# Robust Structure From Motion Based on Relative Rotations and Tie Points

X. Wang, F. Rottensteiner, and C. Heipke

## Abstract

*In this article, we present two new approaches for image orientation with a focus on robustness, starting with relative orientations of available image pairs, an incremental and a global one, and compare their performance. For the incremental approach, we first choose a suitable initial image pair, and we then iteratively extend the image cluster by adding new images. The rotations of these newly added images are estimated from relative rotations by single rotation averaging. In the next step, a linear equation system is set up for each new image to solve the translation parameters with triangulated tie points that can be viewed in that new image, followed by a resection for refinement. Finally, we refine the orientation parameters of the images by a local bundle adjustment. We also present a global method that consists of two parts: global rotation averaging, followed by setting up a large linear equation system to solve for all image translation parameters simultaneously; a final bundle adjustment is carried out to refine the results. We compare these two methods by analyzing results on different benchmark sets, including ordered and unordered image data sets from the Internet and two other challenging data sets to demonstrate the performance of our two approaches. We conclude that while the incremental method typically yields results of higher accuracy and performs better on the challenging data sets, our global method runs significantly faster.*

## Introduction

In recent years, surveying and mapping showed a lot of interest in automatic 3D modeling of architectural and urban areas from images. The determination of image orientation via automatically determined tie points (also called structure from motion [SfM]) is a prerequisite to realize this task. Various methods have been suggested to solve this problem (Snavely *et al.* 2006; Agarwal *et al.* 2009; Wu 2013). SfM can be divided into three categories: incremental SfM, hierarchical SfM, and global SfM. Incremental SfM (Snavely *et al.* 2006; Wu 2013; Schönberger and Frahm 2016; Wang *et al.* 2018) is the earliest idea. Two images or triplets are initially chosen according to some specific requirements; their relative orientation parameters are computed, new images are iteratively added by space resection (also called the perspective-n-point problem [PnP]) and triangulation, and a robust bundle adjustment is typically adopted to obtain reliable results. The above procedure is repeated until no more images can be added. Incremental SfM is relatively robust against outliers because these can be detected and removed incrementally when adding new images. However, due to the repeated use of bundle adjustment, it is rather slow. To overcome this problem, hierarchical SfM (Farenzena *et al.* 2009; Havelena *et al.* 2009; Mayer 2014; Toldo *et al.* 2015) was proposed. The basic idea is to divide the whole data set into several overlapping subsets that are reconstructed independently using incremental methods. Finally, all reconstructions are merged and optimized by bundle adjustment. Global SfM (Govindu 2001; Martinec and Pajdla 2007; Jiang *et al.* 2013; Moulon *et al.* 2013; Ozyesil and Singer 2015; Arrigoni *et al.* 2016; Reich and Heipke 2016; Goldstein *et al.* 2016; Wang *et al.* 2019) considers this problem from a different perspective. Global SfM draws on the well-known idea that rotation and translation estimation (i.e., the computation of the 3D coordinates of the projection center) can be separated. Accordingly, these methods consist of two main steps: global rotation averaging and global translation estimation. Global rotation averaging simultaneously estimates the rotation matrices of all images in a consistent (global) coordinate system (Hartley *et al.* 2013). Given global rotations, global translation estimation aims at simultaneously solving the translation parameters of all images. The advantage of global SfM is that it can solve both rotations and translations without using intermediate bundle adjustment, only a final one is necessary. However, it is more sensitive to outliers than the other methods.

We are most interested how incremental and global methods compare with respect to robust and time-efficient solutions; to this end, we propose and investigate novel incremental and global SfM approaches in this article. Figure 1 shows the work flows of our methods. We first extract features from all images and perform relative orientation of all image pairs; for unordered sets, we first determine image similarity using the method described in (Wang *et al.* 2017). Then, for the incremental approach, an initial image pair is chosen, and clusters of new images are iteratively added and oriented by single rotation averaging and linear translation estimation (see pointed box in Figure 1). Subsequently, new scene points are triangulated, and a local bundle adjustment is used to refine the results. The global method uses the two steps of global rotation estimation and global translation estimation (see dashed box in Figure 1), both making sure that blunders are detected and eliminated beforehand.

The main contribution of this article is threefold. First, as part of the incremental approach, we adopt single rotation averaging to estimate the new image rotation matrix. Second, again for the incremental approach, we set up a linear equation system with only two tie points that can be seen in the new images to calculate the translation parameters. Finally, inspired by the second contribution, if the global rotation matrices can be provided in one way or another, we set up a linear equation system that solves all image translation parameters simultaneously. The L1 norm (minimization of the sum of the absolute values of the residuals) is chosen to solve the above optimization, as it is more robust than the L2 norm (least squares). We evaluate the performance of our approaches' w.r.t. accuracy and time efficiency using various

Wang, Rottensteiner, and Heipke are with the Institute of Photogrammetry and GeoInformation, Leibniz Universität Hannover,Nienburger Str. 1, D-30167 Hannover, Germany (wang@ipi.uni-hannover.de).

Photogrammetric Engineering & Remote Sensing
Vol. 85, No. 5, May 2019, pp. 347–359.
0099-1112/18/347–359

doi: 10.14358/PERS.85.5.347

benchmark data sets. Additional experiments on large data sets, including unordered images from the Internet, demonstrate further capabilities of our approaches.

The remainder of this article is structured as follows. The second section discusses related work. In the third section, we introduce our method for estimating rotation matrices and translation parameters by single rotation averaging and solving a linear equation system, respectively. The fourth section describes our global translation estimation by using the tie points only after global rotation averaging. In the fifth section, we report the results of experiments on a number of data sets to evaluate our methods. Finally, the sixth section concludes our work.

## Related Work

In this section, we review related work on SfM. We discuss the classical space resection (PnP)  problem, rotation averaging, and translation estimation.

### Space Resection

Space resection aims at determining the rotation and translation for one perspective image of known interior orientation parameters from $n \geq 3$ points given their 3D coordinates in object space and their corresponding 2D image coordinates. The most common approach is, of course, based on the collinearity equations and iterative least-squares adjustment; it needs three 3D points if the interior orientation parameters are given. Kilian (1955) gives a direct solution requiring four 3D points, using the additional observations to linearize the problem without requiring initial values. The direct linear transformation (DLT) (Abdel-Aziz and Karara 1971) is another well-known solution. DLT solves for interior and exterior orientation simultaneously, needing six 3D points. Using projective geometry, Hartley and Zisserman (2003) suggested a two-stage procedure for calibrated cameras. They first applied the calibration matrix to the image coordinates, which turns the projection matrix into a so-called image pose matrix. The exterior orientation parameters are then calculated from the pose matrix. Zheng *et al.* (2013) revisited the problem by applying Gröbner bases. Both these methods were demonstrated to give accurate results provided that at least three non-collinear 3D points are given.

### Rotation Averaging

Rotation averaging attracted the attention of vision researchers since the work of Govindu (2001). There are two basic approaches: single rotation averaging and global rotation averaging (Hartley *et al.* 2011). Single rotation averaging computes the mean rotation of a set of rotations, while in global rotation averaging (Govindu 2001; Chatterjee and Govindu 2013; Reich *et al.* 2015, 2017), relative rotations $\boldsymbol{R}_{ij}$ are given for a set of images, and the global rotations of all images are computed simultaneously, satisfying all constraints $\boldsymbol{R}_{ij}\boldsymbol{R}_i = \boldsymbol{R}_j$. Govindu (2001) used quaternions to compute the global rotations by constrained least-squares optimization. Martinec and Pajdla (2007), Arie-Nachimson *et al.* (2012), and Moulon *et al.* (2013) studied this problem by considering the properties of rotation matrices; singular value decomposition was used to solve the corresponding linear equation system. Hartley *et al.* (2011) compared L1 and L2 averaging and demonstrated that the L1 norm performed better than the L2 norm by using the Weiszfeld algorithm. Chatterjee and Govindu (2013) started by propagating initial rotation values using a minimum spanning tree. The initial results were then optimized using the Lie algebra, taking advantage of the fact that rotation matrices form the special orthogonal group *SO(3)* (Hartley *et al.* 2013). Reich *et al.* (2015, 2016, 2017) solved the problem based on a convex relaxed semi-definite program that yields a more robust result. However, due to a breadth-first search, the method is rather computationally intensive.

### Translation Estimation

A number of approaches have recently been proposed for determining image translations from relative orientations and a set of globally consistent rotation values. They can be divided into two categories: (1) the combined use of tie points and relative translation information and (2) the exclusive use of 3D coordinates of tie points only. In the first group, Jiang *et al.* (2013) proposed a linear global approach using tie points of images' triplets to unify the scale factors of the related image pairs and then propagated these scale factors to the connected triplets. Given the relative translations, they set up and solved a global linear homogeneous equation system. They normally recover fewer images than other methods (Moulon *et al.* 2013; Wilson and Snavely 2014) because the triplets are required to be well connected. Wilson and Snavely (2014) presented a method

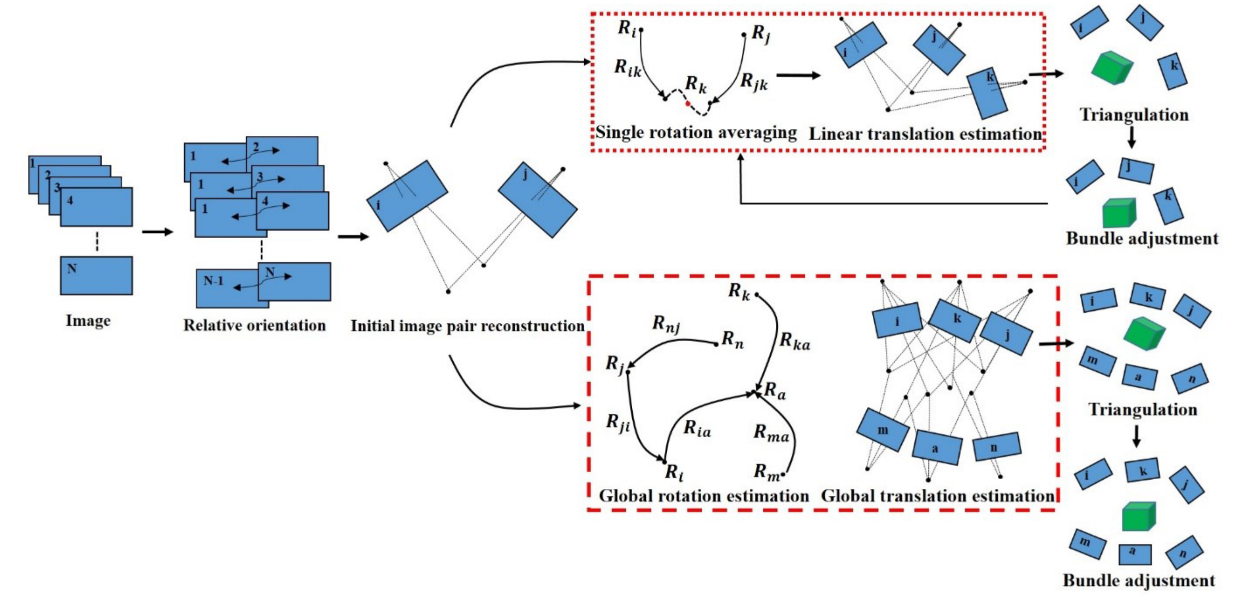


Figure 1. Workflow of image orientation approaches. The pointed box denotes the incremental method, and the dashed one presents the global method (Note that we only show an example of six images for simplicity).

called 1DSfM. They developed an outlier filter by projecting 3D information into different 1D spaces; the inliers of relative translations are then considered to constrain the translation parameters. Cui and Tan (2015) first used several strategies to detect outliers of relative orientations, solved the scale factors by unifying the tie points of all image pairs, and, finally, built a global linear equation system for translation estimation. Cui *et al.* (2015) used the constraint that the 3D coordinates of tie points visible in different image pairs must be identical to compute the translations in a unified coordinate system. The second group of methods, in which only the 3D coordinates of the tie points are used, is not as well studied because detecting outliers from tie points only is normally more difficult than detecting outliers of relative orientations. Cui *et al.* (2017) presented a hybrid SfM method combining a global and an incremental strategy; they estimated the rotation matrices by global rotation averaging, followed by an incremental linear translation estimation in which the rotation matrices remain constant.

