Global Fusion of Relative Motions for Robust, Accurate and Scalable Structure from Motion

Pierre Moulon[1,2], Pascal Monasse[1], Renaud Marlet[1]
[1]Université Paris-Est, LIGM (UMR CNRS), ENPC, F-77455 Marne-la-Vallée. [2]Mikros Image.
firstname.lastname@enpc.fr

Abstract

Multi-view structure from motion (SfM) estimates the position and orientation of pictures in a common 3D coordinate frame. When views are treated incrementally, this external calibration can be subject to drift, contrary to global methods that distribute residual errors evenly. We propose a new global calibration approach based on the fusion of relative motions between image pairs. We improve an existing method for robustly computing global rotations. We present an efficient a contrario trifocal tensor estimation method, from which stable and precise translation directions can be extracted. We also define an efficient translation registration method that recovers accurate camera positions. These components are combined into an original SfM pipeline. Our experiments show that, on most datasets, it outperforms in accuracy other existing incremental and global pipelines. It also achieves strikingly good running times: it is about 20 times faster than the other global method we could compare to, and as fast as the best incremental method. More importantly, it features better scalability properties.

1. Introduction

Photogrammetry, SLAM (simultaneous localization and mapping) and SfM (structure from motion) reconstruct a model of a scene given a set of pictures. They compute both a 3D point cloud (the structure) and camera poses, i.e., positions and orientations (the calibration). Methods for that can be divided into two classes: sequential and global.

Sequential SfM pipelines start from a minimal reconstruction based on two or three views, then incrementally add new views into a merged representation. The most widely used incremental pipeline is Bundler [31]. It performs multiple bundle adjustments (BA) to rigidify the local structure and motion. As a result, it is a rather slow procedure. Yet, some parts of the problem can be solved more efficiently. Image matching can be made more scalable, e.g., thanks to vocabulary tree techniques [24]. Bundle adjustment can be optimized with sparse matrices [1] or using GPU [36]. The number of variables can also be reduced by eliminating structure from the bundle adjustment [26]. Finally some approaches use a divide-and-conquer approach on the epipolar graph to reduce computations [32, 9, 30, 22].

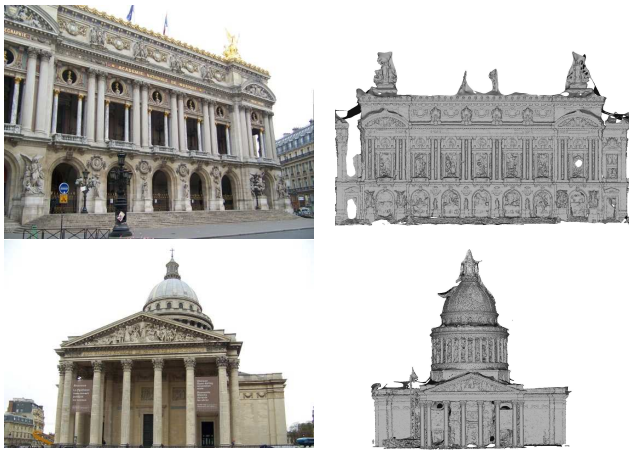


Figure 1. Dense mesh obtained with our global SfM pipeline on the monument datasets (top: 160 images, bottom: 100 images).

However, incremental approaches are known to suffer from drift due to the accumulation of errors and to the difficulty to handle cycle closures of the camera trajectory. An additional weakness is that the quality of the reconstruction depends heavily on the choice of the initial image pair and on the order of subsequent image additions.

Most global pipelines solve the SfM optimization problem in two steps. The first step computes the global rotation of each view and the second step computes the camera translations, together with the structure or not. The interest of separating the two steps is that the relative two-view rotations can be estimated quite precisely even for small baselines, which is not true of relative translations. These approaches take into account the whole epipolar graph, whose nodes represent the views and where edges link views having enough consistent matching points. All cycles of the graph yield multi-view constraints, in the sense that the lo-

cal relative motions in successive nodes of the cycle should compose into the identity when closing the cycle. Enforcing these constraints greatly reduces the risk of drift present in incremental methods. Moreover errors can be evenly distributed over the whole graph, contrary to incremental approaches. But such global approaches suffer from the fact that some two-view geometries, even when they have a large support of point correspondences, may fail to reflect the underlying global geometry, mainly because of mismatches, e.g., due to repetitive structures that create outliers. Additionally, as the minimization is based on the structure and the reprojection errors, the space and time requirements can get very large, even for limited-size datasets of images.

In this paper we present a new robust global SfM method for unordered image sets. The problem complexity is kept low using relative motions that can be merged very fast. We first solve the structure problem at a local scale (2 and 3 views), then merge the resulting relative motions into a common global coordinate frame. We assess the efficiency and precision of our reconstruction pipeline on scenes with ground truth calibration and on challenging datasets with false epipolar geometries. Compared to other approaches, we achieve better or similar precision with significantly shorter running times and better scalability. Figure 1 illustrates meshing [34] after calibrating with our pipeline.

