LABORATORY FOR COMPUTER SCIENCE

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

MIT LCS TR-814

Automatic Recovery of Camera Positions in Urban Scenes

Matthew E. Antone   Seth Teller

Computer Graphics Group

December 20, 2000

545 TECHNOLOGY SQUARE   CAMBRIDGE, MASSACHUSETTS 02139   (617) 253-5851

Automatic Recovery of Camera Positions in Urban Scenes

MATTHEW E. ANTONE   SETH TELLER
Computer Graphics Group
MIT Laboratory for Computer Science
{tone, seth}@mit.edu

Abstract

Accurate camera calibration is crucial to the reconstruction of three-dimensional geometry and the recovery of photometric scene properties. Calibration involves the determination of intrinsic parameters (e.g. focal length, principal point, and radial lens distortion) and extrinsic parameters (orientation and position).

In urban scenes and other environments containing sufficient geometric structure, it is possible to decouple extrinsic calibration into rotational and translational components that can be treated separately, simplifying the registration problem. Here we present such a decoupled formulation and describe methods for automatically recovering the positions of a large set of cameras given intrinsic calibration, relative rotations, and approximate positions.

Our algorithm first estimates the directions of translation (up to an unknown scale factor) between adjacent camera pairs using point features but without requiring explicit correspondence between them. This technique combines the robustness and simplicity of a Hough transform with the accuracy of Monte Carlo expectation maximization. We then find a set of distances between the pairs that produces globally-consistent camera positions. Novel uncertainty formulations and match plausibility criteria improve reliability and accuracy.

We assess our system's performance using both synthetic data and a large set of real panoramic imagery. The system produces camera positions accurate to within 5 centimeters in image networks extending over hundreds of meters.

1 Introduction

A long-standing problem in machine vision is that of external or extrinsic camera pose registration—i.e., determination of the rigid six-degree-of-freedom (6-DOF) Euclidean transformation that describes the scene-relative position and orientation of each camera. Accurate registration is vital to the recovery of geometry, texture, and other 3-D scene properties.

In scenes such as urban landscapes which contain parallel line sets, it is possible to decouple the rotational component of extrinsic pose from the translational component by using position-invariant features (vanishing points). Here we assume that camera orientations are known (obtained using the method of [1]) and focus on robust estimation of the camera positions. In this section, we provide a more precise problem formulation, present a high-level description of our technique, and discuss some relevant past work in the area of camera registration.

1.1 Motivation

The goal of the MIT City Scanning Project [22] is to obtain accurate three-dimensional models of urban landscapes. To this end, a platform equipped with various instrumentation acquires image data in hemispherical configurations (*nodes*) and annotates each image with approximate absolute pose (orientation and position) estimated by on-board sensors such as GPS, accelerometers, and odometry [7]. Nodes may be separated by large distances, and are acquired at different times of day and in different weather and lighting conditions.

Figure 1: Pose Mosaic Data Set
An example configuration of a set of acquired data. Hemispherical nodes, each of which consists of roughly forty 1.5 Mpixel images, are shown overlaid on a campus map.

Intrinsic camera parameters are estimated automatically, and the sets of planar images that comprise each node are rotationally registered to form hemispherical mosaics [11]. A pose refinement stage estimates the scene-relative position and orientation of each node using image features that are automatically detected and manually correlated across mosaics. Finally, registered cameras and various image features are used to reconstruct 3-D geometry.

Currently, extrinsic camera pose refinement is the only system component that requires human input. Manual feature correspondence becomes impractical, and indeed virtually impossible, as the number of acquired nodes increases. We thus wish to develop fully automated, scalable techniques for accurate and robust extrinsic camera registration.

1.2 Overview

The main goal of this work is to determine relative translations among an arbitrarily large set of cameras without requiring human input or explicit feature correspondence. Our system relies on several assumptions:

- *Intrinsic camera parameters are known.* These are estimated in advance by a separate algorithm [11].
- *Cameras are rotationally registered.* Rotations relative to a fixed coordinate frame are also estimated by a separate technique [1], described briefly in Section 3. This technique also provides a classification of observed image features that is used to constrain translational geometry.

- *Approximate translations are known*. We require a rough notion of camera adjacency and motion, both to determine which cameras are likely to view overlapping geometry and to disambiguate potential solutions. The acquisition platform supplies initial pose estimates accurate to within a few meters and a few degrees [7].
- *Images are omnidirectional*. Although not strictly required by our techniques, this assumption helps to resolve motion ambiguity and provides data redundancy for robustness.

We use the above information, as well as 2-D point features observed in the various images, as input to the system. A high-level block diagram is shown in Figure 2. First, a set of adjacent node pairs is established using approximate camera adjacency. The direction of translation (up to an unknown scale factor) is then estimated for each pair in the set via a hybrid Hough transform and Monte Carlo expectation maximization (MCEM) technique which neither requires nor produces explicit feature correspondence. Finally, all translation direction estimates are incorporated into an optimization that determines a globally consistent set of camera positions.

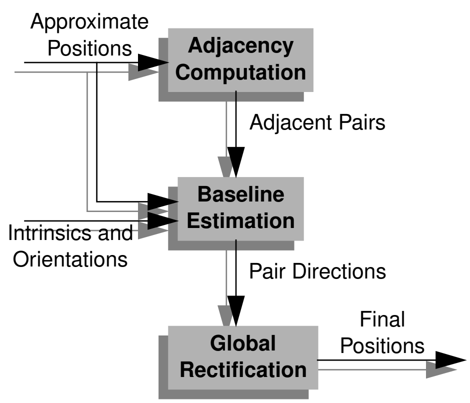


Figure 2: Translation Estimation
Pairs of adjacent cameras are determined using initial camera position estimates. The direction of translation is found for each pair, and these directions are subsequently incorporated into a simultaneous rectification of all cameras.

Hemispherical imagery disambiguates similar camera motions, and knowledge of approximate pose provides good algorithm initialization; consequently, the system is able to handle wide baselines (5–10 meters). In addition, since we use gradient-based image features rather than image texture, the system is largely insensitive to varying lighting conditions. Global optimization over all nodes drastically reduces bias and error propagation effects inherent in purely local techniques. The system has proven to work well even with significant error in initial pose estimates.

1.3 Related Work

There is a large body of work in the area of extrinsic camera calibration. Interactive methods (e.g. [3, 12, 20]) require a human operator to specify strong geometric constraints in 2-D image space or 3-D object space (e.g. point correspondence, parallelism, object primitives, etc.) which facilitate subsequent estimation of scene structure and camera pose. Such methods bypass many difficulties inherent in 3-D vision, such as the tight coupling between various unknown parameters; however, since the operator tends to invest minimal effort, these benefits come at the price of scalability, robustness, and stability.

Automated techniques (e.g. [2, 17, 18]) typically use image texture and geometric constraints to track image features over time in a densely-sampled image sequence such as video. This dense temporal sampling ensures that lighting conditions and feature characteristics do not vary drastically from image to image. One limitation of such techniques is that they assume a single image stream from a single sensor, and thus cannot merge data acquired by different sensors or at different times. Also, since only two or three images are typically considered at a time, these techniques suffer from localization artifacts (e.g. bas-relief ambiguity [4]), error propagation, and asymptotically high running times. Finally, most feature tracking algorithms and optical flow techniques fail when inter-camera baselines or rotations are large, or when there is significant variation in illumination from one image to the next.

Another important class of techniques attempts to circumvent the tracking/correspondence problem by estimating structure and motion using probabilistic correspondence—that is, without relying on explicit one-to-one feature correspondence. Wells [23] formulates an object recognition algorithm that matches new images to a template via probabilistic one-way classification. Rangarajan et al [19, 9] extend this algorithm to handle two-way image-to-image matching, though their formulation infers a somewhat ad-hoc probability distribution over correspondences. Dellaert et al [13] present a probabilistic structure from motion algorithm for multiple cameras that samples from the distribution over all correspondence sets using a Markov chain Monte Carlo algorithm; the technique assumes that the number of 3-D features is known and that all features are available in all images, and thus does not treat outliers or occlusion.

2 Image Features

Since pose images may be acquired at different times of day and under different viewing conditions (e.g. varying illumination and viewpoint), our methods rely on derived image features rather than directly on image texture. Sub-pixel edge features obtained from the image gradient and sub-pixel point features derived from line intersections are insensitive enough to changes in view and acquisition time to allow for reliable registration.

2.1 Lines and Points

Straight lines and points serve as the primary features for pose recovery. Such features are linear subspaces preserved by perspective projection; in addition, they satisfy many elegant geometric relationships (e.g. projective duality), and lend themselves to a wide variety of sophisticated mathematical tools. Since camera intrinsics are known, we represent these features as projective 3-D ray directions on the unit sphere rather than as points on the image plane.

