate are shown in Figs. 7a and 7b. The result of correspondence is shown in Fig. 7c. An experiment involved association across IR and EO cameras was also conducted. Since only motion information is used in discerning association, the modality of the cameras does not affect the viability of the algorithm. In the first set, six objects were recorded by one EO and one IR camera. Although the relative positions of the cameras were fixed in this sequence, no additional constraints were used during experimentation. The vehicles in the field of view moved in a line and one after another performed a U-turn and the durations of observation of each object varied in both cameras. Since only motion information is used, the different modalities did not pose a problem to this algorithm. Fig. 8 shows all six trajectories color coded in their correspondence. Despite the fact that the sixth trajectory (color coded yellow in Fig. 8) was viewed

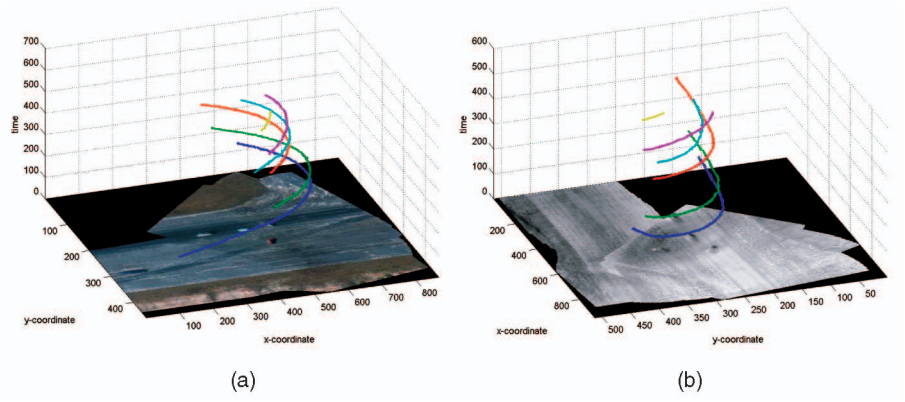

Fig. 8. First UAV Experiment—two cameras, six objects. (a) The EO video. (b) The IR video. Since we are using only motion information, association can be performed across different modalities.

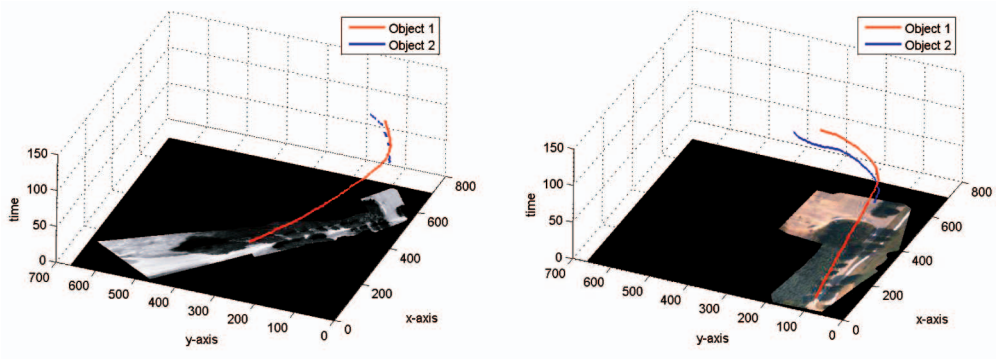

Fig. 9. Repairing broken trajectories. Due to rapid motion of the camera, the object corresponding to the blue trajectory exited and reentered the field of view of the IR camera several times. On the other hand, the same object in the EO camera remained continuously visible. The trajectories were successfully reassociated.