1.1. Related work

Estimating global rotations. Given the relative rotations R_{ij} between views i and j extracted from the essential matrices, computing the global rotation of each view R_i consists in solving the system $R_j = R_{ij}R_i$ for all i, j. This topic is covered by Hartley *et al*. [14].

This rotation averaging task can be performed by distributing the error along all cycles in a cycle basis, as done by Sharp *et al*. [28] for the alignment of range scans. Approximate solution using least square minimization for multi-view registration is proposed by Govindu [10], reused by Martinec *et al*. [18], and extended with semi-definite programming [3]. Alternatively, the averaging can be performed in the SO(3) Lie-group [11, 14]. Crandall *et al*. [5] use a cycle belief propagation, but they rely on known orientations, which do not make it suitable in the general case.

Cycle consistency. As relative R_{ij} estimates may contain outliers, rotation averaging has to be robust. Given the camera epipolar graph, the actual task is to identify both the global rotations and the inconsistent/outlier edges (false essential geometry). Two classes of methods stand out, based on spanning trees or cycles. The spanning tree approaches [12, 25] are based on the classic robust estimator scheme, RANSAC. Random spanning trees are sampled, and global putative rotations are computed by composing relative rotations while walking a spanning tree. The remaining edges, which create cycles, are evaluated based on the rotation angle of $R_j^T R_{ij} R_i$, measuring the discrepancy between the relative motion and the global motion. The solution with the largest cardinal is kept. Angle thresholds of 0.25° [12] or 1° [25] have been used.

Enqvist *et al*. [7] perform cycle removal based on deviation from identity. For this, the graph edges are weighted with the numbers of inlier correspondences and a maximum spanning tree (MST) is extracted. Cycles formed by the remaining edges are considered. A cycle is kept if the deviation from identity over the cycle, normalized by a factor $1/\sqrt{l}$ where l is the cycle lenth, is small enough. The method is highly dependent on the chosen MST; if this tree is erroneous, estimated rotations are wrong.

Zach *et al*. [37] use a Bayesian inference to detect incorrect relative rotation using cycle errors. A limit is set on the number of sampled trees and cycles to keep the problem tractable. The maximal cycle length is set to 6, also to avoid taking into account uncertainties w.r.t. cycle length.

Once global camera rotations R_i are estimated, global translations T_i can be computed. There are two main approaches, finding translations alone or with the structure.

Estimating translations alone. Govindu [10] proposes a method for recovering the unknown translations T_i from the heading vectors t_{ij}, extracted from the estimated essential matrices. He solves a least square problem with linear equations in the unknowns T_i and relative unknown scale factors λ_{ij}: $\lambda_{ij} t_{ij} = T_j - T_i$. Using random sampling, he tries to find the valid set of edges that best represents the global motion [12].

Sim *et al*. [29] propose a solution based on the heading vector extracted from the trifocal tensor that minimizes the angular error between the heading vector and the global camera position. The advantage of such a method is that they use a compact formulation ($3 \times$ number of camera variables) but they are highly dependent on the quality of the initial translation estimates. Arie-Nachimson *et al*. [3] use a least square minimization of the epipolar equation to find the unknown translations. The obvious drawback is the assumption that there is no outlier correspondence as all corresponding point pairs are used. Moreover, Rodríguez *et al*. [26] show that this method can handle neither colinear series of views nor shared optical centers.

Estimating both translations and 3D points. The joint estimation of translations and 3D points can be formulated using second-order cone programming expressing the problem with the l_∞ norm, as proposed by Hartley and Shaffalitzky [15], and later generalized [16]. Such methods rely on upper constraints on the residual error of feature points and rapidly involve a large number of unknowns. They are computationally and memory expensive. The solution is globally optimal thanks to multiple convex optimizations, using bisections of a quasi-convex problem.

Dalalyan *et al*. [6] deal with outliers with formulation using l_1 constraints instead of l_2 cones. It relies on two linear programs, the first one identifying outliers, and the second one solving translations and 3D structure on the selected inliers. It avoids the use of the non-negative slack variables in the single step procedure used by Olsson *et al*. [25] as adding one slack variable per measurement rapidly increases the problem size with the number of images.

Those l_∞ problems can be solved faster. Seo *et al*. [27] find a global solution by using a growing feasible subset while all the residual errors of the measurements are under the precision of the subset. This approach is faster because only a subpart of the data is fed to the l_∞ minimization. However, it is not robust to outliers. Agarwal *et al*. [2] test different bisection schemes and show that the Gugat algorithm [13] converges faster to the global solution. Zach *et al*. [38] use a proximal method to speed up the minimization of such convex problems.

Other approaches. Martinec *et al*. [18] use their global pipeline many times to iteratively discard two-view geometries with largest residuals. To keep good running time, they compute the translation and structure just on a few point pairs: each epipolar geometry is represented by 4 points only. Courchay *et al*. [4] use a linear parametrization of a tree of trifocal tensors over the epipolar graph to solve the camera position. The method is restricted to a single cycle.

1.2. Our global method for global calibration

Our input is an unordered set of pictures $\{I_1, \dots, I_n\}$. The internal calibration parameters K_i are assumed known for each camera: our goal is to robustly recover the global pose of each camera (absolute motion rotation R_i and translation T_i) from relative camera motions (rotation R_{ij} and heading translation vector t_{ij}) between images I_i and I_j.

