MMA: Multi-camera Based Global Motion Averaging

Hainan Cui [1,3*], **Shuhan Shen** [1,2,3]

[1] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[3] CASIA-SenseTime Research Group, China
hncui@nlpr.ia.ac.cn, shshen@nlpr.ia.ac.cn

Abstract

In order to fully perceive the surrounding environment, many intelligent robots and self-driving cars are equipped with a multi-camera system. Based on this system, the structure-from-motion (SfM) technology is used to realize scene reconstruction, but the fixed relative poses between cameras in the multi-camera system are usually not considered. This paper presents a tailor-made multi-camera based motion averaging system, where the fixed relative poses are utilized to improve the accuracy and robustness of SfM. Our approach starts by dividing the images into reference images and non-reference images, and edges in view-graph are divided into four categories accordingly. Then, a multi-camera based rotating averaging problem is formulated and solved in two stages, where an iterative re-weighted least squares scheme is used to deal with outliers. Finally, a multi-camera based translation averaging problem is formulated and a l_1-norm based optimization scheme is proposed to compute the relative translations of multi-camera system and reference camera positions simultaneously. Experiments demonstrate that our algorithm achieves superior accuracy and robustness on various data sets compared to the state-of-the-art methods.

Introduction

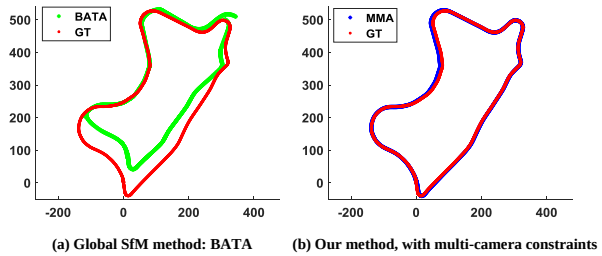
Fully perceive and reconstruct the surrounding environment is a crucial ability for intelligent systems, such as various kinds of robots (Strisciuglio et al. 2018; Chen et al. 2021) and self-driving cars (Heng et al. 2019; Wang et al. 2020). To achieve this goal, multi-camera system is typically used as the sensor platform since it is cheap to maintain, easy to handle and provides 360° high-resolution image data. These data make the reconstruction become more complete and facilitate better camera re-localization (Colledanchise, Malafronte, and Natale 2020) and navigation (Chen et al. 2019). The SfM technology is typically used to realize the reconstruction task and based on the manner of camera poses estimation, its pipeline is divided into two classes: incremental and global.

Incremental SfM selects camera seeds first and then adds cameras one by one as the size of scene model grows up. The quality of incremental scene reconstruction is depending on the selection of camera seeds and the order of camera addition. With the operation of incremental system, the error accumulates gradually, which often leads to significant scene drift (Cornelis, Verbiest, and Gool 2004; Holynski et al. 2020). To mitigate these error, the non-linear bundle adjustments (BA) is performed repeatedly. While accurate, the computational cost increases dramatically. Global SfM has become more popular in recent years, which aims to estimate all camera poses at the same time through motion averaging. Formally, a global SfM method takes a view-graph as input $G = \{V, E\}$, where each node V_i in V denotes an image and each edge E_{ij} in E denotes the connection between image i and j. The relative pose $(\mathbf{R}_{ij}, \mathbf{t}_{ij})$ is estimated for each edge, where $\mathbf{R}_{ij} \in \mathbb{R}^{3\times3}$ denotes a relative rotation matrix, $\mathbf{t}_{ij} \in \mathbb{R}^{3\times1}$ represents a unit vector of the relative translation direction. Ignoring the measurement noise, the following two equations are hold:

$$\mathbf{R}_{ij} = \mathbf{R}_j\mathbf{R}_i^T, \lambda_{ij}\mathbf{t}_{ij} = \mathbf{R}_j(\mathbf{t}_i - \mathbf{t}_j). \quad (1)$$

where λ_{ij} is a scale factor. The motion averaging is to find the camera rotation $\mathbf{R}_i$ via global rotation averaging and camera position $\mathbf{t}_i$ via global translation averaging, such that the observed pairwise relative poses are best explained. The global SfM does not need to carefully select camera seeds and the reconstruction error is uniformly spread to the whole model, avoiding the error accumulation. Moreover, the global BA is only run once, which is more efficient than the incremental system.

Although many global SfM systems (Wilson and Snavely 2014; Dong et al. 2020) have success in reconstructing internet images, they are not suitable to reconstruct the images collected by the multi-camera platform. The reason has three folds. Firstly, the internet images usually cover a single scenic spot, hence the nodes in the view-graph are densely connected. However, the multi-camera platform collects data from the city-scale scene, such as the self-driving car collects images around city blocks. As a consequence, the nodes connection in the view-graph is sparse, which is more challenging for the global motion averaging system. Secondly, the distribution of internet image positions is random, while the multi-camera platform usually moves along straight routes, such as a self-driving car moves along the street, hence many relative translations are collinear, which may ruin the global SfM system (Jiang, Cui, and Tan 2013). Finally, since the cameras in the multi-camera system are



(a) Global SfM method: BATA (b) Our method, with multi-camera constraints

Figure 1: Camera position estimations of a large-scale dataset. 'GT' denotes the ground-truth. Conventional global SfM method BATA (Zhuang, Cheong, and Lee 2018) has large scale drift and can not achieve closed-loop. In this scenario with large loops, our method is more robust.

rigidly mounted, the internal relative poses between cameras in the multi-camera system are considered as fixed. Conventional SfM systems ignore this inherent fixed relative poses, resulting in a large difference between the estimated camera position and the actual camera position.

In theory, if the internal relative poses between cameras in the multi-camera system are known, we only need to estimate the pose of one camera, and the other camera poses can be derived accordingly. These internal relative poses are generally estimated by specific calibration patterns (Zhang and Pless 2004), which needs sufficient large overlapping area of cameras. Although this method achieves high accuracy, it is computationally expensive and incapable of calibrating the multi-camera system with little or even no overlapping fields of view, which is often the case for applications in autonomous vehicles. Some methods (Heng, Furgale, and Pollefeys 2015; Lin et al. 2020) use scene reconstruction to calibrate the internal relative poses, which requires prior knowledge of the scene, hence they are also unsuitable for the reconstruction task and the calibration accuracy of such manner is dependent on the pre-reconstructed scene. In this paper, we propose an adaptive multi-camera based SfM system, where the internal relative poses of the multi-camera system are computed online and a new multi-camera based motion averaging problem is formulated.

