Scalable, Absolute Position Recovery for Omni-Directional Image Networks

Matthew Antone Seth Teller
MIT Computer Graphics Group

Abstract

We describe a linear-time algorithm that recovers absolute camera positions for networks of thousands of terrestrial images spanning hundreds of meters, in outdoor urban scenes, under uncontrolled lighting. The algorithm requires no human input or interaction. For real data, it recovers camera pose globally consistent on average to roughly five centimeters, or about four pixels of epipolar alignment.

This paper's principal contributions include an extension of Markov chain Monte Carlo estimation techniques to the case of unknown numbers of feature points, unknown occlusion and deocclusion, large scale (thousands of images, and hundreds of thousands of point features), and large dimensional extent (tens of meters of inter-camera baseline, and hundreds of meters of baseline overall). Also, a principled method is given to manage uncertainty on the sphere; a new use of the Hough transform is proposed; and a method for aggregating local baseline constraints into a globally consistent pose set is described.

1. Introduction

Extrinsically calibrated imagery is of fundamental interest in a variety of computer vision and graphics applications, including sensor fusion, 3-D reconstruction for model capture, and image-based rendering. In practice, registering imagery can require substantial manual effort—for example, matching common points across multiple images.

We have developed two camera registration algorithms as part of a system for automated model capture in extended urban environments [25, 12]. In our system, a human operator moves a sensor [7] to many viewing positions in and around the scene of interest. The sensor acquires images annotated with a rough estimate of the acquiring camera's position and orientation in absolute (Earth) coordinates.

Images are grouped by optical center into single, wide-FOV mosaics called "nodes" [12]. Each node is subsequently treated as a rigid, super-hemispherical image with a single pose. The sensor's initial pose estimates are accurate only to several meters and several degrees; thus, one critical component of our system is the refinement of these estimates to bring all cameras into pixel-accurate registration. The scale of the dataset rules out interactive techniques, so pose recovery must be fully automated. Solving the general registration problem requires determination of six parameters for each camera—three of rotation, and three of position. Our approach decouples the 6-DOF problem into pure rotation (3-DOF) and pure translation (3-DOF) components. This paper assumes rotationally registered images as inputs [3], and addresses only position recovery.

We make use of a number of existing techniques from computer vision and estimation theory, including: the use of gradient-based features for robustness against lighting variations and strong perspective; probabilistic inference; the Hough transform, for robust and efficient initialization; Markov chain Monte Carlo (MCMC) for sampling high-dimensional spaces; and expectation maximization (EM) methods, for solution of coupled classification and estimation problems.

1.1. Algorithm Overview

The goal of our algorithm is to accurately register every camera (node) to a single, common coordinate system. Our approach exploits the fact that adjacent (nearby) nodes tend to observe overlapping scene structure, namely point features. The algorithm first detects shared structure across pairs of adjacent nodes, estimating a local displacement relating each pair. These local constraints are then propagated throughout the node graph to assign a globally consistent position to each node.

1.2. Input Requirements and Assumptions

Our algorithm requires the following inputs:

- ***Intrinsic calibration.*** Radial distortion has been corrected, and pinhole camera parameters (focal length, principal point, skew) are supplied.
- ***Accurate extrinsic orientations.*** Scene-relative orientations and vanishing point directions are supplied for each node [3].
- ***Rough camera locations.*** Absolute (GPS-based) position estimates for each node are supplied.
- ***Camera adjacency.*** A list of the neighbors of each node is supplied.
- ***Point features.*** Sub-pixel point features, produced by intersecting pairs of image lines, are supplied for each image.

0-7695-1272-0/01 $10.00 © 2001 IEEE

1.3. Paper Overview

The paper is structured as follows. Section 2 reviews projective feature representations and geometric probability. Section 3 describes the position recovery algorithm, and Section 4 reports experimental results. Section 5 reviews related work, and Section 6 summarizes our contributions and results.

2. Preliminaries

A rigid transformation, consisting of a 3-D translation $\boldsymbol{t}$ and orthonormal rotation $\boldsymbol{R}$, expresses points $\boldsymbol{p}^w$ in world space as points $\boldsymbol{p}^c$ in camera space. Formally,

$$\boldsymbol{p}^c = \boldsymbol{R}^\top(\boldsymbol{p}^w - \boldsymbol{t}); \qquad \boldsymbol{p}^w = \boldsymbol{R}\boldsymbol{p}^c + \boldsymbol{t}.$$

Our algorithm assumes that all rotation matrices $\boldsymbol{R}$ are known, and recovers a Gaussian distribution on $\boldsymbol{t}$ for each camera.

2.1. Projective Points

Image points can be represented as planar coordinate pairs (u, v); however, because the field of view in our images is large, we represent point features uniformly as projective rays, or unit vectors on $\mathbb{S}^2$, the *Gaussian sphere*.

2.2. Bingham's Distribution

Features viewed by a single camera are inherently projective, since no depth information is available, and noisy; we wish to represent such features appropriately. Exponential distributions are useful for many inference tasks, but the commonly used Gaussian density is a Euclidean probability measure unsuitable for projective variables. We represent projective image features (points and lines) on $\mathbb{S}^2$ using *Bingham's distribution* [6], a zero-mean Gaussian variable $\boldsymbol{x} \in \mathcal{R}^3$ conditioned on $\|\boldsymbol{x}\| = 1$.

