Automated multi-camera planar tracking correspondence modeling

Chris Stauffer and Kinh Tieu
Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, MA, 02139

Abstract

This paper introduces a method for robustly estimating a planar tracking correspondence model (TCM) for a large camera network directly from tracking data and for employing said model to reliably track objects through multiple cameras. By exploiting the unique characteristics of tracking data, our method can reliably estimate a planar TCM in large environments covered by many cameras. It is robust to scenes with multiple simultaneously moving objects and limited visual overlap between the cameras. Our method introduces the capability of automatic calibration of large camera networks in which the topology of camera overlap is unknown and in which all cameras do not necessarily overlap. Quantitative results are shown for a five camera network in which the topology is not specified.

1. Introduction

If one was given a bank of one hundred video screens showing video that redundantly covered different sections of multiple concourses at a large airport, one would require experience or prior knowledge of the cameras and site to track all objects through all the cameras effectively. For instance, Figure 1 shows a five camera network in which four cameras overlap and one is independent. This model was automatically determined by the method presented in the paper. By understanding how the cameras are related, one could effectively track objects through the entire scene. Figure 2 shows four different types of overlapping relationships a network of cameras could have.

The general problem of automated tracking correspondence is to estimate which tracking data resulted from observation of the same objects. An ideal tracking correspondence model (TCM) would result in only as many tracking sequences as there were independently moving objects in an environment, regardless of the number of sensors or their overlap. For a network of cameras, it would model frame-to-frame correspondence, correspondence in overlapping camera views, correspondence across non-overlapping camera views, and even establish identity when an object returns to the environment after a long period of time has elapsed.

This paper assumes a reliable frame-to-frame tracker ex-
ists for each camera and centers on the subproblem of automated correspondence modeling for overlapping camera networks where objects in each region of overlap lie on or near a ground plane. Planar TCMs have proven their effectiveness in enabling tracking in multi-camera visual surveillance networks, but they still require manual supervision to calibrate.

The goal of this paper is to introduce the first fully automatic method for planar TCM estimation in large camera networks. Specifically, it doesn’t require the scene to be vacated by all but a single moving object, it doesn’t require manual specification of invariant properties of objects in a particular environment to bootstrap the correspondence, it doesn’t require specification of which pairs of cameras have valid correspondence models, and it doesn’t require assumptions about visibility of objects in the region of camera overlap. Our system can automatically determine the topology of camera overlap, determine pair-wise camera correspondence models, estimate visibility constraints for pairs of cameras, and track objects through extended sites. It estimates the TCM directly from the tracking data and is then capable of determining object correspondences. Further, this method could enable continuous validation of the model in the event that the camera positions or the scene are altered.

Section 2 discusses previous work. Section 3 formulates the general tracking correspondence problem. Section 4 describes the planar TCM used in this work. It contains subsections that describe how to determine the likelihood that two sequences correspond, outline our method for estimating the TCM, and outline our method for validating the correspondence model. Section 5 describes calibrating multi-camera networks and tracking in those networks. It shows analysis of our results in a five camera network (twenty camera pairs) over hundreds of tracking sequences. The final section discusses how this work could be integrated in multi-camera tracking systems and draws conclusions about this work.

2. Previous Work

Many early approaches to tracking correspondence have relied on manually calibrated camera networks and complex site models. For instance, Collins et al. [3] coordinated multiple PTZ and static cameras to track many objects over a wide area. This system relied on a 3-d model of the environment and fully calibrated cameras.

With the decreased barrier to setting up cheap multi-camera tracking systems, automated approaches to estimating a correspondence model to enable tracking in extended environments have become increasing interesting to the vision community. Khan et al. [5] automatically estimated the line where the feet of pedestrians should appear in a secondary camera when they exit the reference camera. This assumes that cameras are vertical and pedestrians will leave a scene at a linear occlusion boundary. It also does not exploit tracking data when the object is visible in both cameras to establish correspondence.

Stein et al. [6] did early work in estimating a perspective plane correspondence model of tracking data as well as the temporal alignment. Their system used a robust RANSAC [4] variant that estimates the best homography between two cameras that were known to overlap significantly. Their automated approach has been used and cited by many systems that employ correspondence modeling to do tracking in extended scenes, e.g. Black and Ellis [1]. They extended their approach to calibrate cameras when a triplet of overlapping cameras is available [6]. Recently, Caspi et al. [2] extended Stein’s approach to include a non-planar correspondence model and to take advantage of the tracking sequences rather than just co-occurring object detections resulting in more reliable estimation.