First, images are divided into reference images and non-reference images. For a N-camera system, N images are simultaneously obtained at each moment. We select one camera as the reference camera and the images collected by this camera are denoted as reference images. In this way, only the reference camera poses and the internal relative poses of multi-camera system need to be computed. As a consequence, the number of parameters in our system is about $1/N$ of the number of parameters in the traditional SfM system. The constraints produced by the non-reference images are transformed into the constraints on reference images, which makes the connections of reference images become more dense. According to the different classification attributes of images, the edges in view-graph will generate four different geometric constraints. Based on the constraints on camera rotations, a multi-camera based rotation averaging problem is formulated and solved in two stages: the internal relative rotations of multi-camera system is first estimated and then an iterative re-weighted least squares scheme is further proposed to refine camera rotations. Given global camera rotations, a multi-camera based translation averaging problem is formulated and solved in a global manner, where the internal relative translations of multi-camera system and reference camera positions are simultaneously estimated in a l_1-norm based cost function. As shown in (Wilson, Bindel, and Snavely 2016), the performance of motion averaging is depending on both the accuracy of relative geometries and the number of constraints. However, the two-view relative translation estimation is sensitive to feature match outliers, which means direct computation of the whole view-graph may bring a lot of outliers to the motion averaging system. Thus, we propose an edge selection method to mitigate the impact of relative geometry outliers. Experiments demonstrate that the robustness and accuracy of SfM system are improved by this selection strategy. Fig. 1 shows the camera positions estimation of a large-scale dataset data09, which comes from the odometry benchmark of KITTI (Geiger, Lenz, and Urtasun 2012).

In summary, the main contributions of our work are:

- a multi-camera based motion averaging (MMA) system is proposed, where the internal relative poses between cameras are calibrated automatically. To our best knowledge, we are the first to handle the reconstruction task of multi-camera system in a global manner;
- images are divided into reference images and non-reference images, and in this way, the number of estimated parameters in our system is reduced to $1/N$ of that in the traditional system, where N is the number of cameras in the multi-camera system;
- a multi-camera based rotation averaging (MRA) problem is formulated and solved to simultaneously estimate the internal relative rotations of multi-camera system and camera rotations of reference images;
- a multi-camera based translation averaging (MTA) problem is formulated and solved to simultaneously estimate the internal relative translations of multi-camera system and reference image positions, and an edge selection method is proposed to tackle outliers.

Related Work

Rotation Averaging. RA aims to estimate the absolute rotations from the observations of relative rotations. Govindu and Venu propose to solve this problem by lie-group based averaging (Govindu 2004). In (Hartley, Aftab, and Trumpf 2011), the classical Weiszfeld algorithm is used to update the absolute orientations of each camera iteratively. In (Crandall et al. 2012), a rough rotation initialization is estimated by discrete Markov Random Field (MRF) with loopy belief propagation, and then refined by continuous Levenberg-Marquardt optimization. In (Fredriksson and Olsson 2012),

the RA problem is converted to a dual problem with Lagrangian duality and solved by the Semi-Definite Programming (SDP), which contributes to obtaining the globally minimum solutions. The impact of different cost functions to the performance of RA is summarized in (Hartley et al. 2013). In (Wilson, Bindel, and Snavely 2016), the authors point out that densely connected view-graph and highly accurate geometry can produce a more stable and accurate RA result. Chatterjee and Govindu use an iterative re-weighted least squares (IRLS) formulation to fine-tune the initialization of RA (Chatterjee and Govindu 2017). In (Cui et al. 2017), the view-graph is clustered to improve the connection tightness, but the accuracy of RA relies on the relative geometries between different clusters. Similarly in (Zhu et al. 2018), images are divided into multiple partitions first and a global motion averaging is solved to determines cameras at partition boundaries. In (Cui et al. 2019), the orthogonal maximum spanning tree (Cui et al. 2018) is used to select relative geometry inliers. Shonan-based RA (Dellaert et al. 2020) is proposed to recover globally optimal solutions under mild assumptions on the measurement noise. In (Purkait, Chin, and Reid 2020; Yang et al. 2021), the end-to-end neural network based RA is proposed. In (Chen, Zhao, and Kneip 2021), a hybrid RA framework is represented by leveraging the advantages of global RA and local RA. Besidesin (Dai et al. 2009), the stereo-camera configuration is considered in the RA but it cannot be extended to solve the multi-camera based RA problem. In this work, we are the first to solve the RA problem of multi-camera system by fusing internal relative rotation constraints.

Translation Averaging. TA aims to estimate absolute camera positions from the observations of relative translations. Many linear methods (Rother 2003; Jiang, Cui, and Tan 2013) are proposed to solve TA problem by matrix decomposition. Although efficient, such approaches are sensitive to relative geometry outliers. In openMVG (Moulon, Monasse, and Marlet 2013), a relaxed version of TA problem is proposed and a L_∞ norm based function is used to optimize the estimations. However, the L_∞ norm is sensitive to outliers, which cannot be used in large-scale reconstruction problems. Since the relative geometries are estimated by feature matches, the edge outliers are inevitable. Many algorithms focus on the outliers filtering or robust view-graph construction. Some methods (Zach, Klopschitz, and Pollefeys 2010; Guibas, Huang, and Liang 2019) use loop consistency to remove outliers in the view graph. Some works attempt to directly construct an accurate view-graph (Cui et al. 2021; Barath et al. 2021) or refine the view-graph by loop consistency (Cui and Tan 2015; Sweeney et al. 2015). A least unsquared deviations (LUD) form is proposed in (Ozyesil and Singer 2015) to estimate not only camera positions but also the scale of measurements. Based on the alternating direction method of multipliers (ADMM) (Boyd et al. 2011), a similar cost-function is proposed in (Goldstein et al. 2016), which is called as Shapefit/kick. To desensitize the impact of baseline, a bilinear object function is introduced in (Zhuang, Cheong, and Lee 2018), introducing a variable to perform the requisite normalization. In (Dong et al. 2020), a rank constraint is strengthened to refine the camera positions.