This distribution is parameterized by a symmetric 3×3 matrix $\boldsymbol{M}$, analogous to the information matrix of a zero-mean Gaussian distribution, that encodes a feature's "orientation" (i.e. ray and line directions) and "shape" (i.e. feature type and degree of uncertainty):

$$p(\boldsymbol{x}) = c\exp(\boldsymbol{x}^\top \boldsymbol{M}\boldsymbol{x})$$

where c is a normalizing coefficient that depends only on the shape parameters.

3. Position Recovery Algorithm

Figure 1 depicts the position recovery algorithm. First, translation direction (*baseline*) estimates are initialized for all node adjacencies using a Hough transform. An EM technique refines the baseline by averaging over all possible point feature correspondence sets. Once a baseline estimate has been found for each adjacency, a global optimization step finds the camera positions most consistent with the baselines. Finally, a rigid 3-D to 3-D registration aligns the node set to the raw (initial) pose estimates.

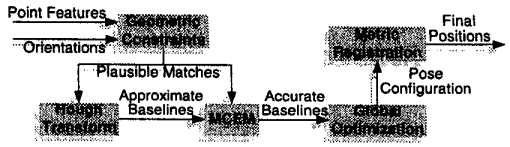

Figure 1: Position Recovery Overview

3.1. Two-Camera Translation Geometry

Given rotationally registered cameras $\mathcal{A}$ and $\mathcal{B}$, and respective point feature sets $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ and $\mathcal{Y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_M\}$, we wish to determine the direction of motion $\boldsymbol{b}$ from $\mathcal{A}$ to $\mathcal{B}$ most consistent with the available data.

3.1.1. Epipolar Geometry Under Known Orientation

An epipolar plane $\mathcal{P}$ contains two camera centers and a 3-D point seen by both cameras. Projections of the 3-D point onto each of the images, $\boldsymbol{x}$ and $\boldsymbol{y}$ respectively, must therefore also lie in $\mathcal{P}$ (Figure 2).

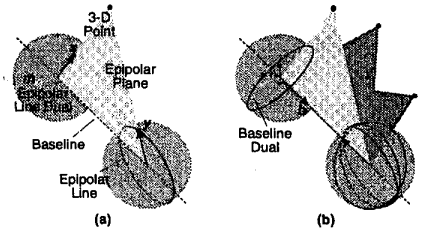

Figure 2: Epipolar Geometry
(a) A single 3-D point lies in an epipolar plane containing the baseline and any projective observations of the point. (b) The epipolar planes induced by a set of 3-D points forms a pencil coincident with the baseline. The normals of these planes thus lie on a great circle orthogonal to the baseline direction.

For rotationally registered cameras, it holds that

$$(\boldsymbol{x} \times \boldsymbol{y}) \cdot \boldsymbol{b} = 0. \tag{1}$$

Define $\boldsymbol{m}_{ij} \equiv \boldsymbol{x}_i \times \boldsymbol{y}_j$. For the correct pairs of i and j—that is, for those (i, j) couplets in which feature $\boldsymbol{x}_i$ truly matches feature $\boldsymbol{y}_j$—the constraint (1) becomes $\boldsymbol{m}_{ij} \cdot \boldsymbol{b} = 0$. If the $\boldsymbol{m}_{ij}$ are viewed as projective epipolar lines, then the baseline $\boldsymbol{b}$ can be viewed as a projective *focus of expansion*, and its antipode the *focus of contraction*, the intersections of all epipolar lines arising from true matches.

3.1.2. Geometric Constraints on Correspondence

The correspondence and baseline are both initially unknown, severely under-constraining the above construction. For images with M and N features, there are MN possible individual feature matches, and $\mathcal{O}((MN)!)$ possible correspondence sets, making the search space enormous. Its dimension can be reduced by eliminating candidate correspondences (Figures 3 and 4).

The 3-D line directions and 2-D line classifications obtained from rotational pose recovery provide strong cues for

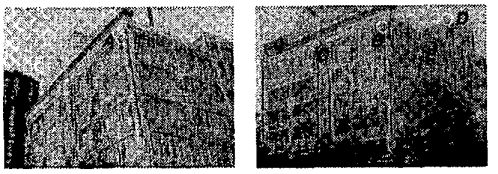

Figure 3: Line Constraints
Two images, and possible matches for a point feature A in the first image. Point B is the true match; C and D are also plausible because they are formed by lines whose directions match those of the lines forming A. Point E is rejected, since its line directions do not match those of A.

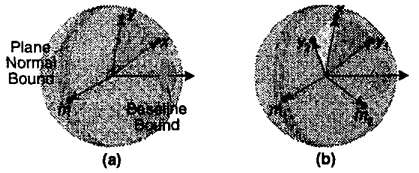

Figure 4: Direction Constraints
(a) Baseline uncertainty induces an equatorial band of uncertainty for epipolar lines. The match between features $\boldsymbol{x}$ and $\boldsymbol{y}$ is plausible because it implies motion in the correct direction. (b) The match between $\boldsymbol{x}$ and $\boldsymbol{y}_1$ is rejected because it implies backward motion; the match with $\boldsymbol{y}_2$ is rejected because its epipolar line does not lie in the uncertainty band.

feature culling and point correspondence rejection. Each candidate correspondence is discarded unless:

- Directions (VPs) of constituent lines match.
- Epipolar line falls within uncertainty bound.
- Inferred depth of 3-D point falls in reasonable range.

