Object Tracking Across Multiple Independently Moving Airborne Cameras

Yaser Sheikh Mubarak Shah

Computer Vision Laboratory,
School of Computer Science,
University of Central Florida,
Orlando, FL 32826

Abstract

A camera mounted on an aerial vehicle provides an excellent means for monitoring 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 tracking objects across multiple moving airborne cameras. Since the cameras are moving and often widely separated, direct appearance-based or proximity-based constraints cannot be used. Instead, we exploit geometric constraints on the relationship between the motion of each object across cameras, to test multiple correspondence hypotheses, without assuming any prior calibration information. There are three novel contributions in this paper. First, we propose a statistically and geometrically meaningful means of evaluating a hypothesized correspondence between two observations in different cameras. Second, since multiple cameras exist, ensuring coherency in correspondence, i.e. transitive closure is maintained between more than two cameras, is an essential requirement. To ensure such coherency we pose the problem of object tracking across cameras as a k-dimensional matching and use an approximation to find the Maximum Likelihood assignment of correspondence. Third, we show that as a result of tracking objects across the cameras, a concurrent visualization of multiple aerial video streams is possible. Results are shown on a number of real and controlled scenarios with multiple objects observed by multiple cameras, validating our qualitative models.

1 Introduction

The concept of a cooperative multi-camera system, informally a 'forest' of sensors [15], 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 high-resolution camera with a wide field of view that monitors a wide 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. In recent literature, several approaches with varying constraints have been proposed, highlighting the wide applicability of the concept. For instance, the problem of tracking across multiple *stationary* cameras with overlapping fields of view has been addressed in a number of papers, e.g. [2], [18], [6], [1], [15] and [13]. Extending the problem to tracking in cameras with non-overlapping fields of view, geometric and appearance based approaches have also been proposed recently, e.g. [12], [4], [10], and [22]. A common assumption shared by these methods is that the camera remains stationary for the duration of sensing. By removing this assumption and allowing the sensors to move, a much wider area can be observed. A limited type of camera motion has been examined in previous work: motion of the camera about the camera center, i.e. pan-tilt-zoom (PTZ) motion. One such work is [16], where Matsuyama and Ukita presented an approach using active cameras, using a fixed point PTZ camera for wide area imaging. In [11] Kang *et al.* proposed a method that involved multiple stationary and PTZ cameras. In this work, it was assumed that the scene was planar and that the homographies *between* cameras were known. A related approach was also proposed in [5], where Collins *et al.* presented an active multiple camera system that maintained a single moving object centered in each view, using PTZ cameras. However, thus far, there has been no work on tracking objects across multiple independently moving cameras, whose *centers* move as well. This is particularly attractive since it allows far wider areas to be monitored by fewer cameras.

In this work, the problem we address is to track objects across multiple independently moving cameras mounted of airborne vehicles, without assuming any calibration of the cameras. To the authors' knowledge this is the first paper to tackle this problem. When using sensors in such a decentralized but cooperative fashion, knowledge of inter-camera relationships become 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 not. In the scenario under study in this paper, obtaining calibration information usually requires sophisticated equipment, such as a global positioning system (GPS) or an inertial navigation system (INS), perhaps with a geodetically aligned elevation map. Thus approaches that do not require prior calibration information and are based only on video data are particularly attractive as an alternative. Furthermore, when cameras are moving independently, the fields of view of different cameras can alternatively move in and out of overlap and as a result the problem of correspondence becomes considerably more complicated than that of the stationary camera case. It is useful to think of the problem in terms of *spatio-temporal overlap* of fields of view (FoV), analogous to spatial overlap in the case of stationary cameras, i.e. for some duration of time, the FoV of each camera overlaps (spatially) with the FoV of another camera while observing the moving objects. In terms of spatio-temporal overlap, we identify four possible cases of the

problem,

1. Each object is simultaneously visible by all cameras, all the time: In this instance, there is continuous spatial and temporal overlap between the fields of view. This rarely occurs in practice, especially over extended sequences, since aerial vehicles usually move continuously.

2. Each object is simultaneously visible by some cameras, all the time: This is the instance of limited spatial overlap, where all objects are within the 'collective' field of view of all the cameras all the time (but not necessarily within *each* camera's field of view). This situation occurs most often when each UAV is in pursuit of a separate target.

3. Each object is simultaneously visible by some cameras for a limited duration of time: This is the general case (within the context of this work), where all objects are visible in some subset of cameras simultaneously. This is the case most often encountered in extended runs.

4. Each object is visible by some cameras, but not necessarily simultaneously: In this case, *temporal* overlap does not necessarily occur between any two cameras, while objects are visible in their field of view. Without making some strong assumptions about object or camera motion it is difficult to address this case. This case is the spatio-temporal analog of the problem of tracking across stationary cameras with non-overlapping fields of view. This is the only case we do not address.

In this work, we require at least limited spatio-temporal overlap between the fields of view of the cameras (Case 3) to discern the relationship of observations in the (uncalibrated) moving cameras, i.e. the proposed approach addresses Cases 1, 2 and 3. For moving cameras, particularly airborne ones where large swaths of areas may be traversed in a short period of time, coherent visualization is indispensable for applications like surveillance and reconnaissance. In fact, the underlying concept of co-operative sensing is to give global context to 'locally' obtained information at each camera. Thus, we show that as a result of tracking objects across multiple moving cameras with spatio-temporal overlap of FoVs, the collective field of view of all the airborne cameras can be simultaneously visualized using a *concurrent* mosaic.