The internal relative translations are not considered in the formulations of those above methods. In this paper, we propose a new formulation for multi-camera based TA problem and compute the internal relative translations and reference camera positions simultaneously.

Multi-camera based Motion Averaging

In this section, we show the pipeline of our global motion averaging. The view-graph is first constructed and then we define the reference camera and non-reference camera. Based on the belonged camera of images, the relationship between the reference camera poses and the internal relative poses of multi-camera system is introduced. Next, we investigate the multi-camera based rotation averaging problem and show a two-stage solution. Finally, the multi-camera based translation averaging problem is formulated and the corresponding l_1 norm based cost function is solved in a global manner.

Construction of View-graph

Given images collected by the multi-camera system, the scale-invariant image features are first detected. Since the multi-camera system is usually equipped on the continuously moving platform, the images are collected sequentially. In our work, the image matching is divided into two stages. In the first stage, the sequential match is performed on the adjacent images. In the second stage, the loop detection is performed, where each image is matched with the images searched by image retrieval (Schönberger et al. 2016). Given feature matches, the 5-point algorithm (Nistér 2004) is used to estimate the two-view relative geometry $(\mathbf{R}_{ij}, \mathbf{t}_{ij})$. The node in view-graph denotes images. When two images have a sufficient number of feature matches, they are connected in view-graph and the corresponding edge records the two-view relative geometry measurement.

Reference Images Definition

Let N be the number of cameras in the multi-camera system. We select one camera as the reference camera and the images collected by this camera are defined as reference images. The other images are defined as non-reference images. During the image data collection, N images are simultaneously obtained at each moment and we denote them as a "rigid set". In each rigid set, there is only one reference image and the number of internal relative poses between cameras in the multi-camera system is $N-1$. All rigid sets share the same internal relative poses.

If all images are fully connected, no matter which camera is selected as the reference camera, the number of the unknown parameters, including the camera pose of reference images and the internal relative poses of multi-camera system, is fixed. However, in the 360° scene, some cameras may usually collect textureless scene, such as indoor walls or outdoor sky. If they are selected as the reference camera, most of images may be left uncalibrated. In our work, the maximum spanning tree (MST) of view-graph is first extracted, where edge weight is set to the number of feature correspondences, and then the images in the MST are divided into N classes based on their belonged camera. The class that has

the most number of images is selected and the images inside are set as the reference images. If image i is a reference image, then its camera pose is the reference camera pose; otherwise, we find the reference image in the rigid set containing image i and set the corresponding camera pose as the reference camera pose of image i.

Let $\{\mathbf{R}_i^{ref}, \mathbf{t}_i^{ref}\}$ be the reference camera pose of image i. $\{\mathbf{R}_i^{rel}, \mathbf{t}_i^{rel}\}$ is the relative pose between images i and its corresponding reference image. For image i, the following two equations are hold:

$$\mathbf{R}_i^{rel} = \mathbf{R}_i \mathbf{R}_i^{ref^T}; \tag{2}$$

$$\mathbf{t}_i^{rel} = \mathbf{R}_i^{ref}(\mathbf{t}_i - \mathbf{t}_i^{ref}). \tag{3}$$

Then, Eq. 1 is transformed into:

$$\mathbf{R}_{ij} = \mathbf{R}_j^{rel} \mathbf{R}_j^{ref} \mathbf{R}_i^{ref^T} \mathbf{R}_i^{rel^T}; \tag{4}$$

$$\lambda_{ij}\mathbf{t}_{ij} = \mathbf{R}_j^{rel}\mathbf{R}_j^{ref}(\mathbf{R}_i^{ref^T}\mathbf{t}_i^{rel} + \mathbf{t}_i^{ref} - \mathbf{R}_j^{ref^T}\mathbf{t}_j^{rel} - \mathbf{t}_j^{ref}). \tag{5}$$

For Eq. 4 and Eq. 5, the left side is the relative geometry measurement, and the right side contains the reference camera poses and the internal relative poses between cameras.

Multi-camera based Rotation Averaging

Since images are distinguished as reference and non-reference images, the edges in view-graph are divided into four classes: $r-r$, $r-n$, $n-r$, $n-n$, where r denotes reference image and n denotes non-reference image. Based on Eq. 4, the formulations of four different classes of edges are derived as follows. For the edge $r-r$, the referenced images of both connected images are themselves, hence the relationship between global camera rotation and relative rotation is transformed into:

$$\mathbf{R}_{ij} = \mathbf{R}_j^{ref} \mathbf{R}_i^{ref^T}. \tag{6}$$

For the edge $r-n$, the transformed relationship is:

$$\mathbf{R}_{ij} = \mathbf{R}_j^{rel} \mathbf{R}_j^{ref} \mathbf{R}_i^{ref^T}. \tag{7}$$

For the edge $n-r$, the transformed relationship is:

$$\mathbf{R}_{ij} = \mathbf{R}_j^{ref} \mathbf{R}_i^{ref^T} \mathbf{R}_i^{rel^T}. \tag{8}$$

For the edge $n-n$, the transformed relationship is:

$$\mathbf{R}_{ij} = \mathbf{R}_j^{rel} \mathbf{R}_j^{ref} \mathbf{R}_i^{ref^T} \mathbf{R}_i^{rel^T}. \tag{9}$$

Let $D1, D2, D3, D4$ be the difference between left side and right side of Eq. 6, Eq. 7, Eq. 8 and Eq. 9, respectively. The multi-camera based rotation averaging is defined as:

$$\min_{\mathbf{R}^{ref}, \mathbf{R}^{rel}} \quad \sum_{ij \in E} \|D1\|_p + \|D2\|_p + \|D3\|_p + \|D4\|_p, \tag{10}$$

where the variable $p = 1, 2$ chooses l_1 or l_2 norm.