3.2. Estimation of Translation Direction

If true correspondence between the feature sets $\mathcal{X}$ and $\mathcal{Y}$ is known, then in the absence of measurement noise, a set of constraints (1) can be used to estimate $\boldsymbol{b}$ by minimizing the objective function

$$E = \sum_{(i,j)\in\mathcal{F}} (\boldsymbol{m}_{ij} \cdot \boldsymbol{b})^2.$$

Here, $\mathcal{F}$ is the set of F pairings (i,j) that represent the true matches. The optimal $\boldsymbol{b}$ is the eigenvector associated with the minimum eigenvalue of the matrix

$$\boldsymbol{A} = \sum_{(i,j)\in\mathcal{F}} \boldsymbol{m}_{ij}\boldsymbol{m}_{ij}^{\top}.$$

In reality, every point feature is a noisy observation arising from the intersection of two noisy image lines, each represented by a Bingham variable. The Bingham uncertainty of the intersection can be determined by fusing the two lines. Similarly, correspondence between point features $\boldsymbol{x}_i$ and $\boldsymbol{y}_j$ induces an uncertain epipolar line $\boldsymbol{m}_{ij}$. Estimation of the baseline must therefore incorporate the uncertainty of the $\boldsymbol{m}_{ij}$, and produce a distribution on $\boldsymbol{b}$.

3.2.1. Baseline Inference from Noisy Data

The Bingham parameters $\boldsymbol{M}_{\boldsymbol{b}}$ of the baseline distribution can be inferred by projective fusion as

$$\boldsymbol{M}_{\boldsymbol{b}} = \sum_{(i,j)\in\mathcal{F}} \boldsymbol{M}_{ij} + \boldsymbol{M}_0$$

where $\boldsymbol{M}_{ij}$ represents the uncertainty of the epipolar line $\boldsymbol{m}_{ij}$, $\boldsymbol{M}_0$ is the prior distribution on $\boldsymbol{b}$, and the sum is taken only over indices associated with true matches. Equivalently, inference can be performed by associating a binary-valued variable b_{ij} with *every possible* correspondence:

$$b_{ij} = \begin{cases} 1, & \text{if } \boldsymbol{x}_i \text{ matches } \boldsymbol{y}_j \\ 0, & \text{otherwise.} \end{cases}$$

The Bingham parameters of $\boldsymbol{b}$ are then determined by

$$\boldsymbol{M}_{\boldsymbol{b}} = \sum_{i=1}^{M}\sum_{j=1}^{N} b_{ij}\boldsymbol{M}_{ij} + \boldsymbol{M}_0,$$

where summation occurs over every possible (i,j) pairing.

Due to feature noise and occlusion, explicit correspondence cannot always be determined. Thus, we employ continuous variables $w_{ij} \in [0,1]$ to represent the *likelihood*, or probability, that feature $\boldsymbol{x}_i$ matches feature $\boldsymbol{y}_j$. In this weighted formulation, inference of $\boldsymbol{b}$ becomes

$$\boldsymbol{M}_{\boldsymbol{b}} = \sum_{i=1}^{M}\sum_{j=1}^{N} w_{ij}\boldsymbol{M}_{ij} + \boldsymbol{M}_0.$$

The b_{ij} represent the deterministic limit of the w_{ij}.

3.2.2. Feature Match Weights

A feature observed in one image has at most one true match in the other image. A true match exists only if the observation corresponds to a real 3-D point, and if its counterpart in the other image is visible and detected. Formally, we write

$$\sum_{k=0}^{N} b_{ik} = \sum_{k=0}^{M} b_{kj} = 1 \quad \begin{matrix} i \in [1,M] \\ j \in [1,N] \end{matrix},$$

where b_{i0} and b_{0j} are binary *slack variables* [10], taking value one if $\boldsymbol{x}_i$ (resp., $\boldsymbol{y}_j$) matches no feature, and zero otherwise. In the probabilistic case, we write analogously

$$\sum_{k=0}^{N} w_{ik} = \sum_{k=0}^{M} w_{kj} = 1 \quad \begin{matrix} i \in [1,M] \\ j \in [1,N] \end{matrix}. \qquad (2)$$

This condition enforces a symmetric (two-way) constraint over all correspondences: each feature in either image can match a set of possible features in the other image with some probability. The weights can be represented by an $(M+1)\times(N+1)$ matrix $\boldsymbol{W}$ (or $\boldsymbol{B}$, in the binary case), whose rows represent the features $\mathcal{X}$, whose columns represent the features $\mathcal{Y}$, and whose individual entries are the match weights (Figure 5). The condition in (2) states that the weight matrix be *doubly stochastic* [23], i.e. that both its rows and its columns sum to one.

