Global Motion Estimation from Point Matches

Mica Arie-Nachimson*, Shahar Z. Kovalsky*, Ira Kemelmacher-Shlizerman†, Amit Singer‡ and Ronen Basri*
*Weizmann Institute of Science, Rehovot, Israel
Email: {mica.arie-nachimson,shahar.kovalsky,ronen.basri}@weizmann.ac.il
†University of Washington, Seattle, WA
Email: kemelmi@cs.washington.edu
‡Princeton University, Princeton, NJ
Email: amits@math.princeton.edu

***Abstract*—Multiview structure recovery from a collection of images requires the recovery of the positions and orientations of the cameras relative to a global coordinate system. Our approach recovers camera motion as a sequence of two global optimizations. First, pairwise Essential Matrices are used to recover the global rotations by applying robust optimization using either spectral or semidefinite programming relaxations. Then, we directly employ feature correspondences across images to recover the global translation vectors using a linear algorithm based on a novel decomposition of the Essential Matrix. Our method is efficient and, as demonstrated in our experiments, achieves highly accurate results on collections of real images for which ground truth measurements are available.**

***Keywords*-structure from motion; 3D reconstruction; camera motion estimation; convex relaxation; linear estimation**

I. Introduction

Given a collection of images, recovering the exterior orientation parameters (i.e., the location and orientation) of the cameras that capture the images is an important step toward 3D shape recovery. This, multiview *Structure from Motion (SfM)* problem [1], [2], is a fundamental and well studied problem in computer vision. Recent decades have seen persistent progress in both problem formulation and the development of suitable algorithms for SfM. This progress, along with the advancement of computational power, has led to systems that are now capable of reconstructing large scale scenes from hundreds of thousands of images (e.g., [3], [4], [5], [6], see review in [7]).

This paper addresses the problem of finding the exterior camera parameters of n cameras given a collection of point correspondences. SfM systems typically use point correspondences to recover epipolar constraints between pairs of images. The camera motion parameters are over-constrained by these epipolar constraints. A consistent recovery of the exterior camera parameters from a partial set of noisy epipolar constraints is a challenging and well-studied problem.

Until recently, practical SfM systems have approached the problem of motion estimation in large, multiview settings using a sequential, greedy strategy that scans the image set in paths that produce a spanning tree [3], [5], [6]. Several recent studies cast the problem in a global optimization framework that accounts simultaneously for all cameras [8], [4], [9], [10], [11], [12], [13], [14], [15]. In this paper we propose an approach that solves for all motion parameters simultaneously in a framework which is both efficient and accurate.

We follow a common pipeline of SfM methods [8], [10]. Given n input images, we use standard software to recover point correspondences and Essential Matrices (relating some of the image pairs). Our method then proceeds by recovering the orientations of the corresponding n cameras and subsequently the n camera locations. For the recovery of camera orientation we use an objective similar to the one proposed by [10] and propose two new methods to solve this problem using (i) eigenvector decomposition and (ii) semidefinite programming (SDP). Subsequently, we introduce a new way to solve for camera locations by casting the problem as a homogeneous linear system of equations. To this end we introduce a novel expression for the Essential Matrix in terms of the global motion parameters and derive a method to recover the camera locations directly from point correspondences. Our method is very efficient: camera parameters are recovered by applying the MATLAB *'eigs'* command on two $3n \times 3n$ matrices. Our method achieves accurate estimates of the motion parameters, overcoming errors in the initial estimates of pairwise Essential Matrices. We demonstrate our method in experiments with standard collections of real images.

II. Background

We begin with a brief summary of the relevant concepts in multiview geometry. A thorough treatment of this subject can be found in [2]. Let $I_1, I_2, ..., I_n$ denote a collection of images of a stationary scene, and let $\mathbf{t}_i \in \mathbb{R}^3$ and $R_i \in SO(3)$ $(1 \leq i \leq n)$ respectively denote the focal points and orientations of the n cameras in some global coordinate frame. Let f_i denote the focal length of the i'th camera. To produce the i'th image a scene point $\mathbf{P} = (X, Y, Z)^T$ is transformed to $\mathbf{P}_i = R_i^T(\mathbf{P} - \mathbf{t}_i)$ and projected to $\mathbf{p}_i = (x_i, y_i, f_i)$ in I_i with $\mathbf{p}_i = (f_i/Z_i)\mathbf{P}_i$, where Z_i is the depth coordinate of $\mathbf{P}_i$.

For a pair of images I_i and I_j $(1 \leq i, j \leq n)$ we define $R_{ij} = R_i^T R_j$ and $\mathbf{t}_{ij} = R_i^T(\mathbf{t}_j - \mathbf{t}_i)$. It can readily be

verified that $\mathbf{P}_j = R_{ij}^T(\mathbf{P}_i - \mathbf{t}_{ij})$. Therefore, R_{ij} and $\mathbf{t}_{ij}$ are the rotation and translation that relate the coordinate frame of image j with that of image i. Clearly, $R_{ji} = R_{ij}^T$ and $\mathbf{t}_{ji} = -R_{ij}^T\mathbf{t}_{ij}$.

By a standard construction, the Essential Matrix E_{ij} is defined as $E_{ij} = [\mathbf{t}_{ij}]_\times R_{ij}$, where $[\mathbf{t}_{ij}]_\times$ denotes the skew-symmetric matrix corresponding to the cross product with $\mathbf{t}_{ij}$. This construction ensures that for every point $\mathbf{P} \in \mathbb{R}^3$ its projections onto I_i and I_j, denoted $\mathbf{p}_i$ and $\mathbf{p}_j$, satisfy the epipolar constraints

$$\mathbf{p}_i^T E_{ij} \mathbf{p}_j = 0. \tag{1}$$

A proper Essential Matrix E_{ij} can be decomposed into a rotation R_{ij} and a translation t_{ij}. This provides two possible rotations and a scale (and sign) ambiguity for the translation, which are determined by the chirality constraint.