Our contributions are the following:

1. We show that an iterative use of the Bayesian inference of Zach *et al*. [37], adjusted with the cycle length weighting of Enqvist *et al*. [7], can remove most outlier edges in the graph, allowing a more robust estimation of absolute rotations R_i (Section 2).
2. We present a new trifocal tensor estimation method based on l_∞ norm, resulting in a linear program, which, used as minimal solver in an adaptive RANSAC algorithm, is efficient and yields stable relative translation directions t_{ij} (Section 3).
3. We propose a new translation registration method, that estimates the relative translation scales λ_{ij} and absolute translations T_i, based on the l_∞ norm, resulting also in an efficient linear program (Section 4).
4. We put together these ingredients into an SfM pipeline (Section 5) that first cleans up an epipolar graph from outliers, then computes the global motions from the relative ones. Our experiments show its robustness, accuracy and scalability (Section 6)[1].

2. Robust estimation of global rotations

For matching points X and X' in images I_i and I_j respectively, the two-view epipolar constraint can be written

$$(K_i^{-1}X)^T E_{ij} (K_j^{-1}X') = 0. \quad (1)$$

The five-point algorithm of Nistér [23] inserted as minimal solver in a RANSAC procedure robustly estimates the essential matrices $E_{ij} = [t_{ij}]_\times R_{ij}$, from which R_{ij} can be extracted, together with the *direction* t_{ij}, since the scale is arbitrary. Four different motions (R_{ij}, t_{ij}) actually have to be tested; the one yielding the largest count of points satisfying the cheirality constraint (positive depth of the 3D point) is retained. It is important to note that the rotation accuracy is nearly insensitive to the baseline [7], contrary to the translation direction. Besides, although the camera rotations between connected views can be chained, the relative translations cannot since they are available up to a differing unknown scale factor λ_{ij}.

We identify inconsistent relative rotations in the graph using the edge disambiguation of Zach *et al*. [37]. As preliminary experiments showed that a number of outlier rotations could pass Zach *et al*.'s test, we made two improvements. First, we adapted the cycle error probability using the results of Enqvist *et al*. [7], weighting errors by a factor $1/\sqrt{l}$ where l is the length of the cycle. Second, we iterate Zach *et al*.'s algorithm until no more edge is removed by the Bayesian inference procedure. Finally, we check all the triplets of the graph and reject the ones with cycle deviation to identity larger than 2°. Experiments in Table 1 show that half of the outliers can remain after the first Bayesian inference, which motivates our iterated elimination.

Global rotations are computed as done by Martinec *et al*. [18], with a least-square minimization that tries to satisfy equations $R_j = R_{ij} R_i$, followed by the computation of the nearest rotation to cover the lack of orthogonality constraint during minimization.

Dataset \ #Iterations	1	2	3	2° check
Orangerie (Fig. 5)	8	4	1	9
Opera (Fig. 1 top)	7	3	—	125
Pantheon (Fig. 1 bottom)	9	2	—	7

Table 1. Number of edges rejected by Bayesian inference iteration.

3. Relative translations from trifocal tensors

To improve robustness and accuracy when computing the relative motion between cameras, we consider triplets

[1] More extensive experiments are provided as supplementary material.

of views instead of pairs as usual. We show in Section 3.2 that this yields a precision jump of an order of magnitude in the estimated translations.

3.1. Robust trifocal tensor with known rotations

Given estimated global rotations R_i, as computed in Section 2, we estimate a "reduced" trifocal tensor using an adaptive RANSAC procedure to be robust to outlier correspondences. Rather than minimizing an algebraic error having a closed form solution as Sim *et al*. [29], we minimize the l_∞ reprojection error of 3D points X_j compared to the observed points $\{(x_j^i, y_j^i)\}_{i\in\{1,2,3\}}$ in the three images:

$$\rho(t_i, X_j) = \left\| (x_j^i - \frac{R_i^1 X_j + t_i^1}{R_i^3 X_j + t_i^3}, y_j^i - \frac{R_i^2 X_j + t_i^2}{R_i^3 X_j + t_i^3}) \right\|, \tag{2}$$

where t_i is the translation of view i and t_i^m its components. The tensor is found by the feasibility of this linear program:

$$\begin{aligned} \underset{\{t_i\}_i, \{X_j\}_j, \gamma}{\text{minimize}} \quad & \gamma \\ \text{subject to} \quad & \rho(t_i, X_j) \leq \gamma, \quad \forall i, j \\ & R_i^3 X_j + t_i^3 \geq 1, \quad \forall i, j \\ & t_1 = (0, 0, 0). \end{aligned} \tag{3}$$

The second constraint ensures that all 3D points are in front of the cameras and the third one defines an origin for the local coordinate system of the triplet of views.

In general, using a linear program can lead to two issues. First, as the number of variables increases, the solving time grows polynomially [27]. Second, robustness to outliers is typically achieved with slack variables [25], which makes the problem even bigger.