The previously mentioned works determine the exterior orientation parameters of images, mostly without initial values. Some restrictions apply, however. For spatial resection, at least three non-collinear ground control points are needed, which may not always be available. Rotation averaging is relatively sensitive to outliers of relative rotations, and the same is true for the first category of translation estimation methods with respect to outliers in relative translation. Moreover, both groups of translation estimation methods can be negatively influenced by errors in the tie points.

Our incremental method needs only two 3D points for translation estimation. To reduce the impact of outliers during rotation determination, we present an iterative method to detect and eliminate them in our single rotation averaging scheme. We refine our rotation matrices by iteratively using space resection and local bundle adjustment, and we also use RANSAC (Fischler and Bolles 1981) to make the choice of tie points as robust as possible. To make our global method more robust, we present a triplet loop closure constraint to detect outliers of relative rotations, and tie points are robustly selected by considering tie point distribution in image space and the reprojection error of the corresponding individual local spatial intersection.

## Incremental Rotation and Translation Estimation

In this section, we present our strategy of choosing a good initial image pair, explain our procedure for calculating the rotation matrices of newly added images, and show how we compute the translation parameters via a linear equation system by minimizing the L1 norm.

### Overview of the Developed Incremental Procedure

As is well known, the selection of the initial pair can have a significant influence on the subsequent reconstruction. To obtain a good initial pair, we introduce two indicators: the number of matched features, which should be large, and the intersection angle, which should be close to 90°.

Given the individual images, we first derive SIFT features (Lowe 2004). Then, for each image pair ($i$, $j$), we compute the relative orientation parameters based on the five-point algorithm of Nistér (2004) with RANSAC, record all inliers, and compute the intersection angle for each matched feature pair ($\boldsymbol{p}_i$,$\boldsymbol{p}_j$). We choose the median of these angles as the intersection angle for the considered image pair. We keep all image pairs that fulfill two conditions: (1) they need to have more matches than a threshold (we use 50 in our work), and (2) at least a certain number of the matches must be inliers (we use 80%). Note that these two thresholds that are set empirically and work well in our experiments can determine the photogrammetric block's size; a small threshold will normally increase the size at the cost of including incorrect image pairs. Among the remaining pairs, the one with an intersection angle closest to 90° is the final choice for the initial image pair.

As long as the base line is not too small, the most obvious way to select the image to be added to the block next is to find the candidate that shares the largest number of correspondences with the images employed so far. As it is not very efficient to add only one image each time, we simultaneously add all images that fulfill two conditions: (1) a certain percentage (we use 60%) of the features extracted from the image have matches to already computed 3D points, and (2) the number of these features is above a threshold (we use 30 here). All images fulfilling these two conditions together with the images processed already are called a *cluster*. The size of the cluster is related to these two thresholds that are set empirically; the cluster size will enlarge with small thresholds, which, on the other hand, may be very dangerous because it might result in adding images with incorrect relative orientations into the cluster. Then, as discussed in the following sections, rotation averaging, translation estimation, and resection refinement are performed independently for each new image of the cluster. Recently added object points' coordinates are initially calculated by triangulation using the estimated exterior parameters (dashed box in Figure 2), followed by a local bundle adjustment using the whole cluster. In the next step, the cluster is extended by selecting images from the image set that is still remaining, and the procedure starts again. The procedure is visualised in Figure 2.

### Robust Rotation Estimation by Single Rotation Averaging

Given the rotation matrices of images that have already been added to the block and the relative rotations between those images and a newly added image, we calculate several

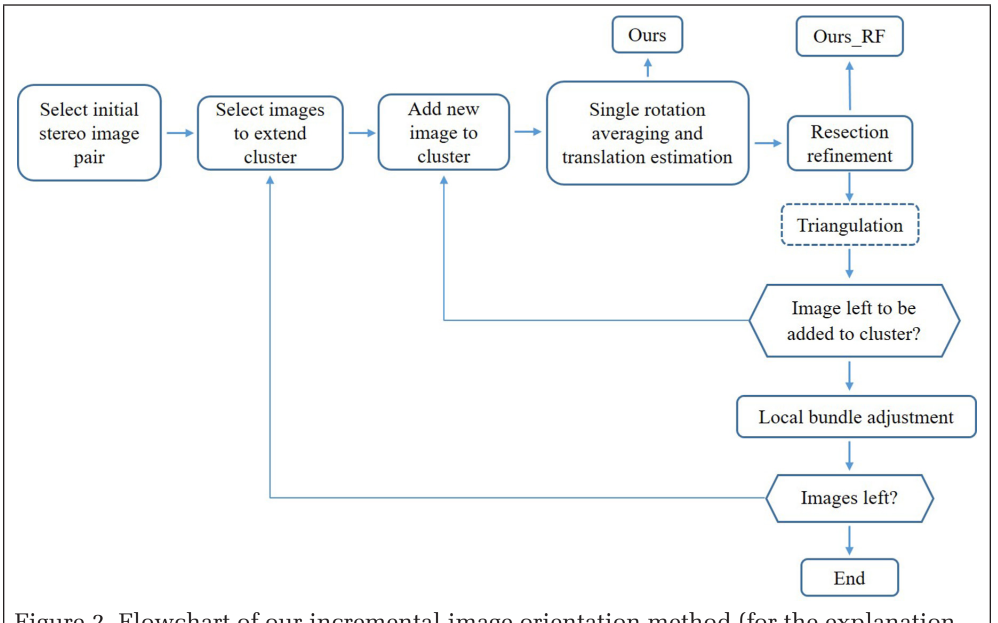

Ours
Ours_RF
Select initial stereo image pair
Select images to extend cluster
Add new image to cluster
Single rotation averaging and translation estimation
Resection refinement
Triangulation
Image left to be added to cluster?
Local bundle adjustment
Images left?
End

Figure 2. Flowchart of our incremental image orientation method (for the explanation of "Ours" and "Ours_RF," see the section "Evaluation of Three Benchmark Data Sets").

rotation matrices for the new image. With reference to Figure 3, let $\boldsymbol{R_a}$ be the rotation matrix of the new image that we want to estimate. $\boldsymbol{R_i}$, $\boldsymbol{R_j}$, $\boldsymbol{R_k}$, $\boldsymbol{R_m}$, $\boldsymbol{R_n}$ are the rotation matrices previously estimated for images *i*, *j*, *k*, *m*, and *n*, and $\boldsymbol{R_{ia}}$, $\boldsymbol{R_{ja}}$, $\boldsymbol{R_{ka}}$, $\boldsymbol{R_{ma}}$, $\boldsymbol{R_{na}}$ contain the relative rotations with respect to image *a* calculated by the five-point algorithm. We propagate the given rotations along these relative rotations to obtain different solutions for $\boldsymbol{R_a}$, namely $\boldsymbol{R_a^i}$, $\boldsymbol{R_a^j}$, $\boldsymbol{R_a^k}$, $\boldsymbol{R_a^m}$, $\boldsymbol{R_a^n}$. We want to average these rotation matrices and obtain a robust result.

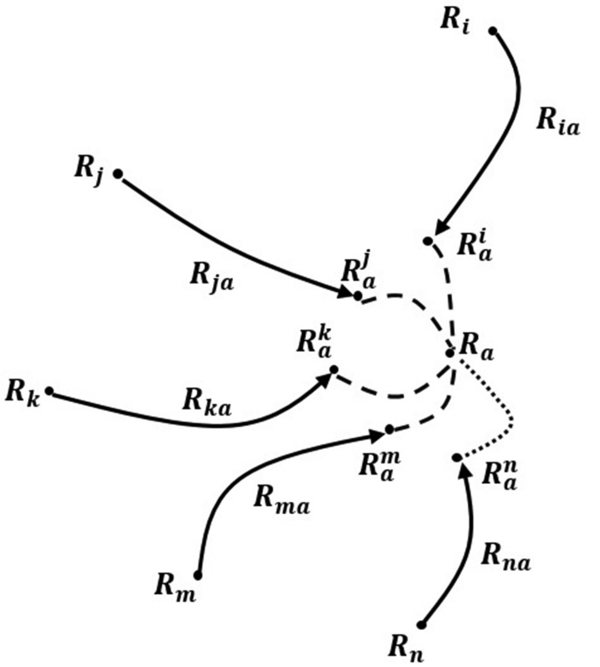


Figure 3. Single rotation averaging.

A 3D rotation can be represented as a rotation by an angle around an axis represented by unit 3-vector $\tilde{\mathbf{v}}$, $\mathbf{v} = \alpha\tilde{\mathbf{v}}$ subject to $\|\tilde{\mathbf{v}}\|_2$=1. Also, rotation matrices form a differentiable manifold that is inherent in every Lie group. Following Hartley *et al.* (2013), we project the rotation matrices from *SO(3)* to their Euclidean tangent space (the Lie algebra *so(3)*) using the logarithm log(·): *SO(3) so(3)*:

$$\log(\mathbf{R}) \;=\; [\mathbf{v}]_\times, \mathbf{v} = \;\arcsin(\|\mathbf{w}_2\|)\frac{\mathbf{w}}{\|\mathbf{w}_2\|}, \mathbf{w} = \frac{\mathbf{R}-\mathbf{R}^{\mathrm{T}}}{2} \qquad (1)$$

$$\text{with } [\mathbf{v}]_\times = \begin{pmatrix} 0 & -v3 & v2 \\ v3 & 0 & -v1 \\ -v2 & v1 & 0 \end{pmatrix} \qquad (2)$$

The inverse transformation projects *so(3)* back into *SO(3)* using the exponential map:

$$R = \;\exp([\tilde{\mathbf{v}}])_\times \;= \mathbf{I} + \sin(\alpha)\;[\tilde{\mathbf{v}}]_\times + (1-\cos(\alpha))[\tilde{\mathbf{v}}]_{\times^2} \qquad (3)$$

We now want to estimate our rotation matrix $\boldsymbol{R_a}$ from the different observations (matrices $\boldsymbol{R_a^i}$, $\boldsymbol{R_a^j}$, $\boldsymbol{R_a^k}$, $\boldsymbol{R_a^m}$, $\boldsymbol{R_a^n}$ with reference to Figure 3) by averaging the observations in tangent space:

$$\boldsymbol{R_a} = \arg\min_{\boldsymbol{R_a}} \sum_{\boldsymbol{d}\in\boldsymbol{M}} \boldsymbol{d}(\boldsymbol{R_a}, \boldsymbol{R_a^d}), \boldsymbol{M} = \{\boldsymbol{i}, \boldsymbol{j}, \boldsymbol{k}, \boldsymbol{m}, \boldsymbol{n}\} \qquad (4)$$

where $d(\boldsymbol{R_a}, \boldsymbol{R_a^d}) = d(\boldsymbol{R_a}, \boldsymbol{R_a^d})_{\text{geod}} = \|\log(\boldsymbol{R_a}, \boldsymbol{R_a^{d^T}})\|_1$ (geodesic distance).

