MASSACHUSETTS INSTITUTE OF TECHNOLOGY
ARTIFICIAL INTELLIGENCE LABORATORY

A.I. Memo No. 1655 April 1999

Monitoring Activities from Multiple Video Streams: Establishing a Common Coordinate Frame

G. Stein R. Romano L. Lee

{gideon, romano, llee}@ai.mit.edu

This publication can be retrieved by anonymous ftp to publications.ai.mit.edu.

Abstract

Passive monitoring of large sites typically requires coordination between multiple cameras, which in turn requires methods for automatically relating events between distributed cameras. This paper tackles the problem of self-calibration of multiple cameras which are very far apart, using feature correspondences to determine the camera geometry. The key problem is finding such correspondences. Since the camera geometry and photometric characteristics vary considerably between images, one cannot use brightness and/or proximity constraints. Instead we apply planar geometric constraints to moving objects in the scene in order to align the scene's ground plane across multiple views. We do not assume synchronized cameras, and we show that enforcing geometric constraints enables us to align the tracking data in time.

Once we have recovered the homography which aligns the planar structure in the scene, we can compute from the homography matrix the 3D position of the plane and the relative camera positions. This in turn enables us to recover a homography matrix which maps the images to an overhead view. We demonstrate this technique in two settings: a controlled lab setting where we test the effects of errors in internal camera calibration, and an uncontrolled, outdoor setting in which the full procedure is applied to external camera calibration and ground plane recovery. In spite of noise in the internal camera parameters and image data, the system successfully recovers both planar structure and relative camera positions in both settings.

Copyright © 1999, by Massachusetts Institute of Technology

This report describes research done at the Artificial Intelligence Laboratory of the Massachusetts Institute of Technology. Support for the laboratory's artificial intelligence research is provided in part by ONR grant N00014-97-0363.

1 Introduction

This paper presents a system for automatically building a global model of the activity in a large site using video streams from multiple cameras. In a typical outdoor urban monitoring scenario, multiple objects such as people and cars move independently on a common ground plane. The ground plane is thus a convenient 3D structure for anchoring a global coordinate system for activity and scene modeling. Transforming the activity captured by distributed individual video cameras from local image plane coordinates to a common coordinate frame then sets the stage for global analysis of the activity in a scene.

In a related paper [9], we focus on classifying activities as recorded by a distributed set of sensors by considering patterns of activity in a common global coordinate frame. Here, we focus on the problem of coordination of distributed sensors. In particular, we consider the following problems: given a set of cameras viewing multiple objects moving in a predominantly planar scene, first determine the areas of overlap, mosaic the views together into a single coordinate frame, and track individual objects as they move between camera views. Second, recover the 3D position of the dominant plane as well as the 3D positions and orientations of the cameras. Both problems rely on knowing correspondences between features in the cameras, but in general finding feature correspondences between very different views is hard. To overcome this difficulty we detect objects moving simultaneously in cameras with partially overlapping views[10]. We then use the object centroids as possible point correspondences in order to recover the planar projective transformations (homographies) between camera pairs. The structure of the plane and relative motion of camera pairs are recovered from each homography, up to a two-fold ambiguity. Using multiple camera pairs, we find a unique solution up to a scale factor for the 3D camera configuration and ground plane position and orientation.

This work presents a fully automated system for viewing activity in an extended scene with multiple inexpensive cameras, only nominally calibrated, at unknown locations. Thus, it is an important component in a visual surveillance and monitoring system such as [4, 9]. In particular, a distributed monitoring system needs to: record common patterns of activity, count statistics on commonly occurring events, detect unusual events compared to normal activity, detect specific events or people, all the while coordinating processing in distributed sensors. To support such processing, we transform individual camera events to a common frame. After recovering the 3D configuration of the cameras and the 3D position and orientation of the activity plane, we warp the planar parts of the scene to an overhead view. This new image can then be used for activity understanding in metric space: the size and speed of two objects in different parts of the scene can be compared, something which cannot be done in the foreshortened camera view (see Figure 13). We can also use the overhead view for registration with aerial photographs.

1.1 Overview of the Paper

In Section 2, we give a general overview of our approach to recovering the 3D configuration of multiple cameras using tracking data from each camera's video stream. Section 3 reviews the mathematical background to our approach. We describe the details of our system in section 4.

We then present experimental results for both laboratory scenes and challenging outdoor scenes. Section 5.1 shows experiments on homography estimation, and Section 5.2 demonstrates how the system can be used to track objects (cars, people) over multiple views.

There are some important practical issues which we address. Our ultimate intent is to use a large number of camera surveillance units. To that end we would like to use off-the-shelf, mass produced components without having to laboriously calibrate each unit. In section 5.3 we test whether the method described by [12] is robust to errors in the calibration of internal camera parameters. We find that it is the variance in parameters between cameras that has a major negative influence on accuracy of the recovered estimates (in this regard our problem is more difficult than with a single moving camera), but the variance of standard cameras is within the acceptable limits.

Finally Section 5.4 demonstrates the system in its entirety on an outdoor scene viewed from three stationary cameras. In this situation many of our theoretical assumptions do not hold in practice: the ground is not perfectly planar, the centroids of objects in multiple images do not correspond to exactly the same point in space, and the cameras are not perfectly calibrated. In spite of these difficulties, we find that good structure and motion estimates can be obtained. The recovered estimates of relative camera and plane position are illustrated in this section.

2 Overview of our Method

Our system assumes the following input and output:

- **Input:** video sequences from n fixed cameras at unknown positions and orientations, and approximate values of internal camera parameters.
- **Output:**
  - Locations and orientations of cameras and the ground plane in a global reference frame, up to scale.
  - Mosaic of the multiple cameras views into a single planar coordinate frame, either the image plane of one of the cameras, or an overhead view.
  - Unique global identifiers for all objects moving in the scene.