III. Estimating Camera Orientation

Our formulation is based on recent estimation methods shown in the context of 3D structure determination of macromolecules in cryo-electron microscopy (EM) images [16], [17]. That work has assumed that pairwise rotations are known for every image pair. We focus on the SfM case where many of the pairwise rotations are missing. Our objective function is similar to that in [10], with our objective function allowing to additionally derive a tighter SDP relaxation. See Section III-A for discussion and comparison with [10].

Other works that take a global approach for recovery of camera rotation include using a reference plane (e.g. [11]) or quaternions [8], [12], which were shown in [10] to be inferior to Frobenious-based techniques. Hartley [18] uses a non-linear L1 minimization for rotation estimation, see discussion in Section VI-B.

We estimate camera orientation for a set of n images $I_1, ..., I_n$. Suppose we are given estimates of some of the $\binom{n}{2}$ Essential Matrices, $\hat{E}_{ij}$. (Below we use the hat accent to denote measurements inferred from the input images.) We factorize each Essential Matrix and obtain a unique pairwise rotation denoted $\hat{R}_{ij}$. We can further use [19] to detect motion degeneracies, in which case $\hat{E}_{ij}$ is ignored. Our aim is to recover camera orientation in each of the n images, $R_1, ..., R_n$, based on the pairwise rotations $\hat{R}_{ij}$.

A. Spectral Decomposition

Suppose all $\hat{R}_{ij}$ are known. Then, following [10], we can cast this problem as an over-constrained optimization:

$$\min_{\{R_1,...R_n\}} \sum_{i,j=1}^{n} \|R_i^T R_j - \hat{R}_{ij}\|_F^2, \tag{2}$$

where $\|.\|_F$ denotes the Frobenius norm of a matrix. We further require each matrix R_i ($1 \le i \le n$) to be a rotation, obtaining seven constraints for each of the rotations – six orthonormality constraints of the form $R_i^T R_i = I$ and one for the determinant, $\det(R_i) = 1$ (to distinguish it from reflections).

We solve the optimization problem (2) using the following observation. Let G be a $3n \times 3n$ symmetric matrix constructed by concatenating the pairwise rotation matrices, namely,

$$G = \begin{pmatrix} I & R_{12} & ... & R_{1n} \\ R_{21} & I & ... & R_{2n} \\ ... & & & ... \\ R_{n1} & R_{n2} & ... & I \end{pmatrix}. \tag{3}$$

Let R be a $3 \times 3n$ matrix constructed by concatenating rotations relative to a universal coordinate system $R = \begin{bmatrix} R_1 & R_2 & ... & R_n \end{bmatrix}$. Then,

Claim 1. *G has rank 3 and its three eigenvectors of nonzero eigenvalues are given by the columns of R^T.*

Proof By definition $R_{ij} = R_i^T R_j$, and so $G = R^T R$ with rank 3. Since $RR^T = nI$, $GR^T = R^T RR^T = nR^T$, and hence the three columns of R^T form the eignevectors of G with the same eigenvalue, n. □

Usually, in SfM problems some of the pairwise rotations are missing. We then modify G to contain zero blocks for the missing rotations. Let d_i denote the number of available rotations R_{ij} in the i'th block row of G, and let D be the $3n \times 3n$ diagonal matrix constructed as

$$D = \begin{pmatrix} d_1 I & 0 & ... & 0 \\ 0 & d_2 I & ... & 0 \\ ... & & & \\ 0 & 0 & ... & d_n I \end{pmatrix}. \tag{4}$$

It can be readily verified that $GR^T = DR^T$, and so the columns of R^T form three eigenvectors of $D^{-1}G$ with eigenvalue 1.

More generally, the construction of G and D can be modified to incorporate weights $0 \le w_{ij} \le 1$ that reflect our confidence in the available pairwise rotations R_{ij}.

In practice, however, the relative rotations $\hat{R}_{ij}$ that are extracted from the estimated Essential Matrices may deviate from the ground truth underlying R_{ij}. This is both because of mismatched corresponding points and errors in their estimated location. Similarly to G, we define $\hat{G}$ as the $3n \times 3n$ matrix containing the *observed* pairwise rotations $\hat{R}_{ij}$.

Claim 2. *An approximate solution to* (2)*, under relaxed orthonormality and determinant constraints, is determined by the three leading eigenvectors of the $3n \times 3n$ matrix $\hat{G}$.*

Details and a proof are provided in the appendix. Note that, in general, the noisy input reduces the spectral gap between the top three eigenvalues of $\hat{G}$ and the rest of its eigenvalues.

To extract the rotation estimates, we denote by M the $3n \times 3$ matrix containing the eigenvectors as in Claim 2. M comprises n submatrices of size 3×3, $M = [M_1; M_2; ...; M_n]$.

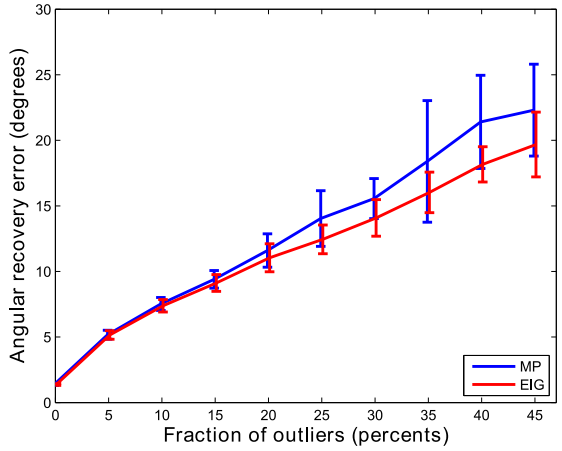

Figure 1. Comparison of our spectral method (in red) to [10]'s (blue). The figure shows angular recovery error as a function of fraction of outliers. Of the pairwise rotations 15% are true rotations perturbed by Gaussian noise of 20DB, corresponding to a mean angular error of 5°, and the rest of the input rotations are either missing or drawn uniformly from $SO(3)$, simulating outliers.