The notation we use in the paper is as follows: there are N cameras, observing a scene with K objects. An object k present in the field of view of camera n is denoted as O_k^n. The imaged location of O_k^n at time t is $\mathbf{x}_{k,t}^n = (x_{k,t}^n, y_{k,t}^n, \lambda_{k,t}^n)^\mathsf{T} \in \mathbb{P}^2$, the homogenous coordinates of the point[1] in sequence n. The trajectory of O_k^n is the set of points $\mathcal{X}_k^n = \{\mathbf{x}_{k,i}^n, \mathbf{x}_{k,i+1}^n, \ldots \mathbf{x}_{k,j}^n\}$, where Δt is the duration from frame i to frame j (for the remainder of the paper, we will drop this notation of time unless explicitly required). For two cameras, a 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. For more than two cameras, a correspondence $c_{i,j,k,\ldots l}^{m,n,o,\ldots p}$ is a hypothesis defined by the tuple $(O_i^m, O_j^n, O_k^o, \ldots O_l^p)$. 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 correspondences C such that $c_{i,j,k,\ldots l}^{m,n,o,\ldots p} \in C$ if and only if $O_i^m, O_j^n, O_k^o, \ldots O_l^p$ are images of the same object in the world. The terminology of graph theory allows us to more clearly represent these different relationships (Figure 1(a)). We abstract the problem of tracking objects across cameras as follows. Each observed trajectory is modeled as a node and the graph is partitioned into N partitions, one for each of the N cameras. A hypothesized correspondence, c, between two observed objects (nodes), is represented as an edge between the two nodes. In Figure 1(a), Object 1 is visible in all cameras, and the correspondence 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 partition corresponding to Camera 2.

The rest of the paper is organized as follows: In Section 2, we present a means to estimate the likelihood of a correspondence hypothesis. In Section 3, the Maximum Likelihood assignment is computed for multiple cameras in a graph-theoretic framework. Results in several controlled and real scenarios are shown in Section 4, with conclusions and a summary in Section 5.

2 The Likelihood of a Correspondence Hypothesis

In this section, we describe how the likelihood that two trajectories, observed by two different cameras, originated from the same world object is estimated - the use of this, in turn, for multiple objects assignment across multiple cameras is described in Section 4. It should be mentioned at the outset that lower level processing, such as object detection and tracking *within* each sequence is outside the scope of this work. Instead, we assume that tracks have already been obtained and frame-to-frame estimation of homography is available for each camera[2]. Our focus is to investigate methods on how to best use this data for correspondence *across* cameras. Thus, at a certain instant of time, we have K_i trajectories for the i-th camera corresponding to the objects visible in that camera. Since the frame-to-frame homography is available, the position of each point in trajectory $\mathcal{X}_k^n$ is transformed to a reference coordinate (e.g. the first frame) of the sequence. The measured image positions of objects, $\mathcal{X}_k^n = \{\mathbf{x}_{k,i}^n, \mathbf{x}_{k,i+1}^n, \ldots \mathbf{x}_{k,j}^n\}$ are described in terms of the true image positions, $\bar{\mathcal{X}}_k^n = \{\bar{\mathbf{x}}_{k,i}^n, \bar{\mathbf{x}}_{k,i+1}^n, \ldots \bar{\mathbf{x}}_{k,j}^n\}$, with independent normally distributed measurement noise, $\mu = 0$ and variance σ^2, that is

$$\mathbf{x}_{k,i}^n = \bar{\mathbf{x}}_{k,i}^n + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma). \tag{1}$$

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 ([7]). This assumption of planarity dictates that a homography $\mathbf{H}_{k,l}^{n,m}$ must exist between any two trajectories that correspond, i.e. for any correspondence 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 $p(c_{k,l}^{n,m}|\mathcal{X}_k^n)$. By assuming conditional independence between each correspondence c, and the probability of a candidate solution C given the trajectories is,

$$p(C|\{\mathcal{X}\}) = \prod_{c_i \in C} p(c_i|\mathcal{X}_i) \tag{2}$$

[1] The abstraction of each object is as a point corresponding to the object centroid.

[2] Several methods have been proposed to achieve this for airborne cameras, and for representative literature on this problem the reader is directed to [9], [19] and [3].

We are interested in the maximum likelihood solution,

$$C^* = \arg\max_{C \in \mathcal{C}} p(C|\{\mathcal{X}\}), \tag{3}$$

where $\mathcal{C}$ is the space of solutions. We now describe how to compute the likelihood that two trajectories observed in two cameras orignated from the same real world object. Using these pair-wise estimates, we describe how to the maximization of Equation 3 in Section 3.

2.1 Correspondence Estimators

Two cost functions that can be used to evaluate a correspondence hypothesis are the algebraic distance and the geometric distance. To compute the algebraic distance $d_{alg}(\mathcal{X}_k^n, \mathcal{X}_l^m)$ associated with a correspondence $c_{k,l}^{n,m}$, the Direct Linear Transform (DLT) algorithm can be used [7], minimizing the norm $||\mathbf{Ah}||$, where

$$\mathbf{A} = \begin{bmatrix} \mathbf{0}^\top & -\lambda_{l,1}^q \mathbf{x}_{k,1}^{p\top} & y_{l,1}^q \mathbf{x}_{k,1}^{p\top} \\ \lambda_{l,1}^q \mathbf{x}_{k,1}^{p\top} & \mathbf{0}^\top & -x_{l,1}^q \mathbf{x}_{k,1}^{p\top} \\ -y_{l,1}^q \mathbf{x}_{k,1}^{p\top} & x_{l,1}^q \mathbf{x}_{k,1}^{p\top} & \mathbf{0}^\top \\ & \vdots & \\ \mathbf{0}^\top & -\lambda_{l,n}^q \mathbf{x}_{k,n}^{p\top} & y_{l,n}^q \mathbf{x}_{k,n}^{p\top} \\ \lambda_{l,n}^q \mathbf{x}_{k,n}^{p\top} & \mathbf{0}^\top & -x_{l,n}^q \mathbf{x}_{k,n}^{p\top} \\ -y_{l,n}^q \mathbf{x}_{k,n}^{p\top} & x_{l,n}^q \mathbf{x}_{k,n}^{p\top} & \mathbf{0}^\top \end{bmatrix}$$