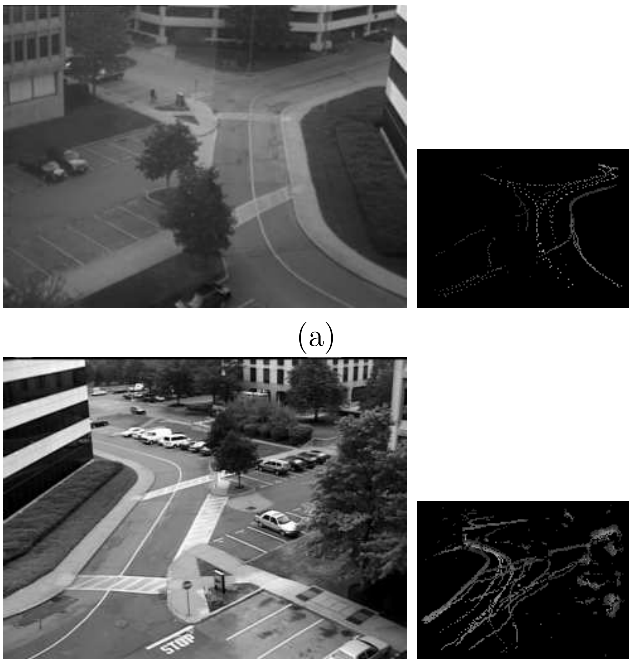
The complete system has four principal steps for taking raw individual video streams and building a global representation of the scene:

1. **Activity Tracking:** Track moving objects in each video camera and record image locations of their centroids.
2. **Ground Plane Alignment:** Robust recovery of homographies between camera pairs with respect to the common plane containing the scene motion, typically the ground plane.
3. **Plane and Camera Structure:** 3D recovery of the unique camera and ground plane locations.
4. **Overhead View Recovery:** Transformation of image data from multiple camera-dependent coordinate frames into a single Euclidean coordinate frame.

The first step is performed using the tracking technique described in [8]. This paper presents the steps 2, 3, and 4. We now describe the general ideas behind the methods; the system details are described in Section 4.

2.1 Ground Plane Alignment

The geometry of multiple views is well understood. For two views there exist geometric constraints that relate corresponding points in the two views to the 3D camera geometry. For a set of 3D points in general position, these take the form of the epipolar constraints. For a set of coplanar points the constraints take the form of a homography. In either case, given a non-degenerate set



(a)

(b)

Figure 1: (a)(b) Two views of a scene from different locations together with the tracks of cars and people over a six minute period.

of point correspondences we can solve for the constraints and thus recover the camera geometry, in particular the relative positions of the cameras.

The hard problem remains of how to find these correspondences. The views of the scene from the various cameras might be very different, so we cannot base the decision solely on the color or shape of objects in the scene. Figure 1 shows two views of a parking lot. One image was taken from inside a building through tinted glass. The other image was taken from an open air parking garage located at the opposite side of the intersection using a different make of camera with different geometric and photometric properties.

The scene in Figure 1 has a dominant plane, the ground plane, with many non-planar structures, thus we might consider methods for robust alignment of dominant planar patches [5]. However, these methods assume the view points are close enough so that gradient based techniques using constant brightness constraints can be used.

To get around these problems we use the centroids of moving objects in the images as features. Objects that appear to be moving simultaneously in two camera views are likely to correspond to the same 3D object. We determine the homography in two steps which are described in detail in sections 4.2 and 4.4:

1. *Rough alignment:* Using moving objects tracked in both views, we determine a rough alignment of two views of the ground plane.

2. *Fine Alignment:* The initial alignment does not perfectly register the ground plane since the centroids of 3D objects being tracked (people, cars) lie on a plane about 1 meter above the ground. Starting with the rough alignment, we use robust estimation techniques on static features to determine a more accurate registration of the ground plane.

In related work, Cham and Cipolla [2] use images taken from the same location but with large changes in orientation and internal parameters. They use a feature based approach, and use a coarse-to-fine search technique to determine the correct homography to align the images. In our case the camera locations might be far apart (as in Figure 1) and the scene is not planar, so most of the features in the image do not belong to a planar surface. Notice also that at a coarse resolution, the intersection, which is the dominant common feature in the scene, has a four-way symmetry. This means that without a good initial guess, coarse-to-fine search techniques will get stuck in local minima. Zoghlami *et al.* [15] exhaustively search all possible feature correspondences in an image pair to determine a homography. Our work is also related to the work of Azarbayejani and Pentland [1] who track blobs (two hands and a face) in two views of an indoor environment and derive the epipolar geometry. They assume that the corresponding blobs can be uniquely identified and have an initial guess for the camera geometry.

2.2 Plane and Camera Structure Recovery

Now suppose three cameras have overlapping fields of view. We choose one camera to be the base camera and compute the homographies mapping the base camera's view of the ground plane to each of the other two views. For each homography of this type, it is known that the planar structure and relative camera locations may be recovered up to a two-fold ambiguity, and that an additional camera resolves the ambiguity. The mathematical foundations of planar structure and motion recovery from image point correspondences were initially presented by Tsai and Huang in [13] and further developed by Weng, Ahuja, and Huang in [14]. The latter work thoroughly develops the recovery of two possible solutions for the camera locations and plane parameters from two cameras and the recovery of a unique, closed-form solution when three cameras are available.

In our setting, multiple camera pairs with overlapping views provide enough redundancy for us to eliminate all but one solution. By enforcing consistency between the ground plane normals recovered from each camera pair, we arrive upon a unique solution for the relative camera positions and ground plane position. These key mathematical results are presented in Section 3 using the notation of [3]. The same results have been used by Murray and Shapiro [6] and Sull and Ahuja [11] among others, to recover structure and motion from an image sequence of a planar structure taken from a single moving camera.

2.3 Determining the Overhead View

Given the 3D position and orientation of the ground plane in the coordinate frame of one of the cameras, we can construct homographies that map the image planes from each camera into a common 2D Euclidean coordinate system aligned with the ground plane. Now planar activity observed from multiple video streams can be merged and globally analyzed in a single Euclidean coordinate frame.

3 Mathematical Background

We represent points as elements of projective spaces using homogeneous coordinates. An image point $\mathbf{x} \cong (x, y, 1)$ is an element of the projective space $\mathbb{P}^2$ and a scene point $\mathbf{X} \cong (X, Y, Z, 1)$ is an element of the projective space $\mathbb{P}^3$, where $\cong$ denotes equality up to a scale factor.

It is known that when a set of 3D points are coplanar, their images under two perspective projections are related by a planar projective transformation or homography, i.e., for all scene points X lying on the plane $\mathbf{\Pi}$,

$$\mathbf{x}_2 \cong \mathbf{H}\mathbf{x}_1, \tag{1}$$

where $\mathbf{x}_1$ and $\mathbf{x}_2$ are the two images of X, and $\mathbf{H}$ is the 3×3 homography matrix corresponding to $\mathbf{\Pi}$.