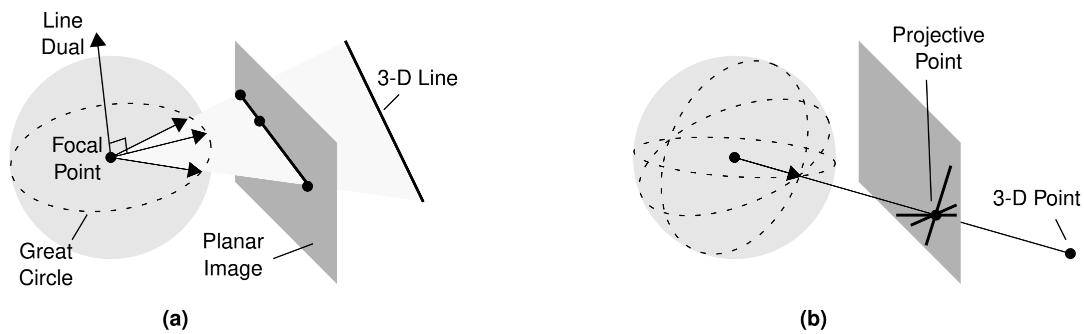


Figure 3: Projective Features
In (a), projections of a 3-D line are shown on the Euclidean image plane and on the sphere. The dual of the line is the projective direction orthogonal to all points on the line. In (b), projections of a 3-D point are shown; a point can be represented as the intersection of a pencil of lines either in the planar image or on the sphere.

Features are detected in each image before any higher-level processing is performed. A Canny edge detector [8] produces a binary image of edge pixels, which are chained and grouped using a connected components method. Least-squares optimization is then performed to find straight lines through the chains, and point features are obtained by the actual or extrapolated sub-pixel intersection of two or more adjacent lines. Although direct detection of corners from image pixels can also be used, we have found experimentally that segment intersections are more reliable.

2.2 Uncertainty

In this work, all geometric entities and inference tasks are represented on the unit sphere, which is a closed, compact, symmetric space. Thus, although features are detected in the Euclidean space of a planar image, they are treated as projective quantities determined solely by their 3-D ray directions. A Euclidean uncertainty model such as a Gaussian distribution cannot be applied to such quantities, so we utilize Bingham's distribution [5, 10], which exhibits antipodal symmetry and can describe a wide variety of shapes on the sphere (e.g. uniform, bipolar, and equatorial).

Bingham's distribution is given by

$$p(\boldsymbol{x}) = c(\boldsymbol{M})\exp(\boldsymbol{x}^T\boldsymbol{M}\boldsymbol{x}) \tag{1}$$

where $\boldsymbol{M}$ is a real symmetric 3×3 parameter matrix, $c(\boldsymbol{M})$ is a normalizing coefficient, and $\boldsymbol{x}$ is a 3-D vector with $\|\boldsymbol{x}\| = 1$. Many analogies can be drawn between this distribution and the Gaussian distribution; the mathematical forms are nearly identical, and in fact (1) is obtained by conditioning an ordinary trivariate Gaussian random variable to have unit length. The parameter matrix $\boldsymbol{M}$ can be further decomposed into

$$\boldsymbol{M} = \boldsymbol{U}\boldsymbol{K}\boldsymbol{U}^T, \tag{2}$$

where $\boldsymbol{U}$ is a unitary matrix describing the orientation of the distribution and $\boldsymbol{K}$ is a singular diagonal matrix whose entries describe the shape of the distribution.

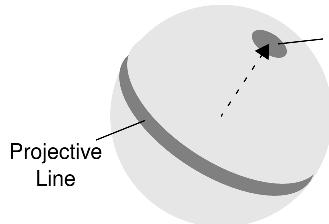


Figure 4: Feature Uncertainty
Uncertain projective lines and points can be represented by equatorial and bipolar Bingham distributions, respectively. Dual distributions are obtained by simple transformations of the parameter matrix.

We represent uncertain projective lines as equatorial Bingham random variables, with dual distribution defined by the parameter matrix $\boldsymbol{M}_d = -\boldsymbol{M}$. We represent uncertain projective points as bipolar Bingham variables. Uncertainty in line intersections, cross products, and fusion of uncertain measurements can also be easily obtained using this distribution, which is closed with respect to Bayesian inference.

3 Rotational Alignment

One of the main assumptions made in this work is that rotational pose is known *a priori*—that is, the orientations of all cameras are expressed relative to the same global coordinate system. If a scene contains sufficient geometric structure (namely a minimum of two parallel line sets viewed in per image), then the rotational component of extrinsic registration can be performed independently of the cameras' positions. An existing technique [1] correlates translationally invariant features among a set of cameras, then registers them in a globally-consistent optimization.

3.1 Method

Rotational registration relies on the detection, refinement, and correlation of 3-D line directions (or vanishing points). These directions are invariant to camera position and, if correlated across multiple cameras, can be used to determine relative orientations. In essence, scene-relative structure in the vicinity of each camera serves as a fixed reference to which the cameras can be rotationally aligned.

Vanishing points (VPs) are derived per hemispherical image using projective representations of 2-D image lines. The approximate number and locations of VPs in each node are first determined using a Hough transform technique that finds statistically significant peaks in a discretization of the space of all line features in a given node. These approximations are then used to initialize an expectation maximization (EM) algorithm which refines the VP directions according to a probabilistic mixture-model formulation.

When accurate vanishing point estimates have been determined for every node, a matching process finds correspondences between VPs in neighboring nodes; neighbors are determined using an adjacency graph of all approximate node positions. Finally, an iterative technique incorporates all vanishing point directions and correspondences, alternately solving for optimal rotations and optimal global vanishing points.

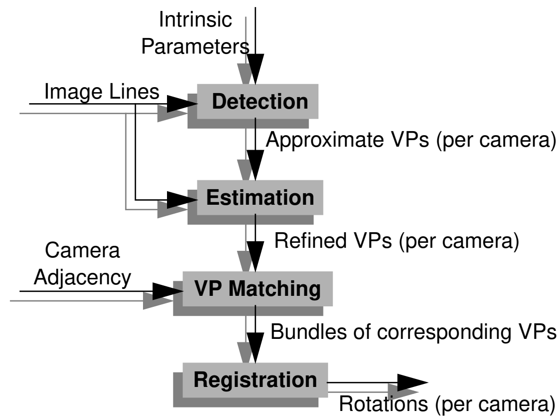


Figure 5: Rotational Registration
Vanishing points, or 3-D line directions, are detected and estimated for each hemispherical image. They are then matched across images, and finally rotationally registered.

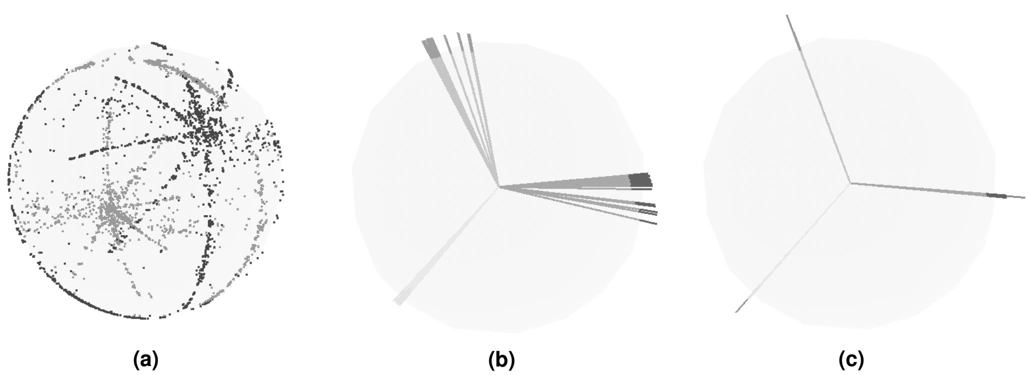


Figure 6: Real Rotational Data
An example of line features and VP cluster bands for a single node is shown in (a). VPs from 41 different nodes are shown before and after alignment in (b) and (c).

3.2 Recovered Information

The system described above produces globally-consistent orientations across a large set of cameras, typically to within about $0.1°$ (2 milliradians); as a result, only 3 of the 6 DOF (namely 3-D position) per node remain unknown. A by-product of rotational registration is a classification of 2-D line features: the mixture model and outlier rejection methods assign each line feature a probability of belonging to each 3-D direction. Thus, most of the 2-D line features have an accurately-known 3-D direction; the remainder can be classified as outliers and discarded. Although one-to-one correspondence between lines in different cameras is not known or required, this technique produces correspondences between *classes* of lines, which can be used to impose constraints when estimating the node positions (Section 4.2).