and $\mathbf{h}$ is the inter-camera homography in row major form. The similarity between two trajectories can be measured by observing the condition number of $\mathbf{A}$, i.e. the ratio of the largest singular value to the smallest singular value of $\mathbf{A}$. Thus, the algebraic distance between two trajectories can be computed using,

$$d_{alg}(\mathcal{X}_k^p, \mathcal{X}_l^q) = \frac{\sigma_1^{\mathbf{A}}}{\sigma_9^{\mathbf{A}}}, \tag{4}$$

where $\sigma_n^{\mathbf{A}}$ is the n-th largest singular value of $\mathbf{A}$. The advantage of the algebraic cost function is that it provides a non-iterative solution and can evaluate the similarity between two trajectories without the need to explicitly estimate the inter camera homography. However, the cost function does not have any geometrical interpretation and it does not allow us to incorporate the measurement error model explicitly. Instead, as will be seen presently, this algebraic approach is useful as a stable initialization for a more geometrically and statistically meaningful formulation.

Since measurement errors are expected in both trajectories, we use the *re-projection* distance as a cost function. This re-projection distance is an alternative cost function that explicitly minimizes the *transfer* error between the trajectories. This approach evaluates the probability of a correspondence hypothesis by estimating a homography, $\mathbf{H}_{k,p}^{l,q}$ and two new trajectories $\bar{\mathcal{X}}_k^p$ and $\bar{\mathcal{X}}_l^q$, related *exactly* by $\mathbf{H}_{k,p}^{l,q}$ that minimize the geometric distance,

$$d_r(\mathcal{X}_k^p, \mathcal{X}_l^q) = \sum_t d(\mathbf{x}_{k,t}^p, \bar{\mathbf{x}}_{k,t}^p) + \sum_t d(\mathbf{x}_{l,t}^q, \bar{\mathbf{x}}_{l,t}^q), \tag{5}$$

where $d(\cdot)$ is a distance metric, like the Euclidean distance. In particular, it is an attractive choice since it can be shown that the re-projection error is related to the Maximum Likelihood estimate of both the homography and the object trajectories, [7]. Since the errors at each point are assumed independent, the conditional probability of the correspondence given the trajectories in the pair of sequences can be estimated,

$$p(\mathcal{X}_k^p, \mathcal{X}_l^q |\, c_{l,q}^{k,p}; \mathbf{H}, \bar{\mathcal{X}}_k^p) = \prod_i \frac{1}{2\pi\sigma^2} e^{-d_r(\mathcal{X}_k^p, \mathcal{X}_l^q)/(2\sigma^2)}. \tag{6}$$

Thus, to estimate the data likelihood, we compute the optimal estimates of the homography and exact trajectories (that minimize Equation 5) and use them to evaluate Equation 7. However, when attempting to obtain globally optimal assignment of multiple objects the lengths of trajectories may vary considerably from object to object, since the effective length of trajectories depends on the duration of temporal overlap between the FoVs of the two cameras. Therefore, to obtain a normalized estimate with respect to the duration of overlap between each trajectory,

$$p(\, c_{l,q}^{k,p} | \mathcal{X}_k^p, \mathcal{X}_l^q) = \prod_i \left(\frac{1}{2\pi\sigma^2} e^{-d_r(\mathcal{X}_k^p, \mathcal{X}_l^q)/(2\sigma^2)} \right)^{1/\Delta t}, \tag{7}$$

where Δt is the duration of overlap between the two trajectories. Finally, taking the log we have,

$$\log p(\, c_{l,q}^{k,p} | \mathcal{X}_k^p, \mathcal{X}_l^q) \propto -\frac{1}{\Delta t} \sum_i \log d_r(\mathcal{X}_k^p, \mathcal{X}_l^q). \tag{8}$$

3 Maximum Likelihood Assignment of Global Correspondence

In the previous section, we developed a model to evaluate the probability of correspondence between two trajectories. Generally, however, when several objects are observed simultaneously by multiple cameras we require an optimal *global* assignment of object correspondences. We show that within the proposed formulation, this global optimality can be described in terms of a Maximum Likelihood estimate. As mentioned earlier, the problem of establishing correspondence between trajectories can be posed within a graph theoretic framework. Consider first, the straightforward case of several objects observed by *two* moving cameras. This can be modeled by constructing a complete bi-partite graph $G = (U, V, E)$ in which the vertices $U = \{u(\mathcal{X}_1^p), u(\mathcal{X}_2^p) \ldots u(\mathcal{X}_k^p)\}$ represent the trajectories in Sequence p, and $V = \{v(\mathcal{X}_1^q), v(\mathcal{X}_2^q) \ldots v(\mathcal{X}_k^q)\}$ represent the trajectories in Sequence q, and E represents the set of edges between any pair of trajectories from U and V. The bi-partite graph is complete because any two trajectories may match hypothetically. The weight of each edge is the probability of correspondence of Trajectory $\mathcal{X}_l^q$ and Trajectory $\mathcal{X}_k^p$, as defined in Equation 7. By finding the maximum matching of G, we find a unique set of correspondence C', according to the maximum likelihood estimate,

$$C' = \arg\max_{C \in \mathcal{C}} \sum_{c_{l,q}^{k,p} \in C} \log p(c_{l,q}^{k,p} |\mathcal{X}_k^p). \tag{9}$$

where $\mathcal{C}$ is the solution space. Several algorithms exist for the efficient maximum matching of a bi-partite graph, for instance [14] or [8] which are $O(n^3)$ and $O(n^{2.5})$ respectively.