The homography $\mathbf{H}$ may be expressed up to a scale factor in terms of the cameras' internal parameter matrices, the parameters of the plane $\mathbf{\Pi}$ and the cameras' relative positions and orientations. Let $\mathbf{M}_1$ and $\mathbf{M}_2$ be the internal camera matrices of camera 1 and camera 2. Let $(\hat{\mathbf{n}}, d)$ be the parameters of $\mathbf{\Pi}$ in the coordinate frame of camera 1, i.e. $\hat{\mathbf{n}}^T\mathbf{X} = d$ for all points $\mathbf{X} \in \mathbf{\Pi}$. Following the convention in [3], express $(\mathbf{R}, \mathbf{t})$, the 3D rotation and translation of camera 1 with respect to camera 2, in the coordinate frame of camera 2. Tsai and Huang showed in [13] that the homography $\mathbf{H}$ may be decomposed as

$$\mathbf{H} \cong \mathbf{M}_2(d\mathbf{R} + \mathbf{t}\hat{\mathbf{n}}^T)\mathbf{M}_1^{-1}. \tag{2}$$

Furthermore, they showed that given $\mathbf{H}$, $\mathbf{M}_1$, and $\mathbf{M}_2$, in general it is possible to recover two physically plausible solutions for $(\mathbf{R}, \mathbf{t}, \hat{\mathbf{n}}, d)$, up to a scale factor. Finally, they showed in [12] that three cameras can serve to disambiguate the two solutions: given a reference image and homographies to two distinct images with respect to the same plane, there is a unique solution for the relative camera positions and the geometry of the plane.

4 The System in Detail

4.1 Tracking and Pre-filtering

Our system tracks moving objects in multiple cameras using the tracking system developed by Stauffer (see [8]). Since we are dealing with static cameras but real world lighting conditions, the program uses adaptive background subtraction to detect moving foreground objects. For each camera, the tracking system is run on a separate computer and delivers a low-level description of each object tracked over multiple frames until it disappears from that camera's view. Each tracked object is given a unique identifier.

For homography estimation, spurious foreground motion is filtered out by discarding objects that disappear after only a few frames and objects that do not move a minimum distance in the image. This removes distracting motion such as trees blowing in the wind. For each salient tracked object, only its centroid in the image and a time stamp generated from the computer clock for the frame in which it was detected are used for homography estimation.

4.2 Rough Homography Estimation

The input to the homography estimation is two lists of triplets, $\{(x, y, t)\}$ and $\{(x', y', t')\}$, from cameras 1 and 2 respectively. For each moving object at each time step, a triplet includes the image coordinates (x, y) of the object's image centroid and the time stamp t of that frame. Each list is sorted by time.

Let us first assume that we know the offset between computer clocks and hence the time stamps from the two tracking sequences and that we have compensated for this offset. We create a list of all possible point pairings of for which $|t - t'| < t_0$, where t_0 is a small time window, typically the frame processing time of the slower computer. We now have M pairs of possibly corresponding image points: $\{(\mathbf{x}_i, \mathbf{x}'_i)\}_{i=1}^{M}$. Of course we will also have generated many false pairs. For example if we have two moving objects in each scene we will have four pairs when only a maximum of two pairs can be correct.

Most of the objects in the scene are moving on the ground plane, and therefore a homography from the coordinates in image 1 to image 2 is a good model. We now proceed with the LMS (least median of squares) algorithm:

1. From the M possible pairs, randomly pick N pairs (We use $N = 4$), $\{(\mathbf{x}_j, \mathbf{x}'_j)\}_{j=1}^{N}$, and use these to compute a homography matrix $\mathbf{H}$ from image 1 to image 2 by computing

$$\hat{\mathbf{H}} = \underset{\mathrm{H}}{\operatorname{argmin}} \sum_{j=1}^{N} ||\mathcal{N}(\mathbf{H}\mathbf{x}_j) - \mathcal{N}(\mathbf{x}'_j)||^2,$$

where $\mathcal{N}$ is the normalization operator so that the third homogeneous coordinate is 1.

2. For each of the M pairs $(\mathbf{x}_i, \mathbf{x}'_i)$, use the homography $\hat{\mathbf{H}}$ from step 1 to project the point $\mathbf{x}_i$ from image 1 to the point $\hat{\mathbf{H}}\mathbf{x}_i$ in image 2. The error for this pair is $||\mathcal{N}(\hat{\mathbf{H}}\mathbf{x}_i) - \mathcal{N}(\mathbf{x}'_i)||^2$.

3. From the M error terms computed in step 2 we find the lowest 20% of the errors and pick the largest of these to be the "LMS score" for this test. We choose 20% as a threshold because we expect less than half the possible point pairings to be correct. (Typically a least median of squares method would choose the threshold to be 50%.)

4. Repeat steps 1 through 3, K times, saving the random choice of N pairs and the corresponding homography that give the best (*i.e.*, smallest) LMS score.

5. After K tests we assume that the choice of $\hat{\mathbf{H}}$ that gave the smallest LMS score was computed from N correct and non-degenerate point pairings. We assume that the 20% of the points that gave the smallest error for this choice also represent correct point pairings. We now recompute the homography as in step 1, but using all of the top 20% point pairs. The resulting homography gives us the *rough alignment*.

4.2.1 How accurate are centroids of silhouettes?

A natural question to ask is whether the centroids of silhouettes in two views actually correspond to the same point in space. If the segmentation is perfect and the objects are spheres (or even ellipsoids) then the centroids of silhouettes do in fact correspond to the same point in space. In the case of people and cars this is not exactly true. Of course as an upper bound on the error we know that the true centroid of the object will be inside the convex hull of the silhouette but the situation is in fact better than that. Our simulation results show that the error is less than 10% of the size of the silhouette. Thus for an object which appears to be 10×10 pixels in the image the error in feature detection is about 1 pixel.

This does not take into account errors due to photometric effects. From one view point the object color might merge with the background, for example. Nor does it take into account occlusions.

4.3 Time Calibration

Until now we have assumed that the offset of the time stamps is known. Let us first observe that in general, if the time offset is incorrect then the pairs will no longer obey the homography constraint even if they come from the same 3D object (since the object would have moved). When we apply the LMS algorithm we will not get a small score. This observation provides us with a method for determining the correct time offset. We perform a one-dimensional search for the time offset that gives us the smallest LMS score. Since the trough is very narrow (see Figure 2) this search requires testing at every 1 second interval.

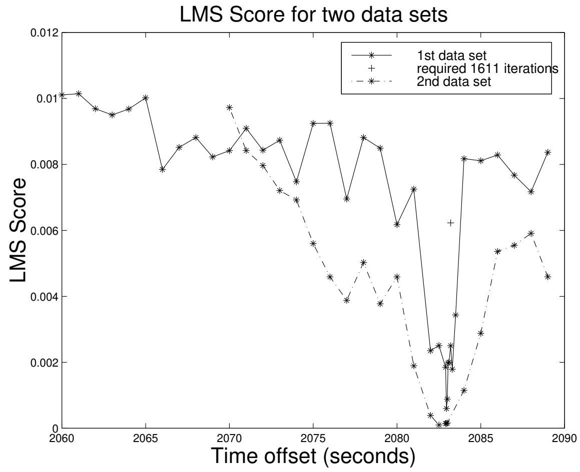


Figure 2: Least median of squares score for different time stamp offsets

There are clearly some motion patterns that will defeat this method. For example, if the 3D objects move on two straight lines at constant velocities then it will fail. This situation rarely occurs in practice since even traffic down a straight road is not always at a constant velocity.