4 Pair-wise Translation

Accurate knowledge of camera orientation simplifies the problem of translational registration by decoupling highly nonlinear constraint equations. We now consider the translational registration of node pairs, i.e. determination of the two parameters (translation up to unknown scale) that describe the position of an offset node relative to a reference node. We use point features derived from image line intersections to determine these parameters without assuming explicit correspondence, although probabilistic correspondence is a by-product of our technique. All estimated pair-wise translations are then incorporated into a global registration step, described in Section 5.

We first discuss the geometric relationships that form the basis of our technique, then present a method which exploits these relationships to find the translation direction relating a given pair of cameras.

4.1 Geometry

For any given camera pair, translation can only be determined up to a scale factor, as shown in Figure 7. The use of 2-D feature observations introduces an inherent ambiguity that forces us to arbitrarily fix the global scale, say to unity. There are thus only 2 DOF to be determined, which can be represented by a 3-D unit vector in the direction of motion from the reference node to the offset node.

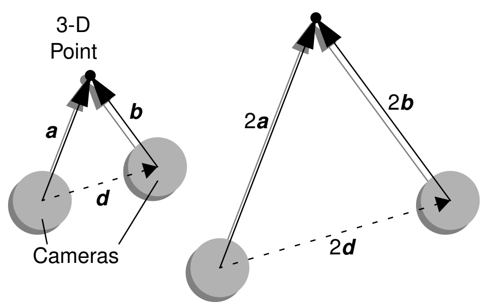


Figure 7: Scale Ambiguity
If observations consist only of 2-D point features (e.g. ray directions $\boldsymbol{a}$ and $\boldsymbol{b}$), only the translation direction $\boldsymbol{d}$ can be recovered; the distance between the cameras along $\boldsymbol{d}$ is arbitrary and simply imposes an isotropic scaling on the entire configuration.

We now examine the process by which point feature observations are generated in rotationally registered hemispherical images. Consider a reference camera at $\boldsymbol{c}_0$ and an offset camera at $\boldsymbol{c}_1$ separated by a pure rigid translation. A given 3-D point $\boldsymbol{p}$ produces projection rays $\boldsymbol{r}_0$ and $\boldsymbol{r}_1$ with respect to the two cameras (Figure 8a). Two projections taken alone, if they correspond to the same 3-D point, define an *epipolar plane* on which that 3-D point must lie; the normal to this plane is $\boldsymbol{m} = \boldsymbol{r}_1 \times \boldsymbol{r}_0$. For any pure translation between the two cameras and set of 3-D points, the set of all such epipolar planes forms a pencil whose intersection is coincident with the camera baseline (Figure 8b). The dual representation constrains all unit plane normals to lie on a great circle whose normal, in turn, is parallel to the baseline. Thus the translation direction $\boldsymbol{d}$ can be deduced solely from two or more corresponding feature ray pairs using the constraint $\boldsymbol{m} \cdot \boldsymbol{d} = 0$, for example by minimizing an error function of the form

$$E = \sum_i \boldsymbol{m}_{\mathrm{i}} \cdot \boldsymbol{d} \tag{3}$$

or by projective fusion of many uncertain Bingham random variables $\boldsymbol{m}_i$.

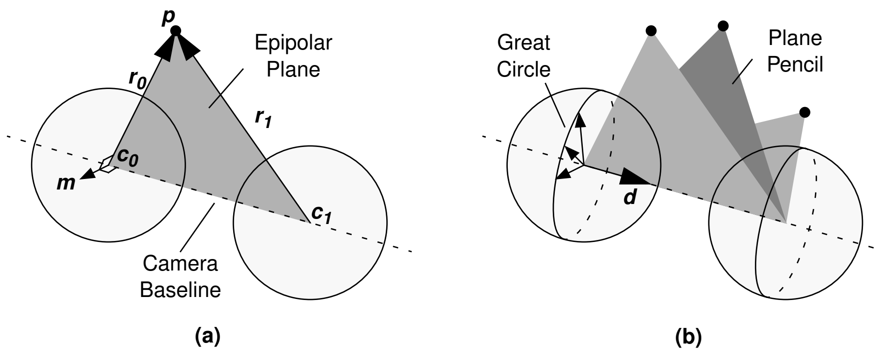


Figure 8: Epipolar Geometry
Geometry of the projection of a 3-D point onto two hemispherical images is shown in (a); projections of several points form a pencil of planes and a set of coplanar unit normals, depicted in (b).

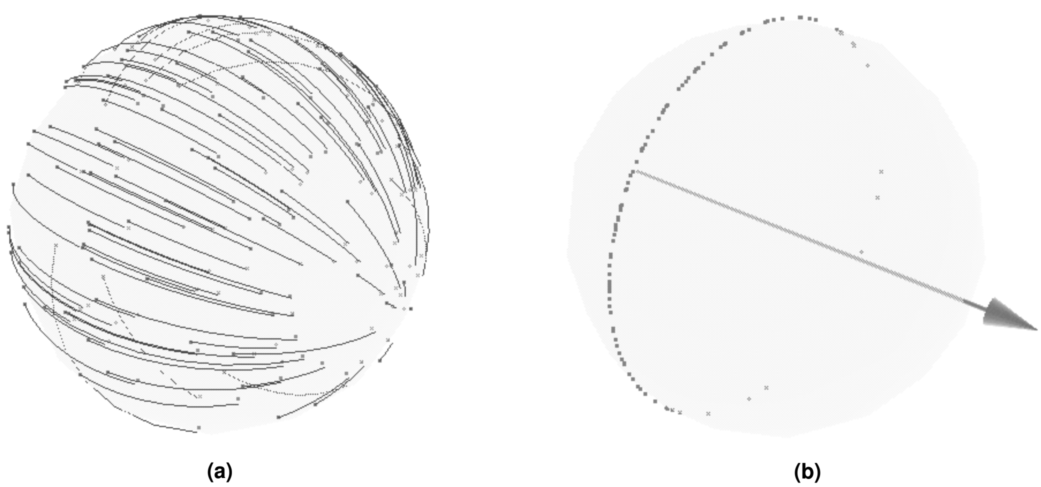


Figure 9: Arcs and Cross Products
The epipolar planes of true feature matches form a set of great circular arcs on the Gaussian sphere, shown in (a). Their duals form a coplanar set of vectors, shown in (b), whose normal is the direction of translation.

4.2 Match Constraints

The above formulation allows the motion direction to be estimated only if one-to-one correspondence between feature rays is available. If F represents the number of point features detected in each image, then there should be a single valid correspondence set containing $O(F)$ matches. However, in reality correspondence is unknown; there are $O(F^2)$ possible matches, and more importantly $O(F!)$ valid correspondence *sets*, a prohibitively large number. We therefore wish to reduce the number of matches considered while preserving a significant portion of the true matches.

A pre-processing step can also be performed which, though taking $O(F^2)$ time itself, reduces all subsequent computations to $O(F)$ and reduces the value of F itself. We utilize several geometric constraints that exploit three facts: first, that point features are generated by the intersection of two or more constituent image lines; second, that we have available a classification of image lines into parallel sets; and third, that we know the approximate camera configuration and can thus bound the uncertainty in the baseline direction. We can reduce F by rejecting any point feature $\boldsymbol{r}$ that was formed by any observed image line with unclassified 3-D orientation or with length smaller than a threshold. We then examine all possible matches, and reject a given match if:

- The 3-D orientations of the point features' associated lines do not match, e.g. $\boldsymbol{r}_0$ is formed by x and z lines, and $\boldsymbol{r}_1$ is formed by y and z lines.
- The "parities" of the point features' associated lines do not match. Each line is assigned a parity which is either 0 (meaning the line represents a dark-to-light transition in texture) or 1 (meaning the line represents a light-to-dark transition).
- $\boldsymbol{r}_1$ is closer to the translation direction than $\boldsymbol{r}_0$, i.e. this match would imply "backward" motion.
- The cross product $\boldsymbol{r}_0 \times \boldsymbol{r}_1$ does not lie in the uncertainty band around the equator. Such a match would imply motion in a direction outside the cone of uncertainty (see Section 4.4).
- The angle between $\boldsymbol{r}_0$ and $\boldsymbol{r}_1$ is greater than a specified threshold. This implies 3-D scene points that are unreasonably close to the cameras.

This pre-processing results in a roughly constant number of plausible matches per observed feature point. The majority of false matches are eliminated from consideration, and determination of the best matches is reduced from $O(F^2)$ to $O(F)$ complexity, which greatly improves performance.