Our approach consists in computing the tensor using a small-size linear program as minimal solver with four tracked point across the three views, in conjunction with the AC-RANSAC framework [21] to be robust to noise and outliers. This variant of RANSAC relies on *a contrario* (AC) methodology to compute an adaptive threshold for inlier/outlier discrimination: a configuration is considered meaningful if its observation in a random setting is unexpected. While global l_∞ minimization aims at finding a solution with the lowest γ value, found by bisection, AC-RANSAC determines the number of false alarms (NFA):

$$\text{NFA}(M, k) = (n-4) \binom{n}{k} \binom{k}{4} e_k(M)^{k-4} \tag{4}$$

where M is a tested trifocal tensor obtained by the minimal solver using four random correspondences, $\gamma = 0.5$ pixel, n is the number of corresponding points in the triplet, and where $e_k = \epsilon_k / \max(w, h)$ depends on the k-th error:

$$\epsilon_k = k^{\text{th}} \text{ smallest element of } \{\max_i \rho(t_i(M), X_j)\}_j. \tag{5}$$

#3D Points	Running time (s)		Angle accuracy (°)	
	Slack variables	AC	Slack variables	AC
200	1.37	0.09	0.07	0.03
400	4.06	0.11	0.06	0.03
600	7.94	0.13	0.04	0.02
800	13.1	0.15	0.03	0.02
1000	19.6	0.16	0.03	0.02

Table 2. Required time and accuracy (average angle of translation directions with ground truth) in robust estimation of trifocal tensor with the global formulation using slack variables [25] and our *a contrario* method (linear program combined with AC-RANSAC).

In these formulas, w and h are the dimensions of the images and e_k is the probability of a point having reprojection error at most ϵ_k. X_j is obtained by least-square triangulation of the corresponding points $\{(x_j^i, y_j^i)\}_{i\in\{1,2,3\}}$. k represents a hypothesized number of inliers. In (4), $e_k(M)^{k-4}$ is therefore the probability that the $k-4$ minimal reprojection errors of uniformly distributed independent corresponding points in the three images (our background model) have error at most ϵ_k, playing the role of the optimal γ of (3) for the inliers. The other terms in (4) define the number of subsets of k inliers among the $n-4$ remaining points. Thus $\text{NFA}(M, k)$ is the expectation of random correspondences having maximum error $\gamma = \epsilon_k$. The trifocal tensor M is deemed meaningful (unlikely to occur by chance) if:

$$\text{NFA}(M) = \min_{5\leq k\leq n} \text{NFA}(M, k) \leq 1. \tag{6}$$

In practice, we draw at most $N = 300$ random samples of 4 correspondences and evaluate the NFA of the associated models. As Moisan *et al*. [19], as soon as a meaningful model M is found, we stop and refine it by resampling $N/10$ times *among the inliers* of M. If no sample satisfies (6), we discard the triplet. Finally, we refine the translations and the k inlier 3D points by bundle adjustment.

Table 2 evaluates the computation time and accuracy of our robust *a contrario* trifocal estimation compared to the equivalent global estimation with slack variables [25] on synthetic configurations. A uniform 1-pixel noise is added to each perfect correspondence and 2% outliers are introduced. We evaluate the accuracy of the results (angular error between ground truth and computed translation) and the required time to find the solution. The global solution finds a solution that fits the noise of the data, but AC-RANSAC is able to go further and find a more precise solution.

3.2. Relative translation accuracy

Following experiments of Enqvist *et al*. [7] concerning two-view rotation precision, we demonstrate that using a trifocal tensor can lead to substantial improvement in the relative translation estimation. To assess the impact of small baseline, a simple synthetic experiment is performed. A set

of fifty 3D points are randomly generated in a $[-1,1]^3$ cube and 3 cameras are placed on a circle at distance 5, at angles $0°$, α and 2α respectively (see Figure 2, left). We vary α from $1°$ to $20°$ to simulate small to medium baselines. A uniform 1-pixel noise is added to image projections. The relative translation of the camera is estimated using AC-RANSAC with a 5-point solver (essential matrix) and with the AC-RANSAC trifocal tensor with known rotations. We compare the angular error w.r.t. the ground truth (since the reconstruction scale is arbitrary) for the directions of the three relative translations and average the results over 50 runs (see Figure 2, right). With an increasing baseline, the accuracy improves with both methods. However, our trifocal estimation performs much better, with good results even at small baselines.

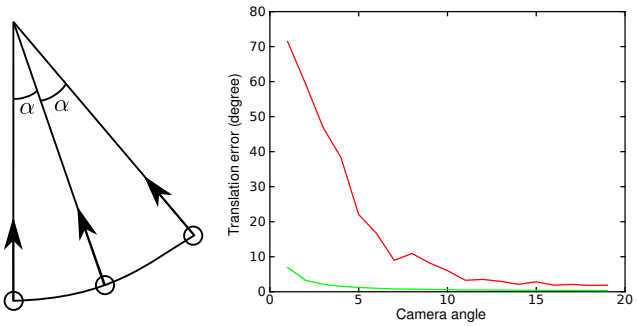


Figure 2. Left: camera position (circle) relative to synthetic scene (top) in $[-1,1]^3$ cube at distance 5. Right: translation direction error relative to ground truth as a function of camera angle α in degrees: using a standard essential matrix estimation (red) and using our trifocal tensor (green).