4.4 Fine Ground Plane Alignment

Using the homography derived from the tracked 3D objects, we can warp image 1 towards image 2 (or vice versa). Since the centroids of the 3D objects do not lie exactly on the ground plane we find that the ground plane features do not align exactly (see Figure 4). Although not perfect, the alignment is close enough so that we can apply robust direct methods for planar alignment such as those in [5].

The general idea is to search for a homography matrix $\mathbf{H}$ that minimizes the sum of squared differences (SSD) between pixels in image 1 and the warped image 2, where image 2 is warped using the homography $\mathbf{H}$. The sum is typically over all the pixels of overlap between the images. This procedure is performed on a Gaussian pyramid for coarse to fine processing. Since there are many surfaces not on the ground plane and therefore are not expected to match well, we follow [5] and iteratively mask out regions where the difference is very large. To compensate for the large variation in brightness we perform histogram equalization. One might also consider using a high pass filter. An alternative is to use feature based methods.

4.5 Ground Plane and Camera Positions

Now let us assume that we have three cameras with optical centers C_1, C_2, and C_3 and internal camera matrices $\mathbf{M}_1$, $\mathbf{M}_2$, and $\mathbf{M}_3$. We choose camera 2 to be the base camera. Let $\mathbf{H}_{21}$ and $\mathbf{H}_{23}$ be homographies from image 2 to image 1 and from image 2 to image 3 corresponding to the scene's ground plane, computed using the above technique. Recall from Equation 2 in Section 3 that each homography can be decomposed in terms of the parameters of the ground plane $(\hat{\mathbf{n}}, d)$ and the internal parameters of the two cameras:

$$
\begin{aligned}
\mathbf{H}_{21} &\cong \mathbf{M}_1(d\mathbf{R}_1 + \mathbf{t}_1\hat{\mathbf{n}}^T)\mathbf{M}_2^{-1} \\
\mathbf{H}_{23} &\cong \mathbf{M}_3(d\mathbf{R}_3 + \mathbf{t}_3\hat{\mathbf{n}}^T)\mathbf{M}_2^{-1},
\end{aligned}
$$

where $(\mathbf{R}_1, \mathbf{t}_1)$ denotes the 3D rotation and translation from camera 1 to camera 2, and $(\mathbf{R}_3, \mathbf{t}_3)$ the 3D transformation from camera 3 to camera 2.

Although in theory three cameras yield a unique solution for the plane parameters $(\hat{\mathbf{n}}, d)$, in practice the image data is imperfect, so no single solution will satisfy both equations. However, the use of three cameras does serve to disambiguate between the two physically plausible solutions recovered from only a single pair of cameras, even in the presence of noise. In practice, there is a particular solution from the first camera pair that is closest to a particular solution from the second camera pair in terms of the angle of the ground plane normal, while all other pairings are significantly farther apart.

Once the multiple solutions for each camera pair have been disambiguated, there are two "nearby" solutions, one for each camera pair. Let $(\mathbf{R}_1, \mathbf{t}_1, \hat{\mathbf{n}}_1, d_1)$ denote the camera and plane solutions recovered from the first camera pair (camera 2 and camera 1) and $(\mathbf{R}_3, \mathbf{t}_3, \hat{\mathbf{n}}_3, d_3)$ those found using the second camera pair (camera 2 and camera 3). To decide on a final solution, the normal recovered from one pair is used to reconstruct the homography estimated for the second camera pair and vice versa. Let $\mathbf{A}_{23}$ be the homography from image 2 to image 3 reconstructed using $(\mathbf{R}_3, \mathbf{t}_3, \hat{\mathbf{n}}_1, d_3)$ and let $\mathbf{A}_{21}$ be the homography from image 2 to image 1 reconstructed using $(\mathbf{R}_1, \mathbf{t}_1, \hat{\mathbf{n}}_3, d_1)$:

$$\begin{aligned}
\mathbf{A}_{23} &\cong \mathbf{M}_3^{-1}(d_3\mathbf{R}_3 + \mathbf{t}_3\hat{\mathbf{n}}_1^T)\mathbf{M}_2^{-1} \\
\mathbf{A}_{21} &\cong \mathbf{M}_1^{-1}(d_1\mathbf{R}_1 + \mathbf{t}_1\hat{\mathbf{n}}_3^T)\mathbf{M}_2^{-1}.
\end{aligned}$$

We define an error measure on these reconstructed homographies that measures the sum of squared distances between image points projected using each homography, i.e.,

$$\epsilon_{21} = \sum_{\mathbf{x}} ||\mathbf{H}_{21}\mathbf{x} - \mathbf{A}_{21}\mathbf{x}||^2$$

and

$$\epsilon_{23} = \sum_{\mathbf{x}} ||\mathbf{H}_{23}\mathbf{x} - \mathbf{A}_{23}\mathbf{x}||^2$$

where $\mathbf{x}$ ranges over all pixels in image 2. The ground plane normal and distance from the camera pair with the smallest error are chosen as the unique solution for $(\hat{\mathbf{n}}, d)$. Section 5.4 presents the results of ground plane camera recovery from viewing activity in an outdoor scene.

4.6 Transformation to Overhead View

Finally, we would like to transform the image points from all three cameras into a single 2D Euclidean coordinate system aligned with the ground plane. In other words, we would like to find for each image plane, a homography mapping it to the image from a virtual overhead camera.

We will focus on finding $\mathbf{G}_2$, the homography from the second camera's image plane to the overhead view. The homographies $\mathbf{G}_1$ and $\mathbf{G}_3$, from cameras 1 and 3 to the overhead view, may then be formed by simply composing the image to image homographies with $\mathbf{G}_2$: $\mathbf{G}_1 \cong \mathbf{G}_2\mathbf{H}_{12}$ and $\mathbf{G}_3 \cong \mathbf{G}_2\mathbf{H}_{32}$, where $\mathbf{H}_{12} \cong \mathbf{H}_{21}^{-1}$ and $\mathbf{H}_{32} \cong \mathbf{H}_{23}^{-1}$.

We will define the homography $\mathbf{G}_2$ in terms of the recovered plane parameters $(\hat{\mathbf{n}}, d)$ using Equation 2. Using this decomposition, $\mathbf{G}_2$ can be thought of as the image transformation of scene points lying in the ground plane when a virtual rotation and translation are applied to camera 2.