This paper outlines an improved method for estimating the planar correspondence model given time-aligned tracking data that takes full advantage of the unique characteristics of tracking correspondences in video streams. This paper also introduces a mechanism for validating the model to determine if the best correspondence model is valid. This is required to enable automated multi-camera TCM estimation without manually specifying the topological arrangement of the cameras (as seen in Figure 2). Finally, the primary goal of this paper is establishing tracking correspondence rather than building pleasing site models or reconstructing camera geometry. As a result, this is the first paper to include quantitative analysis of the effectiveness of estimation and employment of the tracking correspondence model over hundreds of tracking sequences.

3. General Tracking Correspondence

This section outlines the general tracking correspondence problem. It defines a tracking correspondence model (TCM) and explains how it is employed to estimate correspondence. The next section describes the multi-camera planar TCM.

3.1. Formulation

The ultimate goal of tracking correspondence modeling is to estimate which tracking observations resulted from the same objects. The corresponding tracking observations can then be merged across multiple sensors giving a more complete description of object activities in an environment.

An ideal tracking correspondence model (TCM) would result in only as many tracking sequences as there are independently moving objects in an environment regardless
of the number of sensors or their overlap. For a network of cameras, it would model frame-to-frame correspondence, correspondence in overlapping cameras, correspondence between non-overlapping cameras, and even correspondence for objects returning to the environment after a long period of time has elapsed.

The result of successful frame-to-frame tracking is a set of tracking sequences in each individual camera. Each sequence $S_i$ is comprised of $N_i$ discrete tracking observations, $\{s_i(t_0), ..., s_i(t_{N_i})\}$, indexed by the absolute times they occurred. Each observation includes a description of the object from a particular sensor at a particular time.

$$ s_i(t) = (x_i(t), y_i(t), dx_i(t), dy_i(t), s_i(t), image_i(t), ...) $$

The goal of tracking correspondence is to estimate the ideal correspondence matrix $\Gamma^*$ for all sequences. Each element of this matrix is

$$ \gamma^*_{ij} = \begin{cases} 1 & \text{if } S_i \text{ and } S_j \text{ correspond to the same object} \\ 0 & \text{otherwise} \end{cases} $$

3.2 Direct Correspondence

Unfortunately, it is unclear how to estimating the ideal correspondence matrix directly. An alternate form of the ideal correspondence matrix is the direct correspondence matrix, denoted $\Gamma$ in this work. In the direct correspondence matrix, $\gamma_{ij} = 1$ iff $S_i$ and $S_j$ correspond and $S_j$ is the first corresponding observation to follow $S_i$. In the case of overlapping cameras, we can effectively model the likelihood of direct correspondence. This enables us to estimate $\Gamma$. The true estimate of $\Gamma$, can be used to determine $\Gamma^*$.

A tracking correspondence model $\tau$ refers to the model and parameter values used to estimate correspondence between to tracking sequences. Our TCM defines the probability that two sequences are in direct correspondence, or

$$ p(\gamma_{ij} = 1|S, \tau) $$

where $S$ is the set of all sequences.

To determine if two sequences are the same object in the scene ($\gamma_{ij} = 1$), we need to maximize the likelihood of

$$ p(\Gamma|S, \tau) = \prod_{S_i, S_j \in S} p(\gamma_{ij}|S_i, S_j, \tau)^{\gamma_{ij}} $$

where $\sum_i \gamma_{ij}$ and $\sum_j \gamma_{ij}$ are zero or one. This type of assignment problem has been covered in other work (as in [7]) and will not be discussed further here. In fact, for the multi-camera direct correspondence modeling discussed in this paper, a simple likelihood threshold is often sufficient because an object in a mutually observable region is usually in good correspondence to itself and in poor correspondence to any other object (because two objects cannot occupy the same space). Also, objects that are coincidentally co-located, often do not tend to stay that way.

This paper investigates a particular subproblem of tracking correspondence—automated planar correspondence modeling for overlapping camera networks.

We begin by considering a pair of cameras $a$ and $b$. In a given interval, the cameras track $k^a$ and $k^b$ objects respectively, resulting in two sets of tracking sequences, $\{S_{a_1}, \ldots, S_{a_{k^a}}\}$ and $\{S_{b_1}, \ldots, S_{b_{k^b}}\}$.

Since this work concerns only overlapping camera correspondence, we are only considering direct correspondence. If two sequences in two cameras, $S_{a_i}$ and $S_{b_j}$, are observations of the same object in the scene over the same interval of time, they are in direct correspondence. The goal of this work is to estimate the $k_{a_{i}xk_{b_{j}}}$ direct correspondence matrix, $\Gamma$. The estimate of this matrix could trivially be combined with matrices from other correspondence measures to approximate $\Gamma^*$.