We use the L1 norm, as it is more robust than the L2 norm, and apply the Weiszfeld algorithm to iteratively obtain the solution of (4). The pseudocode for single rotation averaging is presented in Algorithm 1. The observations of relative orientations may not be accurate if the relative rotations are not accurate (such as the dotted line between $\boldsymbol{R_a}$ and $\boldsymbol{R_a^n}$ in Figure 3).To overcome this problem, we present a robust method to eliminate the inaccurate relative rotations in Algorithm 1.

**Algorithm 1** Robust single rotation averaging

```
Input a number of observations , $d$ $i, j, k, m, n$}.
Output mean rotation $\overline{\boldsymbol{R_a}}$
1.      Initialize a rotation matrix $\boldsymbol{R_a^0}$ by randomly choosing a rotation from all observations, $\boldsymbol{R_t^0} = \boldsymbol{R_a^0}$. Iteration number $t$ = 0.
2.      Repeat
    {
      2.a Do
          {
          For d = $(i, j, k, m, n)$ { $xd = log(\boldsymbol{R_a^d}\cdot\boldsymbol{R_a^{0^T}})$; }
          $\delta = \Sigma_d^{(i,j,k,l,m,n)}(xd/\|xd\|)/\Sigma_d^{(i,j,k,l,m,n)}(1/\|xd\|)$;
          $\boldsymbol{R_{t+1}^0} = \exp(\delta)\cdot\boldsymbol{R_t^0}$;
          $t = t+1$;
           } while (d($\boldsymbol{R_{t+1}^0}, \boldsymbol{R_t^0}$) ≥0.0001or t <50)

      2.b If d($\boldsymbol{R_{t+1}^0}, \boldsymbol{R_a^d}$) > 0.001
   {
              discard $\boldsymbol{R_a^d}$
   }
    } until (no $\boldsymbol{R_a^d}$ is discarded anymore and step 2a converges)
3.      $\overline{\boldsymbol{R_a}} = \boldsymbol{R_{t+1}^0}$.
```

### Linear Translation Estimation for Each New Image

Based on the rotation matrix that is determined as described in the previous section, image translation parameters can be estimated for the new image using only two 3D points: using the collinearity equations, each 3D point yields two equations (5), and each image has three translation parameters ($\mathrm{X}_0$, $\mathrm{Y}_0$, $\mathrm{Z}_0$). As two 3D points with given image coordinates yield four equations, these two 3D points are sufficient to determine the three unknowns:

$$x = -f\,\frac{r_{11}(\mathrm{X}-\mathrm{X}_0) + r_{21}\;(Y-Y_0) + r_{31}(Z-Z_0)}{r_{13}(\mathrm{X}-\mathrm{X}_0) + r_{23}\;(Y-Y_0) + r_{33}(Z-Z_0)} + x_0$$

$$y = -f\,\frac{r_{11}(\mathrm{X}-\mathrm{X}_0) + r_{21}\;(Y-Y_0) + r_{31}(Z-Z_0)}{r_{13}(\mathrm{X}-\mathrm{X}_0) + r_{23}\;(Y-Y_0) + r_{33}(Z-Z_0)} + y_0 \qquad (5)$$

$$\text{with } \boldsymbol{R_i} = \begin{pmatrix} \mathrm{r}_{11} & \mathrm{r}_{21} & \mathrm{r}_{31} \\ \mathrm{r}_{12} & \mathrm{r}_{22} & \mathrm{r}_{32} \\ \mathrm{r}_{13} & \mathrm{r}_{23} & \mathrm{r}_{33} \end{pmatrix}$$

where (X, Y, Z) are the 3D coordinates of the *j*th object point $\boldsymbol{X_j}$, which is assumed to be viewed by the *i*th image, and (*x*, *y*) are the corresponding 2D image coordinates $\boldsymbol{x_{ij}}$. ($x_0$, $y_0$) are the principal point coordinates of *i*th image, *f* is its focal length, ($X_0$, $Y_0$, $Z_0$) are the coordinates of the unknown projection center $\boldsymbol{T_i}$ (equivalent to the image translation vector), and $r_{mn}$ (*m* = 1, 2, 3; *n* = 1, 2, 3) are the entries of the rotation matrix $\boldsymbol{R_i}$. Note that for the sake of simplicity, we omit the indices *i* and *j* in Equation 5.

To obtain a form that is linear in ($X_0$, $Y_0$, $Z_0$), we multiply Equation 5 by the denominator:

$$\begin{aligned}
&[(x-x_0)r_{13} + fr_{11}]X_0 + [(x-x_0)r_{23} + fr_{21}]Y_0 + [(x-x_0)r_{33} + fr_{31}]Z_0 \\
&\qquad = [(x-x_0)r_{13} + fr_{11}]X + [(x-x_0)r_{23} + fr_{21}]Y + [(x-x_0)r_{33} + fr_{31}]Z \\
&[(y-y_0)r_{13} + fr_{11}]X_0 + [(y-y_0)r_{23} + fr_{21}]Y_0 + [(y-y_0)r_{33} + fr_{31}]Z_0 = [(y-y_0)r_{13} + fr_{11}]X \\
&\qquad + [(y-y_0)r_{23} + fr_{21}]Y + [(y-y_0)r_{33} + fr_{31}]Z
\end{aligned} \qquad (6)$$

Finally, we obtain the linear equation system (7) and the optimization problem (8):

$$\mathbf{v} = \mathbf{Ax} - \mathbf{b} \tag{7}$$

$$\arg\min \left\| \mathbf{Ax} - \mathbf{b} \right\|_1 \tag{8}$$

Here, **x** and **b** are vectors constructed by concatenating the unknown translation parameters and the right part of Equation 6, respectively; **A** is the coefficient matrix; and **v** is the vector of residuals. Equation 8 is based on the L1 norm. Normally, there are abundant 3D points that can be chosen to solve Equation 6. We use RANSAC to find the best image translation. Then two constraints are further checked: the number of inliers should be larger than 30, and the corresponding inlier ratio should be larger than 75%.

### Refinement of Rotation and Translation by Space Resection

For each new image, the rotation estimated in the section "Robust Rotation Estimation by Single Rotation Averaging" and translation estimated in the section "Linear Translation Estimation for Each New Image" are regarded as an initial input for a space resection refinement to compute a more accurate result:

$$\underset{\boldsymbol{R}_i, \boldsymbol{T}_i}{\text{minimize}} \sum_{j=1}^{M} f_{Huber}\left( \left| \boldsymbol{x}_{ij} - \boldsymbol{\varphi}\left( f_i, x_{0i}, y_{0i}, \boldsymbol{R}_i, \boldsymbol{T}_i, \boldsymbol{X}_j \right) \right|, r_{max} \right) \tag{9}$$

where $i$ is the ID of the new image, $M$ is the number of scene points that can be viewed in the $i$th image and $\boldsymbol{x}_{ij}$ denotes their 2D image coordinates and $\boldsymbol{X}_j$ their 3D point coordinates. $\boldsymbol{R}_i$, $\boldsymbol{T}_i$ are the parameters that we aim to optimize. The parameters of interior orientation are the focal length $f_i$ and the coordinates of the principal point ($x_{0i}$, $y_{0i}$), and $\boldsymbol{\varphi}$ represents the collinearity equations (5). The 2D and 3D points are assumed to be inliers as determined in the sections "Robust Rotation Estimation by Single Rotation Averaging" and "Linear Translation Estimation for Each New Image." The Huber loss function $f_{Huber}$ (10) is used in our refinement procedure because it is less sensitive to observations with large residuals than standard least-squares estimation. In this context, the squared error loss ($0.5r^2$) is used only if the absolute value of residuals is smaller than  (we use $r_{max}$ = 2 pixels):

$$f_{Huber}\left(r, r_{max}\right) = \begin{cases} 0.5 \cdot r^2 & if |r| \le r_{max} \\ 2 \cdot \left( |r| - 1 \right) & otherwise \end{cases} \tag{10}$$

### Local Bundle Adjustment

After having added all images selected according to the section "Overview of the Developed Incremental Procedure" and before extending the cluster, we perform a bundle adjustment to reduce block deformation:

$$\underset{f_i, x_{0i}, y_{0i}, \boldsymbol{R}_i, \boldsymbol{T}_i, \boldsymbol{X}_j}{\text{minimize}} \sum_{i=1}^{N} \sum_{j=1}^{M} a_{ij} \cdot f_{Huber}\left( \left| \boldsymbol{x}_{ij} - \boldsymbol{\varphi}\left( f_i, x_{0i}, y_{0i}, \boldsymbol{R}_i, \boldsymbol{T}_i, \boldsymbol{X}_j \right) \right|, r_{max} \right) \tag{11}$$

where $N$ is the number of images and $M$ is the number of scene points and $a_{ij} = 1$ if object point $j$ is viewed in image $i$ and $a_{ij} = 0$ otherwise. $\boldsymbol{R}_i$, $\boldsymbol{T}_i$, and $\boldsymbol{X}_i$ and the interior parameters ($f_i$, $x_{0i}$, $y_{0i}$) are the parameters that we want to refine in adjustment. Note that in principle, each image may have its own set of interior parameters. However, images taken from the same camera with the same settings (i.e., same focal length according to the EXIF headers) are modeled to have identical interior parameters. Again, $\boldsymbol{\varphi}$ represents the collinearity equations (5), and $\boldsymbol{x}_{ij}$ are the 2D image coordinates. We use the Huber loss function (10) with $r_{max}$ = 2 and eliminate observations with absolute values of the residuals $|r| > 4$ pixels. In addition, to avoid numerical problems, any object point is eliminated if the corresponding largest intersection angle of all rays generating that point is smaller than 10°.

## Global Rotation and Translation Estimation

In this section, we introduce our ideas for global rotation and translation estimation. We present the procedure of detecting outliers of relative rotation for global rotation estimation, then show how we robustly select tie points that can connect all the available images; finally, the translation parameters are solved simultaneously in a linear equation system.

### Overview of the Developed Global Procedure

Figure 4 shows the flowchart of our global procedure. After relative orientation estimation, there are two main steps in our global method: global rotation averaging and global translation estimation. Both steps are preceded by outlier detection.

First, we choose an initial stereo image pair, using the selection procedure described in the section "Overview of the Developed Incremental Procedure." In order to eliminate errors in relative rotation, we use a triplet loop closure constraint of rotation (see the section "Rotation Outlier Detection") and then perform global rotation averaging on the cleaned relative rotations. Again examining Equations 6 and 7, it is possible

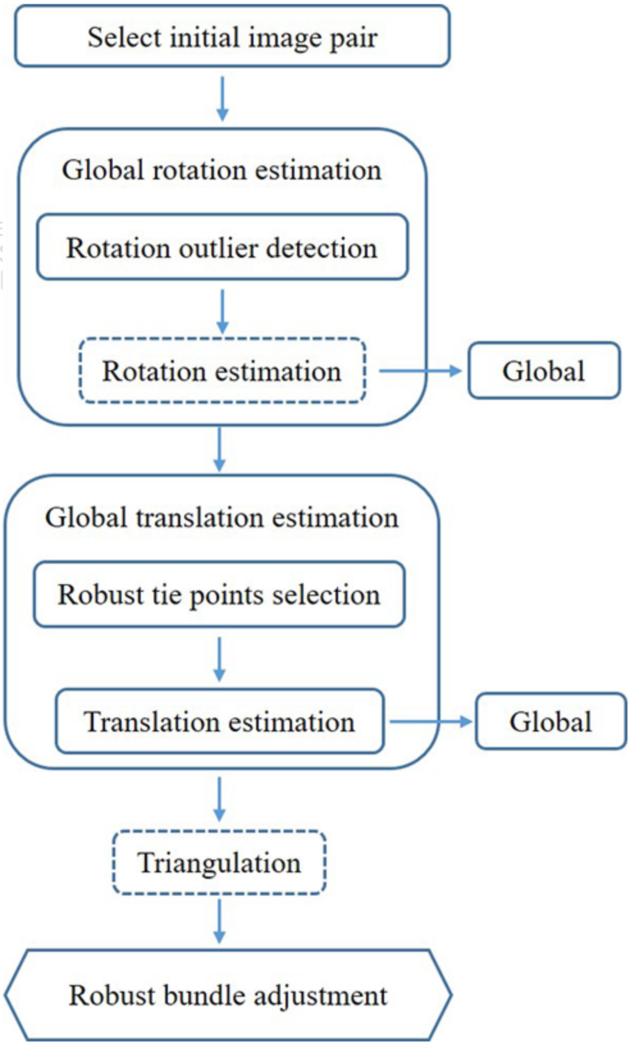

Select initial image pair
Global rotation estimation
Rotation outlier detection
Rotation estimation
Global
Global translation estimation
Robust tie points selection
Translation estimation
Global
Triangulation
Robust bundle adjustment

Figure 4. Flowchart of our global image orientation method (for the explanation of "Global," see the section "Evaluation of Three Benchmark Data Sets").

to set up a common linear equation system to compute the image translation parameters of all available images simultaneously. For this step to be successful, we must carefully select tie points in a robust manner to avoid outliers (see the section "Robust Tie Point Selection"). Given the global rotation matrices and selected tie points, a large linear equation system is then built to solve for the translation parameters (see the section "Translation Estimation"). Similar to our incremental method, the object points' coordinates are calculated by triangulation using the estimated exterior parameters. Finally, a robust bundle adjustment is applied to optimize the results.