4. Translation registration

Given a set of relative motions pairs (R_{ij}, t_{ij}) (rotations and translation directions), we want to find the global location $(T_1, \ldots, T_n)$ of all cameras, as illustrated on top of Figure 3. We are thus looking for n global translations and $\#t_{ij}$ scale factors λ_{ij} that reconcile the different translation directions into a global coordinate frame:

$$\|T_j - R_{ij}T_i - \lambda_{ij}t_{ij}\| = 0, \quad \forall i,j. \tag{7}$$

Due to noise, the set of equations (7) cannot be satisfied exactly but the solution of the linear set of equations can be optimized in the least square sense [10]. The problem is that, with this formulation, the λ_{ij} cannot be constrained to be positive to respect cheirality.

Our approach consists in optimizing equations (7) under the l_∞ norm. As the solution is invariant under translation and scaling, the degrees of freedoms are removed by adding positivity constraints over the λ_{ij} (scale ambiguity) and setting the first camera at origin (translation ambiguity). The

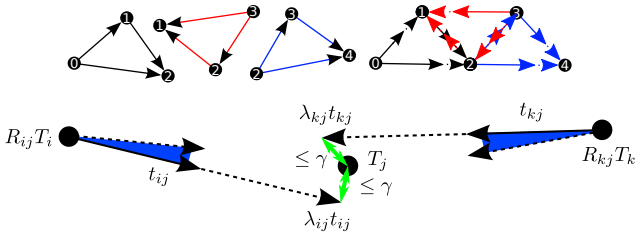


Figure 3. Top left: 3 local tensors. Top right: Merged translations. Bottom: Our approach minimizes Euclidean distances (green arrows) while Sim and Hartley [29] minimize the (blue) angles.

following linear program yields the global optimal solution:

$$\begin{aligned} \underset{\{T_i\}_i, \{\lambda_{ij}\}_{ij}, \gamma}{\text{minimize}} \quad & \gamma \\ \text{subject to} \quad & |T_j - R_{ij}T_i - \lambda_{ij}t_{ij}| \leq \gamma, \quad \forall i,j \\ & \lambda_{ij} \geq 1, \quad \forall i,j \\ & T_1 = (0,0,0). \end{aligned} \tag{8}$$

In our case, we may have different translation directions t_{ij} for a given (i,j) if it belongs to several triplets (see Figure 3, top). We thus consider t_{ij}^τ for a tripet τ containing (i,j). Besides, the relative scales are per triplet (λ^τ) rather than per edge (λ_{ij}). We actually solve the following problem:

$$\begin{aligned} \underset{\{T_i\}_i, \{\lambda_\tau\}_\tau, \gamma}{\text{minimize}} \quad & \gamma \\ \text{subject to} \quad & |T_j - R_{ij}T_i - \lambda_\tau t_{ij}^\tau| \leq \gamma, \quad \forall \tau, \forall (i,j) \in \tau \\ & \lambda_\tau \geq 1, \quad \forall \tau \\ & T_1 = (0,0,0). \end{aligned} \tag{9}$$

Compared to Govindu's approach [10], that does not enforce cheirality conditions, we are sure to find a global optimum. Also, we minimize here a linear program, which is much faster than the SOCP of Sim and Hartley [29]. They use angular errors, whereas our method involves simpler constraints using Euclidean distance (see Figure 3, bottom).

5. The global reconstruction process

We now show how to use these elements in a pipeline to perform robust and accurate global calibration. Our method consists in the following steps: (1) build the epipolar graph of matching image pairs and estimate the essential matrices; (2) check rotation consistency, performing Bayesian inferences on the graph and computing global rotations on the resulting graph; (3) compute relative translations from trifocal tensors; (4) register translations in a global coordinate frame; (5) compute coarse triangulated structure and refine rotations, translations and structure via bundle adjustment.

Step 1: relative pairwise rotation. We use SIFT matching [17] and robust *a contrario* essential matrix estimation [21]. An upper bound of the possible *a contrario* pre-

cision is set to 4 pixels. The epipolar graph is actually split into 2-edge-connected components, for which separate global calibrations can be computed. The resulting poses and structures are later merged (rotated, translated, scaled) based on standard pose estimation/resection.

Step 2: rotation consistency. We use our adapted Zach *et al.*'s Bayesian inference [37] to remove outlier rotations (see Section 2). As in Step 1, the graph is checked and cleaned if necessary. We list the cycles of length 3 in the graph as possible triplets. Those having a chained rotation whose angular discrepancy to identity is greater than 2° are discarded. Finally, the global rotations are computed using a sparse eigenvalue solver as done by Martinec *et al*. [18], which has been shown [8] to be as efficient as constraining the orthogonality during minimization.