Each M_i is an estimate for the rotation of the i'th camera. Due to the relaxation, each M_i is not guaranteed to satisfy $M_i^T M_i = I$. Therefore, we find the nearest rotation (in the Frobenius norm sense) by applying the singular value decomposition $M_i = U_i \Sigma_i V_i^T$ and setting $\hat{R}_i^T = U_i V_i^T$ [20]. We further enforce $\det(\hat{R}_i) = 1$ by negating if needed. Note that this solution is determined up to a global rotation, corresponding to a change in orientation of the global coordinate system.

Additionally, observe the particular structure of G. Note that G has the form of a block adjacency matrix for the graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ constructed by placing an edge $(i, j) \in \mathcal{E}$ for every available Essential Matrix E_{ij}; G includes a 3×3 block of rotation for every entry 1 in the adjacency matrix and a zero block for every entry of zero. The matrix $D^{-1}G$, therefore, is tightly related to the graph Laplacian of $\mathcal{G}$. Consequently, it can be shown that all the eigenvalues of $D^{-1}G$ are in the range $[-1, 1]$. In practice, however, the rank of the *estimate* matrix $\hat{G}$ may exceed 3, and the spectral gap between the three leading eigenvectors and the rest of the eigenvectors is often smaller than 1. Numerical experiments conducted in [16] in the case that $\hat{G}$ is full and theoretical analysis based on random matrix theory demonstrate the robustness of this approach in the context of reconstruction of macromolecules from cryo-EM readings. Our experiments below show similar behavior in typical SfM problems in which $\hat{G}$ is sparse.

Comparison to [10] : Our objective function (2) is equal to [10]'s, who minimized $\sum_{i,j=1}^{n} \|R_j - R_i \hat{R}_{ij}\|_F^2$, and the two solution methods differ in the normalizations applied to account for the missing pairwise rotations. Moreover, our formulation further allows for casting the problem in an SDP framework (section III-B). The following simulation demonstrates that our spectral method is more robust to errors in the estimation of the Essential Matrices. We sampled $N = 100$ rotation matrices representing true camera orientations. We next perturbed 15% of the $N(N-1)/2$ pairwise rotations with Gaussian noise with SNR of 20DB and projected the noisy matrices to $SO(3)$. (Similar results were obtained with other fractions of perturbed rotations and SNR values.) The rest of the pairwise rotations was either considered missing or drawn uniformly from $SO(3)$ (simulating outliers). Figure 1 shows the angular recovery error obtained with our spectral method compared to [10]'s, as a function of the fraction of outliers. It can be seen that our spectral method consistently achieved more accurate estimation of orientation, particularly as the number of outliers increases. We note that in our real experiment with the Notre-Dame sequence (Section VI) the fraction of missing pairwise rotations was 72%.

B. Estimation with SDP

The formulation above, which leads to a solution by spectral decomposition, can be used also to derive a semidefinite programming (SDP) relaxation. Our minimization (2) can be cast as a maximization of

$$\max_{\{R_1,\ldots R_n\}} \sum_{i,j=1}^{n} \operatorname{trace}\left(\hat{R}_{ij}^T R_i^T R_j\right), \tag{5}$$

since assuming $R_i \in SO(3)$, $\|R_i^T R_j\|_F^2 = \|\hat{R}_{ij}\|_F^2 = \operatorname{trace}(I) = 3$. Using our notation above this equation can be written as

$$\max_{G} \operatorname{trace}\left(\hat{G}^T G\right), \tag{6}$$

where the unknown matrix G should be decomposable to $R^T R$. To allow such a decomposition G is required to be positive semidefinite and to have 3×3 identity matrices along its diagonal. In addition, G should have rank 3 and the determinant of each 3×3 block of G should be 1. The problem is analogous to the problem of matrix completion [21], as we seek to complete a rank 3 matrix G from noisy observations $\hat{G}$, with the additional constraint that G is positive semidefinite and composed of rotation matrices.

To obtain a semidefinite program we drop the rank and determinant requirements and solve

$$\begin{array}{lll} \max_G & \operatorname{trace}\left(\hat{G}^T G\right), & \\ \text{s.t.} & G \succeq 0 & \forall k \\ & G_{ii} = I & 1 \le i \le n, \end{array} \tag{7}$$

where G_{ii} denotes the i'th 3×3 block along the diagonal of G. Once the optimal G is found we use SVD to recover the set of n rotation matrices as we did in the previous section.

Notice that the SDP relaxation can be made slightly tighter. If G is indeed rank 3, and since its block diagonal is paved with identity matrices, then every 3×3 block G_{ij} must be either a rotation or reflection, i.e., $G_{ij} \in O(3)$. To discern between these two possibilities we would like to require $\det(G_{ij}) = 1$; these are, however, non-linear constraints that cannot be incorporated into the SDP formulation. Instead, we replace the determinant constraints by equivalent linear inequality constraints. To that end we have the following claim:

Claim 3. *There exists a finite set of rotations $A_1, A_2, \ldots, A_l \in SO(3)$ such that $G_{ij} \in SO(3)$ iff $\mathrm{trace}\,(A_k G_{ij}) \geq -1$ for all $k = 1, \ldots, l$.*

The proof, which proceeds by constructing an ϵ-net over $O(3)$, is provided in the appendix. As a consequence, adding the linear inequalities $\mathrm{trace}\,(A_k G_{ij}) \geq -1$ tightens the SDP relaxation. The number of linear inequalities might be relatively large in practice, and therefore could be randomly sparsified. Constructing the optimal (i.e., minimal) design $A_1, \ldots, A_l$ is beyond the scope of this paper.