The key to solve MRA problem is to provide a good initialization for Eq. 10. In view-graph, the edge weight is set to the number of feature correspondences. The MST of view-graph is extracted and the camera rotation initializations can be computed by the chain rule. Based on these initializations, the reference camera rotation is obtained and each rigid set will produce one internal relative rotation of multi-camera system. Since these estimated internal relative rotations may be inconsistent, we use the RANSAC method (Hartley and Zisserman 2003) to find the best one. In this way, the initialization of Eq. 10 is obtained. Considering the robustness, the initialized camera rotations are refined first by minimized Eq. 10 with l_1 norm and then further optimized by minimizing Eq. 10 with l_2 norm. To handle the gross relative rotation measurements, the l_2 norm based optimizing process is performed in an iterative re-weighted least squares (IRLS) way, where a weighted least squares problem is solved in each iteration. The re-weighting function is set to the $\Phi(\varepsilon) = \alpha^2/(\varepsilon^2 + \alpha^2)$, where ε denotes the residual for each observation and α is the loss width.

If the corresponding reference image of image j is as the same as that of image i, Eq. 7 will be transformed as:

$$\mathbf{R}_{ij} = \mathbf{R}_j^{rel}, \tag{11}$$

Eq. 8 will be transformed as:

$$\mathbf{R}_{ij} = \mathbf{R}_i^{rel^T}, \tag{12}$$

and Eq. 9 will be transformed as:

$$\mathbf{R}_{ij} = \mathbf{R}_j^{rel} \mathbf{R}_i^{rel^T}. \tag{13}$$

Based on these transformations, a relative rotation averaging is proposed to refine the initial internal relative rotation $\mathbf{R}^{rel}$ of multi-camera system by minimizing:

$$\sum_{ij \in E} \|\mathbf{R}_{ij} - \mathbf{R}_j^{rel}\|_1 + \|\mathbf{R}_{ij} - \mathbf{R}_i^{rel^T}\|_1 + \|\mathbf{R}_{ij} - \mathbf{R}_j^{rel}\mathbf{R}_i^{rel^T}\|_1. \tag{14}$$

This occurs when cameras in the multi-camera system have overlapping areas.

Overall, the multi-camera based rotation averaging problem defined in Eq. 10 is solved in two stages. The first stage is to estimate the initialization of camera rotations. When the cameras in the multi-camera system are not overlapped, the initialization is estimated by the MST of view-graph first, and then refined by a l_1 norm based optimization. When the cameras in the multi-camera system are overlapped, the initialization is estimated by the MST of view-graph, and the internal relative poses between cameras are refined by minimizing Eq. 14 with l_1 norm. The second stage is to refine the initialization by minimizing Eq. 10 in an IRLS way.

Multi-camera based Translation Averaging

Similar to MRA problem, edges in view-graph are also divided into four classes in MTA problem. Let $\mathbf{v}_{ij} = \mathbf{R}_j^T \mathbf{t}_{ij}$. For the edge $r-r$, the relative translation relationship in Eq. 1 is transformed into:

$$\lambda_{ij}\mathbf{v}_{ij} = \mathbf{t}_i^{ref} - \mathbf{t}_j^{ref}. \tag{15}$$

For the edge $r-n$, the transformed relationship is:

$$\lambda_{ij}\mathbf{v}_{ij} = \mathbf{t}_i^{ref} - \mathbf{R}_j^{ref^T}\mathbf{t}_j^{rel} - \mathbf{t}_j^{ref}. \tag{16}$$

For the edge $n-r$, the transformed relationship is:

$$\lambda_{ij}\mathbf{v}_{ij} = \mathbf{R}_i^{ref^T}\mathbf{t}_i^{rel} + \mathbf{t}_i^{ref} - \mathbf{t}_j^{ref}. \quad (17)$$

For the edge $n-n$, the transformed relationship is:

$$\lambda_{ij}\mathbf{v}_{ij} = \mathbf{R}_i^{ref^T}\mathbf{t}_i^{rel} + \mathbf{t}_i^{ref} - \mathbf{R}_j^{ref^T}\mathbf{t}_j^{rel} - \mathbf{t}_j^{ref}. \quad (18)$$

When cameras in the multi-camera system are overlapped, the referenced camera of image j may be as the same as that of image i. In this case, Eq. 16 is transformed into:

$$\lambda_{ij}\mathbf{R}_j^{ref}\mathbf{v}_{ij} = \mathbf{t}_j^{rel}, \quad (19)$$

Eq. 17 is transformed into:

$$\lambda_{ij}\mathbf{R}_i^{ref}\mathbf{v}_{ij} = \mathbf{t}_i^{rel}, \quad (20)$$

and Eq. 18 is transformed into:

$$\lambda_{ij}\mathbf{R}_i^{ref}\mathbf{v}_{ij} = \mathbf{t}_i^{rel} - \mathbf{t}_j^{rel}. \quad (21)$$

However, since the observed relative translation $\mathbf{t}_{ij}$ is up to scale and sensitive to feature match outliers, the local averaging on internal relative translations is not performed.

Let $F1, F2, F3, F4$ be the right side of Eq. 15, Eq. 16, Eq. 17 and Eq. 18, respectively. The multi-camera based translation averaging is defined as:

$$\begin{aligned} \min_{\mathbf{t}^{ref},\mathbf{t}^{rel},\lambda_{ij}} \quad & \sum_{ij\in E} \|F1 + F2 + F3 + F4 - \lambda_{ij}\mathbf{v}_{ij}\|_1 \\ \text{s.t.} \quad & \lambda_{ij} > b, \sum_{i\in V} \mathbf{t}_i^{ref} = 0, \end{aligned} \quad (22)$$

where λ_{ij} is a non-negative variable (in our paper, b is set to 1.0). The first constraint on λ_{ij} is to remove the scale ambiguity and the second constraint on $\mathbf{t}_i^{ref}$ is to remove the inherent positional ambiguity. We solve this problem by the ADMM solver (Boyd et al. 2011).

Since the two-view relative translation estimation is more sensitive to feature match outliers, hence many gross relative translation constraints may exist in the averaging system. For the robustness concern, we propose a simple yet effective edge selection strategy to improve the performance of translation averaging. Consider the completeness of scene reconstruction, the motion averaging should be performed on all the connected images in view-graph. To achieve this goal, the MST of the view-graph is extracted, where the weight of edge is set to the number of feature correspondences. Then for each image, its connected edges are ranked by the number of feature correspondence inliers and we select the top-K edges into motion averaging system.