Step 3: relative motion computation. We compute a relative translation estimate for each edge of the graph, knowing its rotation. For this, we first buld a list of all graph edges. Then for each edge, we try to solve as described in Section 3 the triplet with the largest support of tracks to which the edge belongs. Tracks are computed by using the fast union-find approach [20]. If the triplet is solved, we validate the three edges that belong to the tensor and remove them from the list. If trifocal tensor estimation fails, we continue with other triplets containing this edge, if any, in decreasing order of the number of tracks. The process stops when the list of edges is empy. This step not only finds relative translations but also determines coherent 3D structures per triplet. One advantage of this approach is that triplets can be computed in parallel. This method requires having a graph covered by contiguous triplets, which might not always apply. However, it is often the case in practice, in part thanks to ever-improving feature detector repeatability.

Step 4: translation registration. We integrate the relative translation directions and compute global translations using the l_∞ method of Section 4.

Step 5: final structure and motion. The preceding steps provide a good estimation of the motions, as well as structures per triplet. We link points in these structures by feature tracking [20] and compute 3D point positions per track by triangulation. This global structure and the translations are then refined by bundle adjustment (BA) using the Ceres solver [1]. Interestingly, the BA converges in few iterations, which assesses the quality of our initial estimate. A final BA is used to refine the structure and camera rotation and translation to handle noisy rotation estimates of Step 2. Proceedings in two such steps, first with a partial BA with fixed rotations, then with a global BA, is inspired by Olsson and Enqvist's approach [25]. The idea is to prevent compensation of translation errors in the first step by rotation adjustment, since rotations are more reliable. According to our experiments, the two-stage BA improves the precision.

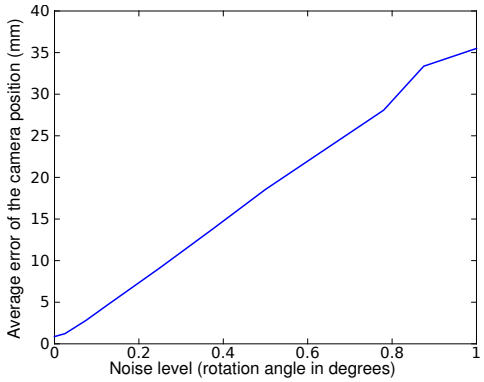


Figure 4. Global translation accuracy as a function of noise in the global rotations.

6. Results

To assess our method, we used Strecha *et al.*'s benchmark [33], which provides ground-truth camera poses. We also experimented with challenging datasets having multiple false epipolar geometries, mainly due to repeated parts. We compared with incremental methods, Bundler [31] and VisualSfM [35], as well as global methods, that of Olsson and Enqvist [25] and Arie-Nachimson *et al*. [3]. All reported figures have been obtained with the authors' software on an 8-core 2.67 GHz machine, except for Arie-Nachimson *et al*., for which only published results are available [3].

Sensitivity to rotation noise. To study the sensitivity of our global translation estimation w.r.t. noisy global rotations, we feed the translation registration process with the ground truth rotations of the fountainP11 dataset [33] altered by a small random rotation whose axis is uniformly sampled on a sphere and whose angle is uniformly drawn between 0 and a given maximum angle. Figure 4 shows that the final error is almost linearly dependent on the noise level. Similar results are shown by Sim and Hartley [29].

Accuracy. Table 3 (left) shows the average baseline error of several incremental and global methods on Strecha *et al.*'s dataset [33]. A 3D similarity registers the ground truth and the computed coordinate frame. While our accuracy is slightly better or comparable to the top performer on the first four datasets, it is remarkably better on the Castle datasets, which feature a loop in a courtyard. One of the reasons is a good rejection of outlier data (wrong point correspondences and false epipolar geometry).

Running time. Table 3 (right) reports the running time of calibration (estimation of camera poses and 3D structure) after epipolar graph computation. Our global method is 5 to 11 times faster (16 to 26 times when parallelized) than that of Olsson and Enqvist [25] (Matlab code with time-critical parts in C++). It is even competitive with the fastest incremental method, which is GPU- and multicore-optimized.

	Accuracy (mm)					Running times (s)						
Scene	Ours	Bundler [31]	VSfM [35]	Olsson [25]	Arie [3]	Ours	OursP	Bundler [31]	VSfM [35]	Olsson [25]	Ratio [25]/Ours	Ratio [25]/OursP
FountainP11	2.5	7.0	7.6	**2.2**	4.8	12	5	36	3	133	11.1	26
EntryP10	**5.9**	55.1	63.0	6.9	N.A.	16	5	16	3	88	5.5	17
HerzJesusP8	**3.5**	16.4	19.3	3.9	N.A.	6	2	10	2	34	5.6	17
HerzJesusP25	**5.3**	21.5	22.4	5.7	7.8	47	10	100	12	221	4.7	22
CastleP19	**25.6**	344	258	76.2	N.A.	20	6	78	9	99	4.9	16
CastleP30	**21.9**	300	522	66.8	N.A.	55	14	300	18	317	5.7	22

Table 3. Left: Average position error, in millimeters, w.r.t. ground truth for different incremental [31, 35] and global [25, 3] SfM pipelines, given internal calibration. Right: running times in seconds and speed ratio. OursP means our parallel version.