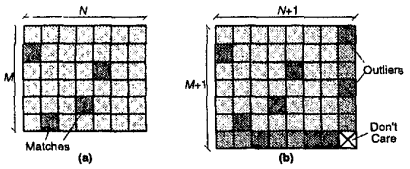

Figure 5: Match Matrix
The match matrix encodes pairwise feature correspondences. (a) A binary match matrix. (b) The matrix augmented with slack variables.

3.2.3. Initialization: Obtaining a Prior Distribution

Motion direction and correspondence are tightly coupled and under-constrained. However, we can use initial pose and match constraints to approximate $\boldsymbol{b}$ *without* explicit correspondence.

Let $\mathcal{M}$ represent the set of *all* plausible correspondences (epipolar lines) between $\mathcal{X}$ and $\mathcal{Y}$, and let the subset $\mathcal{M}' \subset \mathcal{M}$ contain only the F true matches. If all lines in $\mathcal{M}$ are drawn on $\mathbb{S}^2$ (Figure 6), those in $\mathcal{M}'$ (in the absence of noise) will intersect at the motion direction $\boldsymbol{b}$. The remainder, representing false matches, will intersect at uncorrelated locations. The point of maximum incidence on $\mathbb{S}^2$ is the most likely direction of motion; this point can be located by discretizing $\mathbb{S}^2$ and accumulating all (weighted) candidate epipolar lines $\mathcal{M}$ in a Hough transform.

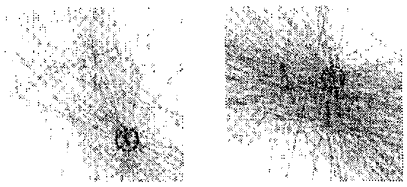

Figure 6: Hough Transform for Baseline Estimation
Two examples of Hough transforms for baseline estimation. Epipolar lines for all plausible matches are accumulated; the transform peak represents the baseline direction.

Because there are MN possible matches and only F (at most $\min(M, N)$) true matches, the true baseline peak may be obscured by spurious peaks. For example, a point feature lying close to the motion direction can plausibly match many features in the other image (Figure 7). Enforcing (2) through iterative normalization [23] dramatically improves the coherence of the true motion direction.

Although the discrete Hough transform has limited accuracy, the resulting motion direction estimate $\boldsymbol{b}_0$ can be used as a strong prior (with parameters $\boldsymbol{M}_0$ in the notation of Section 3.2) for subsequent inference.

3.3. Monte Carlo Expectation Maximization

This section describes refinement of the baseline estimate solely from observed point features, *without* explicit correspondence, by employment of an EM algorithm in which the posterior distribution is discretely sampled. Using maximum likelihood notation,

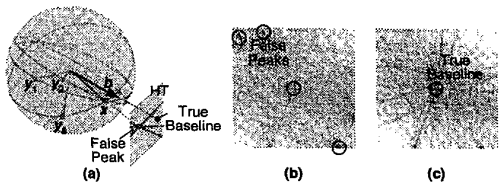

Figure 7: False Hough Transform Peaks
(a) False peaks in the Hough transform can be caused by features too close to the direction of motion, which have many matches. (b) An example. (c) The same example after Sinkhorn normalization.

$$\boldsymbol{b}^* = \text{argmax}_{\boldsymbol{b}}[p(\boldsymbol{b}|\mathcal{M})]$$

Bayes' rule yields

$$\begin{aligned} p(\boldsymbol{b}|\mathcal{M}) &= \sum_{\boldsymbol{B}} p(\boldsymbol{b}, \boldsymbol{B}|\mathcal{M}) \\ &= \sum_{\boldsymbol{B}} p(\boldsymbol{b}|\boldsymbol{B}, \mathcal{M}) p(\boldsymbol{B}|\mathcal{M}) \end{aligned}$$

where $\boldsymbol{B}$ is a valid binary correspondence matrix, and $p(\boldsymbol{B}|\mathcal{M})$ is the prior distribution on the correspondence set. This prior can be uniform, or can incorporate geometric match constraints like those in Section 3.1.2.

3.3.1. Structure from Motion without Correspondence

An optimal estimate for $\boldsymbol{b}$ can be found by maximizing $p(\boldsymbol{b}|\mathcal{M})$, treating correspondence sets as nuisance parameters in a Bayesian formulation, and evaluating over all possible matrices $\boldsymbol{B}$ [14]. We use an EM algorithm, alternating between an M-step, in which a log likelihood function is maximized to estimate $\boldsymbol{b}$ given a distribution on $\boldsymbol{B}$, and an E-step, in which this distribution is estimated given the current baseline. Convergence on the global optimum is virtually guaranteed by the Hough transform initialization.

We relate continuous to binary weights by defining w_{ij} as the marginal probability of match b_{ij} regardless of the other matches; that is,

$$w_{ij} \equiv p(b_{ij} = 1|\boldsymbol{b}, \mathcal{M}) = \sum_{\boldsymbol{B}} \delta(i, j) p(\boldsymbol{B}|\boldsymbol{b}, \mathcal{M});$$

The next sections describe a method for efficiently evaluating the w_{ij} by Monte Carlo sampling.