This formulation generalizes to *multiple* moving cameras by considering complete k-partite graphs instead of the bi-partite

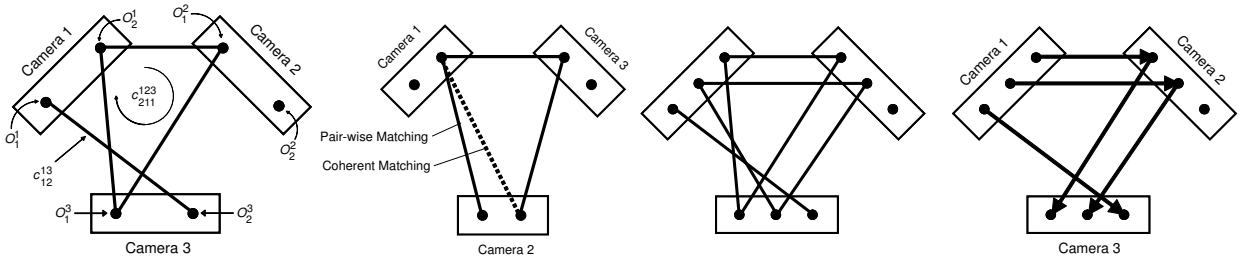


Figure 1: Graphical representation. (a) Graphical Notation. Each partition corresponds to one camera input and each node corresponds to an observed object in that sequence. An edge is a correspondence between objects seen in two cameras. (b) An impossible matching. Transitive closure in matching is an issue for matching in three of more cameras. The dotted line shows the desirable edge whereas the solid line shows a possible solution from pairwise matching. (c) A possible solution in three cameras. (d) The digraph associated with correspondence in (c).

Objective

Given object trajectories from all cameras for $\Delta t > 4$, estimate globally optimal correspondence of object across cameras.

Algorithm

While $p(C|\ \{\mathcal{X}_m^p\ \mathcal{X}_n^p \dots \mathcal{X}_o^p\}, \{\mathcal{X}_i^q \mathcal{X}_j^q \dots \mathcal{X}_k^q\}) > \gamma\sigma$ do,

1. **Number cameras arbitrarily**
2. **For all pairwise** c**, compute** $p(c_{k,l}^{n,m}|\mathcal{X}_k^n, \mathcal{X}_l^m)$
   - **Normalization of** $\mathcal{X}_k^n$: Compute a similarity transform, T_k^n, transforming the mean of the points to the origin and making the average distance of the points from the origin equal to $\sqrt{2}$. This should be done separately for each trajectory.
   - **Initialization**: Use the DLT algorithm to compute an initial estimate of $\tilde{\mathbf{H}}_{k,l}^{n,m}$ for each correspondence hypothesis. Denormalize the computed homography, $\mathbf{H}_{k,l}^{n,m} = \mathsf{T}_k^{n-1}\tilde{\mathbf{H}}_{k,l}^{n,m}\mathsf{T}_k^n$.
   - **Minimize Re-projection Error**: Minimize the re-projection error of Equation 5 using the Levenberg-Marquardt non-linear minimization algorithm. For large number of points sparse minimization methods are recommended (see [20]).[a]
3. **Construct Split Graph** G^*: Find the maximum matching of the split of the acyclic directed graph described in Section 3.
4. **Evaluate confidence in solution**: Using the estimated maximum matching, compute the confidence in solution according to Equation 12.

Figure 2: Algorithm for object tracking across moving cameras

[a] This step is optional. Estimates provided by the DLT algorithm usually suffice.

graphs considered previously, shown in Figure 1. Each hyperedge represents the hypothetical correspondence $c_{i,j,k,\dots l}^{m,n,o,\dots p}$ between $(O_i^m, O_j^n, O_k^o, \dots O_l^p)$. However, it is known that the k-dimensional matching problem is NP-Hard for $k \geq 3$. A possible approximation that is often used is pairwise, bi-partite 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 Figure 1(b). Instead, to address the computational complexity involved while accounting for consistent tracking, we construct a weighted digraph $D = (V, E)$ such that $\{V_1, V_2, \dots V_k\}$ 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). 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 correspondence problem is then equivalent to finding the edges of maximum matching of the split G^* of the digraph D (for a proof see [21]). By forbidding the existence of edges against the ordering of the cameras, D is constructed as an acyclic digraph, and therefore this approach can be used to obtain correspondence can be obtained efficiently. Figure 1(c) shows a possible solution and it's corresponding digraph, Figure 1(d).

Finally, we need to ensure that 'left-over' objects are not assigned correspondence. For instance, consider the case when all but one object in each of two cameras have been assigned correspondence. Now, although the 'left-over' objects in each camera correspond to the two different objects in the real world (each that did not appear in *one* of the camera FOVs), they would be assigned correspondence. In order to avoid this we introduce a new partition in the graph corresponding to a *null camera*, with a node correpsonding to each object in each camera. The null camera is

assigned to be the last in the camera ordering so a directed edge exists between every node in the rest of the graph and a node in the null partition. The weight of each such edge discounts 'left-over' correspondences if the data likelihood is too low, computed as

$$p(c_{k,0}^{p,0}|\mathcal{X}_k^p) = -\frac{1}{\Delta t}\sum_i \log d_r(\mathcal{X}_k^p, \mathcal{X}_{k,0}^{p,0}) \tag{10}$$

where

$$\mathcal{X}_{k,0}^{p,0} = \mathcal{X}_k^p + \gamma \cdot \sigma \tag{11}$$

and σ is the standard deviation of the noise model of 1 and γ is an empirical constant (set to 2 for all experiments reported in this paper). In a meaningful way this ensures that correspondence hypotheses with low likelihoods are ignored.