## Global Rotation Estimation

### Rotation Outlier Detection

To obtain a reliable and robust solution, we first strive to eliminate relative rotation outliers. We can hope to do so when considering more than two images simultaneously. We use three images; one such triplet is shown in Figure 5. Obviously, in the ideal case for each such triplet, Equation 12 holds (we call it the *triplet loop closure constraint*):

$$\boldsymbol{R}_{12}\boldsymbol{R}_{23}\boldsymbol{R}_{31} = \boldsymbol{I} \qquad (12)$$

where, $\boldsymbol{I}$ is the 3 × 3 identity matrix. We use the constraint as follows. For each triplet, we determine an angular error as the arccosine of the average of the main diagonal elements of matrix $\boldsymbol{R}_{12}\boldsymbol{R}_{23}\boldsymbol{R}_{31}$. Note that this value will be 0° if the constraint (12) is fulfilled. If this error is smaller than a threshold (we use 5° in our experiments) for a triplet, the three relative rotations are regarded as inliers; otherwise, they are considered as potential outliers. This procedure is repeated for every triplet that we can find in the data set. Then each relative rotation is examined separately: only if all triplets the relative rotation is part of show a value above the threshold, this relative orientation is considered to be an outlier; otherwise, it is considered an inlier.

### Rotation Estimation

After detecting outliers of relative rotations, we estimate global rotation matrices by global rotation averaging (see dashed box in Figure 4). The problem has been studied by many researchers; we use the method of Chatterjee and Govindu (2013) because their work is considered capable of providing a reliable result for large numbers of images.

## Global Translation Estimation

### Robust Tie Point Selection

Given the global rotations of all images and the results of relative orientation, we now turn to translation estimation. First, we examine the tie points used for relative orientation in image space. As they were derived from a feature extraction step, points exist that tie together more than two images, and we concentrate on these points (on how to efficiently determine these points from the set of all tie points, see Moulon and Monasse 2012). In principle, we could set up a coefficient matrix **A** along the lines of Equation 6; note that since the matrix is linear in the unknowns, initial values are not needed. However, solving this equation system would be an enormous task due to the large number of available tie points; at the same time, potential outliers will deteriorate the results. Therefore, we select a subset of points that can still yield a reliable result. We take the following considerations into account: (1) points should have a small reprojection error, and (2) they should be evenly distributed in image space.

In order to take into account the first criterion, we calculate the reprojection error of the individual local spatial intersection for each image pair. This reprojection error corresponds to the deviation of the extracted point from the epipolar line of its conjugate point in the other image. As Figure 6 shows, let tie point $\boldsymbol{P}$ be observed by image $\boldsymbol{S}$, $\boldsymbol{S}_1$, $\boldsymbol{S}_2$, …, $\boldsymbol{S}_n$. Then the reprojection error of $\boldsymbol{P}$ on image $\boldsymbol{S}$ by individual local spatial intersection will have different values: err($\boldsymbol{SS}_1$) is the reprojection error by considering image pair $\boldsymbol{S}$ and $\boldsymbol{S}_1$, err($\boldsymbol{SS}_2$) is generated by image pair $\boldsymbol{S}$ and $\boldsymbol{S}_2$, and so on. Among these reprojection errors, the median is further considered.

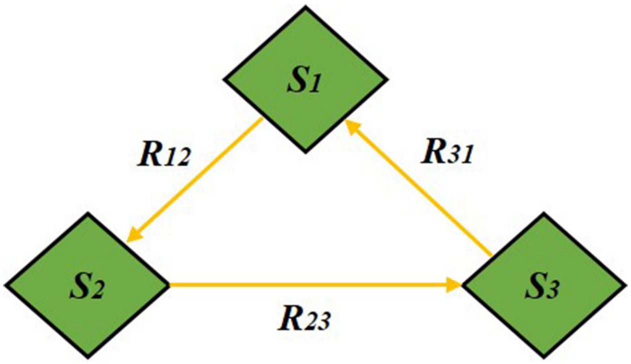


Figure 5. A closed loop of relative rotations from three connected images forming an image triple.

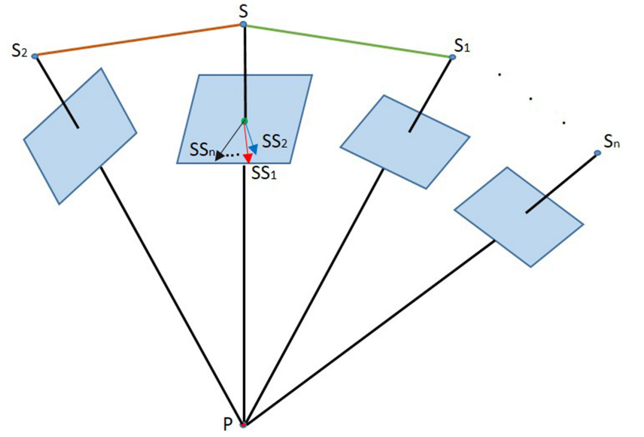


Figure 6. Individual local spatial intersection of tie point.

To take into account the tie point distribution in image space, each image is uniformly divided into *p*×*p* patches. Typically, there will be some tie points in each patch. Figure 7 shows an example of dividing an image into 3×3 patches, where the length of the arrow indicates the value of reprojection error. For each patch, we keep only the tie point with the smallest median reprojection error according to the procedure explained in the previous paragraph, such as the green points.

In our experiments, we choose to divide the images into 3×3 patches, which means that for each image, at most, we keep nine tie points. We find that some patches in some images are empty, but as we require only two tie points to obtain a solution, in most cases we still obtain a solution. If only fewer than two points are available, the corresponding image is eliminated from the block.

### Translation Estimation

We can now proceed in a manner very similar to the way we did in the section "Linear Translation Estimation for Each New Image." The only unknowns to be estimated are the 3D coordinates of the selected tie points and the projection centers, and Equation 6 is linear in these unknowns. We again use one initial image pair to define the datum (we use the same one as for global rotation averaging). Subsequently, a large linear equation system is set up (similar to Equation 7, where **x** contains all unknowns, and all unknowns are solved for simultaneously) and solved. In the end, a final bundle adjustment is used to refine the results.

## Experiments

In this section, we present a detailed evaluation of our approaches. The experiments are conducted on various real-word data sets, including three benchmark data sets (Strecha *et al.* 2008) consisting of 11 to 30 images and several unordered data sets from the Internet with up to more than 1000 images. Two challenging data sets with 175 and 1040 images are also tested, and the results are compared to two widely used software packages.

In the following section, we discuss the results of the three benchmark data sets for single rotation averaging (see the section "Evaluation of Rotation Results"), followed by evaluation of our translation estimation methods based on the ground truth (see the section "Evaluation of Rotation Results"). To further explore the potential of our approaches, we orient the unordered image data sets with many more images (see the section "Evaluation of Translation Result") by comparing with two global SfM methods (Cui and Tan 2015; Cui *et al.* 2015). The section "Experiment on Two Challenging Data Sets" deals with the two additional challenging data sets. All reported experiments are conducted on a computer with four 3.2-GHz Intel Core i5-6500 processors and eight threads. We use the open-source Ceres-solver (Agarwal *et al.* 2017) for bundle adjustment. The parameters of our method (e.g., thresholds) were set in the way described in sections "Incremental Rotation and Translation Estimation" and "Global Rotation and Translation Estimation." These values were found empirically and were not changed in the experiments. While an evaluation of the impact of changing parameters is left for future work, we note that the parameters we used achieved good results for a very heterogeneous set of test data sets without further tuning. We take this as an indication that the chosen values are reasonable and can be applied under rather different circumstances.

### Evaluation of Three Benchmark Data Sets

#### Evaluation of Rotation Results

For the investigation of rotation accuracy, the three benchmark data sets *fountain-P11*, *Herz-Jesu-P25*, and *castle-P30*, which have known ground-truth exterior orientation, are investigated. The interior orientation parameters are taken from the EXIF information provided with the data. Figure 8 shows results for different methods; the abscissa denotes the image ID, and it is in the order of the images that are oriented so that the first two images are the initial stereo image pair chosen by our method. The red triangles indicate stages in which the incremental method carried out a local bundle adjustment. The ordinate is the angle error (in degrees), that is, the difference between the rotation computed by the corresponding methods and ground-truth rotation (see the appendix for the precise definition of that error metric). "Ours" is the result by just using the single rotation averaging (see the section "Robust Rotation Estimation by Single Rotation Averaging"). Applying resection refinement (see the section "Refinement of Rotation and Translation by Space Resection") yields the results of "Ours_RF." Note that while for all clusters but the last one bundle adjustment has also been run as part of the procedure (see Figure 2), the results used for "Ours" are those obtained directly after rotation averaging for all images. The same holds for the other experiments accordingly. "Res_RF" uses the DLT (Hartley and Zisserman 2003), where the projection matrix is first estimated by six points and is then decomposed to obtain rotation and translation parameters; the result is refined by resection refinement (see again the section "Refinement of Rotation and Translation by Space Resection"). "BA" denotes the results after final bundle adjustment. "Global" indicates the results of the global rotation averaging method (see the section "Rotation Estimation"). The initial image pairs of these data sets are (4, 9), (4, 9), and (4, 12), determined by the method described in the section "Overview of the Developed Incremental Procedure," so the rotation of the fourth image is selected as the original one. The relative rotation between the fourth image and the corresponding ground truth is used to project all remaining rotation matrices into the coordinate system of the reference. We calculate the error $\theta$ by comparing the ground-truth rotations to the computed ones in the way described in the appendix. In this way, the error of the fourth image is always zero.