Unlike the spectral decomposition in Section III-A, the SDP approach enables introducing constraints which drive a tighter convex relaxation of the optimization problem; thus, explicitly promoting a solution which is less sensitive to noise and mismatches. In practice, this becomes significant for large-scale SfM problems, in which noise, inaccuracies and outliers are prominent. In our experiments on the small benchmark sets (see Section VI-A), no significant improvement over the spectral methods was observed.

IV. Estimating Camera Location

Once camera orientations $\hat{R}_1, ..., \hat{R}_n$ are recovered we turn to recovering the camera location parameters, $\mathbf{t}_1, ..., \mathbf{t}_n$. We do this using an efficient linear approach.

Previous approaches for estimating camera locations typically exploit the pairwise translations derived from the Essential Matrices (1) to construct a system of equations in the unknown translation parameters, and often also in the unknown depth coordinates. Such methods commonly involve a large excessive number of unknowns either involving 3D point positions for all feature points or additional pairwise scaling factors. Solving such systems can be computationally demanding and sensitive to errors.

For example, Govindu [12] uses the pairwise translations $\mathbf{t}_{ij}$ to estimate the camera location using $\mathbf{t}_{ij} = \gamma_{ij} R_i^T (\mathbf{t}_i - \mathbf{t}_j)$ where γ_{ij} are unknown scale factors separate for each pair of images. He then shows that eliminating these scaling factors lead to unstable results, and so he estimates them using an iterative reweighting approach. Crandall *et al.* [4] uses an MRF to solve simultaneously for camera locations and structure, but relies on prior geotag locations and assumes 2D translations. Kahl & Hartley [13] and subsequently also [8], [10] define a nonlinear, quasiconvex system of equations in the translations and point locations and use SOCP to solve the system under the l_∞ norm. Rother [11] proposes a linear system for solving simultaneously for both camera and 3D point locations. This adds a large number of unknowns to the equations. For example, for the large collection described in Section VI-B, the method in [11] will have 120K unknowns.

Below we propose an alternative approach to solving for the translation parameters. Our approach is based on a simple but effective change of coordinates, which leads to a linear system with a large number of linear equations – an equation for every pair of corresponding points – in a *minimal* number of unknowns, the sought translations $\mathbf{t}_1, ..., \mathbf{t}_n$.

Claim 4. *The Essential Matrix can be expressed in terms of the location and orientation of each camera:*

$$E_{ij} = R_i^T (T_i - T_j) R_j, \tag{8}$$

where $1 \leq i, j \leq n$, and $T_i = [\mathbf{t}_i]_\times$, $T_j = [\mathbf{t}_j]_\times$. This expression generalizes over the usual decomposition of the Essential Matrix; if we express the Essential Matrix in the coordinate frame of the i'th image then $\mathbf{t}_i = 0$ and $R_i = I$, and we are left with $E_{ij} = [\mathbf{t}_{ij}]_\times R_{ij}$.

Proof We derive an expression for the Essential Matrix in terms of a global coordinate system. The construction is similar to the usual derivation of the Essential Matrix. Let $\mathbf{P}$ denote a point in $\mathbb{R}^3$. Let $\mathbf{P}_i = R_i^T(\mathbf{P} - \mathbf{t}_i)$ and $\mathbf{p}_i = (x_i, y_i, f_i)$ denote its projection onto the image I_i $(1 \leq i \leq n)$. For a pair of images I_i and I_j we eliminate P to obtain:

$$R_j \mathbf{P}_j - R_i \mathbf{P}_i = \mathbf{t}_i - \mathbf{t}_j. \tag{9}$$

Taking the cross product with $\mathbf{t}_i - \mathbf{t}_j$ and the inner product with $R_i \mathbf{P}_i$ we obtain

$$\mathbf{P}_i^T R_i^T ((\mathbf{t}_i - \mathbf{t}_j) \times R_j \mathbf{P}_j) = 0, \tag{10}$$

and, due to the homogeneity of this equation, we can replace the points with their projections

$$\mathbf{p}_i^T R_i^T ((\mathbf{t}_i - \mathbf{t}_j) \times R_j \mathbf{p}_j) = 0. \tag{11}$$

This defines the epipolar relations between I_i and I_j. Consequently,

$$E_{ij} = R_i^T (T_i - T_j) R_j.$$

□

The advantage of this representation of the Essential matrix (8) is that it includes only the location and orientation of each camera; pairwise information (R_{ij}, $\mathbf{t}_{ij}$) is no longer required. Let $p_i^{(1)} \ldots p_i^{(M_{ij})}$ and $p_j^{(1)} \ldots p_j^{(M_{ij})}$ be M_{ij} corresponding image points from images I_i and I_j respectively. Then, the expression in (8) defines a homogenous epipolar

line equation for every pair of corresponding points $\mathbf{p}_i^{(m)}$ and $\mathbf{p}_j^{(m)}$, $m = 1 \dots M_{ij}$:

$$\mathbf{p}_i^{(m)^T} R_i^T (T_i - T_j) R_j \mathbf{p}_j^{(m)} = 0. \quad (12)$$

This equation is linear in the translation parameters.