3.1 Evaluating the Matching

For the correspondence of objects to be meaningful, the object must be observed in both cameras simultaneously for some short duration. The minimum number of observations required to discern correspondence for two objects is five observations, i.e. both objects are observed in the field of view for four (not necessarily consecutive) frames, since four correspondences are the minimum required to estimate a homography. Of course, since the motion of cameras is smooth, the duration of overlap is usually significantly greater than four and this allows numerically stable computation of correspondence. However, in real world scenarios, objects tend to move in straight lines, displaying more variant (non-collinear) motion only over larger durations of observation. This can often cause degenerate estimates of the homographies during short durations of observation. Thus, if matching is only performed between objects, as described earlier, the approach becomes dependent on the degree of non-collinearity of the object motion. What needs to be specified is a termination criteria: at what point in time can a set of correspondence C' (see Equation 9) be made confidently?

The base case in the online tracking assumes that some simultaneous observations have been observed (> 4). To evaluate confidence in the solution we have,

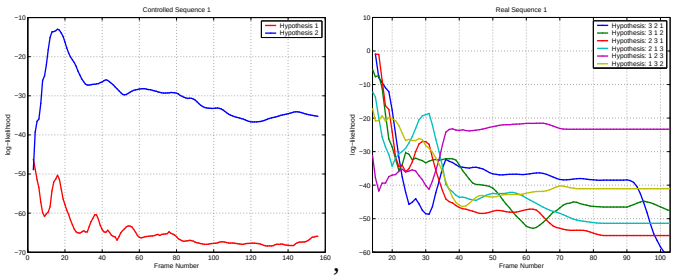
Proposition 1 All homographies mapping pairs of corresponding tracks in Sequences p and q are equal (up to a scale factor), and are, in turn, the same homography that maps the reference coordinate of Sequence p to that of Sequence q.

Since all the objects lie on the same plane, the homography relating the image of the (global-motion compensated) 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 homography are equal to $\mathbf{H}^{p,q}$, the homography that related the images of Sequence p to the images of Sequence q.

Proposition 1 provides us with a termination criteria. By using all the points simultaneously to evaluate the confidence in the solution at each time instance, the spatial separation of different trajectories enforces a strong non-colinear constraint on correspondence. In this way, even with relatively small durations of observation the correct correspondence of objects can be discerned. Since the cameras are continuously moving, matching with small durations of observation is one basic challenge of this work, and the ability to handle this case is an important advantage of our approach. Since a characteristic of the correct correspondence is that all homographies between each pair of corresponded trajectories is equal, a final assignment of correspondence between two sequences is made based on the probability of *all* objects matching simultaneously,

$$p(C|\ \{\mathcal{X}_m^p\ \mathcal{X}_n^p \ldots \mathcal{X}_o^p\}, \{\mathcal{X}_i^q \mathcal{X}_j^q \ldots \mathcal{X}_k^q\}) \tag{12}$$

where C is the set of correspondence of all objects between two cameras that is found by the matching algorithm. Clearly, Equation 12 can be computed in the same way as Equation 7. Thus, if $p(C|\ \{\mathcal{X}_m^p\ \mathcal{X}_n^p \ldots \mathcal{X}_o^p\}, \{\mathcal{X}_i^q \mathcal{X}_j^q \ldots \mathcal{X}_k^q\}) > \gamma\sigma$ for each pair of cameras, then we commit to the current solution. The final algorithm is summarized in Figure 2.



(a) (b)

Figure 4: Variation of some global correspondence hypotheses. (a) Variation for Controlled Experiment 1. (b) Variation for UAV Experiment 2. Due to colinear motion of the object, ambiguity in correspondence exists initially which is quickly resolved as the object begin to show more non-colinear behavior.

3.2 Concurrent Visualization

The purpose of aerial surveillance is to obtain an understanding of what occurs in an area of interest. While it is well known that video mosaics can be used to compactly represent a single aerial video sequence, they cannot compactly represent several such sequences *simultaneously*. If, on the other hand, the homographies between each of the mosaics (corresponding to each aerial sequence) are known, a *concurrent* mosaic can be created of all the sequences simultaneously. Since each sequence is aligned to a single coordinate frame during the construction of individual mosaics, Proposition 1 provides us with the means to register mosaics from multiple sequences onto one concurrent mosaic. To this end, the known point-wise correspondences from object tracking can be used to compute the 'inter-sequence' homography (e.g. the eigenvector associated with the smallest eigenvalue of the matrix $\mathbf{A}$). Final alignment is then refined using direct (gradient based) registration.

4 Results

In this section, we report the experimental performance of the proposed method in two controlled scenarios, and two real scenarios. In each experiment, we demonstrate the efficacy of the approach to accurately track moving objects across multiple moving cameras.

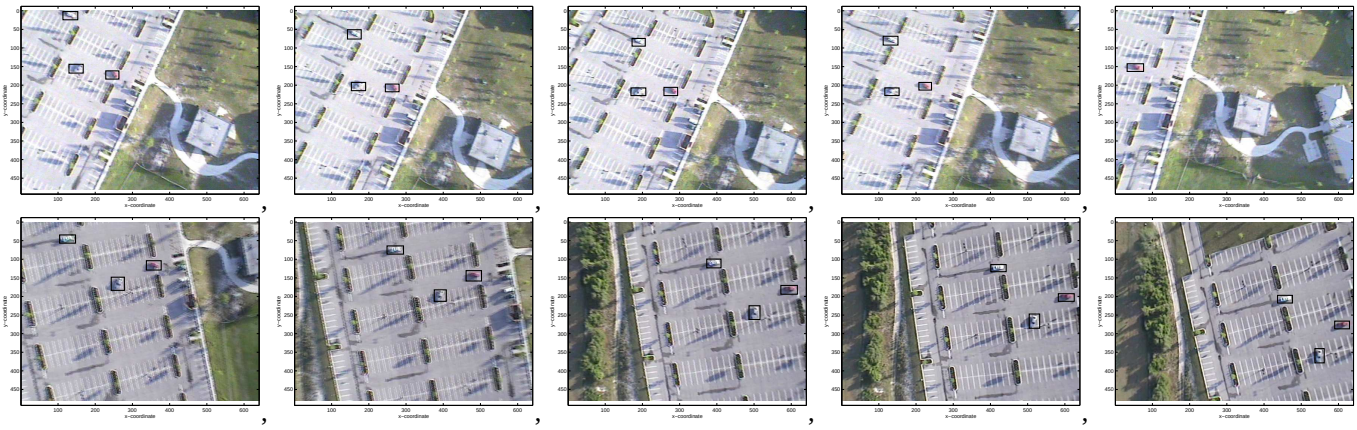


Figure 3: Corresponding frames from two sequences. Both rows show frames recorded from different cameras.