In the case of this paper, our site TCM $(\tau)$ is comprised of a separate TCM for each camera pair,

$$ \tau_{ab} = \{V_{ab}, H_{ab}, R_{ab}\}, $$

where $V_{ab}$ is a binary variable denoting whether a valid model exists between camera $a$ and camera $b$, $H_{ab}$ is the homography relating observations in camera $a$ to camera $b$, and $R_{ab}$ is the region in which valid correspondences are expected in camera $a$.

3.3. The structure of tracking data

One of the most prolific areas of computer vision literature has historically been camera calibration and 3D reconstruction. The vast majority of this literature uses static correspondence points in static scenes to estimate the model, e.g., image mosaics from corners or stereo. The primary goal is usually calibration or reconstruction. Without constraining this problem, there is little hope of inferring the correct feature correspondence from two images with a very wide baseline and no other information available, even for a simple parametric model (e.g., planar homography or mosaic). It would be even more difficult to validate whether the resulting model was the same scene or pictures of a similar scene in two completely different locations.

With tracking correspondences this is not as difficult. In contrast to static correspondences, tracking correspondences

- are very descriptive- In addition to their position, tracked objects can be described by their size, velocity, direction of travel, appearance, color, articulated motion, etc. These characteristics can aid in establishing correspondence.
• are spatially sparse- Usually tracked objects do not densely cover the scene at any point in time. This means that the number of possible corresponding pairs is limited.

• increase linearly with sampling time- By sampling a scene for twice as long, one could collect roughly twice the number of corresponding tracking sequences. These correspondence points will often be much more dense than pixels at some points in the scene. This also allows one to neglect potentially bad correspondence in favor of good ones while still finding a full descriptive model.

• are temporally continuous- The position of a tracked object can be interpolated from adjacent frames, so the tracking doesn’t have to be synchronized on a frame-by-frame basis.

All these properties make the solution of this problem significantly more reliable than trying to mosaic two arbitrary images together in scenes where tracking data is available. And, obviously, a TCM is only required when tracking data is available.

4. Planar TCM with visibility constraints

Often in far-field tracking, most objects lie on or near a single ground plane. This is due to two regularities: in the region of overlap of two cameras the surface on which objects travel is usually fairly flat; and most objects of interest to surveillance systems are constrained by gravity. Thus, this paper assumes objects’ centroids are roughly planar in the regions of overlap, although other models could also be used (see [2]). Under this assumption, the position of an object that is visible in both camera \(a\) camera \(b\) is related by a 3x3 homography, \(H_{ab}\).

\[
H_{ab} p_a(t) \cong \hat{p}_b(t)
\]

where \(p_a(t) = (x, y, 1)\) is the location of the object in camera \(a\) in homogeneous coordinates at time \(t\), \(\hat{p}_b(t)\) is the interpolated position of the tracked object in camera \(b\) at the same time, \(H_{ab}\) is a homography, and \(\cong\) denotes equality up to a scale factor.

As shown in Equation 5, if a pair of cameras has a valid model, the TCM will include valid estimates of a planar correspondence model and a visual occlusion model. A visual occlusion model, \(R_{ab}\), is an estimate of the region of mutual observability for a pair of cameras. If two cameras do not overlap, this region is empty. Otherwise, this region denotes the region in camera \(a\) in which objects should also be visible in camera \(b\).

We currently consider only object position for matches, but any property in one camera that has mutual information with a property in another camera could be used. In principle, a TCM could model any property that can be predicted in camera \(a\) based on a measurement in camera \(b\) including appearance, color, size, velocity, or dynamic properties.

Because one does not generally entertain many competing matching hypotheses, optimal assignment is not generally necessary. Thus, given a valid \(\tau_{ab}\) and a sequence from each camera \(S_a\) and \(S_b\), two sequences correspond if the mean squared error of the predicted sequence is greater than a threshold\(^1\). If a normalization model was available, a per-observation variance estimate could be used rather than a single variance estimate.

4.1. Estimating \(H_{ab}\)

Given the correct values of \(\gamma_{ij}\) and no outlier observations, it is trivial to estimate

\[
H_{ab} = \arg\min_H \sum_{(i,j)\notin \gamma_{ij}} \sum_{t \in ms(i,j)} |s_i(t) - Hs_j(t)|^2
\]

where \(ms(i,j)\) is the time interval of “mutual support” in which sequence \(S_i\) and sequence \(S_j\) were both visible in their respective cameras. Unfortunately, even a few false correspondences or outlier observations will likely result in a poor estimate of \(H_{ab}\).