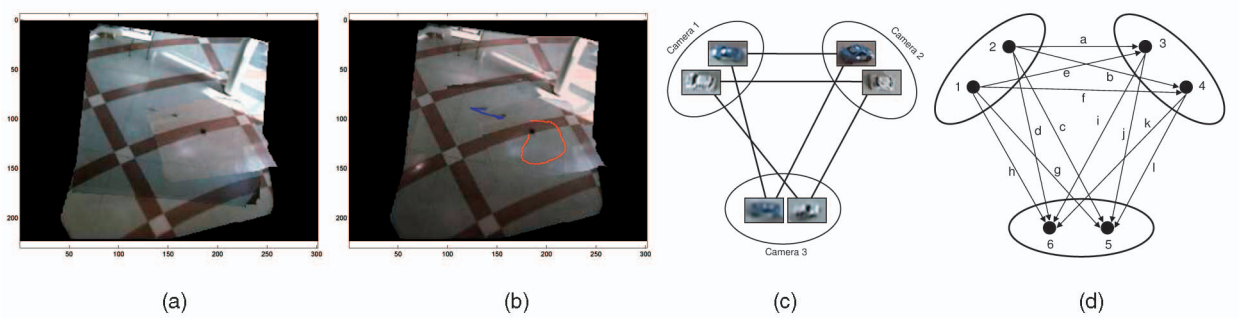

Fig. 10. Concurrent visualization of three sequences. (a) Concurrent mosaic before blending. (b) Blended concurrent mosaic with the track overlayed. Matching in three sequences. (c) Matching of the tripartite graph. (d) The corresponding directed graph.

only briefly in both sequences and underwent mainly collinear motion in this duration due to the matching correct global correspondence was obtained. In the second set, trajectory repairing was tested as two objects were observed by an EO and IR camera, as shown in Fig. 9. Both objects were continuously viewed in the EO camera, but Object 2 repeatedly exited and reentered the FOV of the IR camera, as shown by the fragmented trajectory. Using the trajectory repairing algorithm, the object was successfully reassociated. A final experiment was carried out using more than two cameras, where remote controlled cars were observed by moving camcorders (Sony DCR-TRV 740). Three moving cameras at various zooms observed a scene with two remote controlled cars. Fig. 10c shows the final, correct assignment of correspondence established by our approach. Fig. 10d shows the associated directed graph. The intersequence homographies were estimated and all three mosaics were registered together to create the concurrent mosaic, as shown in Fig. 10a. Fig. 10b shows the tracks of both objects, overlaid after blending each mosaic.

4 Conclusion and Discussion

In this paper, a method to associate objects across multiple airborne cameras was presented. We make two fundamental assumptions about the data: 1) that the altitude of the aerial vehicle upon which the camera is mounted is significantly high with respect to the ground, that a planar assumption is viable, and 2) that at least one object is seen simultaneously between every pair of cameras for at least five frames. Given these assumptions and taking as input the timestamped trajectories of objects observed in each camera, we estimate the intercamera transformations, the association of each object across the views, and "canonical" trajectories, which are the best estimate (in a maximum likelihood sense) of the original object trajectories up to a 2D projective transformation. To that end, we describe an extension to the reprojection error for multiple views, providing a geometrically and statistically sound means of evaluating the likelihood of a candidate correspondence set. We formulate the problem of maximizing this joint likelihood function as a k-dimensional

matching problem and use an approximation that maintains transitive closure. The estimated solution is verified using a strong global constraint for the complete set of associations across all cameras. Using simulations, we tested the sensitivity of the proposed approach to noise strength, the number of cameras, the number of frames viewed, and the "collinearity" of the trajectories. We demonstrated qualitative results on several real sequences, including the standard VIVID data set and the ARDA VACE data, for multiple cameras and between IR and EO video.

REFERENCES

[1] A. Azarbayejani and A. Pentland, "Real-Time Self-Calibrating Stereo Person Tracking Using 3D Shape Estimation from Blob Features," *Proc. Int'l Conf. Pattern Recognition,* 1996.

[2] *Multitarget-Multisensor Tracking: Advanced Applications,* Y. Bar-Shalom, ed. Artech House, 1990.

[3] Q. Cai and J.K. Aggarwal, "Tracking Human Motion in Structured Environments Using a Distributed Camera System," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 21, no. 11, pp. 1241-1247, Nov. 1999.

[4] T. Chang and S. Gong, "Tracking Multiple People with a Multi-Camera System," *Proc. IEEE Int'l Workshop Multi-Object Tracking,* 2001.

[5] O. Chum, T. Pajdla, and P. Sturm, "The Geometric Error for Homographies," *Computer Vision and Image Understanding,* vol. 97, no. 1, pp. 86-102, Jan. 2005.

[6] R. Collins, A. Lipton, H. Fujiyoshi, and T. Kanade, "Algorithms for Cooperative Multisensor Surveillance," *Proc. IEEE,* 2001.

[7] R. Collins, O. Amidi, and T. Kanade, "An Active Camera System for Acquiring Multi-View Video," *Proc. IEEE Int'l Conf. Image Processing,* 2002.