Comparing the angle errors of the different methods shown in Figure 8, "Ours" provided by our single rotation averaging is the worst, probably due to remaining errors and outliers in the relative rotations. "Global" generates errors similar to "Ours" on *fountain-P11*, but, with respect to *Herz-jesu-P25* and *Castle-P30*, which have some repetitive structures, "Global" performs better than "Ours," probably due to the relative rotation outlier detection. It can also be seen that the angle error of some images (e.g., image 10 in Figure 8a and images 10–13 and 20–24 in Figure 8b) are much larger than others. This is probably due to error accumulation. Errors propagate and accumulate when adding images sequentially into a refined block, and we found that the images that were added last have the largest angle

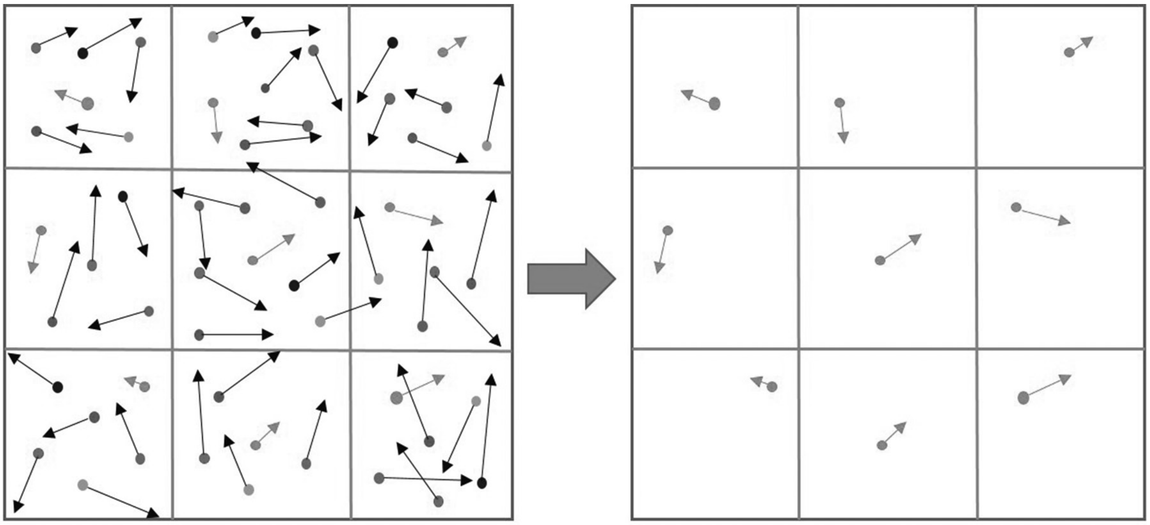
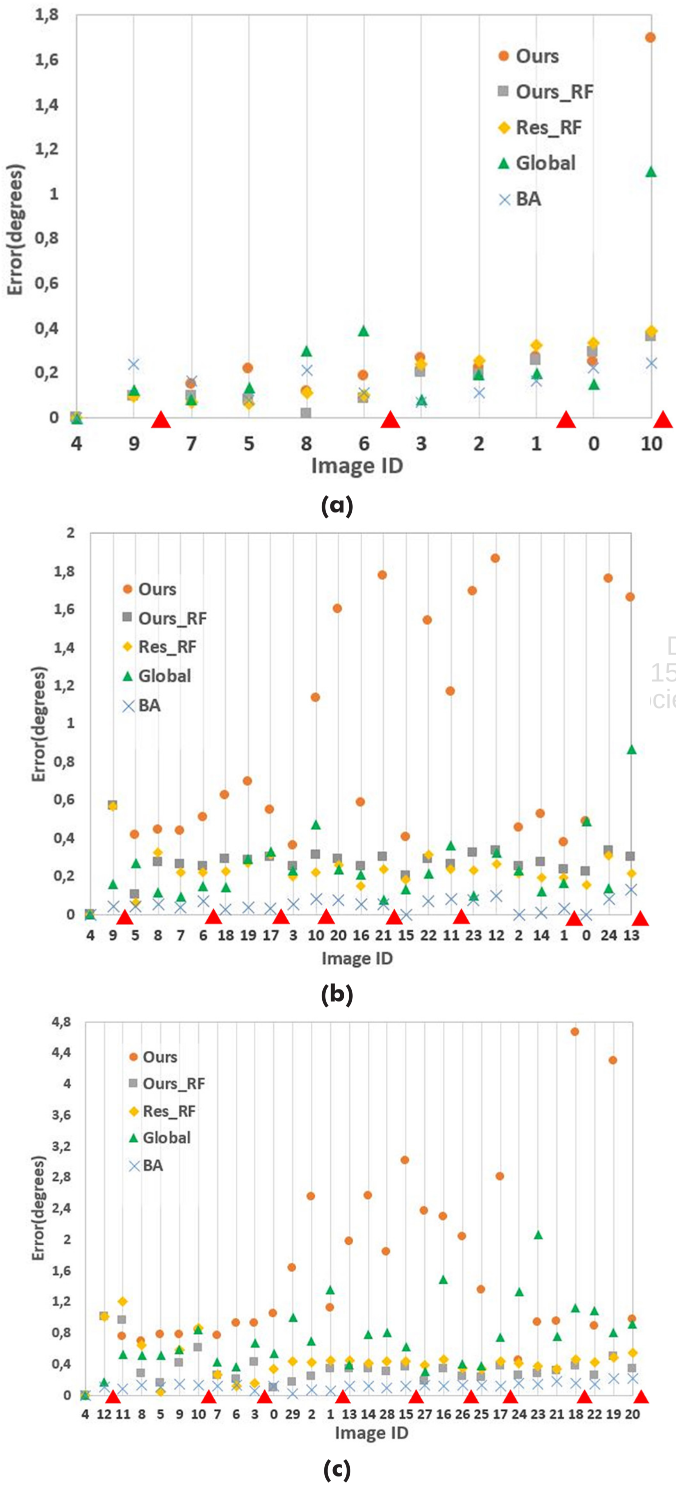


Figure 7. Procedure of tie point selection. Green tie points with the smallest arrow length (i.e., reprojection error) are kept.

Table 1. Mean angle error in degrees for different methods. We compared our results with DLT according to Hartley, Zisserman (2003) including resection refinement (“Res_RF”), Chatterjee and Govindu (2013) (Global), Reich and Heipke (2016) (1), and Jiang *et al.* (2015) (2). “Ours_L2” and “Global_L2” use the L2 norm to solve Equation 4. Note that we cite the results of (1) and (2) from the corresponding papers, and we reimplemented the approach of Chatterjee and Govindu (2013).

| | Before Bundle Adjustment | | | | | | | | After Bundle Adjustment |
|---|---|---|---|---|---|---|---|---|---|
| | Ours | Ours_L2 | Ours_RF | Res_RF | Global | Global_L2 | (1) | (2) | Ours_RF |
| *fountain-P11* | 0.32 | 0.33 | **0.15** | 0.18 | 0.25 | 0.26 | 0.25 | 0.45 | 0.15 |
| *Herz-Jesu-P25* | 0.79 | 0.93 | 0.27 | 0.23 | 0.24 | 0.37 | **0.21** | 0.39 | 0.05 |
| *castle-P30* | 1.24 | 1.57 | **0.34** | 0.44 | 0.75 | 0.95 | 0.58 | 0.96 | 0.12 |



(a)

(b)

(c)

Figure 8. Angle errors of three benchmark data sets. (a) *fountain-P11*; (b) *Herz-Jesu-P25*; (c) *Castle-P30*.

errors. However, if we turn on the resection refinement, these results are significantly improved. Furthermore, by applying resection refinement, both “Ours_RF” and “Res_RF” achieve almost the same accuracy. As expected, the angle errors after final bundle adjustment are the smallest ones: the errors for *fountain-P11* and *castle-P30* are smaller than 0.4°, and for *Herz-Jesu-P25*, they are smaller than 0.2°.

Finally, from the locations of red triangles in Figures 8 and 9, we can see that a cluster of images is composed of two to five images, which indicates that clustering saves a considerable amount of computing time when compared to running a bundle adjustment after each new image.

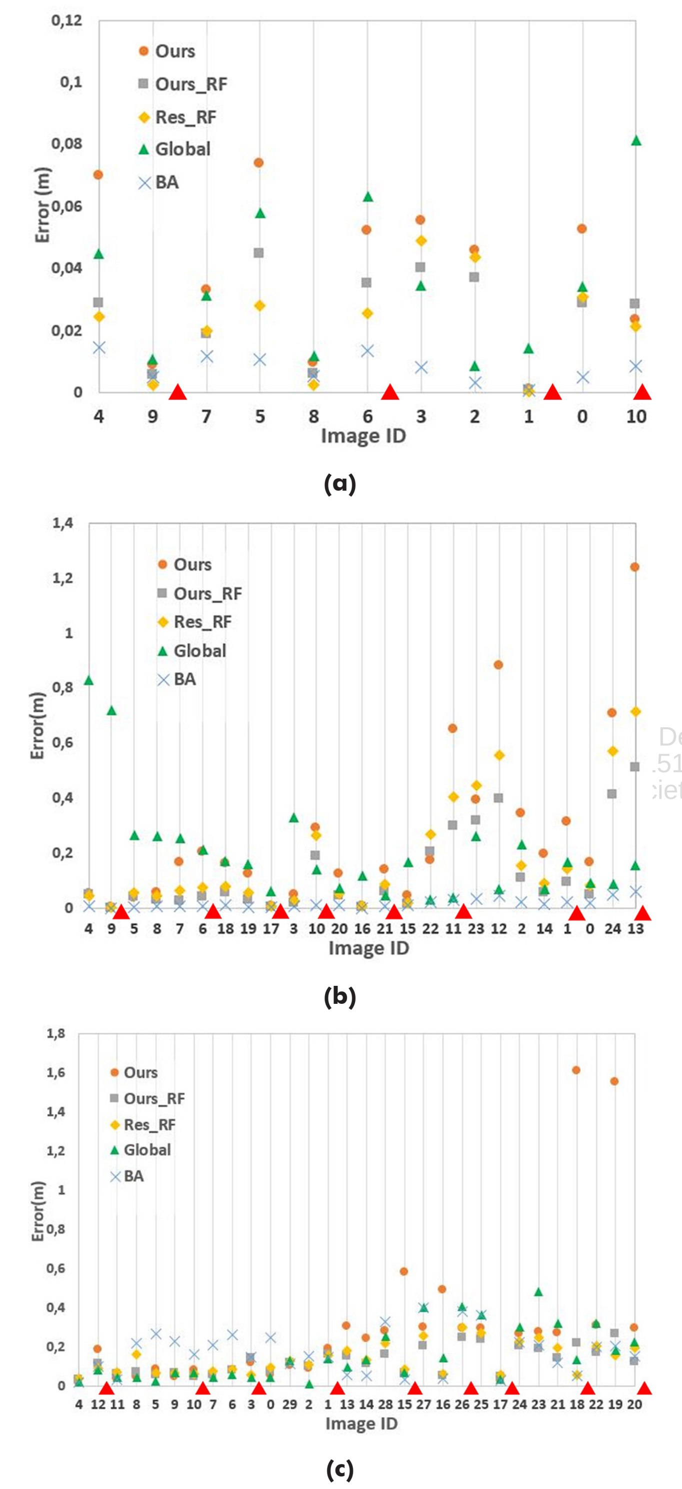
In Table 1, we list the mean angle error of the mentioned methods along with the results of the baseline methods of Chatterjee and Govindu (2013) (“Global”), Jiang *et al.* (2015), and Reich and Heipke (2016). For “Ours_RF” and “Global,” we also list results using the L2 norm for optimization, termed “Ours_L2” and “Global_L2.” One can see that resection refinement has a strong effect on the accuracy of estimating rotations. “Ours_RF” obtains the best results on *fountain-P11* and *castle-P30* before final bundle adjustment and gives a similar accuracy as Reich and Heipke (2016) on *Herz-Jesu-P25*. Nevertheless, after final bundle adjustment, the results are significantly better. When comparing the different error norms, “Ours” (L1 norm) performs almost the same as “Ours_L2” on *fountain-P11*, but on the other two data sets, the L1 norm is much better; similar effects can be seen for “Global” and “Global_L2.”