Experiments

Our experiments are performed on real photo collections provided by different multi-camera platforms, including conventional stereo camera, Insta360 OneX with two fisheye cameras and Insta360 Pro2 with six fisheye cameras. The detailed dataset info is shown in Table 1. The stereo camera dataset comes from odometry benchmark of KITTI (Geiger, Lenz, and Urtasun 2012). Although many data do not contain loops, we test them to verify the limitations of motion averaging technology. We also run experiments on the image projections of panorama to further demonstrate the powerful reconstruction ability of our motion averaging system. The panorama is a composite of two fisheye images from Insta360 OneX and then projected in eight directions to simulate eight virtual cameras in the scene.

Name	N_i	N_c	Sensor	Ground Truth	Loop	Scene
data00	9082	2	Stereo	yes	yes	outdoor
data01	2202	2	Stereo	yes	no	outdoor
data02	9322	2	Stereo	yes	no	outdoor
data03	1602	2	Stereo	yes	no	outdoor
data04	542	2	Stereo	yes	no	outdoor
data05	5522	2	Stereo	yes	yes	outdoor
data06	2202	2	Stereo	yes	yes	outdoor
data07	2202	2	Stereo	yes	yes	outdoor
data08	8142	2	Stereo	yes	no	outdoor
data09	3182	2	Stereo	yes	yes	outdoor
data10	2402	2	Stereo	yes	no	outdoor
data11	2552	2	Insta360 OneX	no	yes	indoor
data12	780	2	Insta360 OneX	no	yes	outdoor
data13	2358	6	Insta360 Pro2	no	yes	indoor
data14	4866	6	Insta360 Pro2	no	yes	outdoor
data15	2312	8	Panorama	no	yes	indoor
data16	2992	8	Panorama	no	yes	outdoor

Table 1: Details of testing dataset. N_i denotes the number of images and N_c denotes the number of cameras.

All datasets are collected sequentially and run on a same computer with 256GB memory. Our method is compared to the state-of-the-art rotation averaging system RRA (Chatterjee and Govindu 2017) and IRA (Gao et al. 2021), and also compared to the state-of-the-art translation averaging system LUD (Ozyesil and Singer 2015) and BATA (Zhuang, Cheong, and Lee 2018). Global rotation averaging results of LUD and BATA are produced by RRA. Both RRA and LUD are implemented in Theia (Sweeney 2015). The code of BATA is provided in (Zhuang 2018). In our work, the parameter K in the translation averaging is set to 8. For the reconstruction implementation, the root-SIFT (Arandjelović and Zisserman 2012) is used to detect scale-invariant features and the feature matching on run on GPU (Schönberger and Frahm 2016). Based on the estimated camera poses, tracks triangulation (Hartley and Sturm 1997) and bundle adjustment (Agarwal, Mierle, and Others 2021) are iteratively performed to get the final reconstruction result.

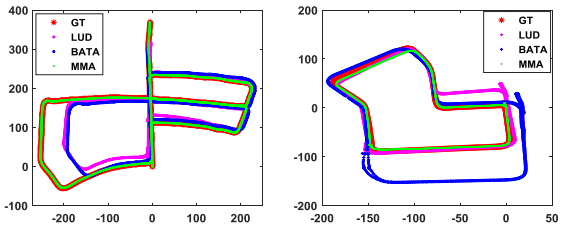
Evaluation of Benchmark Datasets

To evaluate calibration accuracy, our method is run on the large-scale KITTI odometry benchmark dataset, which is captured by a driving car with a stereo camera. Since cars usually travel in a straight line, most of relative translation estimations are collinear, which is challenging for global SfM system. Table 2 shows the results of camera calibration accuracy and system running time. Since the image-based reconstruction is up to scale, the accuracy of internal relative translation is measured by the angle between translation directions.

In most cases, our MRA method is superior than the RRA (Chatterjee and Govindu 2017) and IRA (Gao et al. 2021). Especially for data02 and data06, the results produced by RRA are gross while we produce more accurate camera rotations, which is the key to produce a better trans-

Data	RRA			IRA			MRA				LUD			BATA			MTA			
Name	$\tilde{e_r}$	$\bar{e_r}$	T_r	$\tilde{e_r}$	$\bar{e_r}$	T_r	$\tilde{e_r}$	$\bar{e_r}$	T_r	e_{rr}	$\tilde{e_t}$	$\bar{e_t}$	T_t	$\tilde{e_t}$	$\bar{e_t}$	T_t	$\tilde{e_t}$	$\bar{e_t}$	T_t	e_{rt}
data00	1.3	1.7	20	**0.7**	**0.8**	59	**0.7**	**0.8**	20	0.08	25.7	38.3	4.5	21.7	36.8	7.0	**1.5**	**1.9**	1.4	0.09
data01	1.8	2.8	1	1.2	1.6	15	**0.7**	**1.1**	1	0.07	22.5	41.6	1.3	19.5	29.3	0.3	**16.3**	**23.2**	0.5	0.35
data02	78.3	88.8	30	3.0	5.0	59	**1.8**	**2.3**	32	0.07	105.6	247.0	3.9	104.2	259.5	6.0	**7.0**	**10.1**	5.1	0.75
data03	**0.7**	**0.7**	2	28.8	36.7	2	**0.7**	**0.7**	2	0.08	3.2	7.0	0.4	4.0	10.0	0.3	**0.4**	**1.9**	0.3	0.28
data04	0.4	0.4	1	**0.3**	**0.3**	4	**0.3**	**0.3**	1	0.05	34.3	66.9	0.6	67.4	87.8	0.1	**2.4**	**13.3**	0.6	1.25
data05	1.8	1.8	24	**0.6**	1.3	45	0.7	**0.8**	25	0.04	10.9	26.4	2.8	9.0	19.7	1.7	**0.6**	**1.2**	1.2	0.17
data06	57.8	77.8	3	**0.5**	**0.5**	14	**0.5**	**0.5**	3	0.05	25.9	71.4	1.2	27.5	65.1	0.3	**0.3**	**1.0**	0.2	0.25
data07	2.9	2.9	3	**0.6**	**0.5**	17	**0.6**	0.6	3	0.08	8.3	13.3	1.2	4.9	8.9	0.3	**0.9**	**2.2**	1.0	0.22
data08	0.8	0.9	15	**0.7**	**0.7**	43	0.8	0.8	14	0.03	20.0	32.2	2.7	17.5	24.8	3.5	**2.6**	**3.1**	4.7	0.52
data09	1.4	1.4	6	1.0	1.1	21	**0.6**	**0.6**	6	0.06	28.4	49.9	1.4	20.7	38.5	0.4	**1.7**	**3.5**	1.4	0.38
data10	1.2	1.3	5	**0.6**	**0.7**	16	0.7	0.8	5	0.04	16.3	40.3	1.1	8.3	26.0	0.3	**0.6**	**1.2**	1.1	0.33