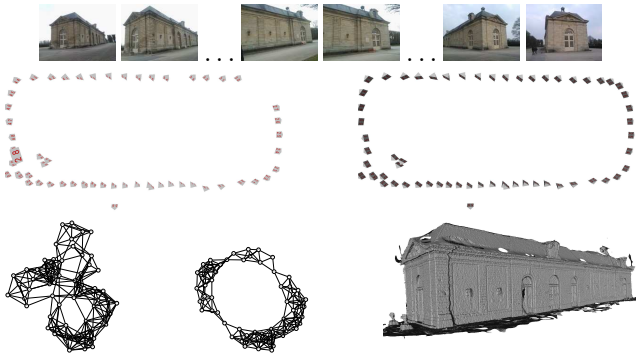


Figure 5. Top: excerpt of the Orangerie dataset. Center: Bundler camera positions (cycle failure), and ours. Bottom: input epipolar graph, our cleaned graph, and mesh obtained from our calibration.

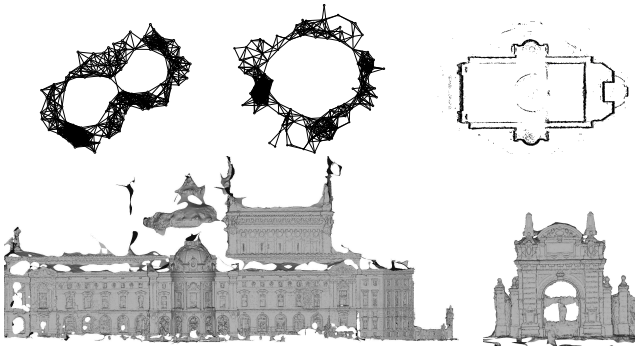


Figure 6. Opera dataset (160 images). Top: input epipolar graph (corrupted by façade symmetries), our cleaned graph, and calibration point cloud. Bottom: orthographic façade and close-up.

Challenging datasets. We tested with datasets featuring repeated or similar scene portions (similar façades, including mirror-imaged), which cause false geometries in the epipolar graph. We show the initial graph, the graph cleaned of false edges thanks to repeated Bayesian inference, and the camera positions. For the Orangerie dataset (61 images, see Figure 5), Bundler [31] is unable to close the loop and misplaces several views, contrary to our method. For the Opera dataset (160 images, see Figure 6), running times (after feature detection and matching) are strikingly different: Bundler runs in 3 hours while we calibrate in 7 minutes (4 minutes for the parallel version). In fact, Bundler spends a lot of time in the repeated bundle adjustments (BA); while they take less than one minute for the first images, they take about ten minutes for the last ones. Our running times and residual information are detailed in Table 4.

7. Conclusion

We have presented a global Structure from Motion system for large-scale 3D reconstruction from unordered image sets. We have shown that the global calibration can be performed by globally merging local relative estimates, preserving robustness and accuracy while ensuring scalability. Moreover we have shown that triplets of relative translations from known rotations can be computed with a good, adaptive accuracy at affordable computation time. These results have been supported by theoretical arguments as well as experimental comparison on synthetic and real datasets.

Our pipeline presents many advantages. It computes a stable structure by merging the tracks at trifocal level and is almost outlier free (less risk of merging false epipolar geometries). As the chain is global, it is not necessary to provide an initial pair — the highly problematic initial seed of incremental methods. Thanks to the good quality of the relative translation estimates, the precision of the global translation provides a fairly good overview of the camera positions even before refinement through bundle adjustment; we do not need to compute the global structure of the scene for that. This is confirmed by the very low number of iterations performed by the bundle adjustment. Our experiments show that the issue that is limiting the precision of our global approach is the precision of the global rotations. We believe that our method could work at city scale even on a standard computer, provided there is enough RAM for the final bundle adjustments, which is optional.

Acknowledgments. This work has been carried out in IMAGINE, a joint research project between École des Ponts

	Triplets			Translation registration				BA_1		BA_2		Total
Dataset	#possible	#solved	time	$\#t_{ij}$	time	γ	$\bar{\rho}$	#iter	$\bar{\rho}$	#iter	$\bar{\rho}$	time
FountainP11	78	28	2	84	< 1	5×10^{-4}	0.75	2	0.26	3	0.25	5
HerzJesusP25	522	102	4	306	< 1	5×10^{-4}	0.85	2	0.47	4	0.46	10
CastleP30	540	103	6	309	1	3×10^{-3}	2.3	2	0.51	3	0.27	14
Opera	3054	588	30	1764	41	1×10^{-2}	5.47	5	1.05	10	0.48	207

Table 4. Running time (s) with our parallel version, and mean reprojection errors $\bar{\rho}$ (pixels) of all 3D points of all cameras.

ParisTech (ENPC) and the Scientific and Technical Centre for Building (CSTB). It was supported by Agence Nationale de la Recherche ANR-09-CORD-003 (Callisto project).

References

[1] S. Agarwal and K. Mierle. *Ceres Solver: Tutorial & Reference*. Google Inc.