Figure 3 illustrates the virtual rotation and translation of camera 2 to a virtual camera with center C_v that gives an "aerial view" of the scene. The ground plane parameters $(\hat{\mathbf{n}}, d)$ are expressed in the coordinate frame with origin C_2, and the rotation and translation of camera 2 relative to the virtual camera, $(\mathbf{R}_v, \mathbf{t}_v)$, are expressed in the coordinate frame with origin C_v. In order to center the image from the virtual camera on the image data viewed by camera 2, we have chosen its origin to lie directly above the point on the ground plane that is intersected by the optical axis of camera

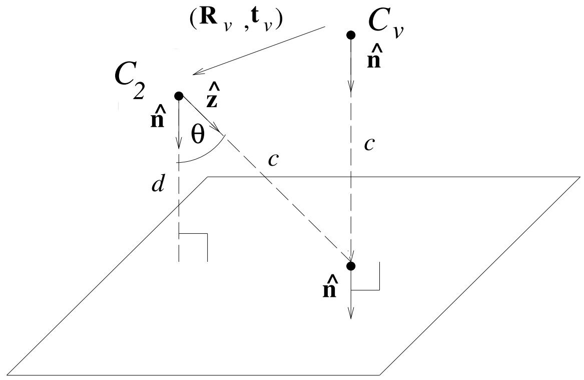


Figure 3: Rotation and translation of a virtual camera centered at C_1 to align the image plane of the camera cetered at C_2 with the ground plane and recover the corresponding homography $\mathbf{G}_2$.

2. We have also chosen the height c of the virtual camera center, C_v, from the ground plane to be equal to the distance from C_2 to the ground plane along the optical axis. Fixing these parameters simply amounts to choosing a translation and scaling within the image plane of the overhead view.

To derive $(\mathbf{R}_v, \mathbf{t}_v)$ in terms of the ground plane parameters in 3D Euclidean space, let $\hat{\mathbf{z}} = (0, 0, 1)^T$ be the optical axis of camera 2, and assume the orientation of the ground plane normal is "downward." The rotation from $\hat{\mathbf{n}}$ to $\hat{\mathbf{z}}$ is a rotation about the axis $\boldsymbol{\omega} = \hat{\mathbf{n}} \times \hat{\mathbf{z}}$ with a rotation angle of $\theta = \cos^{-1}(\hat{\mathbf{n}} \cdot \hat{\mathbf{z}})$. The resulting rotation matrix is $\mathbf{R}_v = e^{[\omega]_\times \theta}$, where $[\omega]_\times$ is the anti-symmetric matrix such that for any vector $\mathbf{v}$, $[\omega]_\times \mathbf{v} = \boldsymbol{\omega} \times \mathbf{v}$ [7]. Note that this 3D rotation implicitly chooses a 2D rotation within the image plane of the virtual camera.

The virtual translation $\mathbf{t}_v$ as expressed in the coordinate frame of camera 2 is $(\hat{\mathbf{n}} - \hat{\mathbf{z}})$. We rotate this vector into the coordinate frame of the virtual camera and scale it by the desired height c to obtain $\mathbf{t}_v = c\,\mathbf{R}_v(\hat{\mathbf{n}} - \hat{\mathbf{z}})$.

The homography from camera 2 to the aerial view image can then be written as

$$\mathbf{G}_2 \cong \mathbf{M}_2(d\,\mathbf{R}_v + \mathbf{t}_v \hat{\mathbf{n}}^T)\mathbf{M}_2^{-1}.$$

Finally, the homographies $\mathbf{G}_1$ and $\mathbf{G}_3$ are constructed from $\mathbf{G}_2$, and all three homographies are used to warp images taken from the cameras into a common overhead view. Sections 5.3 and 5.4 present results of these overhead warps on image streams taken in both laboratory and outdoor settings.

5 Experiments

5.1 Homography Estimation: Outdoor Experiments

Figures 1(a) and 1(b) show two views of a parking lot together with the corresponding tracking data over a period of six minutes. The tracks in Figure 1(b) appear more solid because the camera was connected to a faster computer giving a higher frame rate. For these data sets $M \approx 1300$ possible point pairs were found.

The data in Figure 1(a) was captured live. For technical reasons the data in Figure 1(b) was captured on a video camera and brought back to the lab for processing. The time stamp offset was therefore about 34.5 minutes, and the search algorithm is initialized at this offset value. In general, the search algorithm is initialized to a zero time stamp offset, under the assumption that the computers clocks are correct to within a few seconds. Using the search algorithm, the time stamp offset was found to be 2082.9 secs (34.715 min). Figure 2 shows the 20% LMS score for different time offsets. In all but one case the LMS algorithm found the best score in under 1000 trials. In one case (offset=2083.2 secs) 1611 trials were required. The best score for 1000 trials for this case is marked by a +. The dot-dashed line shows a similar plot for the next 6 minute block of tracking data from the two cameras.

Figure 4(a) shows the image from Figure 1(a) warped to the view in Figure 1(b) using the homography obtained from the tracking data. The results look qualitatively correct. In order to highlight the differences, Figure 4(b) shows the edges from Figure 4(a) overlayed onto Figure 1(b). The alignment is improved by refining the homography using ground plane alignment (Figure 4(c)). The refined alignment compares favorably with that which is achieved using manually selected feature points (Figure 4(d)).

5.2 Application: Combining Tracks from Multiple Views

After we have found the correspondences of the tracked data between multiple views, we can track objects as they move from one camera view to the next. Given the three sets of tracking data in Figure 5, we can align all three images using the estimated ground plane homography. Furthermore, the robust alignment procedure provides correspondences between entire tracks in the various images. Figure 6 shows some examples of tracks over multiple views that the program has determined belong to the same object. Figure 7 shows an example of an error where two cars were traveling close together (about two car lengths apart) and were assigned the same unique identifier.

We have chosen to align the three views with the viewpoint of the middle camera because it gives a clear view of the scene. How to choose such a view automatically is an open research question.

5.3 Camera and Plane Recovery: Laboratory Experiments

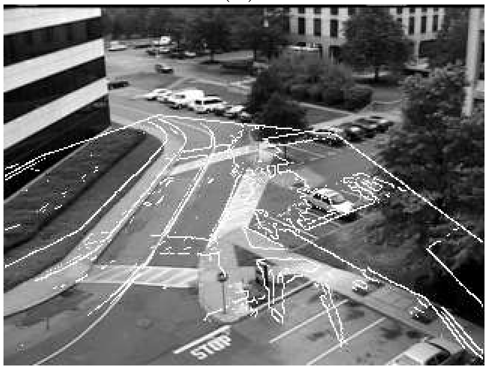
We now test the recovery of the 3D positions of the cameras and the ground plane. This section describes experiments to determine the effect of error in internal camera parameters on the recovered plane and camera positions. A single camera was mounted on the rotating arm of a motion stage (Figure 8). A planar checker board pattern was placed in the camera view close to the axis of rotation of the motion stage. Figure 9 shows two images from the sequence used for the experiments. The images were taken at 5^o intervals. These images show the effects of perspective foreshortening. We chose 3 points known to form a 90^o angle on the checker board. This angle, when measured in the image in Figure 9(a) is 80^o and when measured in Figure 9(b) is 76^o.