### Evaluation of Translation Result

For the three benchmark data sets, a comparison of the translation accuracy is given in Figure 9. The abscissa has the same meaning as in Figure 8, and the ordinate is the translation error (in meters). “Ours” means the method of our incremental linear translation estimation described in the section “Linear Translation Estimation for Each New Image” based on rotations computed according to section 3.2, “Ours_RF” utilizes the resection refinement (see the section “Refinement of Rotation and Translation by Space Resection”), and “Res_RF” and “BA” denote the same methods as in the section “Evaluation of Rotation Results”). “Global” indicates our global translation estimation. The translation errors of *castle-P30* and *Herz-Jesu-P25* are two orders of magnitude larger than those of *fountain-P11* (see ordinate in Figure 9). Inspecting the results in more detail, we found that *castle-P30* and *Herz-Jesu-P25* have lots of repetitive structures and a significant number of image pairs with small intersection angles. Moreover, some images of *castle-P30* are weakly connected; consequently, the block geometry is not as stable as that of *fountain-P11*, which we believe explains the findings. Similar to the angle error shown in Figure 8, “Ours” always provides the worst results. This is probably a consequence of the results of rotation averaging (see the section “Evaluation of Rotation Results”). Comparing Figures 8 and 9, the images with large angle errors are normally those that exhibit large translation errors also. “Global” and “Ours” generate similar accuracy on *fountain-P11*. However, on the other two data sets, “Global” performs better than “Ours”; this can be attributed to the fact that the rotation accuracy of *Castle-P30* and *Herz-jesu-P25* given by

Table 2. Mean translation error in meters for different methods. “Ours_L2” and “Global_L2” use the L2 norm to solve Equation 7.

| | Before Bundle Adjustment | | | | | | | | After Bundle Adjustment |
|---|---|---|---|---|---|---|---|---|---|
| | Ours | Ours_L2 | Ours_RF | Res_RF | Global | Global_L2 | (1) | (2) | Ours_RF |
| *fountain-P11* | 0.039 | 0.036 | **0.023** | 0.027 | 0.035 | 0.040 | 0.035 | 0.072 | 0.008 |
| *Herz-Jesu-P25* | 0.166 | 0.183 | **0.081** | 0.122 | 0.085 | 0.131 | 0.083 | 0.061 | 0.016 |
| *castle-P30* | 0.287 | 0.296 | **0.127** | 0.137 | 0.161 | 0.194 | 1.312 | 1.620 | 0.016 |



(a)

(b)

(c)

Figure 9. Translation errors of three benchmark data sets. (a) *fountain-P11*; (b) *Herz-Jesu-P25*; (c) *Castle-P30*.

“Global” is better. “Ours_RF” and “Res_RF” give very similar results, which are much better than “Ours.” The performance after the final bundle adjustment is always the best; the translation error of *fountain-P11* is 0.08 m; both *Herz-Jesu-P25* and *castle-P30* have translation errors that are smaller than 0.2 m.

Table 2 presents numerical results for the mean translation errors of the different methods. Before final bundle adjustment, “Ours_RF” outperforms all other methods. This means that optimization by resection refinement can improve the accuracy of translation and is a very important step in our processing chain. “Ours” detects and eliminates outliers of tie points, so it is not surprising that “Ours” is much better than the methods of Reich and Heipke (2016) and Jiang *et al.* (2015). In addition, we find that our global translation estimation method significantly improves the accuracy of *Castle-P30* compared to the two approaches just mentioned, which we attribute to our robust strategy for tie point selection. When comparing the error norms, “Ours” (L1 norm) performs almost the same as “Ours_L2” on *fountain-P11*, but on the other two data sets, the L1 norm works again better and comes very close to the results of “Global” and “Global_L2.” This may be due to the fact that *Herz-Jesu-P25* and *castle-P30* have more repetitive structures so that some incorrect correspondences can be generated resulting in large residuals in Equation 7. All our following experiments use the L1 norm. After final bundle adjustment, the best results (those of “Ours_RF”) are improved by nearly one order of magnitude. Visualizations of image orientation results can be seen in Figure 10.

## Experiment on Unordered Data Sets

To further demonstrate the performance of our approaches, we processed several unordered Internet data sets published by Wilson and Snavely (2014). The initial interior and exterior orientation parameters are provided by the authors.[1] To detect mutual overlap of these unordered images for efficient image matching, the approach of Wang *et al.* (2017) based on random k-d trees was applied. For each data set, we evaluate the rotation and translation accuracy and the run time of the different methods.

Table 3 shows the rotation accuracy results of the Internet data sets by comparing the results of the different rotation estimation methods presented in this article. Analyzing the number of images in the largest connected component of images($N_e$) and the number of oriented images ($N_r$), we see that some images that exist in the original set of images are eliminated by the proposed methods. “Global” is able to orient the largest number of images on most data sets, although it typically also eliminates between 10% and 15%. “Ours_RF” and “Res_RF” solve almost the same number of images (a few percent less than “Global”) because they use the same constraint for eliminating outlier images. In addition, some images are eliminated during bundle adjustment. Before bundle adjustment, the rotation accuracies of “Ours_RF” and “Res_RF” are very similar, and both of them perform better than “Global” for most data sets, which is very similar to the results of Table 1. After bundle adjustment, the rotation accuracies of all data sets are improved, and for each individual data set, all three methods achieve similar accuracy with

[1] See http://www.cs.cornell.edu/projects/1dsfm (accessed 15 November 2018).

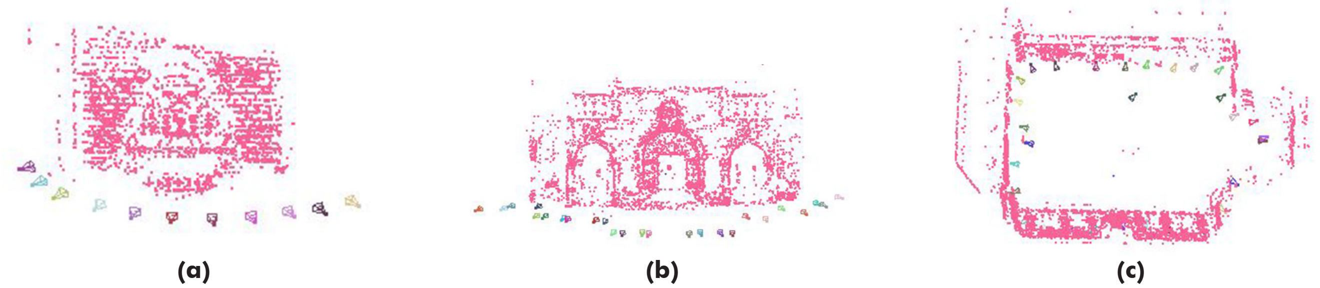


Figure 10. Visualization of the orientation results on benchmark data sets (shown are those of the incremental approach). Red dots are estimated 3D object points, and camera symbols represent the orientation parameters. (a) *fountain-P11*; (b) *Herz-Jesu-P25*; (c) *Castle-P30*.

Table 3. Mean angle error in degrees for different methods. $N_e$ is the number of images in the largest connected component generated from image pairs after checking the epipolar geometry, $N_r$ is the number of the oriented images, and $\tilde{\boldsymbol{v}}$ is the mean angle error. We highlight the best results in terms of mean angle error for each data set before and after bundle adjustment.

| Data | | Before Bundle Adjustment | | | | | | After Bundle Adjustment | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Ours_RF | | Res_RF | | Global | | Ours_RF | | Res_RF | | Global | |
| *Name* | $N_e$ | $N_r$ | $\tilde{\boldsymbol{v}}$ | $N_r$ | $\tilde{\boldsymbol{v}}$ | $N_r$ | $\tilde{\boldsymbol{v}}$ | $N_r$ | $\tilde{\boldsymbol{v}}$ | $N_r$ | $\tilde{\boldsymbol{v}}$ | $N_r$ | $\tilde{\boldsymbol{v}}$ |
| *Alamo* | 513 | 447 | **2.3** | 450 | **2.3** | 459 | 8.3 | 447 | 1.8 | 450 | **1.7** | 450 | 3.7 |
| *Ellis Island* | 209 | 195 | 2.9 | 195 | **2.3** | 200 | 3.3 | 195 | 2.1 | 195 | 2.1 | 191 | **1.9** |
| *Gendarmenmarkt* | 396 | 354 | 25.2 | 354 | **23.9** | 366 | 30.7 | 354 | 16.7 | 354 | **16.4** | 318 | 19.3 |
| *Metropolis* | 293 | 207 | **1.6** | 208 | 2.3 | 262 | 7.6 | 207 | **1.3** | 208 | 1.5 | 249 | 2.7 |
| *Montreal N.D.* | 442 | 417 | **1.0** | 411 | **1.2** | 400 | 1.3 | 417 | 0.8 | 411 | 0.8 | 394 | **0.7** |
| *Notre Dame* | 509 | 423 | 1.3 | 423 | 1.2 | 484 | 4.7 | 423 | **0.9** | 423 | **0.9** | 476 | 1.4 |
| *NYC Library* | 313 | 261 | **3.9** | 268 | 4.3 | 255 | 4.1 | 261 | 1.7 | 268 | 2.5 | 239 | **1.6** |
| *Piazza del Popolo* | 284 | 192 | **1.4** | 192 | 2.0 | 242 | 3.5 | 192 | **1.1** | 192 | **1.1** | 231 | 1.5 |
| *Roman Forum* | 1009 | 831 | 1.4 | 831 | **1.3** | 928 | 5.2 | 831 | 1.1 | 831 | **0.9** | 909 | 1.9 |
| *Tower of London* | 421 | 370 | 2.5 | 370 | **1.6** | 356 | 5.7 | 370 | 2.1 | 370 | **1.2** | 354 | 1.3 |
| *Union Square* | 407 | 260 | **1.5** | 260 | 1.6 | 349 | 6.9 | 260 | 1.3 | 260 | **1.2** | 341 | 2.3 |
| *Vienna Cathedral* | 784 | 638 | **1.8** | 640 | 1.9 | 684 | 7.9 | 638 | **1.4** | 640 | 1.5 | 678 | 2.7 |
| *York Minster* | 396 | 343 | 2.1 | 343 | **1.2** | 348 | 3.1 | 343 | 1.1 | 343 | **1.0** | 344 | 1.5 |

Table 4. Mean translation error in meters for different methods. $N_r$ is the number of the oriented images, and  is the mean translation error. We highlight the best results of $N_r$ and $\bar{\mathbf{e}}$ on each data set. The results of Cui and Tan (2015) and Cui *et al.* (2015) are cited from corresponding papers. "—" denotes that corresponding items are not provided.