4.3 Probabilistic Correspondence

Imposing the constraints described above drastically reduces the number of potential matches, but one-to-one correspondence is still unknown. Ideally, we wish to find $\boldsymbol{d}$ so as to minimize (3), but including only the true matches. In fact if the true matches were known, we could equivalently estimate $\boldsymbol{d}$ by minimizing the error criterion

$$\begin{aligned} E &= \sum_{i,j} b_{ij}(\boldsymbol{x}_i \times \boldsymbol{y}_j) \cdot \boldsymbol{d} \\ &= \sum_{i,j} b_{ij}(\boldsymbol{m}_{ij} \cdot \boldsymbol{d}) \end{aligned} \tag{4}$$

over *all* plausible matches, where $\boldsymbol{x}_i$ and $\boldsymbol{y}_j$ are feature rays from the two cameras and b_{ij} is a binary variable taking value 1 if $\boldsymbol{x}_i$ matches $\boldsymbol{y}_j$ and 0 otherwise.

If there are M point features in the first image and N features in the second image, then the set of binary weights b_{ij} can be represented by a binary $M \times N$ matrix $\boldsymbol{B}$ containing at most one nonzero entry per row and per column (because a single feature cannot have more than one deterministic match). To account for features which do not match any others, we augment $\boldsymbol{B}$ with an extra row and column to form $\tilde{\boldsymbol{B}}$ (Figure 10).

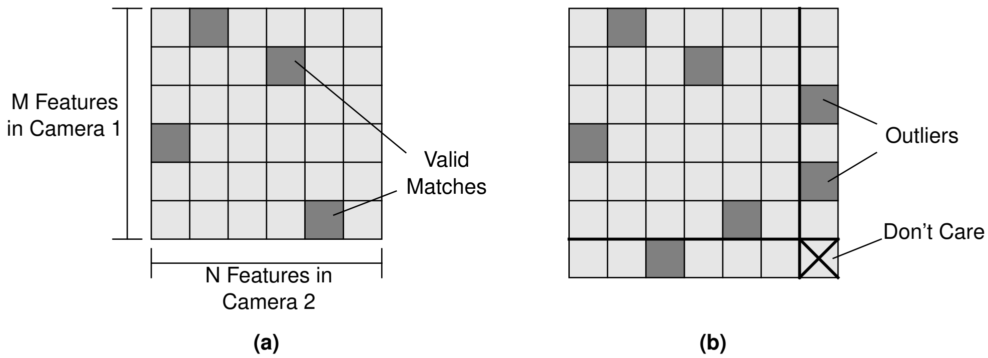


Figure 10: Augmented Match Matrix
In (a), a "degenerate" correspondence matrix containing outliers is shown. In (b), the augmented correspondence matrix treats outliers as correspondences and preserves the property that all rows and columns contain exactly one nonzero entry.

Various factors such as ambiguity of correspondence and uncertainty of feature localization in the image plane make binary correspondence $b_{ij} \in \{0, 1\}$ difficult to obtain, even for accurately known baseline directions. The notion of probabilistic correspondence $w_{ij} \in [0, 1]$ is therefore a sensible alternative. A new match matrix $\boldsymbol{W}$ can be constructed with structure similar to that of $\tilde{\boldsymbol{B}}$, but now each feature $\boldsymbol{x}_i$ can match several features $\boldsymbol{y}_j$, and vice versa. We require only that

$$\sum_i w_{ij} = \sum_j w_{ij} = 1 \qquad \forall (i, j), \tag{5}$$

i.e. that the matrix is doubly stochastic. A given weight w_{ij}, if computed correctly, is an expression of the probability that feature $\boldsymbol{x}_i$ matches feature $\boldsymbol{y}_j$. In this newly weighted formulation, the error to be minimized becomes

$$E = \sum_{i, j} w_{ij} \boldsymbol{m}_{ij} \cdot \boldsymbol{d}. \tag{6}$$

4.4 Discretization

As depicted in Figure 8, our observations consist only of 2-D point features (rays) in each node; true correspondences between them are unknown, as are the locations of their associated 3-D points. However, *any* point feature correspondence, whether correct or not, can be represented by a plane passing through both projection rays, or by the normal to this plane. The special subset of all potential correspondences that includes all true matches forms a pencil of planes or, in the dual sense, a set of coplanar points (Figure 9b). This suggests a method for determining the direction of motion: we can superimpose epipolar planes formed by all plausible feature matches and search the projective space for the point of highest incidence.

This space is two-dimensional and can be represented as a discretization of the surface of the unit sphere. The plane formed by each potential match, when intersected with the sphere, contributes a great circle. The point at which $O(F)$ intersections occur is the translation direction, approximated to the accuracy of the discretization.

In practice, we have initial translation estimates from the acquisition platform and a rough bound on the baseline direction, so it is not necessary to discretize the entire sphere. Instead we use a Hough transform (HT) confined to the solid angle on the sphere corresponding to this bound

(Figure 11a). In order to simplify the implementation and improve performance, the discretization is planar; epipolar planes then intersect the space as lines rather than arcs. The cone of uncertainty induces a band around the equator of the sphere inside which the epipolar plane normals $\boldsymbol{m}_{ij}$ of candidate matches are constrained to lie.

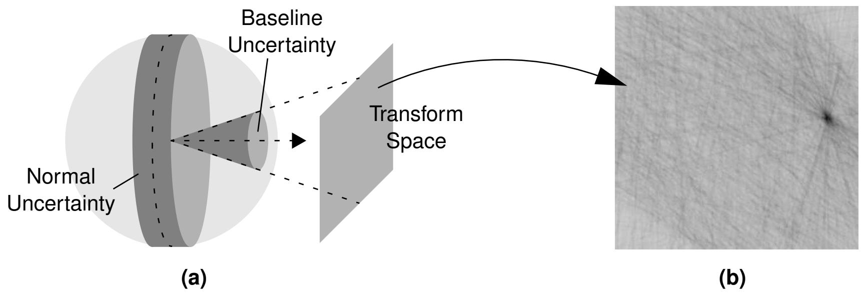


Figure 11: Hough Transform and Uncertainty
The translation direction lies somewhere in a cone of uncertainty that in turn induces an equatorial band on which epipolar plane normals of candidate matches must lie (a). The Hough space is a discretized planar surface corresponding to the intersection of the cone with the unit sphere (b). Each line (in black) arises from a plausible pairing of point features; the dark spot indicates the most likely translation direction.

Each epipolar plane is weighted by the approximate likelihood of its constituent point correspondence. In particular, the entries of a valid $(M+1)\times(N+1)$ match matrix $\boldsymbol{W}$ weight each line's contribution to the HT. The matrix is formed as follows. First, it is initialized to $\mathbf{0}$. Second, a 1 is placed in the matrix for each plausible match (using the criteria of Section 4.2). Third, a 1 is placed in every outlier row and column. Finally, the matrix is reduced to doubly stochastic form, so that all rows and columns sum to unity, by application of Sinkhorn's algorithm [21, 19]. Weighting each line by its approximate likelihood dramatically improves the coherence of HT peaks.

After all epipolar lines have been drawn in the HT, a set of candidate peaks is found by searching the Hough image $h(u, v)$ for relative maxima, i.e. points whose value exceeds all others in a square neighborhood with given size r:

$$h(u, v) \geq h(u+m, v+n) \qquad -r \leq (m, n) \leq r\,. \tag{7}$$

All such peaks are sorted by their magnitude $s(u, v)$, defined as

$$s(u, v) \;=\; \sum_{m,\,n} h(u+m, v+n) \qquad -r \leq (m, n) \leq r \tag{8}$$

and the peak with highest magnitude is chosen as the most probable direction of motion. This direction initializes a refinement method, described in the next section.

4.5 Refinement

The Hough transform quickly and efficiently provides a strong prior on the most likely direction candidate, but two problems remain. The first is accuracy, which is limited by the discrete representation of the transform space, and the second is uncertainty, about which the transform reveals little. To address these problems, the selected direction must be further refined by a continuous-space optimization.

Several authors (e.g. [23, 19, 13]) have proposed the use of EM algorithms for optimization with probabilistic correspondence. Such algorithms alternate between finding expected match likelihoods w_{ij} and estimating parameters (in this case, the baseline direction $\boldsymbol{d}$). Wells' technique [23] is not applicable in our context because it utilizes an asymmetric match function (i.e. matches images to a known template). The methods of Rangarajan [19] do not sample appropriately from the space of correspondence sets, instead relying exclusively on Sinkhorn's algorithm to provide the "correct" distribution on w_{ij}. Dellaert's method [13] samples correctly but makes the limiting assumptions that the number of 3-D features is known and that all features are viewed in all images.