Six corresponding coplanar points were selected in the two views. Using these points, we computed the least squares solution to the homography between the images. Using the procedure described in Section 4.5, the camera motion and plane normal were computed using a set of internal camera parameters derived from the camera specifications (lens focal length = 8mm, CCD diagonal = $\frac{1}{3}''$).

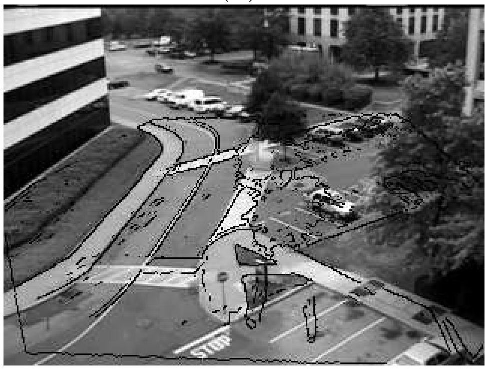
After computing the plane normal, a homography can be computed which brings the image 9(b) to a perpendicular view. The result of warping Figure 9(b) is shown in Figure 9(c). This warp has removed the foreshortening effects of the perspective projection and the angles are now square (90.1^o).

We now explore the effects of error in the internal camera parameters. In particular we will focus on errors in focal length and principal point. It is important to distinguish between the two cases. In the first case, all the cameras are identical but we have an error in the common focal

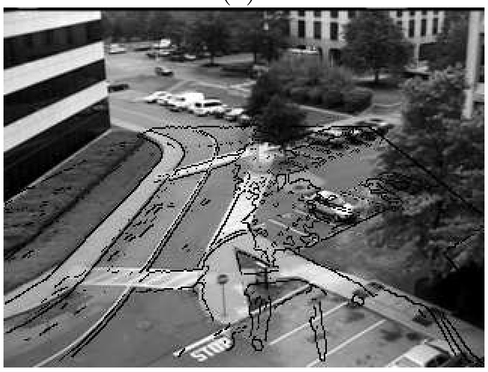
(a)



(b)



(c)



(d)

Figure 4: (a) Figure 1(a) warped towards Figure 1(b). (b)-(d) Edge maps for Figure 1(a) warped to Figure 1(b) and overlayed upon it. (b) Homography determined from tracking data. (c) Refined homography computed from alignment of static features. (d) Homography found using manual correspondences shown for comparison.

Figure 5: Input views and tracking data from three cameras.

(a)

(b)

(c)

Figure 6: Tracks identified as a single object in multiple views. (a), (b) tracks of a car entering top left and exiting bottom right. (c) car pulls out of parking spot in the center of the image and exits top left.

Figure 7: An example of an error. Two vehicles travelling close together (about two cars lengths apart) were assigned the same unique identifier.

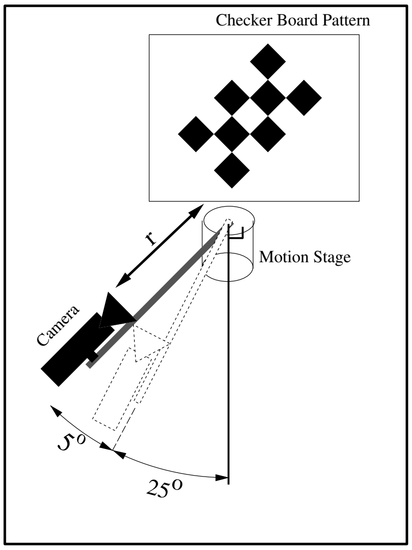
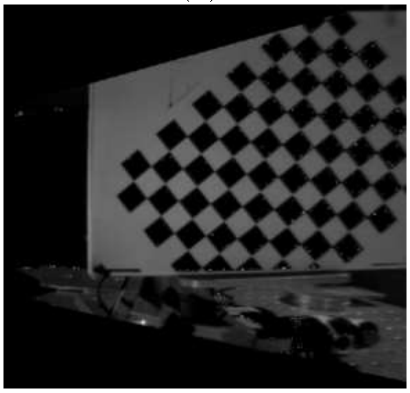


Figure 8: Diagram of the lab setup. The camera is mounted on the rotating arm of a motion stage. The axis of rotation is parallel to the camera Y axis.

(a)

(b)



(c)

Figure 9: (a)-(b) Two images of a checker board pattern. The camera has rotated 5^o between images. The cameras' optical axes are at angles of 25^o and 30^o relative to the plane normal in (a) and (b) respectively. The 6 points overlaid with white squares were used to compute the homography. Of those, the 3 solid squares were used to measure the right angle. (c) Image (b) warped to an overhead view. Note that now the angles of the checker board pattern are rectified to 90^o.

length and principal point as in the case of a single moving camera. In the second case, there is a small variation between the cameras due to the manufacturing process. In off-the-shelf, inexpensive cameras and lenses we can find a variation in focal length of 5-10% and a variation in principal point of up to 10 pixels. In this experiment we have a single camera, but we have simulated the effects of variation among cameras by changing the parameters in only one of the camera matrices.

Figure 10(a) shows the effects of changing the focal length in one or both of the camera matrices on the estimated camera rotation. As we can see, the effect of changing the focal length in only one camera is significantly larger. This pattern repeats itself when we look at the effects of focal length on the estimate of the right angles on the checker board and when we look at the effects of errors in the location of the principal point.

We can conclude that variation in internal camera parameters among the cameras, and in particular the principal point, can have significant impact on the accuracy of the results. However, even with a large error in the principal point of ± 10 pixels, one can achieve useful results.

5.4 Camera and Plane Recovery: Outdoor Experiments

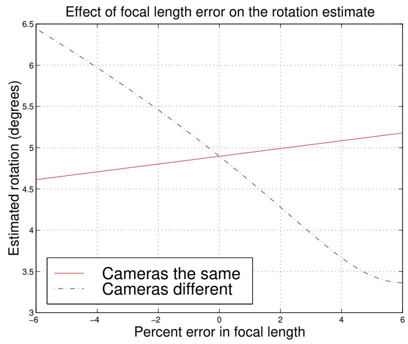
The full system has been tested on video streams captured from three cameras looking out from different rooms of an office building and viewing a busy parking lot. Cameras 1 and 3 are located on the 7th floor of of the building in opposite corners of one face, and camera 2 is located on the 9th floor of the building in the center of the same face (Figure 11). The cameras approximately form an equilateral triangle with a base of $57.4'$ and height of $21.6'$. The base of the triangle is located $114.6'$ above the ground plane. Figure 12 shows a snapshot from each camera. The line of parked cars in Figure 12(c) corresponds to the parking lot labeled in Figure 11. Note that multiple cars and people are moving within each frame.