This epipolar equation system (12) can further be written as follows. Note that the left hand side defines a triple product between the rotated points $R_i\mathbf{p}_i^{(m)}$, $R_j\mathbf{p}_j^{(m)}$ and the translation $\mathbf{t}_i - \mathbf{t}_j$. A triple product is invariant to permutation (up to a change of sign if the permutation is non-cyclic). Consequently, (12) can be written as

$$(\mathbf{t}_i - \mathbf{t}_j)^T (R_i\mathbf{p}_i^{(m)} \times R_j\mathbf{p}_j^{(m)}) = 0. \quad (13)$$

Therefore, *every* point pair contributes a linear equation in six unknowns (three for $\mathbf{t_i}$ and three for $\mathbf{t_j}$). As such, the location of each camera is linearly constrained by each of its feature correspondences. Weighting $w_{ij}^{(m)}$ can be easily incorporated to reflect the certainty of each such equation.

Clearly, $\mathbf{t}_i = (1,0,0)$, $\mathbf{t}_i = (0,1,0)$ and $\mathbf{t}_i = (0,0,1)$ for all $(1,0,0)$ i, are three trivial solutions of this linear system. Therefore, the sought solution is the optimal solution orthogonal to this trivial subspace. This allows recovering the camera locations up to a global translation and a single global scaling factor; these are inherent to the problem and cannot be resolved without external measurements.

Unlike alternative linear methods [12], [11], our linear system is compact: the only unknowns are the camera locations. Thus, employing linear methods for its solution allows for an extremely efficient implementation. Moreover, despite the obvious drawbacks of using an L2 approach, its highly over-constrained formulation plays a main role in promoting its robustness, as is demonstrated in our experiments. Further robustness can be achieved, e.g., by minimizing the L1 norm, e.g, by applying iterative reweighted least squares. This however is left for future research

V. Implementation

Given a collection of images $I_1, ..., I_n$, we follow the common SfM pipeline and apply the following procedure.

- **Obtain matches and pairwise rotations**: Apply a feature detector and seek pairs of corresponding points across images (we used SIFT [22] implementation from [6]), then compute Essential Matrices using the RANSAC protocol. We factor $\hat{R}_{ij}$ from the Essential Matrices and define $\hat{R}_{ij}$ as missing if insufficiently many inliers are found.
- **Rotation estimation**: Use the pairwise rotations $\hat{R}_{ij}$ to form the matrix $\hat{G}$ and to compute the set of global rotations $\hat{R}_1, ..., \hat{R}_n$ using either the spectral decomposition (Section III-A) or the SDP method (Section III-B). This step is based on the computation of the leading three eigenvectors of the sparse $3n \times 3n$ matrix $\hat{D}^{-1}\hat{G}$, where n is the number of images.

Table I
Camera matrix recovery errors for the Fountain-P11 and Herz-Jesu-P25 sequences.

Fountain-P11	Location (meters)	Viewpoint (degrees)	Rotation (Frobenius)
Our method (GT cal.)	**0.0048**	**0.024**	**0.0007**
Our method (Exif)	**0.0270**	**0.420**	**0.0111**
Bundler [6]	0.0072	0.112	0.0044
VisualSFM [24]	0.0099	0.116	0.0046
Sinha *et al.* [9]	0.1317	–	–
Martinec [10] (on R25)	0.0153	–	–

Herz-Jesu-P25	Location (meters)	Viewpoint (degrees)	Rotation (Frobenius)
Our method (GT cal.)	**0.0078**	**0.045**	**0.0012**
Our method (Exif)	**0.0520**	**0.348**	**0.0092**
Bundler [6]	0.0308	0.110	0.0041
VisualSFM [24]	0.0233	0.104	0.0040
Sinha *et al.* [9]	0.2538	–	–
Martinec [10] (on R25)	0.0845	–	–

- **Translation recovery**: Use the corresponding pairs of points and the recovered rotations to form the linear equations and to solve for the global translation vectors $\mathbf{t}_1, ..., \mathbf{t}_n$ (Section IV). Although this step involves an equation for every pair of corresponding points, the set of equations is very sparse containing only 6 unknowns in every equation. Furthermore, as there are $3n$ unknowns, the linear system is solved by finding the four eigenvectors with lowest eigenvalue of a $3n \times 3n$ matrix.
- **Bundle adjustment and dense 3D recovery**: We used the SBA code by [23] for the smaller sets and PBA [24] for the larger sets. We apply dense reconstruction of the scene, e.g., using [25].

VI. Experiments

We tested our method on several real image sequences. We first show results on two benchmark (though small) collections of images for which ground truth rotations and translations are provided [26]. We also tested our method on a common large-scale image set downloaded from the internet (images of the Notre-Dame Cathedral available from [27]). In both experiments, we applied the pipeline from Section V.

A. Benchmark Sets

We show results on benchmark images from [26]. The image collections, called Fountain-P11 and Herz-Jezu-P25, include 11 and 25 images respectively. The images are corrected for radial distortion. For internal calibration we used either calibration parameters supplied as ground truth or (rough) focal lengths extracted from the Exif tags of the raw images.

We further test the stability of our method by repeating each experiment 10 times as the selection of point matches depend on a random protocol (RANSAC). In all runs we were able to recover the camera positions and orientations

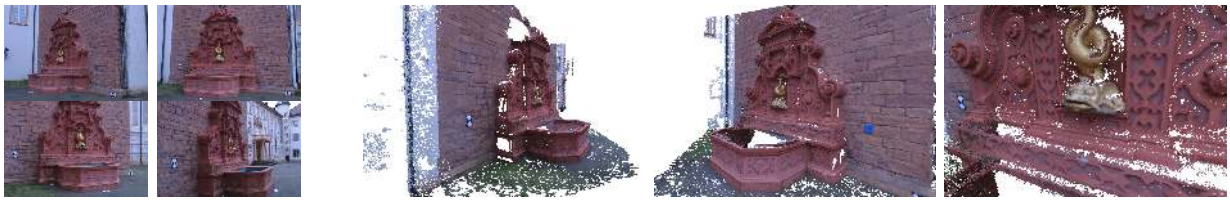

Figure 2. Four of 11 images of the Fountain-p11 sequence (left) and three snapshots of the reconstruction obtained with our method (right).