Here we present an EM formulation that samples correctly from the distribution of all correspondence sets and handles occlusion and outlier features symmetrically for the two-camera case. The M-step consists of estimating $\boldsymbol{d}$ given the current weights w_{ij} according to (6). The E-step, in which the weights are determined given the current direction estimate, is performed using Monte Carlo sampling rather than an explicit analytic distribution, thus making the algorithm a so-called MCEM algorithm.

Sampling must be performed over the space of valid correspondence *sets* (i.e. valid $\boldsymbol{B}$ matrices) rather than individual correspondences, because of the inherent constraints mentioned in Section 4.3. Although this space is combinatorially large, nearly all correspondence sets have very low probability, suggesting that we need sample only in high-likelihood regions of the space. Dellaert proposes a Markov chain Monte Carlo (MCMC) sampler that defines a state as a valid binary correspondence set. The method proceeds as follows:

- Start in any valid state, represented by the matrix $\boldsymbol{B}_i$.
- Compute the likelihood of this state L_i.
- Randomly perturb the system to a new valid state $\boldsymbol{B}_{i+1}$.
- Compute the likelihood ratio $\beta = L_{i+1}/L_i$.
- If the likelihood has increased (i.e. $\beta > 1$), then keep the new state.
- Otherwise, keep the new state with probability β.

This process repeats for N iterations, until "steady state" is reached (typically N is on the order of 10,000 to 100,000). The average over states $\boldsymbol{B}_i$ kept at each iteration produces a valid probabilistic correspondence matrix $\boldsymbol{W}$:

$$\boldsymbol{W} \approx \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{B}_i \tag{9}$$

In order to reach the correct limiting probabilities, the state perturbations must satisfy *detailed balance*, meaning effectively that every valid state is reachable from every other valid state [6]. In the case where all features are available in all images and the number of valid features is known, row and column swaps of the current state $\boldsymbol{B}_i$ suffice, and the likelihood ratios β_{ij} can be computed efficiently due to the Gaussian error model assumed in [13].

However, in the case where outliers are present, we need a technique for introducing new valid matches and discarding current valid matches, as well as an efficient means of computing the β_{ij}. We thus propose two additional complementary state perturbations on the augmented match matrix $\tilde{\boldsymbol{B}}$ described in Section 4.3: the *split* operation converts a valid match into two outliers, while the *merge* operation joins two outliers into a valid match (Figure 12). The addition of these perturbations allows all possible states to be visited, and thus produces the correct steady state probabilities.

Running the MCEM algorithm on each adjacent camera pair produces a set of accurate motion directions, represented as unit vectors. As mentioned above, these directions specify only two of the three DOF in camera position. Aggregation of these vectors into a set of global constraints is thus necessary for determination of a consistent pose configuration.

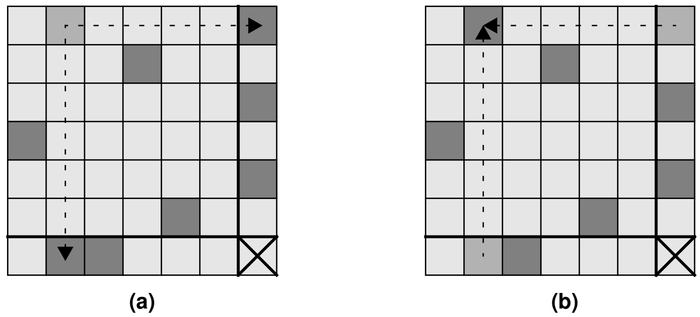


Figure 12: Split and Merge Perturbations

Besides row and column swaps, state perturbations consist of splits (a), in which a valid matrix entry is separated into two outliers, and merges (b), in which two outliers are joined into a valid entry.

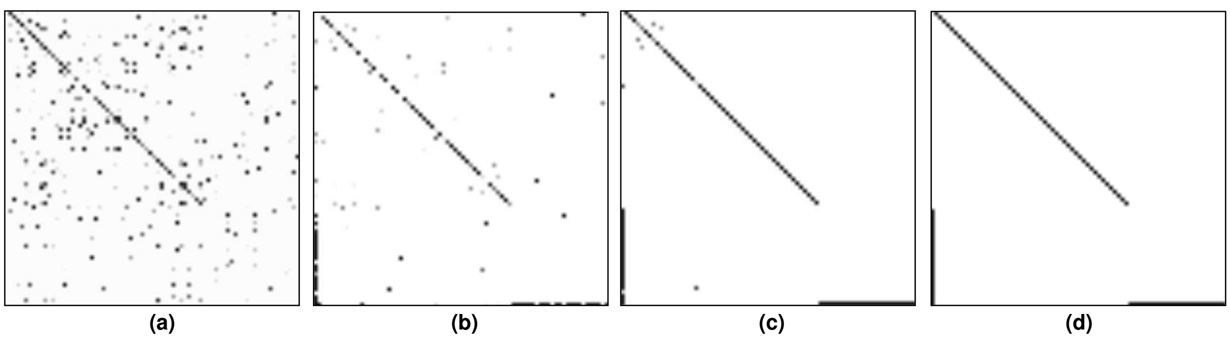


Figure 13: MCEM Convergence

Match matrices from selected iterations of MCEM run on synthetic data are shown above. (a) depicts the initially-estimated match probabilities, while (b) and (c) show intermediate results. The final deterministic matches are found at convergence and shown in (d).

5 Global Rectification

Assigning unit length to the translation vectors suffices for registration of isolated pairs, since the inherent scale ambiguity precludes determination of metric distances. However, a set of N nodes ($N > 2$) requires true relative distances (up to global isotropic scale) between all relevant camera pairs for a unique, consistent solution. The following sections describe a method for determining these distances, and also for expressing the resulting camera configuration in metric coordinates.

5.1 Adjacency

First, the set of all nodes of interest must be examined to obtain a set of pairs to be registered by the technique in Section 4. We use distances between node positions estimated by the data acquisition platform to determine node pairs that are proximal enough to view common scene geometry. We then select the k nearest neighbors of each camera (in practice k is 3–5), resulting in a new adjacency graph whose sites are the camera positions and whose arcs represent the pair-wise translation directions to be estimated.

5.2 Formulation

Let $\boldsymbol{p}_i$ be the 3-D position of node i, and let $\boldsymbol{d}_{ij} = -\boldsymbol{d}_{ji}$ be the unit direction of motion from node i to node j. Further, let $\alpha_{ij} = \alpha_{ji}$ represent the distance along $\boldsymbol{d}_{ij}$ between nodes i and j. Given estimates of $\boldsymbol{d}_{ij}$ and associated 3×3 covariance matrices $\boldsymbol{C}_{ij}$ (determined from the Bingham distributions of these estimates), we wish to determine a set of positions $\boldsymbol{p}_i$ and distances α_{ij} consistent with the $\boldsymbol{d}_{ij}$.

We thus formulate a set of linear vector equations of the form

$$\boldsymbol{p}_j = \boldsymbol{p}_i + \alpha_{ij}\boldsymbol{d}_{ij} \tag{10}$$

for all adjacent node pairs whose translation directions have been estimated. The above equation simply states that the position of node j is obtained by starting at node i and traveling a distance α_{ij} in the direction $\boldsymbol{d}_{ij}$. The covariance matrix $\boldsymbol{C}_{ij}$ can be used to weight the constraints while preserving their linearity:

$$\boldsymbol{C}_{ij}^{-1}(\boldsymbol{p}_i - \boldsymbol{p}_j + \alpha_{ij}\boldsymbol{d}_{ij}) = \boldsymbol{0}. \tag{11}$$

There are P such vector equations (one for each node pair registered), producing $3P$ scalar equations but introducing P additional unknowns α_{ij} for a total of $3N + P$ unknowns. From DOF counting alone, it would seem that a unique solution to this system requires that

$$P \geq \frac{3}{2}N. \tag{12}$$

However, the entire node configuration can be arbitrarily translated and scaled without altering the constraints. There are thus 4 inherent degrees of freedom in the system, regardless of the number of constraints of the form in (11).

In addition, even if an arbitrary global transformation is imposed on the system, determination of the minimal set of distinct pairs needed for a non-degenerate solution is significantly more complicated and depends entirely on the topology of the adjacency graph (i.e. which pairs are chosen and in what configuration they lie). A *sufficient* condition for unique solution is that the graph is fully triangulated; this ensures that the imposed isotropic scale constraint is propagated throughout the graph. This is not, however, a necessary condition, which is in general much more difficult to characterize.