| **Data** | **Without Bundle adjustment** | | | | | | | | **With Bundle adjustment** | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Ours_RF** | | **Res_RF** | | **Global** | | **Cui, Tan, 2015** | | **Ours_RF** | | **Res_RF** | | **Global** | | **Cui, Tan, 2015** | | **Cui *et al.*, 2015** | |
| **Name** | $N_r$ | $\bar{e}$ | $N_r$ | $\bar{e}$ | $N_r$ | $\bar{e}$ | $N_r$ | $\bar{e}$ | $N_r$ | $\bar{e}$ | $N_r$ | $\bar{e}$ | $N_r$ | $\bar{e}$ | $N_r$ | $\bar{e}$ | $N_r$ | $\bar{e}$ |
| *Alamo* | 447 | **0.3** | 450 | 0.3 | 459 | 0.4 | **574** | 2.0 | 447 | **0.2** | 450 | 0.2 | 450 | 0.3 | **574** | 3.1 | 500 | 3.7 |
| *Ellis Island* | 195 | **1.3** | 195 | 1.5 | 200 | 1.7 | **223** | 5.5 | 195 | **0.9** | 195 | 1.1 | 191 | 1.2 | **223** | 4.2 | 211 | 1.8 |
| *Gendarmenmarkt* | 354 | **5.8** | 354 | 6.0 | 366 | 14.2 | **609** | 27.7 | 354 | **4.0** | 354 | **4.0** | 318 | 8.3 | **609** | 27.3 | — | — |
| *Metropolis* | 207 | 1.6 | 208 | **1.4** | 262 | 4.3 | **317** | 10.6 | 207 | **0.8** | 208 | **0.8** | 249 | 1.1 | **317** | 16.6 | — | — |
| *Montreal N.D.* | 417 | **0.6** | 411 | 0.7 | 400 | 1.4 | **452** | 0.7 | 417 | **0.5** | 411 | **0.5** | 394 | 0.7 | **452** | 1.1 | 426 | 1.1 |
| *Notre Dame* | 423 | 2.1 | 423 | 2.0 | 484 | 2.7 | **549** | **0.6** | 423 | 1.7 | 423 | 1.7 | 476 | 1.7 | **549** | **1.0** | 539 | 0.8 |
| *NYC Library* | 261 | **0.7** | 268 | 0.7 | 255 | 1.2 | **338** | 1.9 | 261 | **0.4** | 268 | **0.4** | 239 | 0.7 | **338** | 1.6 | 288 | 6.9 |
| *Piazza del Popolo* | 192 | **0.5** | 192 | 0.6 | 242 | 2.0 | **340** | 2.7 | 192 | **0.3** | 192 | **0.3** | 231 | 1.3 | **340** | 2.5 | 294 | 3.2 |
| *Roman Forum* | 831 | 2.0 | 831 | **1.9** | 928 | 9.3 | **1077** | 9.4 | 831 | 1.8 | 831 | **1.7** | 909 | 2.8 | **1077** | 10.1 | — | — |
| *Tower of London* | 370 | **2.4** | 370 | 2.6 | 356 | 7.6 | **465** | 11.2 | 370 | 2.1 | 370 | **2.0** | 354 | 3.8 | **465** | 12.5 | 393 | 6.2 |
| *Union Square* | 260 | **1.0** | 260 | 1.3 | 349 | 1.9 | **570** | 12.7 | 260 | **0.8** | 260 | 0.9 | 341 | 0.9 | **570** | 11.7 | — | — |
| *Vienna Cathedral* | 638 | **1.4** | 640 | **1.4** | 684 | 4.3 | **842** | 5.9 | 638 | **0.9** | 640 | 1.0 | 678 | 3.7 | **842** | 4.9 | 578 | 4.0 |
| *York Minster* | 343 | **1.1** | 343 | 1.2 | 348 | 4.2 | **417** | 5.7 | 343 | **0.9** | 343 | 1.0 | 344 | 3.4 | **417** | 14.2 | 341 | 14.0 |

differences smaller than 2°. One notable exception is *Gendarmenmarkt*, which we discuss below.

Next, we compare our translation estimation methods with two state-of-the-art global SfM methods (Cui and Tan 2015; Cui *et al.* 2015). Table 4 gives a detailed quantitative comparison of the results (note that those before bundle adjustment are not provided by Cui *et al.* 2015). After carrying out a 3D similarity transformation between ground-truth and our methods' translation estimation results, the mean translation errors  are evaluated. Analyzing the results without bundle adjustment, we find that "Ours_RF" and "Res_RF" perform almost the same: the results of "Res_RF" are better than those of "Ours_RF" for three data sets (*Notre Dame*, *Metropolis*, and *Roman Forum*), while practically the same values are obtained for *Alamo*, *NYC Library*, and *Vienna Cathedral*. As to the remaining seven data sets, "Ours_RF" achieves a better result. Both of these two incremental methods are better than the global methods, including "Global" and the method of Tan and Cui (2015). This was already observed in the section "Evaluation of Translation Result"; the reason is that the repetitive local bundle adjustment reduces the error accumulation.

When comparing our "Global" and Cui and Tan's (2015) methods, the mean translation errors of our global translation method are smaller on most data sets, but the number

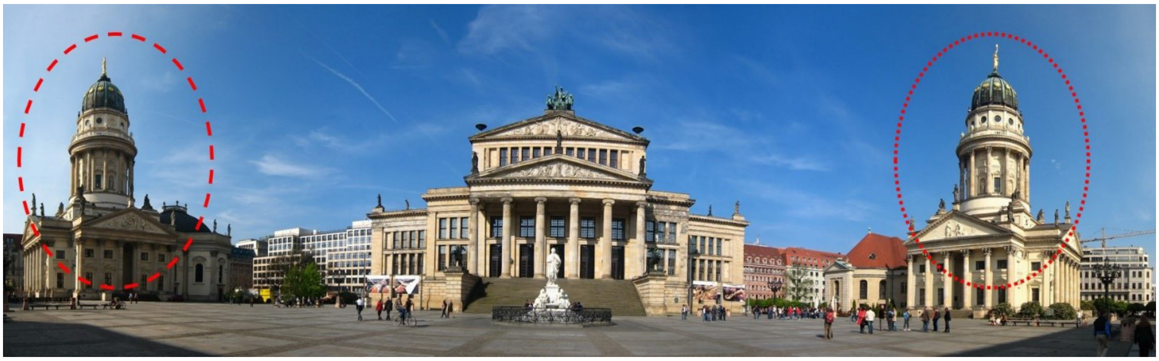


Figure 11. Repetitive structure of *Gendarmenmarkt*. Note that the left church (German Church, dashed circle) and the right church (French Church, pointed circle) are almost identical buildings.

Table 5. Run time in seconds for different methods on internet data. $T_{BA}$ is the time for the final bundle adjustment, $T_{\Sigma}$ denotes the total run time, and $T_R$ and $T_T$ are the run times of our global rotation averaging and translation estimation, respectively. The bold font in brackets of column $T_{\Sigma}$ of "Global" is the factor of speedup when comparing "Ours_RF" and "Global." The most time efficient approach of each row is highlighted.

| **Data** | **Image Orientation Time** | | | | | |
|---|---|---|---|---|---|---|
| | **Global** | | | | **Ours_RF** | **Res_RF** |
| **Name** | $\mathbf{T_R}$ | $\mathbf{T_T}$ | $\mathbf{T_{BA}}$ | $\mathbf{T_{\Sigma}}$ | $\mathbf{T_{\Sigma}}$ | $\mathbf{T_{\Sigma}}$ |
| *Alamo* | 16 | 164 | 299 | **477 (×13.6)** | 6501 | 6862 |
| *Ellis Island* | 5 | 22 | 134 | 161 (**×8.4**) | 1354 | 1291 |
| *Gendarmenmarkt* | 5 | 35 | 369 | 409 (**×4.6**) | 1882 | 2297 |
| *Metropolis* | 3 | 28 | 211 | 242 (**×3.9**) | 934 | 1043 |
| *Montreal N.D.* | 10 | 144 | 562 | 716 (**×6.6**) | 4732 | 4444 |
| *Notre Dame* | 17 | 191 | **466** | 674 (**×13.0**) | 8752 | 8154 |
| *NYC Library* | 2 | 55 | 244 | 301 (**×7.3**) | 2186 | 2563 |
| *Piazza del Popolo* | 5 | 41 | 167 | 213 (**×2.0**) | 427 | 388 |
| *Roman Forum* | 21 | 236 | 1171 | 1428 (**×12.4**) | 17 752 | 17 880 |
| *Tower of London* | 5 | 71 | 299 | 375 (**×12.8**) | 4784 | 4522 |
| *Union Square* | 4 | 42 | 261 | 307 (**×4.0**) | 1225 | 1434 |
| *Vienna Cathedral* | 16 | 286 | 1031 | 1333 (**×7.8**) | 10 374 | 10 098 |
| *York Minster* | 4 | 62 | 483 | 549 (**×7.0**) | 3840 | 4205 |

of oriented images is also smaller. This means that we have probably eliminated some images with outlier observations, which has decreased the mean translation error. After refinement by our robust bundle adjustment, the accuracies of the methods mentioned in this article are all improved, and we are not surprised to see that "Ours_RF" and "Res_RF" generate the best results because many outliers were already eliminated in the previous steps. Analyzing the results of "Global", Cui and Tan (2015), and Cui *et al.* (2015), our global method performs best. We also find that the number of images oriented by our global method is reduced during robust bundle adjustment, whereas the number of images oriented by Cui and Tan (2015) remains constant, and some results after bundle adjustment become worse than before. The reason could be that incorrectly oriented images are kept in their block.

From Tables 3 and 4, we find that the results of *Gendarmenmarkt* are much worse than the others. This data set is very challenging due to the highly repetitive structures (see Figure 11), which generate ambiguous relative orientations and tie points.

Table 5 provides the run time of image orientation for the proposed new methods. We find that the most time consuming step of our "Global" method is the final bundle adjustment. When comparing the run time of the global and the incremental methods, the global method is normally 4 to 10 times faster than the incremental method, depending on the size of data set; for example, the global method only uses one-thirteenth of the time for the incremental method on *Notre Dame*. The reason is that the time of the computationally expensive repetitive bundle adjustment is saved in global image orientation. The run times for "Ours_RF" and "Res_RF" are very close to each other because the most time consuming process of the incremental method is the repetitive intermediate local bundle adjustment, and by using the same strategy to select images to extend the cluster, these two incremental methods should have similar timings. We do not show the run time of the three ordered data sets discussed in the section "Evaluation of Three Benchmark Data Sets" because, not surprisingly, for very small data sets, the run times of our incremental and global methods are almost the same.

## Experiment on Two Challenging Data Sets

To further explore the capability and limitation of our methods, we test two more data sets: *Remond* and *Campus* published by Cohen *et al.* (2012) and Cui and Tan (2015), respectively. *Remond* has 175 street-view images with rather symmetric building facades; in other words, the vertical walls of the houses have very similar structure and texture, which can produce ambiguous correspondences. *Campus* has 1040 images, including a closed loop, with rather small baselines. This means that the initial stereo image pair must be carefully selected, and the loop closure can be checked.

In these two data sets, various problems become apparent when using the different methods (see Figure 12). In this section, we compare the results obtained by our methods to those from the most well established SfM packages, VisualSFM and Colmap.[2] In Figure 12, the rectangles mark problem areas when using VisualSFM: some images are incorrectly oriented in the left part of *Remond*, possibly due to the symmetric facades, and for *Campus*, the block is split into two unconnected parts. Colmap performs better on these two data sets: the *Remond*, results look correct, and *Campus* yields one connected block. However, errors become visible in the circle (Figure 12). Both our methods generate better results; for example, the closure error for "Global" is much smaller than that for the other two approaches, and "Ours_RF" is closed altogether (see the triangle in Figure 12 and Table 6).