Table 2: Calibration accuracy of KITTI datasets. $\tilde{e_r}$ and $\bar{e_r}$ denote median and mean rotation errors in degrees, respectively. $\tilde{e_t}$ and $\bar{e_t}$ denote median and mean position errors in meters, respectively. e_{rr} and e_{rt} denote relative rotation error and relative translation error in degrees, respectively. T_r denotes running time of RA in seconds. T_t denotes running time of TA in minutes.



(a) data05 (b) data07

Figure 2: Comparison of calibrated camera positions. 'GT' denotes the ground-truth camera positions. Since the driving car runs on a flat road, only the top-view of calibrated camera trajectory is shown in the figure.

lation averaging result. Since IRA is an incremental method, it is not as efficient as our global MRA method and it cannot generate an usable result for data03. For the translation averaging, neither LUD nor BATA can fulfil this task. The main reason is that the view-graph contains too many collinear two-view translation measurements, which degenerate the scales estimation. Figure 2 shows the camera positions produced by LUD, BATA and MMA, respectively. Although the loop is achieved by LUD and BATA, the scale error is extremely large, while our result is nearly coincide with the ground-truth. In our work, the camera pose constraints generated by edges containing non-reference images are converted to constraints on reference images. As a consequence, many non-collinear constraints are added to the reference images and the number of constraints on reference camera poses is higher than that of conventional methods. From Table 2, our motion averaging results are more accurate than those produced by LUD and BATA. For the running time, the performances of these three systems are similar.

Since loop closure is vital to mitigate camera calibration errors, the result produced on the data with loops is more accurate than those data without loops. Based on this comparison, it is necessary for mobile robots to surround a circle in the scene in order to better perceive the surrounding scene. This is an inherent limitation of global motion averaging system. Although our system alleviates the limitation, if the image acquisition platform moves linearly for a long time without closing the loop, the system cannot produce accurate results, such as the results of data01 and data02.

Evaluation of Self-collected Datasets

To demonstrate the scalability of system, six self-collected datasets are also tested. The dataset info is shown in Table 1 and the corresponding results are shown in Fig. 3. Since these datasets are collected sequentially, the camera moving trajectory should be continuous.

Among these datasets, only data11 is successfully reconstructed by all three methods. The main reasons include the following two aspects. First, most of the camera trajectory is not along a line, hence the view-graph contains little collinear relative translations, which is more suitable for conventional global motion averaging systems. In addition, since the scene is collected multiple times, the nodes in the view-graph are densely connected. These two good conditions make the scene be reconstructed easily. However, this collecting manner is infeasible for self-driving cars.

For data12 and data13, the scene is a underground garage and the camera moving trajectories are similar. The data12 is collected by Insta OneX with two fisheye cameras and data13 is collected by Insta Pro2 with six fisheye cameras. In the garage, there are many identical cars, signs and diversion lines, which produce many gross relative geometries. Although the corresponding ground truth camera positions are not available, the similar and smooth camera trajectories produced by our motion averaging show that our system performs better than the LUD and BATA. The camera positions calibrated by LUD and BATA are discontinuous, which violates the sequential collecting manner.

For data14 and data16, both BATA and our system reconstruct them successfully. Although the shape of the camera trajectory produced by LUD is similar to that produced by our system, many erroneous camera positions exist in its result. Since the platform just walks around the scene once and the distance between cameras in the multi-camera system is close, the camera trajectory should seems like a single loop. However for LUD, many loops and gross camera positions appear in the result. The data15 is collected in the street and the camera moving path contains many straight lines. Neither BATA nor LUD is capable of reconstructing

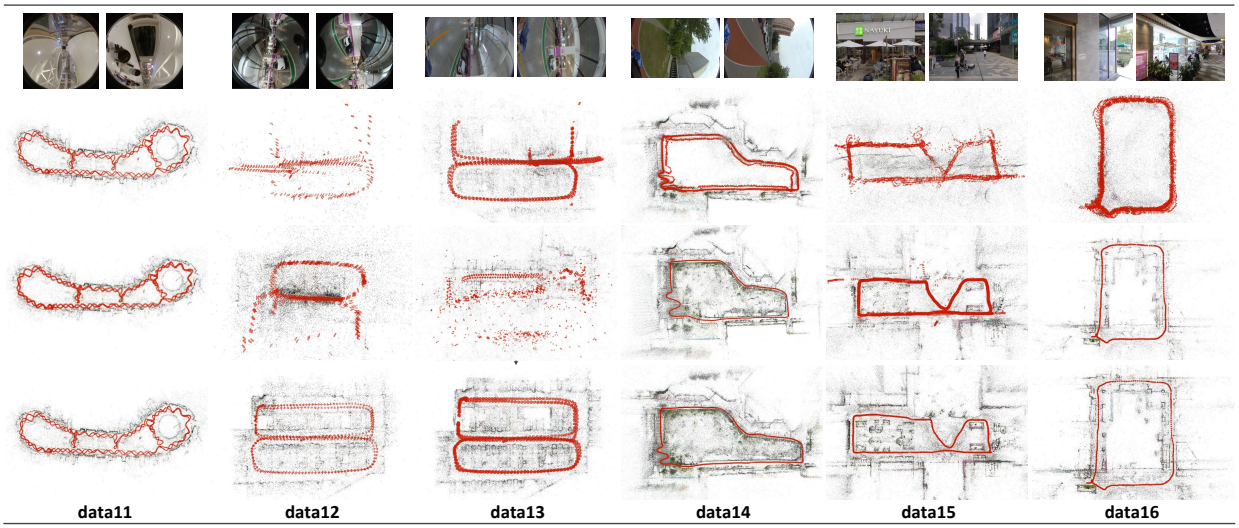
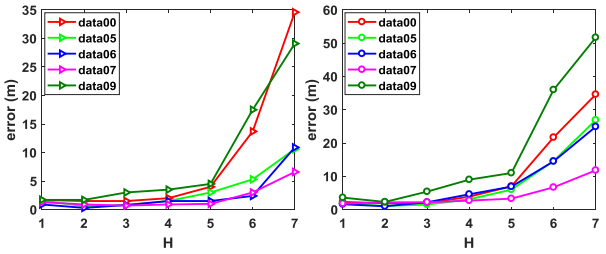


Figure 3: From top to bottom, the reconstruction results are produced by LUD, BATA and our method MMA, respectively. The calibrated camera positions are shown in red.



