2016 23rd International Conference on Pattern Recognition (ICPR)
Cancún Center, Cancún, México, December 4-8, 2016

Robust Global Translation Averaging with Feature Tracks

Hainan Cui*, Shuhan Shen*†, Zhanyi Hu*†
{*hncui, shshen, huzy*}*@nlpr.ia.ac.cn*
* NLPR, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
† University of Chinese Academy of Sciences, Beijing 100049, China

***Abstract*—How to average translations is the single most difficult task in global structure-from-motion (SfM) to fully tap its potentials in terms of reconstruction efficiency and accuracy since usually only noisy translation directions can be factored out from essential matrices due to the inevitable matching outliers. To tackle this problem, this work proposes a two-step strategy. Firstly, a "2-point method" is introduced to refine the epipolar geometry by which a more accurate track set is generated. Then, translation lengths are computed by solving a convex L1 optimization according to the adjacent triangles induced by the selected tracks and translations. Extensive experiments show that our method performs similarly or better than the state-of-art SfM approaches in terms of the reconstruction accuracy, completeness and efficiency.** [1]

I. Introduction

Structure-from-Motion (SfM) technique has been widely used to reconstruct the 3D scene from a set of image collections. Depending on the manner of computing initial camera poses, the SfM pipeline is divided into two classes: incremental and global. Incremental SfM methods [1], [2] reconstruct the "seed" images first, then incrementally add other images into the reconstruction system. This mode usually suffers from scene drift due to the errors accumulation and heavy computation load. Thus, many recent SfM approaches [3]–[5] adopt the global mode, where all the camera poses are computed at the same time.

The pipeline of global SfM consists of four main modules: rotation averaging, translation averaging, triangulation and bundle adjustment. For the rotation averaging, many literatures [6]–[8] have a wide study on its computation and formulation. For example, Chatterjee *et al.* [8] proposed a rotation averaging method called 'L1-IRLS', which solved the rotation averaging problem in the $SO(3)$ Lie space first, then refined by an IRLS(Iterative Re-weighed Least Square) optimization method. For the translation averaging, the problem is more difficult in the global SfM method due to the following two main challenges. Firstly, the translation estimations are noisy due to the features match outliers. Secondly, the pairwise translations generated by factoring the essential matrices only tell the direction of cameras motion, many methods degenerate at collinear camera motions [9], [10]. In the literatures [11]–[13], these erroneous translations are filtered out from epipolar geometry graph (EG) first based on various consistency constraint. However, the filtering is still a non-trival problem, and filtering edges may break the original parallel rigidity of EG [10]. As a result, the reconstruction completeness cannot be guaranteed. In this work, aiming to solve the translation averaging problem, we refine the translations first and then compute the translation scales.

Given the global camera rotations computed by [8], relative rotations on the epipolar edges could be updated. Then, we tackle the translation estimation problem by a robust "2-point" method for each epipolar edge. We find that the translation refining process is very fast and the resulting directions are more accurate than the original directions generated by factoring the essential matrices.

For the translation scale estimation, we solve it by importing the tracks. In this way, the parallel rigidity [10] of EG are enhanced without loss of any cameras. As the reconstruction result is only up to a scale, thus we only need to calculate the ratio between two translation scales. Then, by setting the scale of anyone translation to a fixed constant, the whole translation scales are obtained. To determine the ratio between two translation scales (scale-ratio), we resort to the tracks and translations to construct adjacent triangles. In this way, the scale-ratio could be calculated by the basic triangle principle and a median filtering strategy. Given the ratios of pairwise translation scales, the translation averaging estimation problem could be formulated as two convex L1 problems, which could be rapidly optimized. Extensive results show that our method performs similarly or better than the recent global SfM methods and well-known incremental methods in terms of computational cost and reconstruction accuracy.

II. Related Work

The rotation averaging is a relatively matured problem in the global SfM [6]–[8], [14]. Thus, we use the state-of-art approach proposed in [8] to solve the rotation averaging task, which was also adopted in several recent works [5], [10], [13], [15]. Since our work is focused on the translation averaging, the most relevant work are the following translation averaging approaches.

[1]This work was supported in part by the National High Technology R&D Program of China (863 Program) under Grant 2015AA124102, and in part by the Natural Science Foundation of China under Grants 61333015, 61473292 and 61273280.