We have tested under a diverse set of situations: with multiple cameras, multiple objects, cameras of different modalities and at different zooms. It is recommended that the results be viewed in color. Additional results and videos associated with these results have been included in the supplementary folder.

4.1 Controlled Sequences

Two controlled experiments were carried out, where remote controlled cars were observed by moving camcorders. In the first experiment, two moving cameras were used, along with two remote controlled cars. The cars were operated on a (planar) floor with the two moving cameras viewing their motion from the height of about 12 feet. Figure 5 shows the trajectories of the car on the registered coordinate frame of Sequence 1. With two views and two objects there are five possible hypothesis,

1. $\{(O_1^1, O_1^2), (O_2^1, O_2^2)\}$
2. $\{(O_2^1, O_1^2), (O_1^1, O_2^2)\}$
3. $\{(O_1^1, O_1^2), (O_2^1), (O_2^2)\}$
4. $\{(O_2^1, O_1^2), (O_1^1), (O_2^2)\}$
5. $\{(O_2^1), (O_1^2), (O_1^1), (O_2^2)\}$

According to the first two hypothesis there are at least 2 objects present in the world, according to the third and the fourth hypothesis there are at least three objects in the world, and according to the fifth hypothesis there are at least four objects in the world. The variation of the first two hypotheses with respect to time is shown in Figure 3.2(a). Clearly, the first hypothesis is the correct one. Videos and online results are available in the supplementary material.

The second controlled experiment was carried out to test the performance of the system for more than two cameras. Three moving cameras at various zooms observed a scene with two remote controlled cars. Using successful object tracking results across the moving cameras, the inter-sequence homographies were estimated and all three mosaics were registered together to create the concurrent mosaic, as shown in Figure 6(a). Figure 6(b) shows the correspondence of the three sequence trajectories. The final correspondence of objects can be seen in Figure 6(c). The supplementary video contains the individual videos and some extended images describing the correspondence graph.

4.2 UAV Sequences

In these experiments, two unmanned aerial vehicles (UAVs) mounted with cameras viewed real scenes with moving cars, typically with a smaller duration of overlap than the controlled sequence. In the first experiment, six objects were recorded by one EO and one IR camera[3]. 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 the proposed approach. Figure 7 shows all six trajectories color coded in their correspondence. Final correspondence likelihoods are shown in Table 4.2. Despite the fact that the sixth trajectory (color coded yellow in Figure 7) was viewed only briefly in both sequences and underwent mainly colinear motion in this duration, due to the global spatial constraint of Equation 12 and the maximum matching, correct global correspondence was obtained.

In the next experiment, sequences with very short temporal overlap was used. Since the motion of aerial vehicles is far less controlled than that of controlled sequences, the duration of time in which a certain object is seen in both cameras is smaller. We show that despite the challenge of smaller overlap, object can be successfully tracked across the moving cameras. The variation of the 'goodness' of each hypothesis is shown in Figure 3.2(b). Since the motion of the objects were generally colinear in the beginning of the experiment, the probability of each correspondence fluctuates, but the correct correspondence, (Hypothesis: 1 2 3), is clearly higher as the process reaches an equilibrium. Using this correspondence, the concurrent mosaic of the scene was generated, shown in Figure 8. Examples of frames from the two sequences can be seen in Figure 3.

5 Conclusion and Summary

In this paper, we propose a method to correspond objects across uncalibrated cameras that are mounted on aerial vehicles. By defining an appropriate error model for our measurements, a geometrically and statistically meaningful approach is presented to estimate the likelihood of a correspondence hypothesis. Next,

[3]The relative positions of the cameras were fixed in this sequence but no additional constraints were used during experimentation.

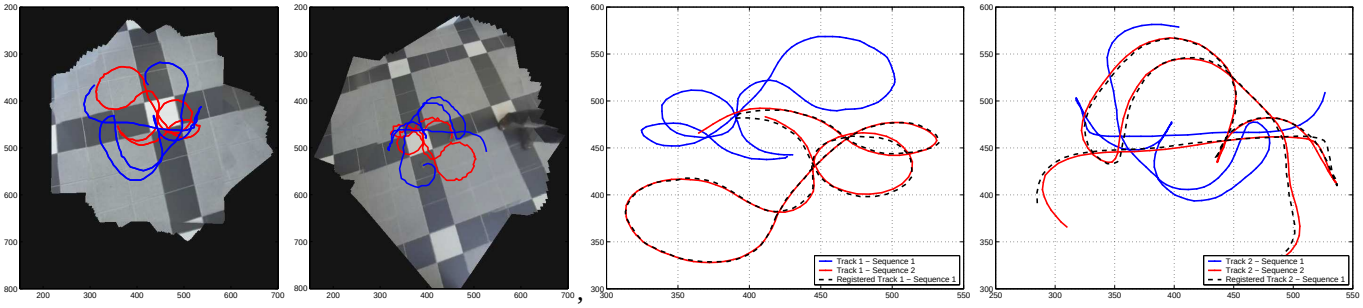


Figure 5: First Controlled Sequence - Two cameras and two objects. The trajectories of each object in Sequence 1 (red) and Sequence 2 (blue) are shown, along with the trajectory of Sequence 2 registered to Sequence 1 (dashed black) using the mosaic-mosaic homography.