In typical real-world data sets, the topology of the camera configuration is underconstrained. One way to ameliorate both the global transformation ambiguity and the degeneracies in graph topology is to utilize the camera positions initially estimated by the acquisition platform as weak constraints. This can be accomplished by appending an additional set of N linear vector equations of the form

$$\varepsilon(\boldsymbol{p}_i - \boldsymbol{a}_i) = \boldsymbol{0}, \tag{13}$$

where $\boldsymbol{a}_i$ is a constant denoting the initial position of node i, and ε is a very small scalar weight (say 10^{-4}), and a single scalar equation of the form

$$\sum_{i,j} \alpha_{ij} = \sum_{i,j} \|\boldsymbol{a}_j - \boldsymbol{a}_i\| . \tag{14}$$

The number of unknowns remains $3N + P$, but the number of constraints rises to $3P + 3N + 1$. Addition of the equations (13) provides constraint on singular or nearly singular modes of the equation system, but leaves the remaining modes unaffected; this effectively sets node positions to their initially estimated values in regions where these positions are otherwise indeterminate. The constraint in (14) prevents trivial zero solutions and also imposes an approximate global scale according to the initial configuration. The system can be solved by ordinary linear least-squares techniques.

5.3 Metric Registration

The configuration that results from the above formulation is expressed relative to a somewhat arbitrary coordinate frame. In fact any rigid Euclidean transformation (translation, rotation, and scale) applied to all cameras preserves self-consistency and thus also yields a valid pose configuration. The City Project requires that cameras be expressed in Earth-relative coordinates so that metric reconstruction (i.e. in the correct units and world positions) is possible; we thus wish to find the best rigid transformation to accomplish this task.

Assuming that the initial pose as estimated by the acquisition platform is unbiased, this problem amounts to an optimal 3-D to 3-D registration of the new configuration with the old configuration. One-to-one correspondence between the two sets of cameras is known, and the optimal transformation can be found using the technique of absolute orientation [16]. First, the new camera set is translated so that its center of mass is coincident with that of the original set. Next, the optimal rotation of the new set about its center of mass is computed and applied. Finally, the global isotropic scale factor that best rectifies distances from the centers of mass is estimated. Each step consists of simple algebraic operations; the result is a consistent camera configuration expressed in world coordinates.

6 Results

Several experiments were designed and run in order to assess the system's performance. Synthetic data with controllable levels of various noise sources was used for quantitative testing. Qualitative tests on a set of real hemispherical images compare the output of this system with that of the previously used semi-automated photogrammetric method.

6.1 Synthetic Data

A set of 50 synthetic nodes viewing 500 3-D features was generated with camera baselines of approximately 10 meters. Noise in initial camera positions was added with standard deviation of 3 meters. Feature location error, relative rotation error, and number of outliers (in the form of spurious 2-D features and 3-D point occlusions) were varied, and the resulting estimates of translation directions and scene-relative positions were compared to the true values.

Several plots are shown below. For a fixed variance in point feature localization error, pose was estimated in the presence of varying numbers of outliers (Figure 14a). For a fixed number of outliers, pose was estimated with varying point feature localization error (Figure 14b). Finally, fixing both the number of outliers and feature error, we estimated the camera positions with varying rotational pose error (Figure 14c).

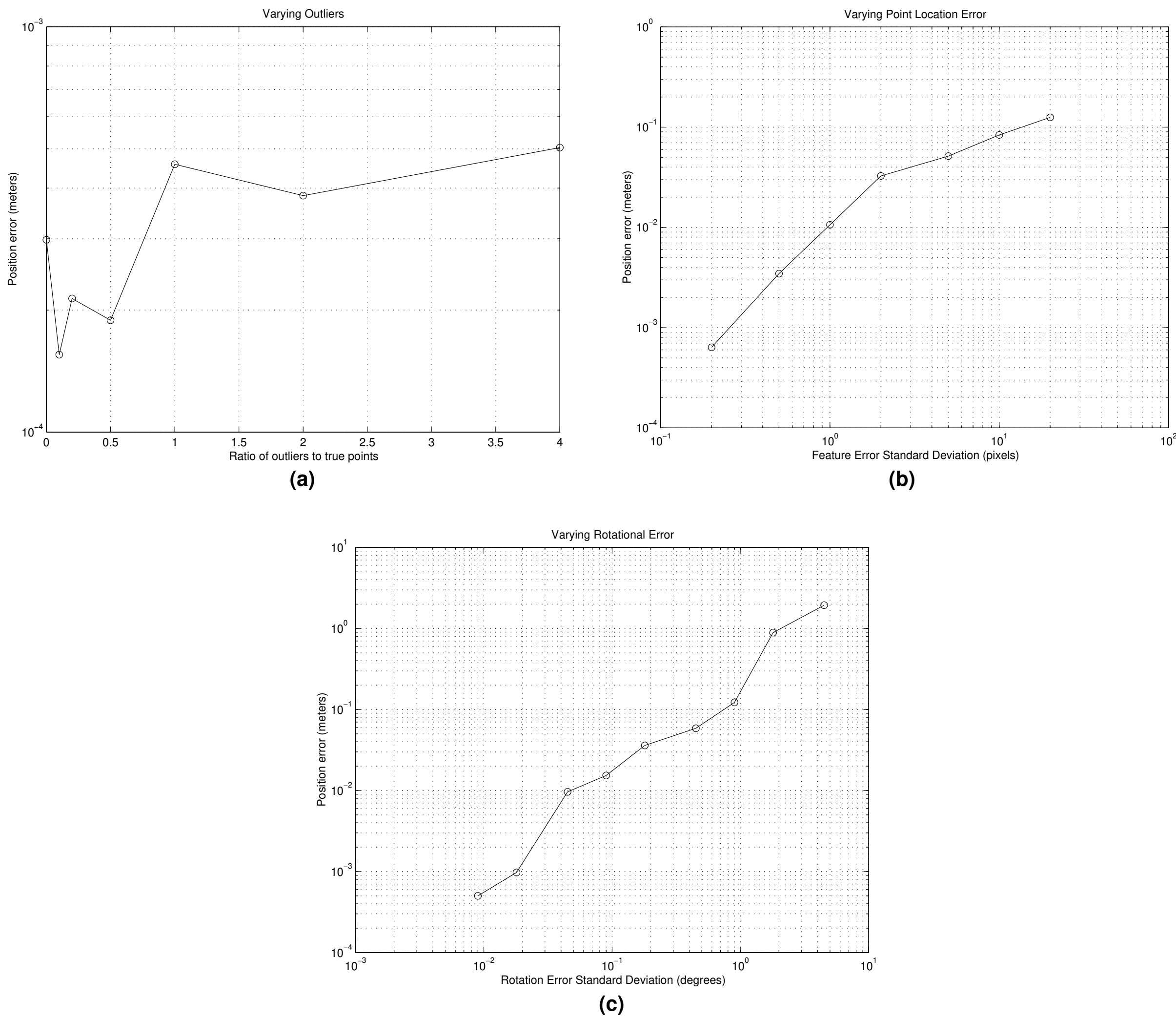


Figure 14: Results on Simulation

The plot in (a) was generated by varying the number of outliers present in a sample of 500 feature points per camera; the true features were perturbed by random noise with standard deviation of 1 pixel. In (b), the feature noise standard deviation was varied, with outlier percentage fixed at 10%. The plot in (c) shows position error as a function of noise in estimated rotational pose.

Qualitatively, we have found that the system to be quite robust against outlier point features. When relative rotations are accurately known and point projection error is small, a 4:1 ratio of outliers to true data points only slightly increases the error in final camera positions. Point projection error standard deviations of roughly 5 pixels (typical of City Project data) produced positions accurate to within 5cm.

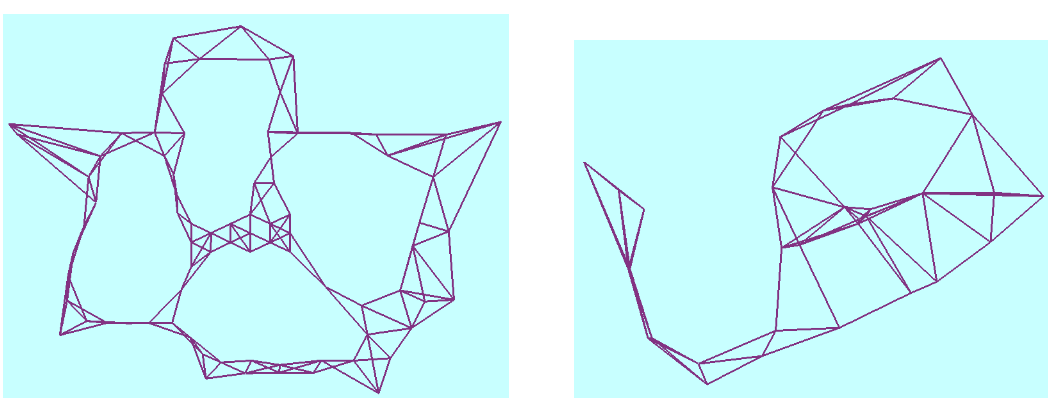
6.2 Real Data

Cameras from two real rotationally-registered data sets were registered using these techniques. The first set (TechSquare) consisted of 75 nodes, which were automatically aligned and compared with cameras registered by manual feature correspondence. Selected epipolar geometry was compared for several node pairs. The second set (GreenBuilding) consisted of 30 nodes. There was significant error in the initial pose estimates (up to six meters and twenty degrees of relative pose misalignment), which was corrected by this technique.