[2] S. Agarwal, N. Snavely, and S. M. Seitz. Fast algorithms for l_∞ problems in multiview geometry. In *CVPR*, 2008.

[3] M. Arie-Nachimson, S. Z. Kovalsky, I. Kemelmacher-Shlizerman, A. Singer, and R. Basri. Global motion estimation from point matches. In *3DIMPVT*, 2012.

[4] J. Courchay, A. S. Dalalyan, R. Keriven, and P. F. Sturm. On camera calibration with linear programming and loop constraint linearization. *IJCV*, 97(1):71–90, 2012.

[5] D. J. Crandall, A. Owens, N. Snavely, and D. Huttenlocher. Discrete-continuous optimization for large-scale structure from motion. In *CVPR*, 2011.

[6] A. Dalalyan and R. Keriven. Robust estimation for an inverse problem arising in multiview geometry. *JMIV*, 43(1):10–23, 2012.

[7] O. Enqvist, F. Kahl, and C. Olsson. Non-sequential structure from motion. In *ICCV Workshops*, pages 264–271, 2011.

[8] J. Fredriksson and C. Olsson. Simultaneous multiple rotation averaging using lagrangian duality. In *ACCV*, 2012.

[9] R. Gherardi, M. Farenzena, and A. Fusiello. Improving the efficiency of hierarchical structure-and-motion. In *CVPR*, 2010.

[10] V. M. Govindu. Combining two-view constraints for motion estimation. In *CVPR*, 2001.

[11] V. M. Govindu. Lie-algebraic averaging for globally consistent motion estimation. In *CVPR*, 2004.

[12] V. M. Govindu. Robustness in motion averaging. In *ACCV*, 2006.

[13] M. Gugat. Fast algorithm for a class of generalized fractional programs. *Man. Sci.*, 1998.

[14] R. Hartley, J. Trumpf, Y. Dai, and H. Li. Rotation averaging. *IJCV*, 103(3):267–305, 2013.

[15] R. I. Hartley and F. Schaffalitzky. l_∞ minimization in geometric reconstruction problems. In *CVPR*, 2004.

[16] F. Kahl and R. I. Hartley. Multiple-view geometry under the l_∞-norm. *IEEE Trans. PAMI*, 30(9):1603–1617, 2008.

[17] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[18] D. Martinec and T. Pajdla. Robust rotation and translation estimation in multiview reconstruction. In *CVPR*, 2007.

[19] L. Moisan and B. Stival. A probabilistic criterion to detect rigid point matches between two images and estimate the fundamental matrix. *IJCV*, 57(3):201–218, 2004.

[20] P. Moulon and P. Monasse. Unordered feature tracking made fast and easy. In *CVMP*, 2012.

[21] P. Moulon, P. Monasse, and R. Marlet. Adaptive structure from motion with a contrario model estimation. In *ACCV*, 2012.

[22] K. Ni and F. Dellaert. HyperSfM. In *3DIM/3DPVT*, 2012.

[23] D. Nistér. An efficient solution to the five-point relative pose problem. In *CVPR*, volume 2, 2003.

[24] D. Nistér and H. Stewénius. Scalable recognition with a vocabulary tree. In *CVPR*, 2006.

[25] C. Olsson and O. Enqvist. Stable structure from motion for unordered image collections. In *SCIA*, 2011. LNCS 6688.

[26] A. L. Rodríguez, P. E. López-de Teruel, and A. Ruiz. Reduced epipolar cost for accelerated incremental SfM. In *CVPR*, 2011.

[27] Y. Seo and R. I. Hartley. A fast method to minimize l_∞ error norm for geometric vision problems. In *ICCV*, 2007.

[28] G. C. Sharp, S. W. Lee, and D. K. Wehe. Toward multiview registration in frame space. In *ICRA*, 2001.

[29] K. Sim and R. Hartley. Recovering camera motion using l_∞ minimization. In *CVPR*, volume 1, 2006.

[30] S. N. Sinha, D. Steedly, and R. Szeliski. A multi-stage linear approach to structure from motion. In *ECCV Workshops (1)*, 2010.

[31] N. Snavely, S. M. Seitz, and R. Szeliski. Photo tourism: exploring photo collections in 3D. *TOG*, 25(3):835–846, 2006.

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

[33] C. Strecha, W. von Hansen, L. J. Van Gool, P. Fua, and U. Thoennessen. On benchmarking camera calibration and multi-view stereo for high resolution imagery. In *CVPR*, 2008.

[34] H.-H. Vu, P. Labatut, J.-P. Pons, and R. Keriven. High accuracy and visibility-consistent dense multiview stereo. *Pattern Analysis and Machine Intelligence*, 2012.

[35] C. Wu. Towards linear-time incremental structure from motion. In *3DV*, 2013.

[36] C. Wu, S. Agarwal, B. Curless, and S. M. Seitz. Multicore bundle adjustment. In *CVPR*, 2011.

[37] C. Zach, M. Klopschitz, and M. Pollefeys. Disambiguating visual relations using loop constraints. In *CVPR*, 2010.

[38] C. Zach and M. Pollefeys. Practical methods for convex multi-view reconstruction. In *ECCV*, 2010.