3.3.2. Sampling the Posterior Distribution

In the MCMC formulation, each possible binary match matrix $\boldsymbol{B}$ represents a distinct state; random transitions between states occur according to the Metropolis criterion [21] until steady state is reached. When the transition likelihoods are appropriately chosen, the steady-state probabilities represent the distribution on correspondence matrices

B. In practice, all visited B^k (where k is a transition index) are averaged to obtain a weight matrix W, which by construction meets the criteria of (2).

The MCMC algorithm requires a valid starting state and random state perturbations that satisfy detailed balance, meaning that every valid state is reachable from every other valid state. Suitable perturbations must thus be defined.

3.3.3. Match Perturbations

Dellaert's method [14] treats the case in which all features are visible in all images, and thus only allows swap perturbations: B^{k+1} is identical to B^k except for a single row (or, equivalently, column) swap. Since simple swapping preserves the number of matches, states with greater or fewer matches than the current state are never reached.

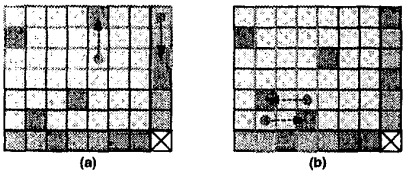

Figure 8: Row and Column Swaps
(a) Two rows of the match matrix, including outliers, are interchanged. (b) Two columns are interchanged.

We generalize Dellaert's technique, in the two-camera case, to handle an unknown number of visible 3-D features, and also to handle outliers and occlusion, by utilizing slack variables and by adding two complementary operations (Figure 9). The *split* perturbation converts a match into two unmatched features; the *merge* perturbation joins two unmatched features into one match.

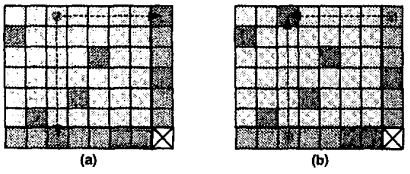

Figure 9: Split and Merge Perturbations
(a) A valid match is split into two outliers, reducing the number of matches by one. (b) Two outliers are merged into a match, increasing the number of matches by one.

3.4. Multi-Camera Method

The baseline estimation method recovers camera motion only up to an unknown scale factor for each baseline. This section describes a method for assembling the pair-wise baseline directions and recovering a globally consistent position for each node.

3.4.1. Baseline Constraints

Only the *directions* (not distances) between adjacent nodes and rough initial camera positions are known. We employ an iterative algorithm that updates each node's position p_i using constraints imposed by its associated baselines.

At each iteration, the list of all nodes is traversed in random order. For a given node i, a set of constraints is assembled by constructing rays originating at the current positions p_j of its neighbor nodes and emanating in the direction of the estimated baselines b_{ji} (Figure 10). The new position p_i' for node i is chosen to minimize the mean-square distance to each baseline ray. In the absence of baseline uncertainty,

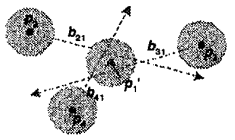

Figure 10: Single Node Baseline Constraints
A node's position is constrained by adjacent positions and baselines.

p_i' can be determined according to

$$p_i' = \left(\sum_j (I - b_{ji} b_{ji}^\top)\right)^{-1} \left(\sum_j (I - b_{ji} b_{ji}^\top) p_j\right). \quad (3)$$

Uncertainty in baseline directions can be incorporated by replacing $b_{ji} b_{ji}^\top$ in (3) with the second-moment matrix of the baseline's Bingham density. Uncertainty in p_i', in the form of a 3×3 Euclidean covariance matrix, is approximated by the inverse matrix in (3).

3.4.2. Metric Registration

A final registration phase finds the optimal rigid translation, rotation, and isotropic scaling [18] that align the refined node positions to the initial (Earth-relative) node positions (Figure 11). This phase also transforms each node's uncertainty estimates into metric Earth coordinates so that end-to-end confidences can be assessed.

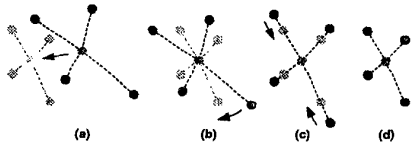

Figure 11: Metric Registration Process
(a) The original configuration is shifted so that the two centroids coincide. (b) Rays from the centroid to each camera are rotationally aligned. (c) The optimal scale is computed and applied. (d) The final configuration.

3.5. Asymptotic Running Time

The number of adjacencies in the node graph is linear in the number of nodes n, since every node has at most a constant number of neighbors. The algorithm's pairwise baseline estimation stage requires $\mathcal{O}(f^2)$ time per adjacency, where f is the maximum number of point features in the two nodes involved; in practice we bound f by a constant. Thus the pairwise stage requires $\mathcal{O}(n)$ time.

The global algorithm of Section 3.4.1 runs for at most a constant number of iterations. At each iteration, n new

node positions are computed, each in $\mathcal{O}(1)$ time. The final metric alignment runs in $\mathcal{O}(n)$ time; thus, the end-to-end algorithm requires $\mathcal{O}(n)$ time.