7 Conclusions

We have described a robust method for globally-consistent translational registration of a large set of images. The method assumes internal calibration, knowledge of rotational pose, and initial estimates of position. It is fully automated and overcomes some limitations of traditional feature correspondence techniques. Thus far, testing has shown that our method finds accurate translation directions between cameras and produces consistent global pose configurations. For real data sets consisting of thousands of images acquired with 5–10 meter baselines over regions hundreds of meters across, this method achieves end-to-end accuracy in position to within 5 centimeters. Because of the Hough transform technique, the system is virtually insensitive to point feature outliers, though as one would expect, reliability suffers drastically as the error in supplied relative rotation becomes significant.

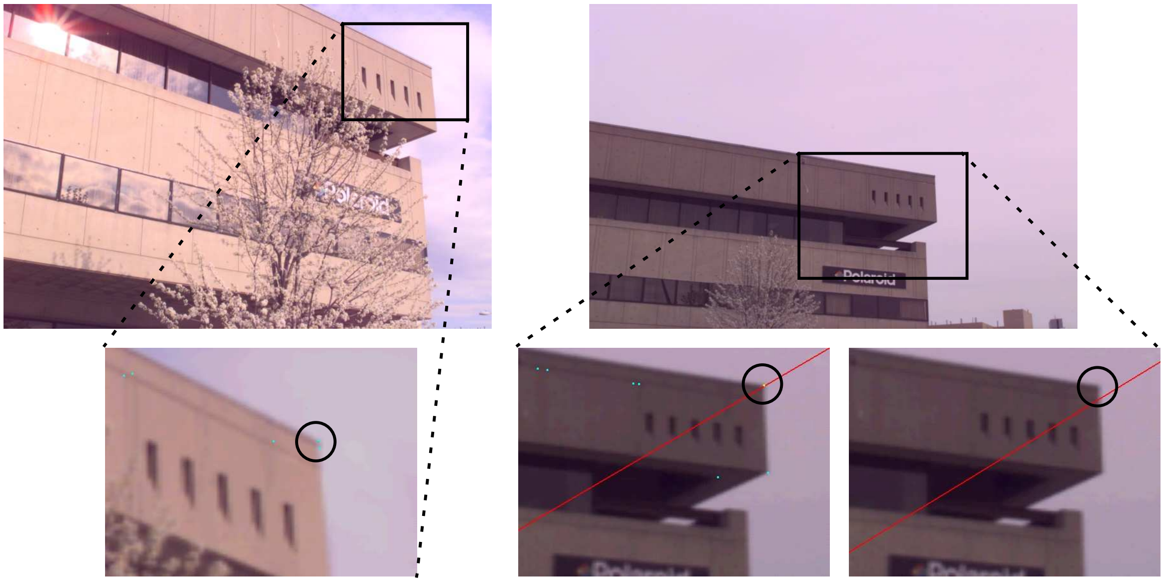
Computations for baseline estimation are $O(MF^2)$ in the number of nodes M and the number of features per node F. This is ameliorated by geometric constraints and match culling. In practice, F was typically on the order of a thousand, and translation direction estimation for real images running on a 250 MHz SGI O_2 required an average of about 55 seconds of computation per pair. Global rectification involves solving a somewhat large linear system, but sparse matrix techniques can be used to dramatically improve performance.



(a) (a)

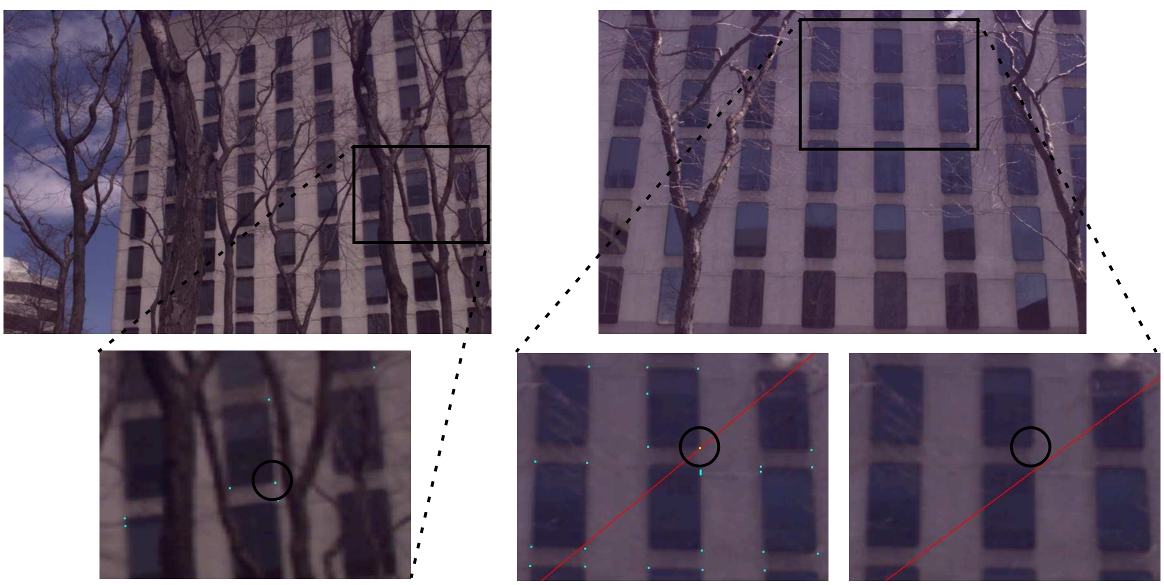
Figure 15: Node Configurations

Initial configurations for two data sets are viewed from above, with lines representing adjacency. A set of 81 nodes is shown in (a), and a set of 30 nodes acquired near a tall building is shown in (b).



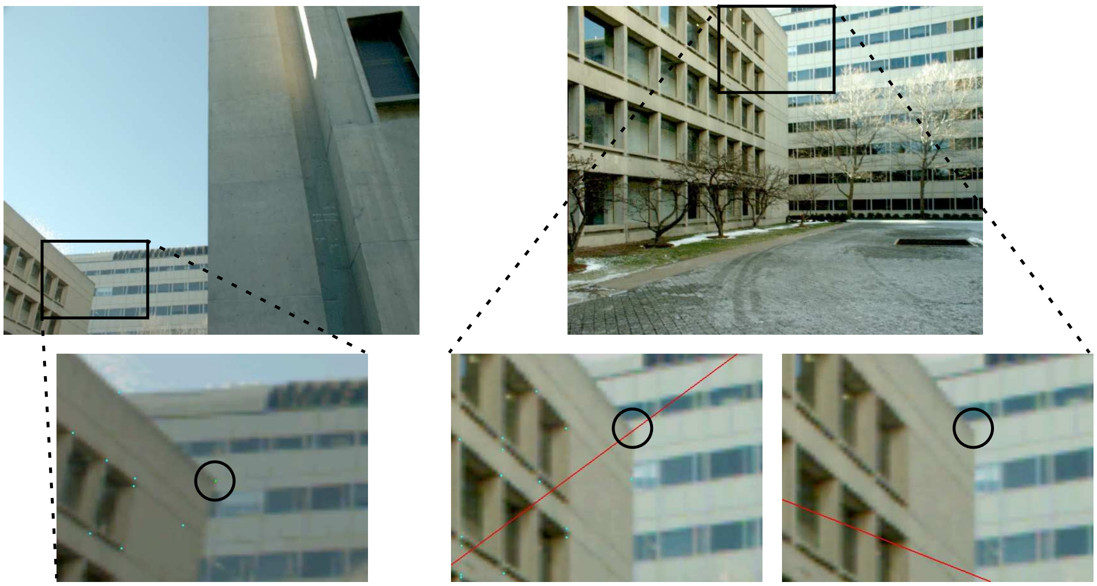
(a) (b) (c)

Figure 16: TechSquare Epipolar Geometry I
A building corner and close-up is shown in (a). The epipolar line corresponding to this corner as seen from a different viewpoint is shown in (b) using automatically corrected cameras, and compared to (c) using cameras generated by manual feature correspondence.