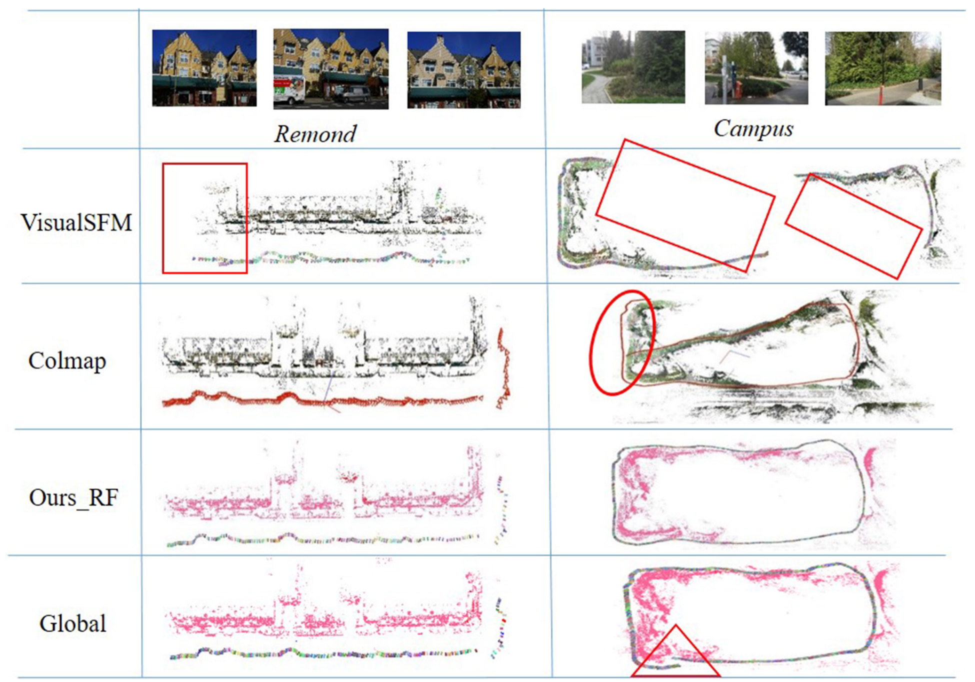


Figure 12. Visualization of results of *Remond* and *Campus* after final bundle adjustment (example images in the first row) by using different methods. The rectangle, circle, and triangle denote problematic areas discussed in the text.

Furthermore, whereas the bundle adjustment does not converge for both data sets using VisualSFM, nor does it for *Campus* when Colmap is employed,[3] both of our methods obtain good results with an *rms* value smaller than 1.5 pixels. Although our global method (Global) gives an incorrect result (a drift) for *Campus* (triangle in Figure 12), this drift is much smaller than the drifts of VisualSFM and Colmap. Our incremental method (Ours_RF) gives the best results, no visual drift exists, and the corresponding *rms* values are the smallest.

Table 6. Evaluations of *Remond* and *Campus*. $N_r$ is the number of the oriented images, and *rms* denotes the root mean square value of the reprojection error after bundle adjustment; the unit is pixel. "—" denotes runs where the bundle adjustment did not converge.

| | VisualSFM | | Colmap | | Ours_RF | | Global | |
|---|---|---|---|---|---|---|---|---|
| | $N_r$ | *rms* | $N_r$ | *rms* | $N_r$ | *rms* | $N_r$ | *rms* |
| *Remond* | 175 | — | 175 | 0.67 | 175 | 0.53 | 175 | 0.59 |
| *Campus* | 1040 | — | 1040 | — | 1040 | 0.56 | 1003 | 1.34 |

## Conclusions

In this article, we present two robust structure-from-motion (or image orientation) methods by combining the information of relative rotations and tie points. First, an incremental method involving single rotation averaging and linear translation estimation is proposed. Second, a global method is introduced as an alternative. This method uses an existing approach for rotation averaging, followed by simultaneous translation estimation for the whole block. In both procedures, we pay special attention to the robustness of the method and take appropriate steps to identify and eliminate outliers in the set of observations. The evaluation using three small benchmark data sets demonstrates that our approaches perform well. Moreover, experiments on some Internet data sets show that it is also possible to orient larger sets of unordered images. By comparing the new incremental method with the global one, we find that the incremental method shows better performance in terms of accuracy and is also more successful when applied to more challenging data sets, but the global method run significantly faster while yielding better results than comparable global approaches from the literature.

In future work, we will focus on further robustifying the results of relative orientation, which are used as input for both new orientation methods, for example, by detecting critical configurations with very small baseline or image pairs with repetitive structures. Also, when selecting tie points (see the section "Robust Tie Point Selection"), we will include a measure for taking into account the number of rays per point.

## Appendix: Rotation Errors

Given two similar rotations $\boldsymbol{R}_i$ and $\boldsymbol{R}_j$, $\theta$ is the angle difference we want to compute. We start by computing a value :

$$\alpha = \mathrm{trace}(\boldsymbol{R}_i \boldsymbol{R}_j^{-1})\ /3 \qquad (13)$$

where $\alpha$ is the average value of the main diagonal elements of $\boldsymbol{R}_i \boldsymbol{R}_j^{-1}$. We than compute the angular difference $\theta$ by

$$\theta = \arccos(\alpha) \cdot 180/\boldsymbol{\pi} \qquad (14)$$

## Acknowledgments

Xin Wang would like to thank the China Scholarship Council for financially supporting his PhD study at Leibniz Universität, Hannover, Germany.

[2] More information about VisualSFM and Colmap can be found at http://ccwu.me/vsfm/doc.html and https://demuc.de/colmap (both accessed 15 November 2018).

[3] When running VisualSFM and Colmap, we used the default settings provided by the authors.

## References

Abdel-Aziz, Y.I.; Karara H.M., 1971. Direct linear transformation from comparator coordinates into object-space coordinates. ASP Symp. Close-Range Photogrammetry, University of Illinois, Urbana, Ill., USA, pp. 1-18.

Agarwal, S., Mierle, K. et al., 2007. Ceres Solver. http://ceres-solver.org (accessed 08.05.2017).

Agarwal, S., Snavely, N., Simon, I., Seitz, S. M., Szeliski, R., 2009. Building Rome in a day. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp.72-79.

Arie-Nachimson, M., Kovalsky, S.Z., Kemelmacher-Shlizerman, I., Singer, A., Basri, R., 2012. Global motion estimation from point matches. In: Proceedings of the International Conference on 3D Imaging, Modeling, Processing, Visualization and Transmission (3DIMPVT), pp. 81–88.

Arrigoni, F., Fusiello A., Rossi B., 2016. Camera motion from group synchronization. In: Proceedings of the IEEE International conference on 3D Vision (3DV), pp.546-555.

Chatterjee, A., Govindu, V.M., 2013. Efficient and robust large-scale rotation averaging. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 521–528.

Cohen, A., Zach, C., Sinha, N.S., Pollefeys, M., 2012. Discovering and Exploiting 3D Symmetries in Structure from Motion. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.1514-1521.

Cui, Z., Tan, P., 2015. Global Structure-from-Motion by Similarity Averaging. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp.864-872.

Cui, Z., Jiang, N., Tang, C., Tan, P., 2015. Linear global translation estimation with feature tracks. In: Proceedings of the British Machine Vision Conference (BMVC).

Cui, H., Gao, X., Shen, S., Hu., Z., 2017. HSfM: Hybrid-Structure-from-Motion. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1212-1221.

Farenzena, M., Fusiello, A., Gherardi, R., 2009. Structure-and-motion pipeline on a hierarchical cluster tree. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshop, pp. 1489-1496.

Fischler, M. A., Bolles, R. C., 1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM 24(6), 381–395.

Govindu, V.M., 2001. Combining two-view constraints for motion estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2, pp. II–218.

Goldstein, T., Hand, P., Lee, C., et al., 2016. Shapefit and shapekick for robust, scalable structure from motion. In: Proceedings of the European Conference on Computer Vision (ECCV). Springer, pp.289-304.

Hartley, R., Zisserman, A., 2003. Multiple View Geometry in Computer Vision. Cambridge University Press.

Hartley, R., Aftab, K., Trumpf, J., 2011. L1 rotation averaging using the Weiszfeld algorithm. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 1, pp. 3041–3048.

Hartley, R., Trumpf, J., Dai, Y., Li, H., 2013. Rotation averaging. *International Journal of Computer Vision*, 103 (3), pp. 267–305.

Havlena, M., Torii, A., Knopp, J., Pajdla, T., 2009. Randomized Structure from Motion Base on Atomic 3D Models from Camera Triplets. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.2874-2881.

Jiang, N., Cui, Z., Tan, P., 2013. A global linear method for camera pose registration. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 481–488.

Jiang, N., Lin,W.-Y., Do, M. N., Lu, J., 2015. Direct structure estimation for 3d reconstruction. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2655–2663.

Kilian, K., 1955. Über das Rückwärtseinschneiden im Raum. Österreichische Zeitschrift für Vermessungswesen 43(6), pp. 97-104.

Lowe, D.G., 2004. Distinctive Image Features from Scale-invariant Keypoints. *International Journal of Computer Vision*, 60(2), pp. 91–110.

Martinec, D., Pajdla, T., 2007. Robust rotation and translation estimation in multiview reconstruction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–8.

Mayer, H., 2014. Efficient hierarchical triplet merging for camera pose estimation. In: Proceedings of German Conference on Pattern Recognition (GCPR). Springer, Berlin, pp. 99–409.

Moulon, P., Monasse, P., 2012. Unordered feature tracking made fast and easy. European Conference on Visual Media Production, CVMP.

Moulon, P., Monasse, P., Marlet, R., 2013. Global fusion of relative motions for robust, accurate and scalable structure from motion. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV).

Nistér, D., 2004. An efficient solution to the five-point relative pose problem. *IEEE Transactions on Pattern Analysis and Machine Intellintelligence*, 26 (6), pp.756–770.

Ozyesil, O., Singer, A., 2015. Robust camera location estimation by convex programming. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2674-2683.

Reich, M., Heipke, C., 2015. Global rotation estimation using weighted iterative lie algebraic averaging. ISPRS Annals of Photogrammetry Remote Sensing and Spatial Information Science, II-3/W5, pp.443–449.

Reich, M., Heipke, C., 2016. Convex image orientation from relative orientations. ISPRS Annals of Photogrammetry Remote Sensing and Spatial Information Science, III-3, pp.107-114.

Reich, M., Yang, M. Y., Heipke, C., 2017. Global robust image rotation from combined weighted averaging. *ISPRS Journal of Photogrammetry & Remote Sensing*, 127, pp.89-101.

Schönberger, J. L., Frahm, J. M., 2016. Structure-from-Motion Revisited. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Snavely, N., Seitz, S. M., Szeliski, R., 2006. Photo tourism: exploring photo collections in 3d. *Acm Transactions on Graphics*, 25(3), pp.835-846.

Strecha, C., von Hansen, W., Van Gool, L., Fua, P., Thoennessen, U., 2008. On benchmarking camera calibration and multi-view stereo for high resolution imagery. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–8.

Toldo, R., Gherardi, R., Farenzena, M., Fusiello, A., 2015. Hierarchical structure-and-motion recovery from uncalibrated images. Computer Vision & Image Understanding, 140, pp.127-143.

Wilson, K., Snavely, N., 2014. Robust global translations with 1DSFM. In: Proceedings of the European Conference on Computer Vision (ECCV). Springer, pp. 61–75.

Wu, C. 2013. Towards Linear-Time Incremental Structure from Motion. In: Proceedings of the IEEE Conference on 3dtv, pp.127-134.

Wang, X., Zhan, Z. Q., Heipke, C., 2017. An efficient method to detect mutual overlap of a large set of unordered images for structure-from. ISPRS Annals of Photogrammetry Remote Sensing and Spatial Information Science, IV-1-W1, pp.191-198.

Wang, X., Rottensteiner, F., Heipke, C., 2018. ROBUST IMAGE ORIENTATION BASED ON RELATIVE ROTATIONS AND TIE POINTS. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, IV-2, pp. 295-302.

Wang, X., Rottensteiner, F., Heipke, C., 2019. Structure from motion for ordered and unordered image sets based on random k-d forests and global pose estimation. *ISPRS Journal of Photogrammetry & Remote Sensing*, 147, pp.19-41.

Zheng, Y., Kuang, Y., Sugimoto, S., Okutomi, M., 2013. Revisiting the PnP Problem: A Fast, General and Optimal Solution. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp.2344-2351.