(a) median error (b) mean error

Figure 4: Camera position accuracy under different parameter settings. To simplify parameter search process, we set $K = 4H, H = 1, 2, ..., 7$. x-axis denotes the value of H and y-axis shows the corresponding camera position error in meters.

this scene. In comparison, our calibrated camera moving trajectory is continuous and the corresponding reconstruction result is more reliable.

Overall, our system is not sensitive to the situation of collinear relative translations and more accurate than the global motion averaging system LUD and BATA. In addition, the number and the distribution of cameras in the multi-camera system does not affect the performance of our system. Since the data collection manner for cars or robots is usually moving straight along the road, our system is more applicable for the related applications of self-driving.

Ablation Study of Parameter K

Since feature match outliers produce many gross relative geometry estimations, our edge selection strategy is proposed to balance the accuracy of observed relative translations and the number of constraints on cameras. The only parameter in our work is the number of selected edges for each node of view-graph, which is defined as K. To guarantee the robustness of scene reconstruction, five datasets containing loops are used to find the optimal parameter.

Fig. 4 shows the performance of our system under different parameter settings. Note that when $H = 7$, all edges of view-graph are selected into our multi-camera based translation averaging system, which means the corresponding result is produced without our edge selection strategy. From this figure, using all edges gets the worst result, indicating the necessity of edge selection. From $K = 4$ to $K = 8$, both median position error and mean position error decrease. However, as more edges are selected, the error of camera position estimations increases progressively. Hence, based on this discovery, the parameter K is set to 8, which means that in our multi-camera based translation averaging, each image is constrained only by the best 8 edges connected to it.

Conclusion

In this paper, we propose a multi-camera based motion averaging system, where the multi-camera based rotation averaging and multi-camera based translation averaging are presented. By embedding the internal relative poses constraint into reconstruction system, both our rotation averaging and translation averaging are superior than the state-of-the-art global systems. Extensive experiments on different multi-camera systems demonstrate the scalability, robustness and accuracy of our system. By equipping our system, both self-driving cars and intelligent robots can better perceive the surrounding environment. Next, on the premise of ensuring efficiency, we will consider adding a small number of 3D points into system to study whether they can further improve the reconstruction performance.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (under grants 62073320, 61873265 and $U1805264$).

References

Agarwal, S.; Mierle, K.; and Others. 2021. Ceres Solver. http://ceres-solver.org.

Arandjelović, R.; and Zisserman, A. 2012. Three things everyone should know to improve object retrieval. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Barath, D.; Mishkin, D.; Eichhardt, I.; Shipachev, I.; and Matas, J. 2021. Efficient Initial Pose-Graph Generation for Global SfM. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Boyd, S. P.; Parikh, N.; Chu, E.; Peleato, B.; and Eckstein, J. 2011. Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. *Found. Trends Mach. Learn.*, 3: 1–122.

Chatterjee, A.; and Govindu, V. M. 2017. Robust relative rotation averaging. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 40(4): 958–972.

Chen, C.; Liu, Y.; Kreiss, S.; and Alahi, A. 2019. Crowd-Robot Interaction: Crowd-Aware Robot Navigation With Attention-Based Deep Reinforcement Learning. *ICRA*.

Chen, X.; Vizzo, I.; Läbe, T.; Behley, J.; and Stachniss, C. 2021. Range Image-based LiDAR Localization for Autonomous Vehicles. *ICRA*.

Chen, Y.; Zhao, J.; and Kneip, L. 2021. Hybrid Rotation Averaging: A Fast and Robust Rotation Averaging Approach. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Colledanchise, M.; Malafronte, D.; and Natale, L. 2020. Act, Perceive, and Plan in Belief Space for Robot Localization. *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 3763–3769.

Cornelis, K.; Verbiest, F.; and Gool, L. V. 2004. Drift detection and removal for sequential structure from motion algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 26(10): 1249–1259.

Crandall, D. J.; Owens, A.; Snavely, N.; and Huttenlocher, D. P. 2012. SfM with MRFs: Discrete-continuous optimization for large-scale structure from motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 35(12): 2841–2853.

Cui, H.; Gao, X.; Shen, S.; and Hu, Z. 2017. HSfM: Hybrid structure-from-motion. In *CVPR*. IEEE.

Cui, H.; Shen, S.; Gao, W.; Liu, H.; and Wang, Z. 2019. Efficient and robust large-scale structure-from-motion via track selection and camera prioritization. *ISPRS Journal of Photogrammetry and Remote Sensing*.

Cui, H.; Shen, S.; Gao, W.; and Wang, Z. 2018. Progressive Large-Scale Structure-from-Motion with Orthogonal MSTs. *International Conference on 3D Vision (3DV)*.

Cui, H.; Shi, T.; Zhang, J.; Xu, P.; Meng, Y.; and Shen, S. 2021. View-graph construction framework for robust and efficient structure-from-motion. *Pattern Recognition*, 114: 107712.

Cui, Z.; and Tan, P. 2015. Global Structure-from-Motion by Similarity Averaging. In *International Conference on Computer Vision (ICCV)*. IEEE.

Dai, Y.; Trumpf, J.; Li, H.; Barnes, N.; and Hartley, R. I. 2009. Rotation Averaging with Application to Camera-Rig Calibration. In *Asian Conference on Computer Vision (ACCV)*. Springer.

Dellaert, F.; Rosen, D. M.; Wu, J.; Mahony, R.; and Carlone, L. 2020. Shonan Rotation Averaging: Global Optimality by Surfing SO (p)$\hat{n}$. In *European Conference on Computer Vision (ECCV)*. Springer.