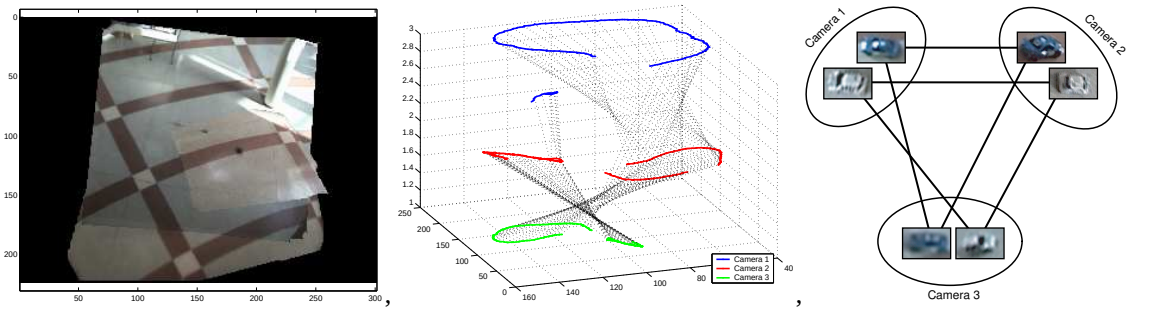


Figure 6: Second Controlled Experiment - Three cameras and two objects. (a) Concurrent visualization of three sequences. Information from all three zooms are simultaneously visible in this view.(b) Correspondence of points in trajectories viewed in each cameras. (c) Correspondence graph of objects in the three views.

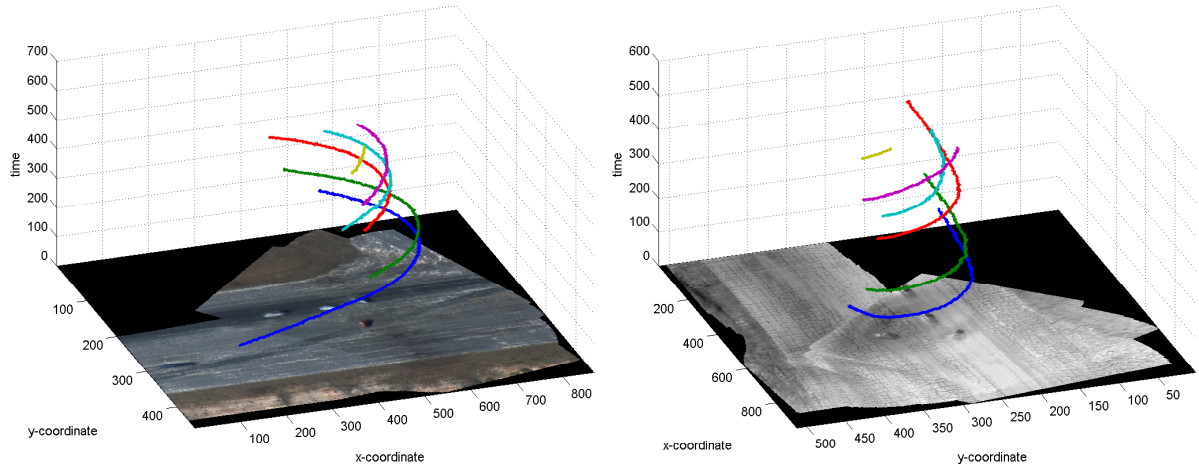


Figure 7: First UAV Experiment - two cameras, six objects. Concurrent visualization of two sequences. The two mosaics were blended using a quadratic color transfer function. Information about the objects and their motion is compactly summarized in the concurrent mosaic.

after computing likelihoods for all pair-wise hypotheses we find the Maximum Likelihood assignment of correspondences. This is done by posing the problem in graph-theoretic terms and using an approximation to k-dimensional matching. A major advantage of such an approach is that the matching is coherent, i.e. transitive closure is maintained in assignment. We define a termination criteria based on the 'goodness' of the solution to avoid committing to degenerate configurations, and finally show that as a result the multiple video streams can be concurrently visualized. There are three important assumptions of the proposed approach. First, we assume that the height of the aerial vehicle allows the scene to be modelled by a plane. It is noted here that for oblique-view aerial vehicles or for aerial vehicles monitoring terrain with significant relief, this assumption may be acceptable. Second, we assume that limited spatio-temporal overlap occurs between the fields of view of each pair of cameras. An interesting future direction would be to investigate the case where such overlap does not necessarily occur, i.e. Case 4 of the introduction. Third, we as-

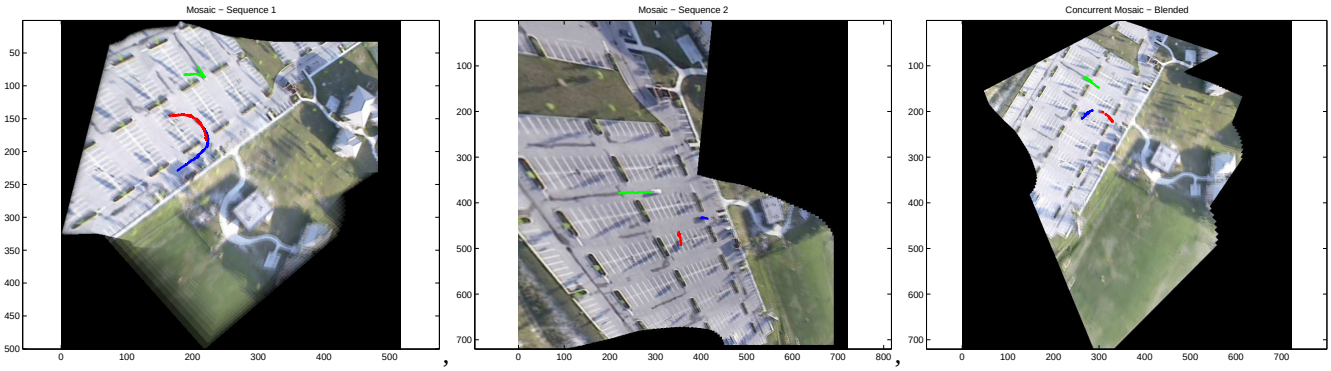


Figure 8: Second UAV experiment - Short temporal overlap. Despite a very short duration of overlap, correct correspondence was estimated. (a) Mosaic of Sequence 1 (b) Mosaic of Sequence 2 (c) Concurrent visualization of two sequences. The two mosaics were blended using a quadratic color transfer function. Information about the objects and their motion is compactly summarized in the concurrent mosaic.