3.6. Limitations

The algorithm has several limitations. It requires point features. It relies on pairwise baseline estimates, so may be unstable for degenerate input configurations, or incorrect node adjacencies. The algorithm's assumption that nearby nodes are likely to have observed overlapping scene structure may be faulty, for example when two nodes lie on opposite sides of a thin structure.

4. Experiments

We implemented the position registration algorithm in roughly 5,000 lines of C++ code, and instrumented its performance on a 250MHz SGI O2 with 1.5 GB of memory. We report the following quantities for each dataset:

- ***Data size.*** The number of omni-directional nodes, constituent rectangular images, detected point features, and adjacent node pairs, as well as the average distance between adjacent nodes.
- ***Computation time.*** Average and total running times for each stage, excluding file I/O.
- ***Positional offsets.*** The difference between each node's initial and refined position, which enables us to assess the algorithm's robustness against poor initial pose.
- ***End-to-End position error.*** Uncertainty estimates for the refined node positions using 99% confidence bounds of the recovered Gaussian densities.
- ***Feature consistency.*** We assessed end-to-end feature consistency by thresholding (at 80%) each MCEM weight matrix to a binary match matrix. For each surviving match, we compute the 3-D distance (in cm) between rays extruded from each node through the point feature, and the 2-D distance (in pixels) between each point feature and its epipolar line in the other node.

4.1. Tech Square Data Set

The Technology Square data set consists of 75 nodes spanning an area of roughly 285 by 375 meters (Figure 12). For this data set, our algorithm corrected initial position errors of nearly seven meters, producing pose consistent on average to 5.6 cm of position and 1.22 pixels. The maximum pose error for any node was 11.0 cm of position and 5.71 pixels. Total CPU time was just under three hours.

4.2. Comparison to Manual Solution

A manually generated pose solution was available for this dataset [12], enabling comparison to our automatic technique. Identifying five or more correspondences by hand for each of roughly 200 adjacencies requires significant human effort; thus the student operator omitted many point matches, producing a convergent, but not over-determined, constraint set. Figure 13 compares epipolar geometry for manual and automated pose recovery.

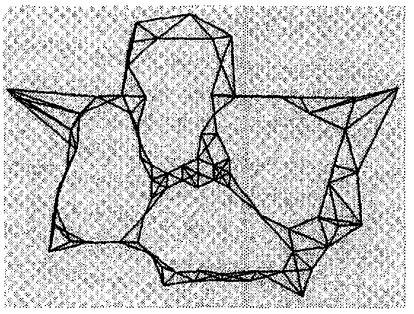

Figure 12: Tech Square Node Configuration
Node positions and adjacencies for the Tech Square data set. The average baseline was 30.88 meters.

Data Type	Per Image	Per Node	Total
Nodes	—	—	75
Images	—	48	3899
Points	227	10,958	887,598
Node Adj	—	—	189

	Per Pair	Total
Baseline Hough	8.1 s	25 m 31 s
Baseline MCEM	45.3 s	2 h 23 m
Global Opt	—	0 m 53 s
Total	53.4 s	2 h 49 m

Table 1: Tech Square Data Size, CPU Times by Stage

	Avg	Max
Position Diff	0.70 m	6.70 m
Position Bound	5.6 cm	11.0 cm

	Avg	Max	Std. Dev.
3-D Ray Error	9.6 cm	12.4 cm	3.3 cm
2-D Epi Error	1.22 pix	5.71 pix	2.33 pix

Table 2: Tech Square Consistency

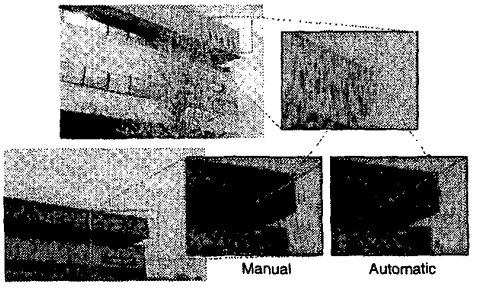

Figure 13: Tech Square Comparison I
A point feature in one image and its epipolar line in another image, as computed using pose generated by manual correspondence (bottom middle) vs. our automatic method (bottom right). Note the error in the manual solution, due to insufficient match constraints.

Figure 14 compares epipolar geometry for a corner from a repeating series of windows obscured by foliage. The

manual solution has poor epipolar geometry, since the human user did not enter this particular match constraint. It is plainly impossible to match these window corners *given only this pair of images*, due to the camera's limited FOV; even given omni-directional images, human operators find it difficult or impossible to match window corners due to the severe clutter obscuring most individual views. Our algorithm succeeds where the human fails by combining many omni-directional observations of many point features.

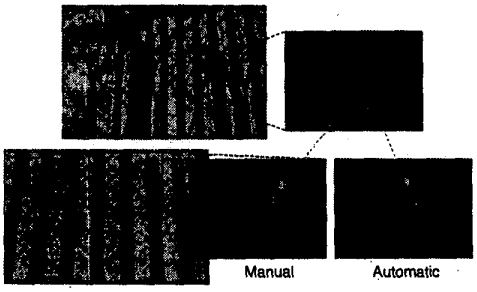

Figure 14: Tech Square Comparison II
A feature whose match is difficult for a human operator to identify. Epipolar geometry is shown for manual (bottom middle) and automated (bottom right) pose solutions. Note the error in the manual solution.