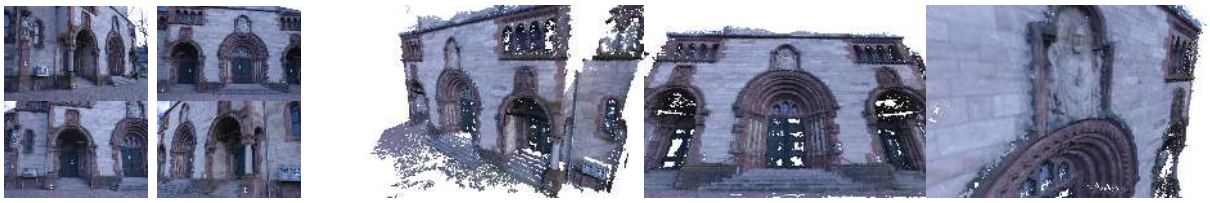

Figure 3. Four of 25 images of the Herz-Jesu-p25 sequence (left) and three snapshots of the reconstruction obtained with our method (right).

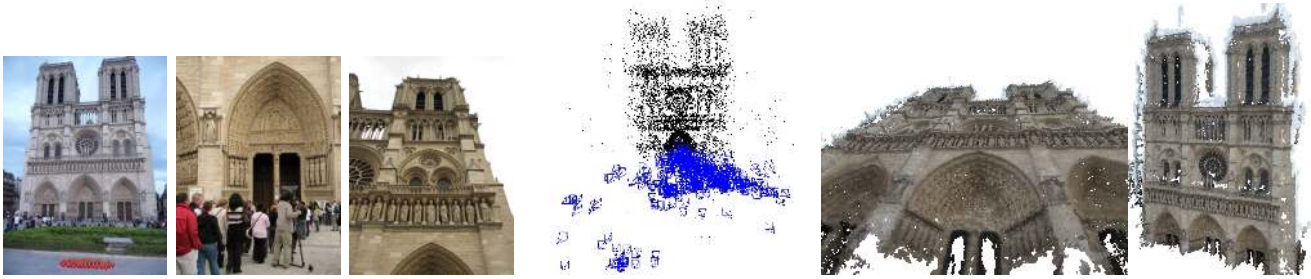

Figure 4. From left to right: three of 420 images of the Notre-Dame sequence, camera locations (in blue) against a sparse reconstruction, and two snapshots of the 3D reconstruction obtained with our method.

accurately and to obtain rich 3D reconstructions of the scenes. Figures 2-3 show several of the input images along with snapshots from the reconstructions we obtained using our method. Table I shows the errors in rotation and camera locations (averaged over the images in each sequence) of our recovery with respect to the supplied ground truth rotations and locations. For the two sequences, with ground truth calibration and after bundle adjustment, our algorithm achieved a very accurate estimation with an average error of only 4.8mm and 7.8mm in camera location and 0.024° and 0.045° in viewpoint orientation. Further comparisons to other recent methods are shown in Table I. We note that Martinec's algorithm [10] was tested in [26] on slightly different sequences of the *same* scenes, the Fountain-R25 and Herz-Jesu-R23. (Unfortunately, the ground truth values for those exact sequences were not made available to us by the authors.)

Our efficient linear setting allows estimating the rotations and translations for these datasets (steps 3 and 4 in our pipeline, Section V) between 0.1 to 0.2 seconds on a regular desktop computer (4 core 3GHz).

B. Large-Scale Image Collection

We next applied our method to the common image set of the Notre-Dame Cathedral available from the Photo Tourism dataset [27]. In the first test we used the focal lengths and radial distortion corrections produced in [6]. We follow other large-scale SfM methods and prune images with noisy or poor point matches. We prune images efficiently by discarding images corresponding to nodes with small degree in the image graph $\mathcal{G}$. Similarly to other global SfM methods, when the image graph is not connected (or very weakly connected), our method is applied separately to each of the connected components (stitching them is beyond the scope of this paper).

After computing the pairwise Essential Matrices we kept the subset of 420 images (out of 715) with degree 10 or more (results were similar to other choices of the minimal degree). Our method succeeded in estimating the camera parameters accurately, achieving low reprojection error (RMSE of 0.83 pixels) and similar rotations to the Bundler software package [6] (differed by 0.7° in viewpoint angle). Unfortunately, [6] does not provide a metric reconstruction and so camera locations could not be compared. The achieved reconstruction can be seen in Figure 4 and in the supplementary material.

Estimating the camera rotations and locations took only 46 seconds with our non-optimized MATLAB code, with additional 73 seconds for the BA. We compare our running time to VisualSFM [24], which is a highly optimized GPU implementation of the Photo Tourism [27]'s sequential algorithm. On a parallel GPU architecture, VisualSFM took 788

seconds on the same desktop PC. (This time includes also the computation of the pairwise Essential matrices, but does not include feature extraction and matching). Our method achieves results that are comparable to state-of-the-art on this sequence in substantially less time.

We repeated the Notre-Dame image set experiment, with 218 images for which Exif tags are available. We assume the focal lengths obtained with the Exif tags are sufficiently accurate for internal calibration, and used only these focal lengths with no further information. Our method estimates the camera rotations and locations correctly, with RMSE of 0.80 pixels, and viewpoint angle different from [6]'s by 0.85°.