	O_1^2	O_2^2	O_3^2	O_4^2	O_5^2	O_6^2
O_1^1	**-4.848**	-15.106	-32.925	-40.430	-69.078	-69.078
O_2^1	-10.434	**-4.484**	-26.017	-57.386	-21.647	-69.078
O_3^1	-35.285	-15.242	**-4.726**	-14.998	-14.634	-63.502
O_4^1	-69.078	-38.826	-15.954	**-4.451**	-3.662	-38.261
O_5^1	-69.078	-19.473	-18.328	-7.690	**-3.879**	-13.840
O_6^1	-69.078	-69.078	-51.216	-42.026	-51.688	**-18.042**

Table 1: Object correspondence log-likelihoods for the first UAV experiment. The values of correct correspondences are shown in bold. Despite some ambiguities (such as correspondence $c_{6,6}^{1,2}$ and $c_{4,6}^{1,2}$) the maximum matching resolves these ambiguities. Clearly, a greedy algorithm would have failed.

sume that the object display sufficiently non-colinear motion. This is not a strong constraint, since as we have demonstrated conclusively through Figure 3.2(b) the degree of non-colinearity does not have to be very large. One of the major strengths of the proposed approach is that we demonstrate that calibration information is not required to discern correspondence of object across the cameras. To our knowledge, this problem has not been tackled before, with calibrated or uncalibrated cameras.

Finally, using multiple aerial vehicles for observing wide areas is an idea of significant applicability. While several algorithms have been proposed for rearranging the positions of the aerial vehicles based on some sensors like GPS or INS for optimal coverage, object correspondence across multiple aerial vehicles presents an interesting option once the 'control loop' is closed, namely that of rearranging multiple sensors using image information and object correspondence. Instead of a cost function of maximum coverage, or maximum overlap between aerial vehicles, more intelligent cost functions based on object positions, proximity or object importance can be autonomously used.

References

[1] A. Azarbayejani and A. Pentland, *Real-Time Self-Calibrating Stereo Person Tracking Using 3D Shape Estimation from Blob Features*, ICPR, 1996.

[2] Q. Cai and J. K. Aggarwal, *Tracking Human Motion in Structured Environments using a Distributed Camera System*, TPAMI, 1999.

[3] I. Cohen and G. Medioni, *Detecting and Tracking Objects in Video Surveillance*, IEEE CVPR. 1999.

[4] R. Collins, A. Lipton, H. Fujiyoshi and T. Kanade, *Algorithms for Cooperative Multisensor Surveillance*, Proceedings of the IEEE, 2001.

[5] R. Collins, O. Amidi, T. Kanade, *An active camera system for acquiring multi-view video*, IEEE ICIP, 2002.

[6] T. Darrell, D. Demirdjian, N. Checka and P. Felzenszwalb, *Plan-view Tracjectory Estimation with Dense Stereo Background Models*, IEEE ICCV, 2001.

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

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

[9] M.Irani, B.Rousso and S.Peleg, *Detecting and Tracking Multiple Moving Objects Using Temporal Integration*, ECCV, 1992.

[10] O. Javed, Z. Rasheed, K. Shafique and M. Shah , *Tracking in Multiple Cameras with Disjoint Views*, IEEE ICCV, 2003.

[11] J. Kang, I. Cohen and G. Medioni, *Continuous Tracking Within and Across Camera Streams*, IEEE CVPR, 2003.

[12] V. Kettnaker and R. Zabih, *Bayesian Multi-Camera Surveillance*, IEEE CVPR, 1999.

[13] S. Khan and M. Shah, *Consistent Labeling of Tracked Objects in Multiple Cameras with Overlapping Fields of View*, IEEE TPAMI, 2003.

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

[15] L. Lee, R. Romano and G. Stein, *Learning Patterns of Activity Using Real-Time Tracking*, IEEE TPAMI, 2000.

[16] T. Matsuyama, N. Ukita, *Real-Time Multitarget Tracking by a Cooperative Distributed Vision System*, Proceedings of the IEEE, 2002.

[17] S. Mann and R. Picard, *Video Orbits of the Projective Group: A Simple Approach to Featureless Estimation of Parameters*, IEEE Transactions on Image Processing, 1997.

[18] A. Mittal and L. Davis, *M_2 Tracker: A Multi-View Approach to Segmenting and Tracking People in a Cluttered Scene*, IJCV, 2003.

[19] R. Pless, T. Brodsky and Yiannis Aloimonos, *Detecting Independent Motion: The Statistics of Temporal Continuity*, IEEE TPAMI, 2000.

[20] W. Press, B. Flannery, S. Teukolsky and W. Vetterling, *Numerical Recipes in C*, Cambridge University Press, 1988.

[21] K. Shafique and M. Shah, *A Noniterative Greedy Algorithm for Multiframe Point Correspondence*, IEEE TPAMI, 2005.

[22] C. Stauffer and K. Tieu, *Automated Multi-Camera Planar Tracking Correspondence Modelling*, IEEE CVPR, 2003.