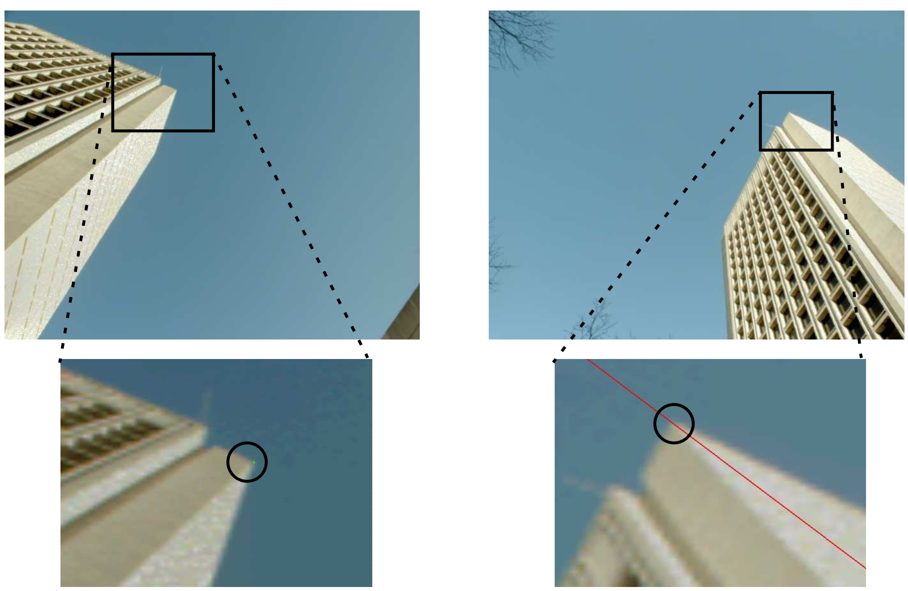
(a) (b) (c)

Figure 17: TechSquare Epipolar Geometry II
Window corners and close-ups shown from two different viewpoints. Ambiguities arising from regular geometry can make manual correspondence difficult. A particular window corner is shown in the original view (a) and in the second view using cameras generated by (b) automatic registration and (c) manual feature correspondence.



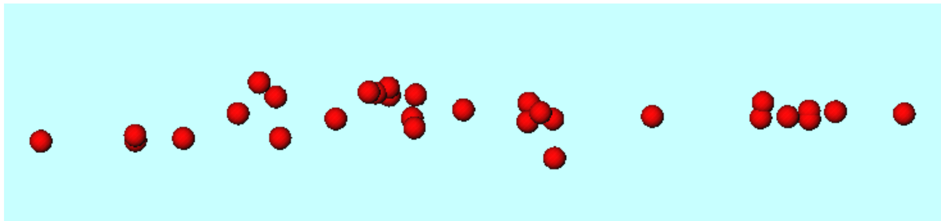
(a) **(b)** **(c)**

Figure 18: GreenBuilding Epipolar Geometry I
A building corner is shown in (a), and its epipolar line in another image as viewed by automatically registered cameras (b) and the initially acquired pose (b).

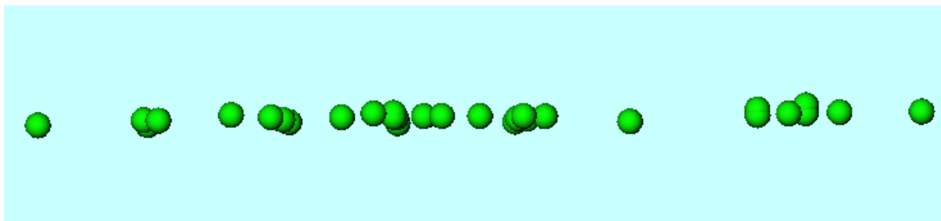


(a) **(b)**

Figure 19: GreenBuilding Epipolar Geometry II
A distant building corner (a) and its corresponding epipolar line as viewed by automatically registered cameras (b).



(a)



(b)

Figure 20: GreenBuilding Pose Correction
Pose configurations for terrestrial imagery should be approximately planar. The camera configuration is viewed from the side, before pose correction in (a) and after in (b). Cameras were moved by an average of 2.86 meters.

References

[1] Antone, M. E. and Teller, S. "Automatic Recovery of Relative Camera Rotations for Urban Scenes". In *Proceedings of CVPR*, June 2000, Vol. 2, pp. 282-289.

[2] Azarbayejani, A. and Pentland, A. "Recursive Estimation of Motion, Structure, and Focal Length". *PAMI*, Vol. 17, No. 6, June 1995, pp. 562-575.

[3] Becker, S. and Bove, V. M. "Semi-Automatic 3-D Model Extraction from Uncalibrated 2-D Camera Views". In *Proceedings of SPIE Image Synthesis*, 1995, pp. 447-461.

[4] Belhumeur, P. N., Kriegman, D. J., and Yuille, A. L. "The Bas-Relief Ambiguity". In *Proceedings of CVPR,* 1997, pp. 1060-1066.

[5] Bingham, C. "An Antipodally Symmetric Distribution on the Sphere". *The Annals of Statistics*, Vol. 2, 1974, pp. 1201-1225.

[6] Bishop, C. *Neural Networks for Pattern Recognition*. Clarendon Press, Oxford, 1995.

[7] Bosse, M. C. and Teller, S. "A High-Resolution Geo-Referenced Pose Camera". Manuscript, 2000.

[8] Canny, J. F. "A Computational Approach to Edge Detection". *PAMI*, Vol. 8, No. 6, November 1986, pp. 679-698.

[9] Chui, H. and Rangarajan, A. "A New Algorithm for Non-Rigid Point Matching". In *Proceedings of CVPR*, June 2000, Vol. 2, pp. 40-51.

[10] Collins, R. T. "Model Acquisition using Stochastic Projective Geometry". PhD thesis, University of Massachusetts, 1993.

[11] Coorg, S., Master, N., and Teller, S. "Acquisition of a Large Pose-Mosaic Dataset". In *Proceedings of CVPR*, 1998, pp. 872-878.

[12] Debevec, P. E., Taylor, C. J., and Malik, J. "Modeling and rendering architecture from photographs: A hybrid geometry- and image-based approach". In *Proceedings of SIGGRAPH*, 1996, pp. 11-20.

[13] Dellaert, F., Seitz, S. W., Thorpe, C. E., and Thrun, S. "Structure from Motion without Correspondence". In *Proceedings of CVPR*, Vol. 2, June 2000, pp. 557-564.

[14] Faugeras, O. *Three-Dimensional Computer Vision: A Geometric Viewpoint*. MIT Press, Cambridge, Massachusetts, 1996.

[15] Grimson, W. E. L. and Huttenlocher, D. P. "On the Sensitivity of the Hough Transform for Object Recognition". *PAMI*, Vol. 12, No. 3, March 1990, pp. 255-274.

[16] Horn, B. K. P. "Closed-Form Solution of Absolute Orientation using Unit Quaternions". *Journal of the Optical Society of America* Vol. A, No. 4, 1987, pp. 629-642.

[17] Liebowitz, D. and Zisserman, A. "Metric Rectification for Perspective Images of Planes". In *Proceedings of CVPR*, 1998, pp. 482-488.

[18] Luong, Q. T. and Faugeras, O. "Camera Calibration, Scene Motion, and Structure Recovery from Point Correspondences and Fundamental Matrices". *IJCV*, Vol. 22, No. 3, 1997, pp. 261-289.

[19] Rangarajan, A., Chui, H., and Duncan, J. S. "Rigid Point Feature Matching using Mutual Information". *Medical Image Analysis*, Vol. 3, No. 4, 1999, pp. 425-440.

[20] Shum H. Y., Han, M., and Szeliski, R. "Interactive Construction of 3D Models from Panoramic Image Mosaics". In *Proceedings of CVPR*, 1998, pp. 427-433.

[21] Sinkhorn, R. "A Relationship Between Arbitrary Positive Matrices and Doubly Stochastic Matrices". *Annals of Mathematical Statistics*, Vol. 35, No. 2, June 1964, pp. 876-879.

[22] Teller, S. "Automated Urban Model Acquisition: Project Rationale and Status". In *Proceedings of the Image Understanding Workshop*, 1998.

[23] Wells, W. "Statistical Approaches to Feature-Based Object Recognition". *IJCV*, Vol. 21, No. 1/2, January 1997, pp. 63-98.