All three cameras are similar Phillips camera modules with $\frac{1}{4}''$ CCD's. Cameras 2 and 3 have 4.8mm lenses. Camera 1 has an 8.5mm lens. Nominal focal lengths were computed using these specifications. The principal point is taken to be the center of the image. No further calibration is performed.

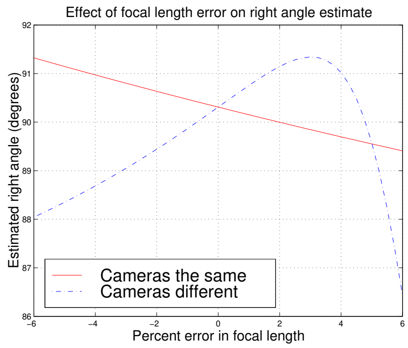
The tracking algorithm processes the video streams from each camera for a period of 10 minues; the resulting tracks are shown in Figure 12. Using our robust homography estimation on the tracked object centroids, the system finds homographies from camera 2, the reference camera, to each of camera 1 and camera 3. These homographies are then decomposed as described in Section 3 and solutions for the plane parameters and relative camera positions are recovered.

To evaluate the success of the ground plane recovery, 14 points in the scene's ground plane were chosen and the actual Euclidean distances between them were measured outdoors. Figure 13(a) displays the chosen points and measured segments in the image plane of camera 2. Several of the measured distances are labeled. Note that there is a significant foreshortening effect in the input images. Figure 13(b) shows the effect of warping these points with the same ground plane homography used to warp the images. The foreshortening effects are drastically reduced, and the new distances between points now resemble the true distances in the Euclidean plane. The mean error of the distances in the warped images is 10%, while the mean error in the unwarped images is 32%. In the worst case, the improvement in distance error is much greater than the mean error, as shown by the example distances in Figure 13.

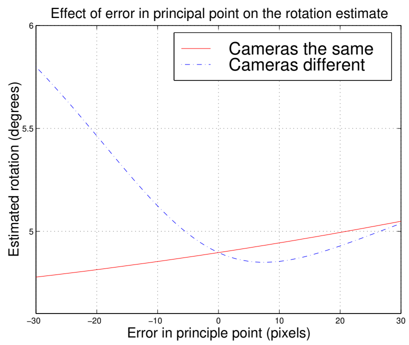
Figure 14 illustrates a sparse 3D reconstruction of the camera locations, ground plane, and measured points and distances in the ground plane. Figure 14(a) displays an overhead view of the 3D model, showing that that the three recovered cameras lie along a line where we expect the side of the building to stand (see Figure 11). Figure 14(b) shows a 3D view of the same model: the relative heights of the cameras are also roughly consistent with their known physical locations: the



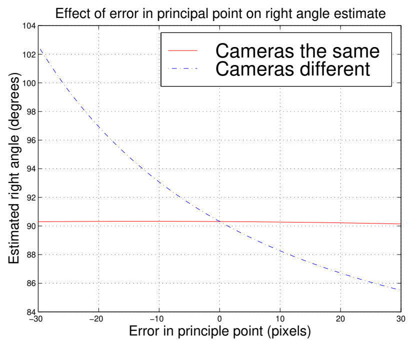
(a)



(b)



(c)



(d)

Figure 10: (a) Effect of errors in focal length parameter on the estimates of the rotation angle (ground truth is 5^o). (b) Effect of focal length error on the estimate of the right angle of the checker board. (c)-(d) Effect of errors in the principal point on the rotation estimate and the right angle, respectively.

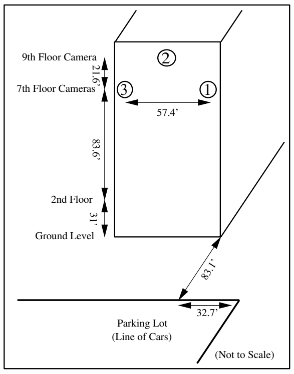
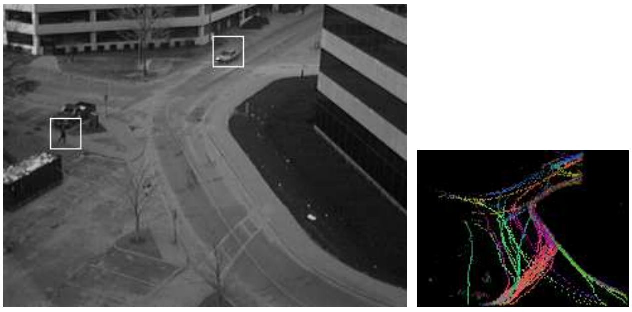
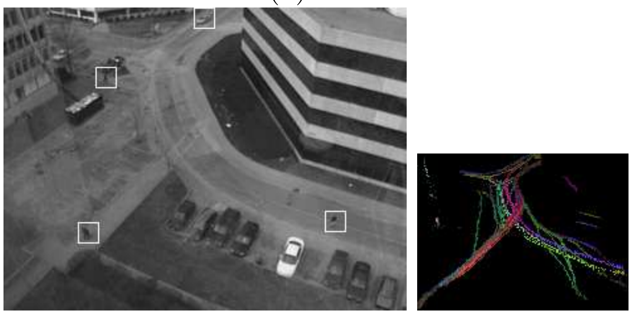


Figure 11: Diagram of the camera setup for the outdoor experiment (see text).



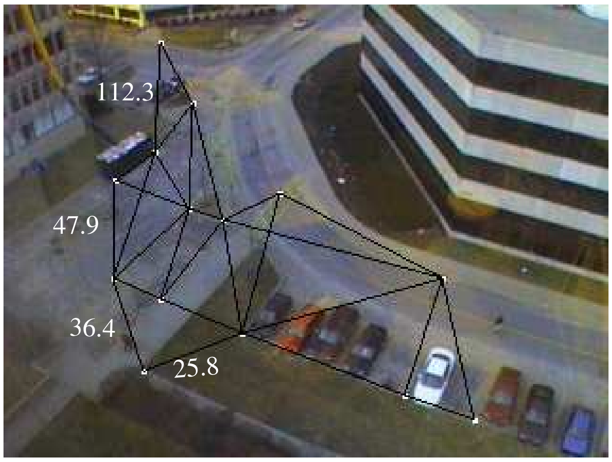
(a) Camera 1



(b) Camera 2

(c) Camera 3

Figure 12: Example snapshots from three cameras viewing an outdoor scene with 10 minutes of tracking data (right). Moving cars and pedestrians are highlighted with boxes.