Stein[6] used a sampling procedure that took \(k\) random samples of four co-occurring observation pairs in two cameras. For sufficiently large values of \(k\), at least one of the sets would result in a reasonable estimate of the true homography. The \(H_{ab}\) that minimized the maximum error of the \(n\%\) percent best matches (e.g., \(n=50\%\) would correspond to the median error) was used to filter the best matches and to refit the model. Unfortunately, even hundreds of thousands of iterations are not likely to produce a good homography in many cases.

Caspi et al. [2] used pairs of tracking sequences sampled from an initial correspondence function estimate, \(\hat{\Gamma}\). The procedure for computing the initial estimate of \(\gamma_{ij}\) is unclear as is the importance of having a good estimate. In the four cases shown in their paper, there were very limited numbers of objects travelling reasonably articulated paths, and one case involved manually specifying that projected size should be used for correspondence.

Unfortunately, in most surveillance applications, all the objects will primarily move in straight lines within the limited regions of overlap. Thus a single pair of tracks (one from each camera) will result in degenerate homography estimation. By finding any two pairs of non-collinear corresponding tracks, one is guaranteed to get a non-degenerate solution. Thus, by always using \(n\) pairs of tracks and sampling sufficiently, one is guaranteed to get a non-degenerate

\(^1\)This corresponds to a hypothesis test on the likelihood ratio of the two observations occurring given that they correspond vs. them occurring given that they don’t correspond.
homography estimation except in the case that ALL objects that pass through the area of intersection are on the SAME line. In this case, a degenerate solution (one that only aligns the data on that line) is still sufficient for tracking correspondence in that scene.

We define the set of all temporally co-occurring sequences as

$$C_{ab} = \{ \{ S_i, S_j \} | \forall \{ i, j \} s.t. ms(i, j) > 0 \}. \quad (8)$$

Unfortunately, for scenes with small regions of overlap, large numbers of objects, or significant occlusions, this set is very large. Thus we implement a method for sampling from $C_{ab}$ non-uniformly. To estimate a homography, we sample two $\{ i, j \}$-pairs from $C_{ab}$ with a likelihood $L_{ij}$ and use them as correspondence point sets. $L_{ij}$ is a heuristic on the likelihood that this pair of tracks correspond to the same object. Our heuristic likelihood included the probability of this particular pair corresponding given: the number of other objects that occurred at the same time$^2$; the probability of a corresponding track lasting a particular time interval; and the probability of matching the observed locations of $S_i$ to $S_j$ directly.

Thus, one is more likely to sample pairs of sequences that occurred when few other objects were present, that were sufficiently long, and that could match each other reasonably well. Other terms that might increase the likelihood of sampling an informative (and valid) pair might be: the probability that the pair isn’t a trivial match (two straight lines); the probability of a color match, the probability of velocity matches, the probability of size match, etc.

The $H_{ab}$ estimate with the maximum number of corresponding tracking sequences is chosen as the model for this camera pair. Figure 3 shows images from two cameras and the resulting homography.

4.2. Estimating $R_{ab}$

Once $H_{ab}$ is estimated and a reasonable set of likely corresponding tracks are available, it is possible to model the region of mutual observability of objects in camera $b$ within camera $a$, $R_{ab}$, $R_{ba}$ can be computed by projecting that region in camera $b$. If the TCM is valid, $R_{ab}$ must not be empty. Also, any sequence $S_{a_i}$ that passes through $R_{ab}$ should have a corresponding sequence in camera $b$, and any sequence $S_{b_j}$ that passes through $R_{ba}$ should have a corresponding sequence in camera $a$.

In unobstructed views, $R_{ab}$ is simply the intersection of the bounding rectangle of camera $a$ with the bounding rectangle of camera $b$ transformed into the coordinates of camera $a$. Some examples of visual occlusions that result in

$^2$This is approximated as the ratio of the number of possible valid pairs to the total number of pairs during that time.

$^3$Using a symmetric measure is an important consideration, because very dense tracking data can match any scene.