Dong, Q.; Gao, X.; Cui, H.; and Hu, Z. 2020. Robust camera translation estimation via rank enforcement. *IEEE Transactions on Cybernetics*.

Fredriksson, J.; and Olsson, C. 2012. Simultaneous multiple rotation averaging using lagrangian duality. In *Asian Conference on Computer Vision (ACCV)*. Springer.

Gao, X.; Zhu, L.; Xie, Z.; Liu, H.; and Shen, S. 2021. Incremental Rotation Averaging. *International Journal of Computer Vision (IJCV)*, 129: 1202–1216.

Geiger, A.; Lenz, P.; and Urtasun, R. 2012. Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Goldstein, T.; Hand, P.; Lee, C.; Voroninski, V.; and Soatto, S. 2016. ShapeFit and ShapeKick for Robust, Scalable Structure from Motion. In *European Conference on Computer Vision (ECCV)*. Springer.

Govindu, V. M. 2004. Lie-algebraic averaging for globally consistent motion estimation. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Guibas, L. J.; Huang, Q.; and Liang, Z. 2019. A condition number for joint optimization of cycle-consistent networks. *Advances in Neural Information Processing Systems*, 32: 1007–1017.

Hartley, R.; Aftab, K.; and Trumpf, J. 2011. L1 rotation averaging using the Weiszfeld algorithm. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE.

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

Hartley, R.; and Zisserman, A. 2003. *Multiple view geometry in computer vision*. Cambridge university press.

Hartley, R. I.; and Sturm, P. 1997. Triangulation. *Computer Vision and Image Understanding (CVIU)*, 68(2): 146–157.

Heng, L.; Choi, B.; Cui, Z.; Geppert, M.; Hu, S.; Kuan, B.; Liu, P.; Nguyen, R.; Yeo, Y. C.; Geiger, A.; et al. 2019. Project autovision: Localization and 3d scene perception for an autonomous vehicle with a multi-camera system. In *International Conference on Robotics and Automation (ICRA)*. IEEE.

Heng, L.; Furgale, P.; and Pollefeys, M. 2015. Leveraging image-based localization for infrastructure-based calibration

of a multi-camera rig. *Journal of Field Robotics*, 32(5): 775–802.

Holynski, A.; Geraghty, D.; Frahm, J.-M.; Sweeney, C.; and Szeliski, R. 2020. Reducing Drift in Structure From Motion Using Extended Features. In *2020 International Conference on 3D Vision (3DV)*. IEEE.

Jiang, N.; Cui, Z.; and Tan, P. 2013. A global linear method for camera pose registration. In *International Conference on Computer Vision (ICCV)*. IEEE.

Lin, Y.; Larsson, V.; Geppert, M.; Kukelova, Z.; Pollefeys, M.; and Sattler, T. 2020. Infrastructure-based multi-camera calibration using radial projections. In *European Conference on Computer Vision(ECCV)*, 327–344. Springer.

Moulon, P.; Monasse, P.; and Marlet, R. 2013. Global Fusion of Relative Motions for Robust, Accurate and Scalable Structure from Motion. In *International Conference on Computer Vision (ICCV)*. IEEE.

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

Ozyesil, O.; and Singer, A. 2015. Robust camera location estimation by convex programming. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Purkait, P.; Chin, T.-J.; and Reid, I. 2020. NeuRoRA: Neural robust rotation averaging. In *European Conference on Computer Vision (ECCV)*. Springer.

Rother, C. 2003. Linear Multi-View Reconstruction of Points, Lines, Planes and Cameras using a Reference Plane. In *International Conference on Computer Vision (ICCV)*. IEEE.

Schönberger, J. L.; and Frahm, J.-M. 2016. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Schönberger, J. L.; Price, T.; Sattler, T.; Frahm, J.-M.; and Pollefeys, M. 2016. A Vote-and-Verify Strategy for Fast Spatial Verification in Image Retrieval. In *Asian Conference on Computer Vision (ACCV)*. Springer.

Strisciuglio, N.; Tylecek, R.; Blaich, M.; Petkov, N.; Biber, P.; Hemming, J.; van Henten, E.; Sattler, T.; Pollefeys, M.; Gevers, T.; et al. 2018. Trimbot2020: an outdoor robot for automatic gardening. In *ISR 2018; 50th International Symposium on Robotics*. VDE.

Sweeney, C. 2015. Theia Multiview Geometry Library: Tutorial & Reference. http://theia-sfm.org.

Sweeney, C.; Sattler, T.; Hollerer, T.; Turk, M.; and Pollefeys, M. 2015. Optimizing the Viewing Graph for Structure-from-Motion. In *International Conference on Computer Vision (ICCV)*. IEEE.

Wang, Y.; Huang, K.; Peng, X.; Li, H.; and Kneip, L. 2020. Reliable frame-to-frame motion estimation for vehicle-mounted surround-view camera systems. In *2020 IEEE International conference on robotics and automation (ICRA)*. IEEE.

Wilson, K.; Bindel, D.; and Snavely, N. 2016. When is Rotations Averaging Hard ? In *European Conference on Computer Vision(ECCV)*. Springer.

Wilson, K.; and Snavely, N. 2014. Robust global translations with 1DSfM. In *European Conference on Computer Vision (ECCV)*. Springer.

Yang, L.; Li, H.; Rahim, J. A.; Cui, Z.; and Tan, P. 2021. Hybrid Rotation Averaging: A Fast and Robust Rotation Averaging Approach. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Zach, C.; Klopschitz, M.; and Pollefeys, M. 2010. Disambiguating visual relations using loop constraints. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Zhang, Q.; and Pless, R. 2004. Extrinsic calibration of a camera and laser range finder (improves camera calibration). In *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE.

Zhu, S.; Zhang, R.; Zhou, L.; Shen, T.; Fang, T.; Tan, P.; and Quan, L. 2018. Very large-scale global sfm by distributed motion averaging. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Zhuang, B. 2018. BATA code. https://bbzh.github.io/.

Zhuang, B.; Cheong, L.-F.; and Lee, G. H. 2018. Baseline desensitizing in translation averaging. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE.