Robustifying relative orientations with respect to repetitive structures and very short baselines for global SfM

Xin Wang[1] Teng Xiao[1,2] Michael Gruber[3] Christian Heipke[1]
[1]Institute of Photogrammetry and GeoInformation, Leibniz Universität Hannover, Germany
[2]School of Geodesy and Geomatics, WuHan University, China
[3]Vexcel Imaging GmbH, Austria

[1]{wang, heipke}@ipi.uni-hannover.de [2]xiaoteng@whu.edu.cn [3]michael.gruber@vexcel-imaging.com

Abstract

Recently, global SfM has been attracting many researchers, mainly because of its time efficiency. Most of these methods are based on averaging relative orientations (ROs). Therefore, eliminating incorrect ROs is of great significance for improving the robustness of global SfM. In this paper, we propose a method to eliminate wrong ROs which have resulted from repetitive structure (RS) and very short baselines (VSB). We suggest two corresponding criteria that indicate the quality of ROs. These criteria are functions of potentially conjugate points resulting from local image matching of image pairs, followed by a geometry check using the 5-point algorithm combined with RANSAC. RS is detected based on counts of corresponding conjugate points of the various pairs, while VSB is found by inspecting the intersection angles of corresponding image rays. Based on these two criteria, incorrect ROs are eliminated. We demonstrate the proposed method on various datasets by inserting our refined ROs into a global SfM pipeline. The experiments show that compared to other methods we can generate the better results in this way.

1. Introduction

In recent years, Structure-from-Motion (SfM) has undergone an impressive development in both computer vision and photogrammetry [1, 29, 30]. The so called incremental SfM has received a notable amount of attention, demonstrated e.g. by the success of the software packages Bundler [26, 27], VisualSFM [33, 34], and COLMAP [23, 24]. The general idea is that one good initial image pair is firstly selected to do stereo reconstruction. Additional images are sequentially chosen based on some criteria to extend the photogrammetric block, and bundle adjustment is repetitively used to refine the results. As [6, 14, 29] demonstrated, this approach is impeded by a long computational time and artefacts such as visual drift. To overcome these drawbacks, [2, 5, 7, 14, 17, 18, 30, 31] presented a global solution. Global SfM is typically separated into two steps, global rotation averaging [8, 9, 10, 11, 20, 21, 32] and global translation estimation [5, 29, 30].

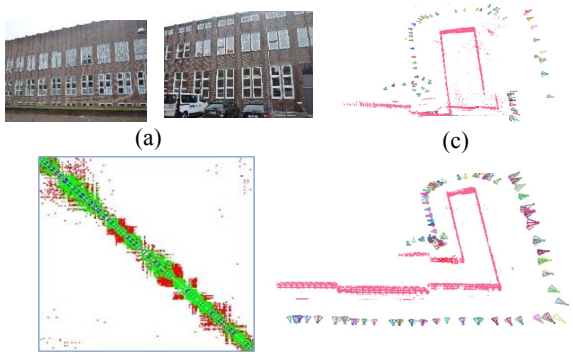


Figure 1. An example scene with repetitive structure and image pairs with very short baselines. (a) Two example images with repetitive structure. (b) Ground truth of overlap graph with the image IDs on the horizontal and vertical axes; green pixels denote overlapping image pairs, red pixels represent non-overlapping pairs with incorrect ROs due to RS, and blue pixels indicate the corresponding VSB image pairs. (c) incorrect reconstruction without eliminating incorrect ROs. (d) accurate reconstruction after eliminating incorrect ROs using the method suggested in this paper.

The exterior orientation parameters of all available images are first simultaneously estimated, followed by only one final bundle adjustment. Compared to incremental SfM, global SfM is more sensitive to outliers in relative orientations (ROs) between image pairs [6, 30, 31].

Many outliers in ROs can be eliminated by using the five-point method combined with RANSAC [7, 19]. However, incorrect ROs typically remain undetected, mainly due to two reasons: 1) repetitive structure (RS), and 2) critical configurations stemming from very short baselines (VSB).

Repetitive structure is a characteristic of a single image and describes the fact that many parts in the image look similar. Typically, the reason is that the 3D structure of the scene is repetitive (this is why we speak about repetitive structure, and not about repetitive texture, as texture refers to the 2D image space). As a consequence, when extracting features, the resulting descriptors are rather similar. Matching images with repetitive structure leads to many ambiguous point pairs and many outliers. In our context an *image pair due to repetitive structure* is a non-overlapping

image pair, for which potentially conjugate points are being found due to these ambiguities. Such non-overlapping, but nevertheless similarly looking images can e.g. stem from a set of façade images, when the façade is somewhat symmetric. If enough such incorrect point pairs are extracted, it is possible that the 5-point algorithm is not able to detect the error, and incorrect relative orientation parameters are derived.

A critical configuration with a very short baseline (VSB) results from improper image acquisition planning, e.g. when images are taken in different directions, but from basically the same projection centre. In addition, crowd source datasets such as images available on the Internet are widely used nowadays. These datasets may contain pairs with critical configurations as well.

In Fig. 1, we show an example with both RS and VSB image pairs. (a) shows repetitive structure of windows and bricks, (b) depicts the ground truth of the overlap graph where green pixels denote the desired overlapping image pairs; (c) and (d) show the reconstruction result when applying the global SfM method of [30] without and with applying our method to eliminate outliers in ROs, respectively. It is obvious that the reconstruction is more reasonable after eliminating the incorrect ROs.

In this paper, we present a novel method to eliminate outliers in ROs which are due to RS and VSB. We suggest two corresponding criteria that indicate the quality of ROs. These criteria are functions of potentially conjugate points resulting from local image matching of image pairs, followed by a geometry check using the 5-point algorithm combined with RANSAC. RS is detected based on counts of corresponding conjugate points of the various pairs, while VSB are found by inspecting the intersection angles of corresponding image rays. The contributions are *threefold*: *First*, we publish three benchmark datasets with repetitive structure and very short baseline image pairs, as well as ground truth ROs as shown in Fig.1(b) for one example. *Second*, we present a method to compute the probability of an image pair to stem from repetitive structure or a very short baseline. *Finally*, based on the related criteria, we propose a method to eliminate incorrect ROs.

This paper is organized as follows: Section 2 outlines relevant related work. In section 3, we introduce the method of computing the mentioned criteria Section 4 describes the algorithm to eliminate incorrect ROs. In section 5, we report experimental results on various datasets. Finally, section 6 concludes our work.

2. Related work

In this section, we review the related work on detecting blunders in ROs. A conventional way based on RANSAC is to use the epipolar geometry constraint, in which the essential (or fundamental) matrix is estimated after image matching. The ROs are only considered correct if a minimum number of point pairs conforms with the model of central perspective [15]. Although many wrong ROs can be eliminated in this way, non-overlapping pairs may still exist resulting from RS and VSB. Many works try to detect these errors. Here, we divide them into three categories: missing correspondences analysis, loop consistency constraint analysis and other methods.

Missing correspondences analysis. Zach et al. [35] first presented the so-called missing correspondences among image triplet to infer incorrect ROs. The main idea is that if a substantial portion of correspondences between two images from the triplet cannot be observed by the third image, then the relative orientation between the two images is potentially incorrect. The authors used a Bayesian framework for all image triplets to check the correctness of the corresponding image pairs. Roberts et al. [22] improved this idea by verifying the incorrect ROs via an expectation-maximization method which integrates the cues of missing correspondence and timestamp information, however, the latter is not available in general, e.g., for unordered images the acquisition sequence is unknown. Jiang et al. [15] extended the missing correspondences idea by minimizing the number of missing correspondences across the entire reconstruction instead of the triplets. Specifically, a spanning tree is first built and then problematic ROs are iteratively detected in a greedy way. As a consequence, the method may get stuck in a local minimum.

Loop consistency constraint analysis. Zach et al. [36] were among the first to adopt the loop consistency constraint to infer the validity of ROs. They first generate cycles in the overlap graph; the relative rotations are then concatenated within each cycle, as a result an identify mapping should be obtained if all ROs in the cycle are correct. Potential errors are indicated by using a Bayesian network. Reich et al. [21] presented a sequential graph optimization method to eliminate incorrect relative rotations. Both [36] and [21] need a long processing time when dealing with a large image dataset where all the relative rotations need to be considered. Shen et al. [25] presented a graph-based consistent method, where a minimum spanning tree is incrementally expanded by checking the loop consistency within a triplet until all available images are included in the tree. In previous work [30] we presented a triplet loop closure constraint based on relative rotations and translations. We eliminate ROs if the closure error of all corresponding triplets is above a predefined threshold and then use [10] and a newly developed method for global rotation averaging and translation estimation, respectively.

Other methods. Wilson and Snavely [31] proposed a 1DSfM approach. Their basic idea is to project the 3D relative translations into different 1D direction vectors. They then used a kernel density estimator to sample these directions, wrong ROs normally stand out clearly in the direction of the 1D vector. However, as the authors write,

their method fails in the presence of repetitive structure. Wang et al. [28] presented a hierarchical ROs selection method for repetitive structure. They first built a minimum spanning tree (MST), and then used a hierarchical scheme for RO selection. Finally, ROs are validated to avoid a structure collapse. The method only selects validated ROs along the MST which may break up the block of images, and image pairs with very short baselines are not dealt with. To solve for artefacts caused by repetitive structure, Cohen et al. [3] recovered various symmetrical structures using geometric and appearance cues to refine their bundle adjustment process. Heinly et al. [13] presented a post-processing step using the SfM result as input for their method. They split the overlap graph into subsets and use conflicting correspondences to identify repetitive structure. The subsets of the overlap graph which are free from conflict are then merged into a correct reconstruction.

Compared with the above-mentioned methods, we propose a pipeline that can deal with RS ROs, with which loop consistency constraint analysis has difficulties, and can also deal with VSB which missing correspondences analysis has problems with.

3. Detecting ROs of repetitive structures and very short baseline

In this section, we first present the method to detect error ROs that are due to repetitive structure, and a criterion that indicates the degree of RS is introduced. Then, the method of detecting incorrect ROs that result from VSB is proposed, a criterion that indicates the degree of VSB is also presented.

3.1. Detecting ROs of repetitive structures

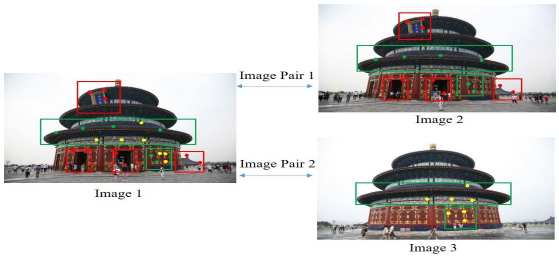


Figure 2. Image pairs of non-repetitive and repetitive structures, green boxes denote RS and red boxed denote non-RS. Red points are the correspondences from non-RS, green ones are the correspondences from RS in image pair 1, yellow points are the correspondences from RS in image pair 2.

To distinguish RS ROs from all ROs, one normally takes advantage of non-repetitive structure, also present in the images. If two images depict a scene with 100% repetitive structure, even interactively we cannot tell overlapping image pairs apart from non-overlapping ones. Fig.2 shows an example of two image pairs. From both image pairs correspondences can be generated by image matching as the red, green and yellow points in Fig. 2 show. Visually, we can easily tell that image pair 1 is a pair with overlap since it contains non-repetitive structure (see the red boxes). In contrast, image pair 2, which is non-overlapping, does not have such non-repetitive structure. We argue that:

- assuming a constant image size (in pixels) the number of features per image is approximately constant (if the image size varies, a normalisation needs to be carried out).
- given a constant overlap, overlapping pairs have more conjugate points than non-overlapping pairs, because the latter do not have any inliers with respect to a central perspective model, as the pair has no overlap.

We use these two hypotheses to detect and subsequently eliminate non-overlapping image pairs which survived the five-point geometry check.

We first construct a set S of feature point correspondences S: S= $\{S_1, S_2, S_3, \ldots, S_n\}$, n is the number of images, S_i is the set of feature points in the i-th image, each represented by an ID (in Fig. 2, S_1 contains the red, green and yellow points). Then, Q_{ij}^i is the set of feature point IDs of the i-th image that have matches between the i-th and the j-th image, such as the red and green points in image pair 1 of Fig. 2. Now, we construct the difference sets between S and Q, denoted by $D_i^j = S_i \backslash Q_{ij}^i$ for image i and $D_j^i = S_j \backslash Q_{ji}^j$ for image j. Since S_i is assumed to be approximately constant, and overlapping pairs are assumed to have more matches than non-overlapping ones (see hypotheses above), the number of IDs in both, D_i^j and D_j^i is small for overlapping pairs, and large otherwise. In addition, we consider the IDs in D_i with respect to the other images which have correspondences with the i-th image and generate a vector $\boldsymbol{g}_{ij}=[g_i^1, g_i^2, g_i^3, \ldots, g_i^n]$, where $g_i^j=0$ and $g_i^k=|\{f \in D_i^j \,\&\&\, f \text{ is a feature matched to the } k\text{-th image}\}|$, $||$ is the operator which returns the number of set elements. We finally use equation (1) to compute a value RS_{ij} representing the degree of repetitive structure of image i and j.

$$RS_{ij} = (|D_i^j| + |D_j^i|)(\boldsymbol{g}_{ij}^T \boldsymbol{g}_{ji})/(|Q_{ij}^i| + |Q_{ji}^j|) \quad (1)$$

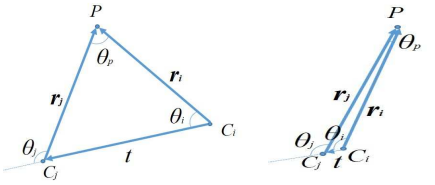
As mentioned, overlapping image pairs are assumed to have a small number of elements in the difference set and the value of $\boldsymbol{g}_{ij}^T \boldsymbol{g}_{ji}$ should be small as well, while the number of correspondences in the denominator of (1) is large. Thus, the smaller RS_{ij} is, the more probable it is that the image pair does overlap and the RO is correct, rather than being solely due to repetitive structure.

Our hypotheses are violated if images only overlap partially, in particular if the overlapping area is small. We argue that ROs of such image pairs are not robust either, thus if there is enough proper overlap between the images, it is reasonable to eliminate pairs with small overlap also.

3.2. Detecting ROs of very short baseline

Critical configurations stemming from very short baselines (VSB) decrease the robustness of SfM both in

structure and motion estimation, because VBS lead to small intersection angles and thus imprecise coordinates of the intersection point during triangulation and global translation estimation.



Standard case Very short baseline

Figure 3. Two-view geometry constrain

In Fig. 3, we show a standard case of a two-view geometry with a relatively wide baseline and a case with VSB. P is the object point, C_i and C_j are the projection centres of images i and j, t represents the baseline vector from C_i to C_j, r_i and r_j are two projection rays, θ_i is the intersection angle of t and r_i, θ_j is the intersection angle of t and r_j, θ_p is the intersection angle of r_i and r_j. In the standard case, we obtain the inequality $0 < \theta_i < \theta_j < \pi$, whereas, for VSB, an approximate equation $0 < \theta_i \approx \theta_j < \pi$ can be set up. We use these two equations to distinguish cases with very short baselines from standard cases. For a standard case, we obtain:

$$cos^{-1}\frac{(Rx_j)^T t}{|x_j|\cdot|t|} > cos^{-1}\frac{x_i{}^T t}{|x_i|\cdot|t|} \quad (2)$$

$$\text{i.e.} \quad \frac{x_i{}^T t}{|x_j|\cdot|t|} > \frac{(Rx_j)^T t}{|x_i|\cdot|t|} \quad (3)$$

where R is the relative rotation and t is the relative translation. x_i and x_j are the image coordinates of conjugate points as predicted from image matching. We can rewrite these equations as

$$(|x_j|\, x_i{}^T - |x_i|\,(Rx_j)^T)\frac{t}{|t|} > 0 \quad (4)$$

$$cc_{ij}(\mathrm{R}) = |\,(|x_j|\, x_i{}^T - |x_i|\,(Rx_j)^T)\frac{t}{|t|}\,| \quad (5)$$

For VSB, we have $0 < \theta_i \approx \theta_j < \pi$, this means $cc_{ij}(R)$ should close to 0. We can also derive the formulae $X_P = X_{Ci} + \lambda_i R_i x_i$ and $X_P = X_{Cj} + \lambda_j R_j x_j$, where X_P denotes the coordinate vector of object point P and X_{Ci}, X_{Cj} are the projection centres C_i and C_i, λ_i and λ_j are the scale factors, R_i and R_j are the corresponding rotation matrices from image to object space.

$$X_P = X_{Ci} + \lambda_i R_i x_i = X_{Cj} + \lambda_j R_j x_j \quad (6)$$

which can be rewritten as

$$x_j = \lambda_{ij}(R\, x_i + v_i t) \quad (7)$$

$$\lambda_i R_i x_i = X_{Cj} - X_{Ci} + \lambda_j R_j x_j \quad (8)$$

where $t = R_j^{-1}(X_{Ci} - X_{Cj})$ is the baseline vector, $\lambda_{ij} = \lambda_i/\lambda_j$, $v_i = 1/\lambda_i$. For VSB $X_{Ci} = X_{Cj}$ and we have $\lambda_i R_i x_i \approx \lambda_j R_j x_j$ and $R = R_i R_j^{-1}$, which also leads to $cc_{ij}(R)$ being close to 0.

As for each pair of correspondences we have one $cc_{ij}(R)$, we use the mean value VSB_{ij} of $cc_{ij}(R)$ in equation (9) as a criterion to quantify the degree of an image pair to have a VSB: the smaller the VSB_{ij} is, the higher the probability that the image pair is a VSB pair.

$$VSB_{ij} = \text{avg}\,(cc_{ij}(R)) \quad (9)$$

Note, there exists an implicit assumption that the length of baselines cannot equal 0 when decomposing the essential matrix into relative rotation and translations [11, 16]. However, relative rotation can obviously be computed for image pairs of 0 baseline as this is the task of transforming images into epipolar geometry, and equation (9) remains correct in this case; the corresponding derivation can be found in the appendix.

In order to investigate, in how far the assumption is relevant for us, we design a simulation experiment to see whether rotation can still be accurately estimated when the baseline is very short or even exactly 0.

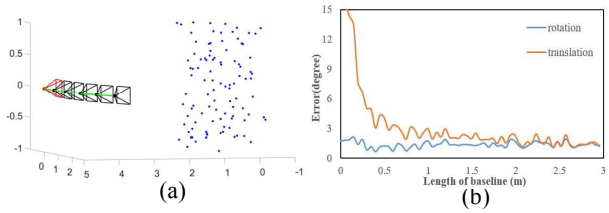


Figure 4. Simulation experiment. (a) shows the poses of the simulated cameras and the position of object points, the red frame is the fixed camera and the black frames denote the different projection centres of the second camera. (b) shows the error in degree of relative rotation and translation for different baseline lengths.

As Fig. 4(a) shows, a set of 100 3D points is randomly generated in a cube of $[-1,1]^3$. We simulate two cameras with focal length 3500 pixels and an image size of 1200× 800 pixels viewing these 3D points. We keep one camera fixed at point (5,0,0) and start to move the second camera from this point along an arc (shown by the green line) with centre of (0,0,0) and 5m radius in small steps, until these two cameras are 3m (arc distance) away from each other. The corresponding rotation matrices are designed by requiring these two cameras to be able to view all 3D points, then, image pairs with known exterior parameters are simulated. The image coordinates of the 3D object points are generated via the collinearity equations with 0.2 (pixel) standard deviation Gaussian noise. The relative orientations of these image pairs are then estimated using the five-point algorithm from the resulting conjugate point coordinates, and they are compared to the simulated exterior parameters. Since the relative translation is normalized and the scale is unknown, we can only compare the translation directions. The arc between two cameras is transferred into baseline length. We obtain results showed in Fig. 4(b); the relative rotation error remains stable, while the relative translation error increases as the baseline decreases, which means that the relative rotation can be robustly estimated, while the relative translation cannot, when the baseline is very short.

4. Eliminating incorrect ROs

In this section, we introduce our method for eliminating incorrect ROs. After calculating the criteria for *RS* and *VSB* for each image pair, a comprehensive method that cannot only filter as many RO outliers as possible, but also keep the photogrammetric block intact as a whole is presented.

The input ROs are computed and filtered using the five-point algorithm [19], we then generate a view graph $G = (\boldsymbol{\nu}, \boldsymbol{\varepsilon})$ with the largest number of connected images (the number is indicated by N_e), where the vertice set $\boldsymbol{\nu}$ represents the images and the edge set $\boldsymbol{\varepsilon}$ denotes the ROs. To keep the block intact we assign each edge in *G* a weight $wmst_{ij}$ calculated by (10) and construct a minimum spanning tree for *G,* described in algorithm 1, where *norm*(.) normalizes the corresponding values into the range (0, 1). We add the selected ROs in the spanning tree into a set *O*.

$$wmst_{ij} = norm(RS_{ij}) + norm(1/VSB_{ij}) \quad (10)$$

Algorithm 1 Constructing a minimum spanning tree

Input The view Graph $G = (\boldsymbol{\nu}, \boldsymbol{\varepsilon})$ and the corresponding $wmst_{ij}$ values.

Output The set *O* with $N_e - 1$ ROs connecting all vertices in $\boldsymbol{\nu}$.

1. Select one RO with smallest $wmst_{ij}$, add this RO into *O* and the vertices into a set *L*, then compute the difference set of *L* and $\boldsymbol{\nu}$ by using $DS = \boldsymbol{\nu} \backslash L$.
2. Figure out vertices from *DS* that have ROs linking to any vertex in *L*, and choose one vertex with RO of smallest $wmst_{ij}$, add this RO into *O* and the corresponding vertex into *L*, *DS* is updated by $\boldsymbol{\nu} \backslash L$.
3. Repeat 2 until all vertices in $\boldsymbol{\nu}$ are included by *L*.

To eliminate incorrect ROs due to RS, we set a_{ij}=1 if image *i* and *j* overlap, otherwise a_{ij}=0. For each image we keep at least N_r ROs. From the input ROs, $M_r \times N_e$ ROs are selected with the smallest sum of corresponding RS_{ij} values; M_r is a free parameter which needs to be set according to experience. Based on the above mentioned requirements, we then do the selection by minimizing equation (11).

$$\min_{a_{ij}} RS_{ij} a_{ij} \quad \text{subject to:}$$
$$\forall i \in \nu, \sum_j a_{ij} \geq N_r, \text{ where } (i,j) \in \boldsymbol{\varepsilon};$$
$$\forall (i,j) \in O, a_{ij} = 1; \forall (i,j) \in \boldsymbol{\varepsilon}, \sum a_{ij} = M_r \times N_e; \quad (11)$$

We apply the same idea to select image pairs with reasonably long baselines, as equation (12) shows. We set b_{ij}=1 if the baseline between images *i* and *j* is long enough, otherwise b_{ij}=0. Then, equation (12) is maximised, note that the total number of selected ROs is $M_v \times N_e$. where M_v is a free parameter similar to M_r.

$$\max_{b_{ij}} VSB_{ij} b_{ij} \quad \text{subject to:}$$
$$\forall i \in \nu, \sum_j b_{ij} \geq N_r, \text{ where } (i,j) \in \boldsymbol{\varepsilon};$$
$$\forall (i,j) \in O, b_{ij} = 1; \forall (i,j) \in \boldsymbol{\varepsilon}, \sum b_{ij} = M_v \times N_e; \quad (12)$$

The *Mosek* library [1] that can solve linear optimization problems is used to obtain solutions of equation (11) and (12). Finally, we eliminate ROs with a_{ij}=0 or $b_{ij} = 0$, while ROs with a_{ij}=1 and $b_{ij} = 1$ are kept for further global SfM.

5. Experiments

In this section, we present a detailed evaluation of our method. The experiments are conducted with various datasets including three image datasets with *ground truth* for the ROs. Ground truth means that we know for each potential pair of images whether the pair overlaps, whether it is non-overlapping with repetitive structure and whether it has a very short baseline, as shown in Fig. 1(b); ground truth has been established manually. Further, we investigated four public datasets with highly repetitive structure and a challenging dataset, which many global SfM methods cannot deal with. We feed the refined ROs to the global SfM pipeline of our previous work [30] consisting of global rotation averaging by [10] and global translation estimation and compare the results with other methods.

5.1. Evaluation on three datasets with RO ground truth

To demonstrate the performance of our outlier elimination method we generate three image datasets with RO ground truth[2]. We choose three buildings with repetitive structure (we use abbreviations B1, B2 and B3 to represent these three datasets henceforth), images are acquired around these buildings. The camera is rotated along the vertical axis several times at each exposure station to obtain VSB image pairs, see Fig. 6, where the VSB ROs (blue pixels in the second column) are all located along the main diagonal of the overlap graph. Tab. 1 offers detailed information on these three datasets. N_p is the number of input ROs. In this experiment, N_r is set to 5 to make sure that each image is reliably connected to the block, M_r and M_v are selected according to how redundant N_p is (for example, in B1, on average each image in N_e has 2011/182=11 overlapping images which means for each image, the corresponding ROs are somewhat redundant when conducting global SfM) and the number of RS and VSB ROs. For B2 and B3 larger values are chosen for M_r and M_v than for B1, because B2 and B3 have many more redundant ROs than B1 does, see Tab. 1. Thus, we can choose more non-RS and non-VSB ROs for each image of B2 and B3.

[1] More information about *Mosek* can be found at www.mosek.com.

[2] These datasets are available on github.com/wx7531774/SFM_results.

	N_e	N_p	Correct ROs	RS	VSB	N_r	M_r	M_v
B1	182	2011	1089	784	138	5	6	10
B2	215	6357	1935	4030	392	5	10	27
B3	342	4956	3202	1422	332	5	10	13

Table 1. Description of the generated RO datasets.

To validate the performance of our method in detecting RS and VSB ROs, and to see how RS and VSB ROs affect global SfM, we conduct experiments using four pipelines: RS and VSB ROs elimination (indicated by "Ours" in the following tables and figures), no ROs elimination, only RS elimination, and only VSB elimination. Based on the ground truth ROs, Tab. 2 provides the precision and recall values for detecting RS, VSB and the correct ROs. We find that most ground truth ROs can be detected (recall is higher than 90%), however, some false positives are detected which leads to a lower precision. In addition, the precision of detecting VSB ROs is lower than that of detecting RS ROs. Thus, our method also eliminates some image pairs with should be kept for further processing. Nevertheless, the precision of "Ours" is higher than that of the other pipelines. Fig. 5 is the visualization of SfM results. We find that the results with our ROs elimination are the best. The other three pipelines generated artefacts shown in the green ellipses. We also conclude that both RS and VSB ROs can negatively affect global SfM. Thus, it is necessary to eliminate both types of errors.

	Detection of RS only		Detection of VSB only		Correct ROs after RS and VSB elimination	
	P	R	P	R	P	R
B1	80	95	67	93	**81**	97
B2	93	97	66	92	**94**	93
B3	81	90	62	96	**91**	93

Table. 2 Precision and Recall value in percent on detecting RS, VSB and correct ROs. P and R denote precision and recall.

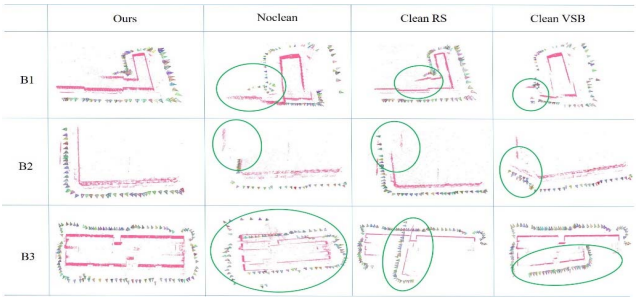


Figure. 5 Visualization of SfM results from different pipelines.

To further investigate our RO elimination method, we compare the results with those of Zach et al. [36], Wilson and Snavely [31] and Wang et al. [30]. Compared with the other methods, Tab. 3 shows that we have the lowest number of ROs attributed as correct; this is also illustrated by Fig. 6 where the overlap graph of "Ours" is filled with less black pixels than that of the other methods. We calculate the precision and recall values based on the ground truth of ROs, see Tab. 4. We find that the recall values of all methods are higher than 90%, which means they are all able to detect most of the correct ROs, whereas, our method has a much higher precision. This indicates that we detect many false positives, and thus consider fewer incorrect ROs as being correct.

	Ours	[36]	[31]	[30]
B1	1303	1684	1569	1846
B2	1918	5839	5391	5066
B3	3278	4349	3690	4776

Table. 3 Comparison of the no. of selected ROs from dif. methods

	Ours		[36]		[31]		[30]	
	P	R	P	R	P	R	P	R
B1	**81**	97	60	93	65	94	56	95
B2	**94**	93	35	99	37	99	40	99
B3	**91**	93	73	99	82	95	67	99

Table. 4 Comparison of precision and recall value in percent of different methods. P and R denote precision and recall.

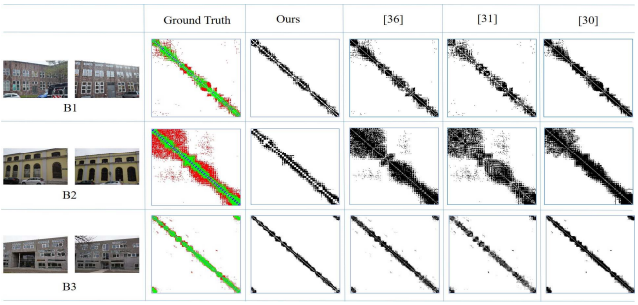


Figure.6 Overlap graphs of the three datasets from different methods. The first column shows sample images. The second column is the RO ground truth; green pixels denote correct ROs, red pixels are RS ROs, blue pixels denote VSB ROs. In the next columns, black pixels indicate that the corresponding ROs are kept and white pixels are the corresponding eliminated ROs.

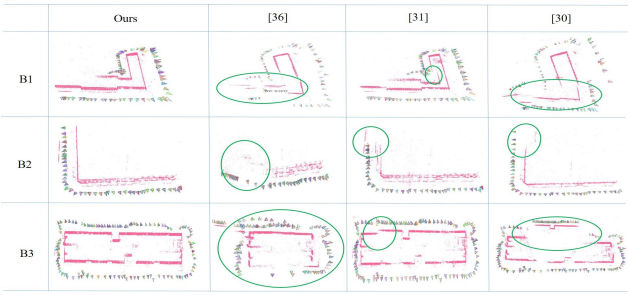


Figure 7. Visualization of SfM results from different methods.

Fig. 6 compares the RO elimination results. Comparing our results and the ground truth overlap graph, it can be seen that most of the red and blue pixels along the diagonal are eliminated, whereas, many red pixels still exist in [36], [31], and [30]. This means that the loop constraints used in [36], and constraints based on relative translation as employed in [31] and [30] are not capable to deal with repetitive structures in these three building datasets in the same way as our method. In addition, many blue pixels also remain in the results of [36], [31], and [30]. Fig. 7 shows a

visualization of the obtained SfM results. We come to similar conclusions as for Fig. 5: the proposed RO elimination method generates the best result, the other three methods generate artefacts shown in the green ellipses.

5.2. Evaluation on four public datasets with highly repetitive structure

In this section, we report on experiments on four public datasets with highly repetitive structure, namely the Temple of Heaven (ToH), Street (Str.) [3], Indoor (Ind.) and Stadium (Sta.) [25]. For these four datasets, there are no VSB ROs, so we only strive to eliminate RS ROs by setting N_r to be 5 and M_r to be 7, respectively, and only RS_{ij} is used in equation (10) to compute the $wmst_{ij}$. Tab. 5 contains the N_e of each dataset, the number of input ROs N_p, and the number of ROs attributed as correct.

	N_e	N_p	Number of correct ROs			
			Ours	[36]	[31]	[30]
ToH	341	56429	2387	34195	48540	48507
Str.	175	5171	1225	4544	4089	3832
Ind.	152	4740	1064	3380	3449	4059
Sta.	156	1733	1092	728	1338	1368

Table 5. Comparison of the no. of ROs from different methods.

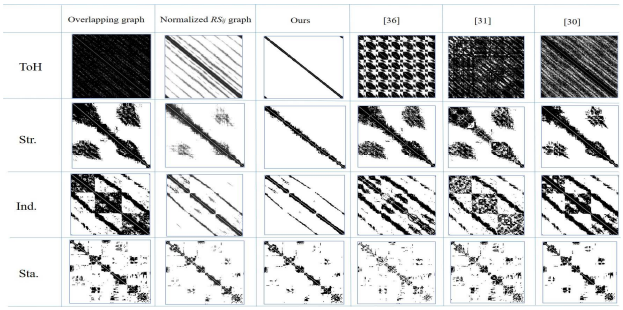


Figure 8. Overlap graph of the four public datasets obtained from the different methods. The second column is the overlap graph from the input ROs, black pixels denote that the corresponding ROs are available. This is also true for the last four columns. The third column is the normalized RS_{ij} graph.

Taking Fig. 8 and Tab. 5 into consideration together, we find that after having applied the five-point algorithms each dataset has a very redundant number of ROs, and incorrect ROs survived due to RS. The third column of Fig. 8 shows the normalized RS_{ij} graph. Specifically, we calculate the RS_{ij} values for each RO by using equation (1) and normalize them into the range (0,1). The brighter the pixel is, the larger is the corresponding RS_{ij} and, thus, the higher the probability that the image pair does not overlap. When comparing with other methods in Tab. 5, for TOH, Str. and Sta., we eliminate many more ROs and the number of ROs after elimination from our method is smallest, because based on M_r, only a limited number is selected by design. The last four columns in Fig.8 show the overlap graphs of the different methods. Our method results in a much cleaner overlap graph. For Sta., we keep more ROs than [36], which, however, reconstructs only part of the image block.

Since no ground truth ROs are provided for these datasets, to validate the quality of the correct ROs, we insert the ROs obtained by different methods into the global SfM pipeline [30] (note that we turn off the ROs elimination process of [30] when inserting the ROs of other methods). The reconstruction results are shown in Fig. 9; green ellipses denote artefacts. Compared to the other three methods, we don't have visual artefacts in our results and our method is able to improve the result of global SfM. For ToH, only part of the temple is reconstructed by [36], [31] and [30]. Probably due to the RS ROs in the Str. dataset, which results in an overlap graph with a pair of wings shown by Fig. 8, [36], [31] and [30] generated a folded reconstruction. As for Ind., many images are orientated into the wrong position by using [36], [31]and [30], and these methods also produce a folded reconstruction result. The reconstruction result of Sta. by [36] does not keep a consistent block, and it has the lowest number of ROs (728); and the original circular stadium is not closed by [31] and [30].

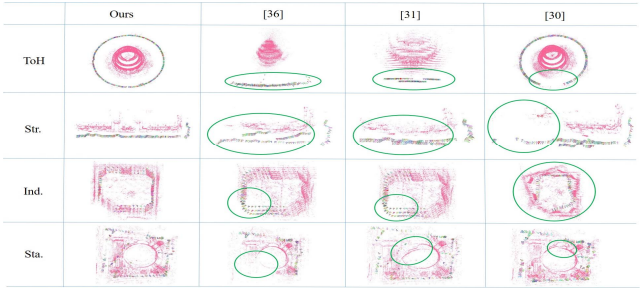
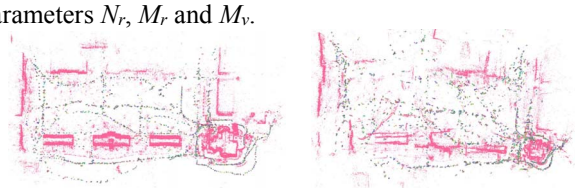


Figure 9. Visualization of SfM results from different methods.

5.3. Evaluation on a challenging dataset and limitation

To further explore the capability and limitation of the proposed method, we test one more challenging dataset, namely Quad [4], which many global SfM methods fail on [6, 30]. [30] claimed that the reasons are due to RS and critical geometric configuration of VSB image pairs, hence, we try our method on this dataset. Compared to the ground truth of reconstruction result [4], we obtain a reasonable result, see Fig. 10(a), by setting N_r =5, M_r = 15 and M_v =22, whereas, we obtain an incorrect reconstruction, see Fig.10(b), by setting N_r =5, M_r =20 and M_v =25. There are many incorrect reconstructions on buildings in Fig. 10(b). The limitation of our method is thus its sensitivity to the parameters N_r, M_r and M_v.



(a) (b)

Figure. 10 Visualization of Quad SfM result by different settings.

6. Conclusion

In this paper, we presented a novel method to eliminate blunders in ROs for conducting robust global SfM. We deal with incorrect ROs that are the result of repetitive structure and very short baselines. Criteria for these two cases are introduced, and incorrect ROs are eliminated based on these two criteria. For evaluation we processed various datasets and compared the results to those of other methods, and thereby demonstrated that our method can improve the robustness of global SfM. Since all the test datasets are from architectures and the current limitation of our new method is the parameter setting, we next plan to investigate different kind of images (for example, UAV images) and the idea of choosing reasonable values for the free parameters when facing different kinds of datasets.

Appendix

Proposition: The elements of the relative rotation matrix R can be accurately estimated from the essential matrix, no matter how short the baseline length is.

Proof. Inspired by the calculation of the essential matrix [11], we use the correspondences to obtain a solution of a 3 $\times$3 matrix L.

$$\boldsymbol{x}_j{}^T L \boldsymbol{x}_i = 0 \Leftrightarrow (\boldsymbol{x}_i{}^T \otimes \boldsymbol{x}_j{}^T)\, vec(L) = 0 \tag{13}$$

where $\otimes$ denotes the Kronecker product, $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are the coordinates of all the correspondences from image i and j, $vec(.)$ is the vectorization of a matrix. To take the relative rotation and translation into consideration, we rewrite (13) using the mixed-product property of $\otimes$ and equation (7),

$$\begin{aligned} &(\boldsymbol{x}_i{}^T \otimes (\lambda_{ij}(R\,\boldsymbol{x}_i + \boldsymbol{v}_i t)^T)\, vec(L) = 0 \\ \Leftrightarrow\ &(\boldsymbol{x}_i{}^T \otimes [\boldsymbol{x}_i{}^T \ \boldsymbol{v}_i])\,(I_3 \otimes [R \quad t]^T) vec(L) = 0 \end{aligned} \tag{14}$$

where λ_{ij} is eliminated as λ_{ij} is always larger than 0, $\boldsymbol{v}_i$ contains all the v_i values from all the correspondences,

$$\begin{Bmatrix} x_1{}^T \otimes [x_1{}^T \ v_1] \\ x_2{}^T \otimes [x_2{}^T \ v_2] \\ \vdots \\ x_m{}^T \otimes [x_m{}^T \ v_2] \end{Bmatrix} (I_3 \otimes [R \quad t]^T)\, vec(L) = 0 \tag{15}$$

For the whole correspondences, m is the number of correspondences, we can get equation (15). Then, we use U which represents the parameter matrix and z representing the unknowns, namely R, t and L

$$U = \begin{Bmatrix} x_1{}^T \otimes [x_1{}^T \ v_1] \\ x_2{}^T \otimes [x_2{}^T \ v_2] \\ \vdots \\ x_m{}^T \otimes [x_m{}^T \ v_2] \end{Bmatrix},\ z = (I_3 \otimes [R \quad t]^T)\, vec(L) \tag{16}$$

$$Uz = 0 \tag{17}$$

We analyse the U matrix, which is shown in (18)

$$\begin{Bmatrix} x_1{}^2 & x_1y_1 & x_1 & x_1v_1 & y_1x_1 & y_1{}^2 & y_1 & y_1v_1 & x_1y_1 & 1 & v_1 \\ x_2{}^2 & x_2y_2 & x_2 & x_2v_2 & y_2x_2 & y_2{}^2 & y_2 & y_2v_2 & x_2y_2 & 1 & v_2 \\ & & & & \vdots & & & & & & \\ x_m{}^2 & x_my_m & x_m & x_mv_m & y_mx_m & y_m{}^2 & y_m & y_mv_m & x_my_m & 1 & v_m \end{Bmatrix} \tag{18}$$

As we can easily find out, columns 2, 3 and 7 are equal to columns 5,9 and 10, so, $rank(U) \leq 9$. Therefore, when $m \geq 9$ the homogeneous equation (17) has three linearly independent basic solutions

$$\begin{aligned} \varepsilon_1 &= (0\,1\,0\,0\ -1\,0\,0\,0\,0\,0\,0\,0\,0) \\ \varepsilon_2 &= (0\,0\,1\,0\,0\,0\,0\,0\ -1\,0\,0\,0) \\ \varepsilon_3 &= (0\,0\,0\,0\,0\,0\,0\,1\,0\,0\ -1\,0\,0) \end{aligned} \tag{19}$$

So, the general solution space of z is

$$z = (I_3 \otimes [R \quad t]^T)\, vec(L) = (k_1\varepsilon_1 + k_2\varepsilon_2 + k_1\varepsilon_3) \tag{20}$$

in which k_1, k_2, k_2 are all real numbers. According to (16),

$$(I_3 \otimes [R \quad t]^T)\, vec(L) = vec([R \quad t]^T L) \tag{21}$$

Replacing (21) in equations (20) yields

$$[R \quad t]^T L = \begin{Bmatrix} R^T L \\ t^T L \end{Bmatrix} = \begin{Bmatrix} 0 & -k_1 & -k_2 \\ k_1 & 0 & -k_3 \\ k_2 & k_3 & 0 \\ 0 & 0 & 0 \end{Bmatrix} \tag{22}$$

Then, we obtain the relationship between R and L, t and L.

$$L = R \begin{Bmatrix} 0 & -k_1 & -k_2 \\ k_1 & 0 & -k_3 \\ k_2 & k_3 & 0 \end{Bmatrix} = R[\boldsymbol{k}]_\times \tag{23}$$

$$t^T L = 0 \tag{24}$$

and $\boldsymbol{k} = (k_1, -k_2, k_3)$. (23) means that the solution L of (10) is just the essential matrix [12]. So, R can be decomposed from L. From (23) and (24), when $t \neq \mathbf{0}$, we get $t = \pm R\boldsymbol{k}$ and R can be correctly estimated using SVD decomposition [11]. For t=0, it is clear that L has no relationship with t, and it is still related to R and $\boldsymbol{k}$. $\boldsymbol{k}$ can never be a zero vector which means t is not related to $\boldsymbol{k}$, because the solution z will be zero if $\boldsymbol{k}$ is a zero vector and this requires that the homogenous equation should have a full rank which means $Rank(U) = 12$, and this can never happen. So, R can still be correctly estimated from L when t=0, and the corresponding solution for t is not the correct relative translation, but the $\boldsymbol{k}$ vector.