We additionally compared our rotation estimation on the Notre Dame to the L1 optimization of Hartley *et al.* [18]. Our L2 objective function achieves comparable results to [18] (geodesic error of 0.66° compared to 0.82° reported in [18]) in substantially less time on a similar laptop platform (less than one second with our method compared to 36 seconds in [18]).

In conclusion, as these experiments demonstrate, our method provides an accurate and efficient means to recover the camera location and orientation in challenging multiview sequences. The method is very fast, with the actual estimation done by a MATLAB *'eigs'* command on two $3n \times 3n$ matrices.

VII. Conclusion

We presented a method for recovering the position and orientation of n cameras given a collection of images. The method is based on finding an optimal fit of the camera rotations to the pairwise rotations deduced from the geometry of point matches, and then, given the estimated rotations, finding the set of translations that is consistent with all the available point correspondences. We show that the camera rotations can be recovered either by spectral or SDP relaxations, and the camera translations can be recovered by a linear least squares. Our experiments demonstrate that our method can achieve fast and highly accurate recovery of the camera locations and orientations. The method is very efficient, yielding equation systems whose size depends only on the number of input images. We demonstrate the effectiveness of our method by experimenting with calibrated as well as internet photo collections. Our method assumes that the input images form a fairly connected graph and that the calibration parameters are given. For many common image sets the latter assumption implies only that focal lengths need to be known. Our current implementation can further be improved by incorporating prior knowledge into the estimation process and by constructing algorithms to stitch estimates obtained for different connected components.

Appendix

Claim. *An approximate solution to* (2)*, under relaxed orthonormality and determinant constraints, is determined by the three leading eigenvectors of the* $3n \times 3n$ *matrix* $\hat{G}$*.*

Proof As shown in Section III-B, assuming $R_i \in SO(3)$, equation (2) can be written as (5). Denote by $\mathbf{r}_1^{iT}$, $\mathbf{r}_2^{iT}$, and $\mathbf{r}_3^{iT}$ the three rows of R_i. Then, (5) can be written as

$$\max_{\{R_1,\ldots R_n\}} \sum_{i,j=1}^{n} \left(\mathbf{r}_1^{iT} \hat{R}_{ij} \mathbf{r}_1^j + \mathbf{r}_2^{iT} \hat{R}_{ij} \mathbf{r}_2^j + \mathbf{r}_3^{iT} \hat{R}_{ij} \mathbf{r}_3^j \right), \tag{14}$$

subject to the orthonormality of the rows of each R_i and the determinant constraint for the triplet of rows. We rewrite this expression in matrix form. Let $\hat{G}$ be a $3n \times 3n$ symmetric matrix constructed by concatenating the pairwise rotation matrices, namely,

$$\hat{G} = \begin{pmatrix} I & \hat{R}_{12} & \ldots & \hat{R}_{1n} \\ \hat{R}_{21} & I & \ldots & \hat{R}_{2n} \\ \ldots & & & \ldots \\ \hat{R}_{n1} & \hat{R}_{n2} & \ldots & I \end{pmatrix}. \tag{15}$$

Then, the above optimization can be expressed as

$$\max_{\{\mathbf{m}_1,\mathbf{m}_2,\mathbf{m}_3\}} \mathbf{m}_1^T \hat{G} \mathbf{m}_1 + \mathbf{m}_2^T \hat{G} \mathbf{m}_2 + \mathbf{m}_3^T \hat{G} \mathbf{m}_3, \tag{16}$$

where $\mathbf{m}_l \in \mathbb{R}^{3n}$, $(l = 1, 2, 3)$ is obtained by a concatenation of n $\mathbf{r}_l^i$ vectors, i.e., $\mathbf{m}_l = [\mathbf{r}_l^1; \mathbf{r}_l^2; ...; \mathbf{r}_l^n]$. The orthonormality and determinant constraints now apply to triplets of the entries of $\mathbf{m}_l$.

To make this optimization tractable we relax it by requiring only the (full) vectors $\mathbf{m}_l$ to be orthonormal (reducing the number of constraints from $7n$ to just 6). The obtained maximization problem is related to the classical problem $\max_{\mathbf{m}} \mathbf{m}^T \hat{G} \mathbf{m}$ subject to $\|\mathbf{m}\|^2 = n$ whose solution is given by the eigenvector of $\hat{G}$ of largest eigenvalue. We therefore expect the solution to (16) to consist of the three leading eigenvectors of $\hat{G}$.□

Claim 5. *There exists a finite set of rotations* $A_1, A_2, \ldots, A_l \in SO(3)$ *such that* $G_{ij} \in SO(3)$ *iff* $\operatorname{trace}(A_k G_{ij}) \geq -1$ *for all* $k = 1, \ldots, l$*.*

Proof In one direction: suppose $O \in SO(3)$ then $\operatorname{trace}(O) \geq -1$. Indeed, any rotation has one eigenvalue equal to 1 (the corresponding eigenvector is the rotation axis), and the magnitude of all its eigenvalues are less than or equal to 1 (since it is an isometry); in particular, the rotation with the minimum trace is $\operatorname{diag}[-1, -1, 1]$. Clearly, $A_k O$ is also in $SO(3)$ and as a result $\operatorname{trace}(A_k O) \geq -1$. In the other direction: first, recall that $SO(3)$ is compact. Therefore, for every $\epsilon > 0$ there exists an ϵ-net, i.e., a cover design consisting of $n = n(\epsilon)$ rotations $A_1, A_2, \ldots, A_n \in SO(3)$ with the property that for every $A \in SO(3)$ there exists an element A_k of the design such that $\|A - A_k\|_F < \epsilon$. Suppose