4.3. Ames Court Data Set

The Ames Court data set spans an area of 315 by 380 meters; the average adjacency baseline was 23.53 meters. Initial pose was corrected by over six meters, achieving average consistency of 5.7 cm and 3.88 pixels. The maximum pose inconsistency was 8.8 cm and 5.02 pixels. Total CPU time was just under four hours.

4.4. Benefit of Omni-Directional Imagery

There is substantial theoretical evidence that wide-FOV (i.e., omni-directional) images are fundamentally more

	Per Image	Per Node	Total
Nodes	—	—	100
Images	—	20	2,000
Points	257	4, 132	413,254
Node Adj	—	—	232

	Per Pair	Total
Baseline Hough	7.8 s	30 m 10 s
Baseline MCEM	52.6 s	3 h 24 m
Global Opt	—	1 m 04 s
Total	60.4 s	3 h 55 m

Table 3: Ames Court Data Size, CPU Times by Stage

	Avg	Max
Position Diff	3.53 m	6.18 m
Position Bound	5.7 cm	8.8 cm

	Avg	Max	Std. Dev.
3-D Ray Error	14.9 cm	20.2 cm	5.6 cm
2-D Epi Error	3.88 pix	5.02 pix	2.10 pix

Table 4: Ames Court Consistency

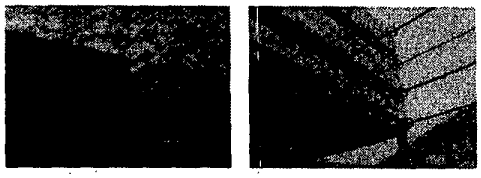

Figure 15: Ames Court Epipolar Geometry
Point features and corresponding epipolar lines for a typical node pair in the Ames Court set.

powerful than narrow-FOV (i.e., planar) images (e.g. [15]). Here, we show some empirical evidence using position (baseline) recovery. We examined the Hough transform, and resulting baseline direction estimate, for a node pair as a function of the FOV. Transform values are plotted in Figure 16. The sharpness of the peak, and the consistency of the resulting baseline estimate, increases with field of view. Moreover, we observed that narrow-FOV images do not provide sufficient feature overlap for convergence in any of our datasets.

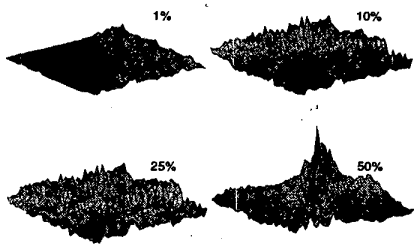

Figure 16: Hough Transform Peak Coherence
The dependence of baseline Hough transform peak coherence on FOV. The transform is shown for increasing percentages of sphere coverage.

5. Related Work

Interactive tools have been used for bundle adjustment [13, 22], but do not scale to our problem size. Algorithmic approaches recover relative pose through the use of known targets [28], or by tracking targets from a moving camera [20, 16, 4]. These methods are sensitive to image noise, illumination variations, and strong perspective or occlusion due to extended baselines. Feature [24] and texture [19] trackers find correspondences only for short baselines and sequences.

Many robust algorithms attempt to discard outliers [16, 27, 1, 8], but do not sample from the space of feature sets in a principled way or account for match ambiguities and feature noise. Several authors have formulated correspondence probabilistically [9, 14, 26, 5], but have not demonstrated their methods for large numbers of features or extended camera motions. Correspondence-free pose estimation techniques [17] have not been demonstrated for scenes with significant occlusion or lighting variation.

Measurement uncertainty has typically been treated with additive Gaussian noise [29], but spherical distributions are

more appropriate [11]. Finally, some authors have proposed EM algorithms for coupled SFM (e.g. [9]), but none have provided a principled treatment of measurement noise and matching ambiguity. Recently, a probabilistic EM formulation has been proposed which handles multiple images and match ambiguity [14], but only when the number of 3-D features is known, and all features are visible in all images.

6. Contributions and Conclusions

This paper makes several contributions. First, we propose the use of *a priori* absolute position estimates. Second, we present evidence that wide-FOV (omni-directional) images are more powerful observations than conventional images. Third, we extend existing probabilistic feature correspondence methods to correctly incorporate projective uncertainty, and to handle unknown numbers of features, occlusion, deocclusion, and outliers (elsewhere [2] we show that the resulting algorithm is robust against up to 80% outliers). Fourth, we combine Hough transform and probabilistic techniques to address the limitations of both methods. Fifth, we demonstrate how to aggregate uncertain pair-wise baseline constraints to produce globally consistent node position estimates.

We also assessed end-to-end error of the 6-DOF pose recovery system. To our knowledge, the resulting datasets are the largest registered terrestrial image datasets in existence, regardless of whether manual or automated calibration algorithms are used. Producing equivalent datasets using manual photogrammetric bundle-adjustment would require prohibitive human effort.

One advantage of working at this scaling regime is that of overconstraint and data redundancy. We emphasize that the image datasets registered with our algorithm were acquired outdoors, over wide baselines, under uncontrolled and varying lighting conditions, and in the presence of significant occlusion and visual clutter. The algorithms described here and in [3] represent a new end-to-end capability for automated, absolute registration of terrestrial images.