[8] A. Criminisi and A. Zisserman, "A Plane Measuring Device," *Proc. British Machine Vision Conf.,* 1997.

[9] T. Darrell, D. Demirdjian, N. Checka, and P. Felzenszwalb, "Plan-View Tracjectory Estimation with Dense Stereo Background Models," *Proc. IEEE Int'l Conf. Computer Vision,* 2001.

[10] S. Dockstader and A. Tekalp, "Multiple Camera Fusion for Multi-Object Tracking," *Proc. IEEE Int'l Workshop Multi-Object Tracking,* 2001.

[11] D. Makris, T. Ellis, and J. Black, "Bridging the Gaps between Cameras," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2004.

[12] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision.* Cambridge Univ. Press, 2000.

[13] J. Hopcroft and R. Karp, "A $n^{2.5}$ Algorithm for Maximum Matching in Bi-Partite Graph," *SIAM J. Computing,* 1973.

[14] T. Huang and S. Russell, "Object Identification in a Bayesian Context," *Proc. Int'l Joint Conf. Artificial Intelligence,* 1997.

[15] O. Javed, Z. Rasheed, K. Shafique, and M. Shah, "Tracking in Multiple Cameras with Disjoint Views," *Proc. IEEE Int'l Conf. Computer Vision,* 2003.

[16] O. Javed, K. Shafique, and M. Shah, "Appearance Modeling for Tracking in Multiple Non-Overlapping Cameras," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition,* 2005.

[17] J. Kang, I. Cohen, and G. Medioni, "Continuous Tracking Within and Across Camera Streams," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2003.

[18] V. Kettnaker and R. Zabih, "Bayesian Multi-Camera Surveillance," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 1999.

[19] S. Khan and M. Shah, "Consistent Labeling of Tracked Objects in Multiple Cameras with Overlapping Fields of View," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 25, no. 10, pp. 1355-1360, Oct. 2003.

[20] J. Krumm, S. Harris, B. Meyers, B. Brumitt, M. Hale, and S. Shafer, "Multi-Camera Multi-Person Tracking for Easy Living," *Proc. IEEE Workshop Visual Surveillance,* 2000.

[21] H. Kuhn, "The Hungarian Method for Solving the Assignment Problem," *Naval Reserach Logistics Quarterly,* 1955.

[22] L. Lee, R. Romano, and G. Stein, "Learning Patterns of Activity Using Real-Time Tracking," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 22, no. 8, pp. 747-757, Aug. 2000.

[23] T. Matsuyama and N. Ukita, "Real-Time Multitarget Tracking by a Cooperative Distributed Vision System," *Proc. IEEE,* 2002.

[24] A. Mittal and L. Davis, "M_2 Tracker: A Multi-View Approach to Segmenting and Tracking People in a Cluttered Scene," *Int'l J. Computer Vision,* 2003.

[25] A. Nakazawa, H. Kato, and S. Inokuchi, "Human Tracking Using Distributed Vision Systems," *Proc. Int'l Conf. Pattern Recognition,* 1998.

[26] C. Papadimitriou, *Computational Complexity.* Addison Wesley, 1994.

[27] A. Rahimi, B. Dunagan, and T. Darrell, "Simultaneous Calibration and Tracking with a Network of Non-Overlapping Sensors," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2004.

[28] K. Shafique and M. Shah, "A Noniterative Greedy Algorithm for Multiframe Point Correspondence," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 27, no. 1, pp. 51-65, Jan. 2005.

[29] Y. Shan, H. Sawhney, and R. Kumar, "Unsupervised Learning of Discriminative Edge Measures for Vehicle Matching between Non-Overlapping Cameras," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition,* 2005.

[30] Y. Shan, H. Sawhney, and R. Kumar, "Vehicle Identification between Non-Overlapping Cameras without Direct Feature Matching," *Proc. IEEE Int'l Conf. Computer Vision,* 2005.

[31] C. Stauffer and K. Tieu, "Automated Multi-Camera Planar Tracking Correspondence Modelling," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2003.

[32] P. Sturm, "Vision 3D Non Calibrée—Contributions à la Reconstruction Projective et Étude des Mouvements Critiques pour l'Auto-Calibrage," PhD thesis, 1997.