$\mathrm{trace}\,(A_k O) \geq -1$ for all $k = 1, \ldots, n$, and assume to the contrary that $O \notin SO(3)$. Then $A = -O^T \in SO(3)$, and there exists A_k for which $\|A - A_k\|_F < \epsilon$. From $A_k O = (A_k - A)O + AO = (A_k - A)O - I$, it follows that $\mathrm{trace}\,(A_k O) = \mathrm{trace}\,((A_k - A)O) - 3$. The Cauchy-Schwarz inequality, the design property $\|A - A_k\|_F < \epsilon$, and the orthogonality of O ($\|O\|_F = \sqrt{3}$) altogether give $\mathrm{trace}\,((A_k - A)O) \leq \|A_k - A\|_F \|O\|_F < \sqrt{3}\epsilon$. Therefore, $\mathrm{trace}\,(A_k O) < \sqrt{3}\epsilon - 3$. Choosing $\epsilon = 2/\sqrt{3}$ yields $\mathrm{trace}\,(A_k O) < -1$, a contradiction. □

ACKNOWLEDGEMENTS

Research was conducted in part while RB was at TTIC. At the Weizmann Inst. research was supported in part by the Israel Science Foundation grant 628/08 and conducted at the Moross Laboratory for Vision and Motor Control. We thank the organizers of the Machine Learning Workshop in the University of Chicago, June 1-11, 2009, where we held first discussions on this topic.

REFERENCES

[1] H. C. Longuet-Higgins, "A computer algorithm for reconstructing a scene from two projections," *Nature*, vol. 293, no. 2, 1981.

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

[3] S. Agarwal, N. Snavely, I. Simon, S. M. Seitz, and R. Szeliski, "Building rome in a day," in *ICCV*, 2009.

[4] D. Crandall, A. Owens, N. Snavely, and D. P. Huttenlocher, "Discrete-continuous optimization for large-scale structure from motion," in *CVPR*, 2011.

[5] X. Li, C. Wu, C. Zach, S. Lazebnik, and J. Frahm, "Modeling and recognition of landmark image collections using iconic scene graphs," in *ECCV*, 2008, pp. 427–440.

[6] N. Snavely, S. M. Seitz, and R. Szeliski, "Skeletal graphs for efficient structure from motion," in *CVPR*, 2008.

[7] N. Snavely, "Scene reconstruction and visualization from internet photo collections: A survey," *IPSJ Trans. on Computer Vision and Applications*, vol. 3, pp. 44–66, 2011.

[8] O. Enqvist, F. Kahl, and C. Olsson, "Non-sequential structure from motion," in *OMNIVIS*, 2011.

[9] S. N. Sinha, D. Steedly, and R. Szeliski, "A multi-stage linear approach to structure from motion," in *RMLE*, 2010.

[10] D. Martinec and T. Padjla, "Robust rotation and translation estimation in multiview reconstruction," in *CVPR*, 2007.

[11] C. Rother, "Linear multi-view reconstruction of points, lines, planes and cameras using a reference plane," in *ICCV*, 2003.

[12] V. Govindu, "Combining two-view constraints for motion estimation," in *CVPR*, 2001.

[13] F. Kahl and R. I. Hartley, "Multiple-view geometry under the l_∞-norm," *PAMI*, vol. 30, pp. 1603–1617, 2008.

[14] P. Sturm and B. Triggs, "A factorization based algorithm for multi-image projective structure and motion," in *ECCV*, 1996.

[15] J. Tardif, A. Bartoli, M. Trudeau, N. Guilbert, and S. Roy, "Algorithms for batch matrix factorization with application to structure-from-motion," in *CVPR*, 2007.

[16] A. Singer, "Angular synchronization by eigenvectors and semidefinite programming," *Applied and Computational Harmonic Analysis*, vol. 30, no. 1, pp. 20–36, 2011.

[17] A. Singer and Y. Shkolnisky, "Three-dimensional structure determination from common lines in cryo-em by eigenvectors and semidefinite programming," *SIAM Journal on Imaging Sciences*, vol. 4, no. 2, pp. 543–572, 2011.

[18] R. Hartley, K. Aftab, and J. Trumpf, "L1 rotation averaging using the weiszfeld algorithm," in *CVPR*, 2011.

[19] P. Torr, A. Zisserman, and S. Maybank, "Robust detection of degenerate configurations while estimating the fundamental matrix," *Computer Vision and Image Understanding*, vol. 71, no. 3, pp. 312–333, 1998.

[20] K. Fan and A. J. Hoffman, "Some metric inequalities in the space of matrices," *Proceedings of the American Mathematical Society*, vol. 6, no. 1, pp. 111–116, 1955.

[21] E. J. Candes and T. Tao, "The power of convex relaxation: Near-optimal matrix completion," *IEEE Trans. Inform. Theory*, vol. 56, no. 5, pp. 2053–2080, 2009.

[22] D. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, vol. 60, no. 2, pp. 91–110, 2004.

[23] M. A. Lourakis and A. Argyros, "SBA: A Software Package for Generic Sparse Bundle Adjustment," *ACM Trans. Math. Software*, vol. 36, no. 1, pp. 1–30, 2009.

[24] C. Wu, S. Agarwal, B. Curless, and S. M. Seitz, "Multicore bundle adjustment," in *CVPR*, 2011.

[25] Y. Furukawa and J. Ponce, "Accurate, dense, and robust multiview stereopsis," *PAMI*, vol. 32, no. 8, pp. 1362–1376, 2010.

[26] C. Strecha, W. von Hansen, L. V. Gool, P. Fua, and U. Thoennessen, "On benchmarking camera calibration and multi-view stereo for high resolution imagery," in *CVPR*, 2008.

[27] N. Snavely, S. M. Seitz, and R. Szeliski, "Photo tourism: Exploring photo collections in 3d," in *SIGGRAPH*, 2006.