Acknowledgements

Support for this research was provided in part by the Office of Naval Research under MURI Award SA 1524-2582386.

References

[1] A. Adam, E. Rivlin, and I. Shimshoni. ROR: Rejection of outliers by rotations in stereo matching. In *Proc. CVPR*, pages 2–9, June 2000.

[2] M. Antone and S. Teller. Automatic recovery of camera positions in urban scenes. Technical Report 814, MIT LCS, Dec. 2000.

[3] M. Antone and S. Teller. Automatic recovery of relative camera rotations for urban scenes. In *Proc. CVPR*, pages 282–289, June 2000.

[4] A. Azarbayejani and A. Pentland. Recursive estimation of motion, structure, and focal length. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(6):562–575, June 1995.

[5] M. Ben-Ezra, S. Peleg, and M. Werman. Real-time motion analysis with linear programming. *CVIU*, 78(1):32–52, April 2000.

[6] C. Bingham. An antipodally symmetric distribution on the sphere. *Annals of Statistics*, 2(6):1201–1225, November 1974.

[7] M. Bosse, D. de Couto, and S. Teller. Eyes of Argus: Georeferenced imagery in urban environments. *GPS World*, pages 20–30, April 1999.

[8] S. Chaudhuri and S. Chatterjee. Robust estimation of 3-D motion parameters in presence of correspondence mismatches. In *Proc. Asilomar Conference on Signals, Systems and Computers*, pages 1195–1199, November 1991.

[9] H. Chui and A. Rangarajan. A feature registration framework using mixture models. In *Proc. IEEE Workshop on Mathematical Methods in Biomedical Image Analysis*, pages 190–197, 2000.

[10] H. Chui and A. Rangarajan. A new algorithm for non-rigid point matching. In *Proc. CVPR*, pages 44–51, June 2000.

[11] R. T. Collins and R. Weiss. Vanishing point calculation as statistical inference on the unit sphere. In *Proc. ICCV*, pages 400–403, December 1990.

[12] S. Coorg, N. Master, and S. Teller. Acquisition of a large pose-mosaic dataset. In *Proc. CVPR*, pages 872–878, June 1998.

[13] P. E. Debevec, C. J. Taylor, and J. Malik. Modeling and rendering architecture from photographs: A hybrid geometry- and image-based approach. In *Proc. SIGGRAPH*, pages 11–20, 1996.

[14] F. Dellaert, S. M. Seitz, C. E. Thorpe, and S. Thrun. Structure from motion without correspondence. In *Proc. CVPR*, pages 557–564, June 2000.

[15] C. Fermüller and Y. Aloimonos. Ambiguity in structure from motion: Sphere versus plane. *IJCV*, 28(2):137–154, 1998.

[16] A. W. Fitzgibbon and A. Zisserman. Automatic camera recovery for closed or open image sequences. In *Proc. ECCV*, pages 311–326, June 1998.

[17] P. Fua and Y. G. Leclerc. Registration without correspondence. In *Proc. CVPR*, pages 121–128, June 1994.

[18] B. K. P. Horn. Closed-form solution of absolute orientation using unit quaternions. *J. of the Optical Society of America A*, 4(4):629–642, April 1987.

[19] B. K. P. Horn and B. G. Schunck. Determining optical flow. *Artificial Intelligence*, 16(1–3):185–203, August 1981.

[20] M.-S. Lee, G. Medioni, and R. Deriche. Structure and motion from a sparse set of views. In *Proc. the International Symposium on Computer Vision*, pages 73–78, November 1995.

[21] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *J. of Chemical Physics*, 21(6):1087–1092, 1953.

[22] H. S. Shum, M. Han, and R. Szeliski. Interactive construction of 3D models from panoramic image mosaics. In *Proc. CVPR*, pages 427–433, 1998.

[23] R. Sinkhorn. A relationship between arbitrary positive matrices and doubly stochastic matrices. *Annals of Mathematical Statistics*, 35(2):876–879, June 1964.

[24] R. Szeliski, S. B. Kang, and H.-Y. Shum. A parallel feature tracker for extended image sequences. In *Proc. International Symposium on Computer Vision*, pages 241–246, 1995.

[25] S. Teller. Automatic acquisition of hierarchical, textured 3D geometric models of urban environments: Project plan. In *Proc. of the Image Understanding Workshop*, 1997.

[26] P. Torr and C. Davidson. IMPSAC: Synthesis of importance sampling and random sample consensus. In *ECCV*, pages 819–833, 2000.

[27] P. H. S. Torr and A. Zisserman. MLESAC: A new robust estimator with application to estimating image geometry. *Computer Vision and Image Understanding*, 78:138–156, 2000.

[28] R. Y. Tsai. A versatile camera calibration technique for high accuracy 3D machine vision metrology using off-the-shelf TV cameras and lenses. *IEEE J. of Robotics and Automation*, 3(4):323–344, 1987.

[29] Z. Zhang. Determining the epipolar geometry and its uncertainty: A review. *IJCV*, 27(2):161–195, 1998.