Acknowledgements

Parts of works were done at Vexcel Imaging GmbH, Graz, financially supported by the EU project "*innoVation in geOspatiaL and 3D daTA — VOLTA"* funded under the Marie-Curie RISE scheme as no. 734687. The author Xin Wang would like to thank the China Scholarship Council (CSC) for financially supporting his PhD studying at Leibniz Universität Hannover, Germany. The author Xiao Teng would like to thank the Graduate Student Exchange Program of Wuhan University for his visiting scientist scholarship for studying in Germany.

References

[1] S. Agarwal, N. Snavely, I. Simon, S. M. Seitz, R. Szeliski. Building Rome in a day. *In Proc. ICCV*, pages 72-79, 2009.

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

[3] A. Cohen, C. Zach, S. N. Sinha, and M. Pollefeys. Discovering and exploiting 3d symmetries in structure from motion. *In Proc. CVPR*, pages 1514-1521, 2012.

[4] D. Crandall, A. Owens, N. Snavely, and D. P. Huttenlocher. Discrete-continuous optimization for large-scale structure from motion. *In Proc. CVPR*, pages 3001–3008, 2011.

[5] Z. Cui, N. Jiang, C. Tang, and P. Tan. Linear global translation estimation with feature tracks. *In Proc. BMVC*, 2015.

[6] Z. Cui, P. Tan. Global Structure-from-Motion by Similarity Averaging. *In Proc. ICCV*, 2015.

[7] M. Fischler, R. Bolles. Random sample consensus: a paradigm for model fitting with application to image analysis and automated cartography. *Commun. ACM,* 24: 381-395, 1981.

[8] V. M. Govindu. Combining two-view constraints for motion estimation. *In Proc. CVPR*, pages 218–225, 2001.

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

[10] A. Chatterjee, V. M. Govindu. Efficient and robust large-scale rotation averaging. *In Proc. ICCV*, pages 521-528, 2013.

[11] R. Hartley and A. Zisserman. *Multiple view geometry in computer vision*. Cambridge University Press, 2003.

[12] R. Hartley, J. Trumpf, Y. Dai, and H. Li. Rotation averaging. *International Journal of Computer Vision (IJCV)*, 103:267–305, 2013.

[13] J. Heinly, E. Dunn, and J.-M. Frahm. Correcting for duplicate scene structure in sparse 3D reconstruction. *In Proc. ECCV*, page 780-795, 2014.

[14] N. Jiang, Z. Cui, and P. Tan. A global linear method for camera pose registration. *In Proc. ICCV*, 2013.

[15] N. Jiang, P. Tan, and L.-F. Cheong. Seeing double without confusion: Structure-from-motion in highly ambiguous scenes. *In Proc. CVPR*, pages 1258-1465, 2012.

[16] H. Longuet-Higgins. A Computer Algorithm for Reconstructing a Scene from Two Projections, *Nature*, 293(10):133-135, 1981.

[17] D. Martinec and T. Pajdla. Robust rotation and translation estimation in multiview reconstruction. *In Proc. CVPR*, pages 1– 8, 2007.

[18] P. Moulon, P. Monasse, and R. Marlet. Global fusion of relative motions for robust, accurate and scalable structure from motion. *In Proc. ICCV*, 2013.

[19] D. Nister. An efficient solution to the five-point relative pose problem. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26:756–777, 2004.

[20] M. Reich, C. Heipke. Global rotation estimation using weighted iterative Lie algebraic averaging. *ISPRS Ann. Photogram., Rem. Sens. Spatial Inf. Sci. 1*, pages 443–449, 2015.

[21] M. Reich, M.Y. Yang, C. Heipke. Global robust image rotation from combined weighted averaging. J*ournal of Photogrammetry and Remote Sensing*, pages 107-114, 2017.

[22] R. Roberts, S. N. Sinha, R. Szeliski, and D. Steedly. Structure from motion for scenes with large duplicate structures. *In Proc. CVPR*, 2011.

[23] J. L. Schonberger. COLMAP: SfM and MVS. https://demuc.de/colmap/, 2016.

[24] J. L. Schonberger and J.-M. Frahm. Structure-from-motion revisited. *In Proc. CVPR*, pages 4104–4113, 2016.

[25] T. Shen, S. Zhu, T. Fang, R. Zhang, and L. Quan. Graph based consistent matching for structure-from-motion. *In Proc. ECCV*, pages 139–155. Springer, 2016.

[26] N. Snavely. Bundler: Structure from Motion for Unordered Image Collections. http://www.cs.cornell.edu/~snavely/bundler/, 2008.

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

[28] F. Wang, A. Nayak, Y. Agrawal, and R. Shilkrot. Hierarchical Image Link Selection Scheme for Duplicate Structure Disambiguation. *In Proc. BMVC*, 2018.

[29] X. Wang, F. Rottensteiner, C. Heipke. Robust image orientation based on relative rotations and tie points. *PE&RS (in print)*, 2019.

[30] X. Wang, F. Rottensteiner, C. Heipke. Structure from motion for ordered and unordered image sets based on random k-d forests and global pose estimation. *ISPRS Journal of Photogrammetry and Remote Sensing*, pages 10-41, 2019.

[31] K. Wilson and N. Snavely. Robust global translations with 1dsfm. *In Proc*. ECCV, 2014.

[32] K. Wilson, D. Bindel, N. Snavely. When is Rotation Averaging Hard? *In Proc. ECCV*, pages 255–270. Springer, 2016.

[33] C. Wu. VisualSFM: A visual structure from motion system. http://homes.cs.washington.edu/ ˜ccwu/vsfm, 2011.

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

[35] C.Zach, A. Irschara, and H. Bischof. What can missing correspondences tell us about 3d structure and motion? In *Proc. CVPR*, 2008.

[36] C. Zach, M. Klopschitz, and M. Pollefeys. Disambiguating visual relations using loop constraints. In *Proc. CVPR*, pages 1426-1433, 2010.