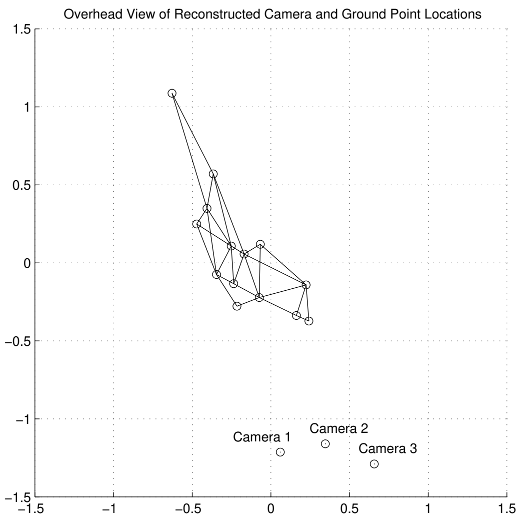


(a)

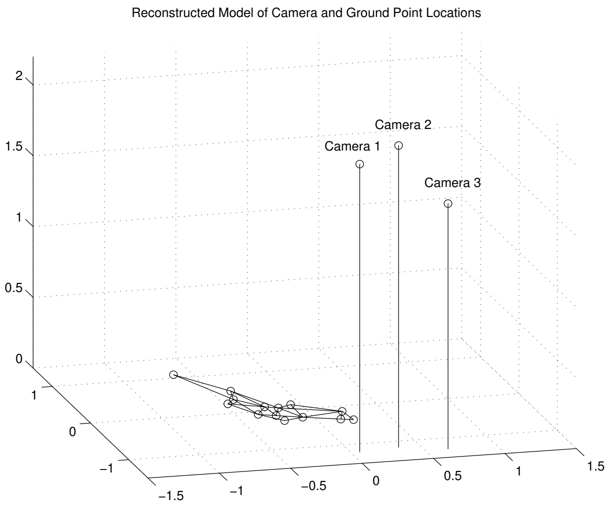
(b)

Figure 13: (a) Lattice of measured distances when viewed in the input image. Note how the four marked line segments appear to be the same length while their real lengths vary significantly. (b) Lattice of measured distances in the warped image.

actual height ratios are 0.8 : 1.0 : 0.8 and the reconstructed height ratios are 0.9 : 1.0 : 0.6.



(a)



(b)

Figure 14: (a) Overhead view of the reconstructed camera locations and lattice of measured points. (b) 3D view of the reconstructed camera locations and lattice of measured points.

6 Conclusions

In this paper, we have demonstrated a method for coordinating the activities detected in a distributed, uncalibrated, set of cameras into a single global coordinate frame. The method uses tracking data from moving objects to solve the correspondence problem between cameras. This allows multiple views to be coordinated to a single view or to a global, groundplane view. This calibration stage is an essential first step for systems [9, 4] that classify activities in sites based on learned patterns of common occurrences.

While the present work is sufficient to support such monitoring of activities, additional opportunites are evident. The next step in this work is to improve its robustness by exploiting redundancy in data to perform more accurate ground plane recovery. Widening the base line between camera pairs should significantly improve the accuracy of the reconstruction. In addition, more cameras will be added to the system, which gives rise to new problems such as performing a global optimization of the various camera positions and rotations and the ground plane geometry. With enough redundancy in the data, this optimization can reduce the weights on ground plane estimates from camera pairs with narrow base lines (which are typically error-prone), while increasing the weights on stable camera pairs with wide base lines, which are more likely to offer good ground plane solutions.

References

[1] Ali Azarbayejani and Alex P. Pentland. Real-time self-calibrating stereo person tracking using 3-d shape estimation from blob features. Technical Report 363, Media Laboratory, Massachusetts Institute of Technology, Cambridge, MA, January 1996.

[2] Tat-Jen Cham and Roberto Cipolla. A statistical framework for long-range feature matching in uncalibrated image mosaicing. In *CVPR*, pages 442–447, Santa Barbara, California, June 1998.

[3] O.D. Faugeras. *Three-Dimensional Computer Vision*, pages 206–208, 289–297. MIT Press, 1993.

[4] W.E.L. Grimson, C. Stauffer, R. Romano, and L. Lee. Using adaptive tracking to classify and monitor activities in a site. In *Computer Vision and Pattern Recognition*, pages 22–29, 1998.

[5] M. Irani, B. Rousso, and S. Peleg. Recovery of ego-motion using image stabilization. In *CVPR*, pages 454–460, Seattle, Washington, June 1994.

[6] D.W. Murray and Shapiro L.S. Dynamic updating of planar structure and motion: The case of constant motion. *Computer Vision and Image Understanding*, 63(1):169–181, January 1996.

[7] R. M. Murray, Z. Li, and S.S. Sastry. *A Methermatical Introduction to Robotic Manipulation.* CRC Press, 1994.

[8] C. Stauffer and W.E.L. Grimson. Adaptive background mixture models for real-time tracking. In *to be published CVPR*, 1999.

[9] C. Stauffer and W.E.L. Grimson. Learning patterns of activity using real time tracking. *under review*, pages ?–?, 1999.

[10] G. Stein. Tracking from multiple view points: Self-calibration of space and time. In *DARPA Image Understanding Workshop*, 1998.

[11] S . Sull and N. Ahuja. Estimation of motion and structure of planar surfaces from a sequence of monocular images. In *CVPR*, 1991.

[12] R.Y. Tsai and T.S. Huang. Analysis of 3-d time varying scene. IBM RC 9479, IBM Watson Research Center, Yorktown Heights, New York, 1982.

[13] R.Y. Tsai and T.S. Huang. Estimating three-dimensional motion parameters of a rigid planar patch, ii: Singular value decomposition. *IEEE Transaction on Acoustics, Speech, and Signal Processing*, 30(4):525–534, August 1982.

[14] J. Weng, N. Ahuja, and T.S. Huang. Motion and structure from point correspondences with error estimation: Planar surfaces. *IEEE Transactions on Signal Processing*, 39(12):2691–2717, December 1991.

[15] I. Zoghlami, O. Faugeras, and R. Deriche. Using geometric corners to build a 2d mosaic from a set of images. In *CVPR*, pages 421–425, San Juan, Puerto Rico, June 1997. IEEE Computer Society Press.