$$\begin{align*}
\text{Figure 3:} & \quad \text{The merged view of cameras (a) and (b) generated from the estimated } TCM_{ab} \text{ is shown in (c). Note that the tracking correspondence model corresponds to the plane of the centroids of the tracked objects and does } \text{NOT} \text{ correspond to the ground plane. Thus, perfect image alignment is not expected (or desirable). The region of mutual observability of objects in (a) and (b), } R_{ab}, \text{ is highlighted in (d).}
\\
RS & \text{ that are smaller or have "holes" are: buildings; fences; trucks; trees; window/lens dirt; or window frames.}
\\
\text{We estimate the probability that a point} & \quad (x, y) \text{ is in the region of mutual observability as}
\\
\delta_i(t) = 1 & \quad (9)
\\
\text{where } Q_{x,y} & \text{ is the set of tracking data in either camera projected to camera } a \text{ coordinates within a distance } d_i \text{ of } (x, y) \text{ and } \delta_i(t) = 1 \text{ if there exists a corresponding sequence in the other camera and } t \text{ is within the period of mutual observability. } p(R_{ab}(x, y) = 1) \text{ is defined as zero if no data is present in that region of the image. } p(R_{ab}(x, y) = 1) \text{ is simply an estimate of the probability of getting a positive match at position } (x, y). \text{ Figure 3(d) shows an example for two cameras.}
\\
\text{4.3. Validating } & \quad T_{ab}
\\
\text{Assuming a particular frame-to-frame tracking success rate, each track in } R \text{ of either camera should have a corresponding track } \rho \text{ of the time. In practice, a rather conservative estimate of } \rho \text{ may be used.}
\\
\text{Thus, the probability of having a valid TCM can be estimated as the probability that drawing a sample from the region of correspondence (in either camera) will result in a match. Thus the condition for a valid TCM is}
\\
V_{ab} = \begin{cases} 
1 & \text{if } E[|\delta_i|] > \rho \\
0 & \text{otherwise.}
\end{cases} \quad (10)
\end{align*}$$
where the expectation is over positions drawn from $R_{ab}$. If no model for $R_{ab}$ was available, tracks that were missed because they were behind visual occlusions (e.g., buildings) would greatly reduce this value. Since, any two random scenes will result in a correspondence model and corresponding tracks, it is important to use $R$ to factor out those types of missed correspondences.

### 5. Calibrating multi-camera networks

Thus far, we have only talked about pairs of cameras. This section explains automatic TCM estimation and tracking correspondence in a multi-camera network. The estimation procedure is

- Estimate best $\tau_{ij}$ for all pairs of cameras $\{c_i, c_j\}$.
  - Estimate $H_{ij}$.
  - Estimate $R_{ij}$.
  - Estimate $V_{ij}$ (validate camera pairs).
- Build camera graph with $c_i$’s as nodes and $V_{ij}$’s as edges.
- (optional) Build visualization for each clique of cameras.

At any point in time, any tracking sequence in camera $i$ which passes into $R_{ij}$ for a valid $\tau_{ij}$ can be tested for correspondence. If they correspond, the tracking IDs can be set to the same value. Because this system is automated, it could be used to continuously validate the correspondence model and detect if the scene or camera positions have changed.

#### 5.1. A five camera example

Given one hour of data from the five cameras shown in the top row of Figure 4, our system learned a site model. Figure 4 shows the second scene warped into the reference scene with the estimated $H_{ab}$ for all camera pairs and Figure 5 shows the resulting $R_{ab}$ estimate highlighted for all camera pairs. Table 1 shows the expectation of a match within the region of correspondence for all camera pairs. Note that models with no overlap tended to result in somewhat random homographies, very small $Rs$, and low validation scores.

The topology is evident. One clique of cameras is $\{c_1 - c_5 - c_2 - c_3\}$ and the other clique is the remaining camera $c_4$. They can be used to build a mosaic of the environment automatically. This is shown in Figure 1. Again, I must caution the reader not to evaluate the results based on how well the ground planes are aligned, because the goal is to align tracking centroids. If the ground planes were aligned in this figure, it would be a poor TCM estimate.

Our results would be more robust if values were normalized prior to estimating our model. Objects at different depths exhibit different noise characteristics. Our estimation would also be more robust if we augmented our matching criterion to include appearance, color, dynamic properties. By using the most unusual objects in an extremely busy environment the search for valid correspondence pairs could be greatly reduced.

### 6. Conclusions

We have described a method for automatically estimating a planar tracking correspondence model (TCM) for large camera networks with overlapping fields of view given a large set of tracking data from the observed area. In addition to recovering the homography between all pairs of cameras, the TCM also includes the corresponding regions of mutual observability. Furthermore, we use these regions to validate true overlapping fields of views as opposed to cameras with disjoint fields of view. The TCM can be used to better track objects between all the camera views.

### References


Figure 4: Estimated homographies for all pairs of cameras. Note that in general, some homography will be found even for non-overlapping fields of view. The region of mutual observability measure is used to validate true homographies, indicated by the boxed images.
The estimated regions of correspondence, $R_{ab}$, for all pairs of cameras are highlighted. These regions signify areas of mutual observability between pairs of camera field of views. They are used to validate true homographies, shown here with large regions of correspondence.