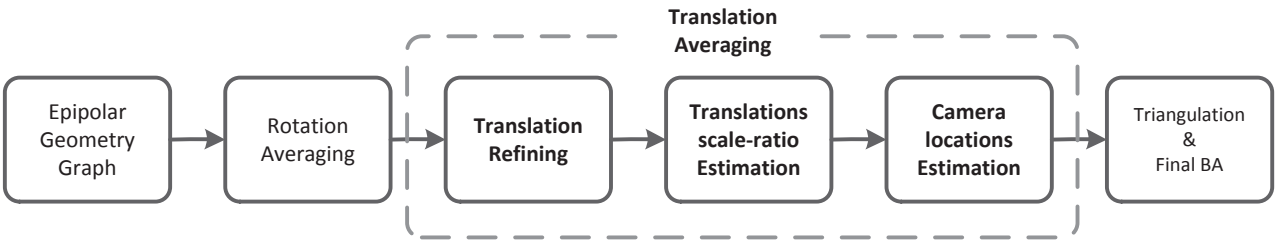

Fig. 1. Pipeline of our global SfM method.

Linear Translation Averaging Many linear methods [9], [11], [15]–[17] are proposed for the translation averaging problem solving. For example, Jiang *et al.* [11] proposed a linear method where accurate pairwise geometries are required to perform the SVD decomposition for camera positions estimation, and Cui *et al.* [15] enforced it with feature tracks. While efficient, such linear methods are rather sensitive to outliers.

Triplet Translation Averaging To increase the reconstruction accuracy, many approaches [12], [18], [19] resort to accurate triple geometries. For example, Moulon *et al.* [12] proposed a contrario trifocal tensor estimation method to extract accurate translations. While the accuracy could be increased, such methods are likely to discard many useful images since EG may be not accurate and dense enough. As a result, the completeness of the reconstruction cannot be guaranteed.

Outlier EG filtering To decrease the impact of noisy translations factorized from poor essential matrices, many methods [13], [20]–[22] adopt an epipolar edge filtering step before the camera translation averaging. For example, for densely matched images, Wilson *et al.* [13] proposed a hash-like method, called 1DSfM, which projected the 3-dimensional translation directions to multiple 1-dimensional subspaces. Based on the orientation consistency on a 1-dimensional axis, the translation outliers are filtered by a voting scheme. However, as proved in [10], essential matrices only determine camera positions in a parallel rigid graph. As a result, inaccurate filtering rule may destroy the original parallel rigidity of EG [10] and make the reconstruction incomplete.

EG refining Instead of filtering translation outliers, some methods [3], [5], [10] prefer to refine the translations on the epipolar edges. Sweeney *et al.* [3] improved the quality of the relative geometries in the epipolar geometry graph by enforcing loop consistency constraints with the epipolar point transfer. Though this assumption works well for Internet images, it fails in cases where the epipolar lines are parallel or in a coincidence. In addition, in order to improve local relative motion estimations, Cui *et al.* [5] performed a local bundle adjustment (LBA) step on each epipolar edge before the translation averaging.

Fusing Auxiliary Information Other work [4], [23]–[25] utilize some auxiliary information to improve the translation accuracy. For example, Crandall *et al.* [23] proposed a discrete-continuous optimization system, where noisy auxiliary info (GPS and vertical vanishing point) is incorporated to the MRF formulation. Cui *et al.* [4] fused the auxiliary imaging information to iteratively distinguish potential inliers. However, since these auxiliary information are vital to the reconstruction accuracy in such methods, they cannot be extended to general cases.

In this work, a two-step translation averaging strategy is introduced. Given the global camera rotations, the translation is first refined by a "2-point" method under RANSAC paradigm. With updated epipolar geometry graph, tracks are generated for the translation scales computation. Then, by calculating the ratio relationship on the adjacent triangles induced by tracks and pairwise translations, we formulate the translation averaging problem as convex L1 optimizing problems which reach the global optimum rapidly [26].

III. Proposed Translation Averaging Algorithm

A. Overview

The input of our system is an epipolar geometry graph (EG), where vertices correspond to images and edges link matched image pairs. For each edge, its essential matrix is obtained by 5-point algorithm [27]. The essential matrix on an edge e_{ij} encodes the relative rotation $\mathbf{R}_{ij}$ and the relative translation $\mathbf{t}_{ij}$. We aim to estimate global camera locations $\mathbf{c}_i$ and rotations $\mathbf{R}_i$ for each view to satisfy following two constraints:

$$\begin{aligned} \mathbf{R}_{ij} &= \mathbf{R}_j \mathbf{R}_i^T \\ \lambda_{ij}\mathbf{t}_{ij} &= \mathbf{R}_j(\mathbf{c}_i - \mathbf{c}_j) \end{aligned} \tag{1}$$

Compared to the rotation averaging problem, translation averaging problem is more difficult since the $\mathbf{t}_{ij}$ factorized from essential matrix only gives the motion direction between two cameras. In this case, only the parallel rigid subgraph of EG could be determined, as demonstrated in [10]. Thus, we utilize the tracks to enhance the parallel rigidity of epipolar graph. In this way, all the camera poses in the epipolar geometry graph could be determined.

With our proposed translation averaging method, Fig. 1 shows the pipeline of our global SfM. In particular, the rotation averaging problem is solved by the L1-IRLS method proposed

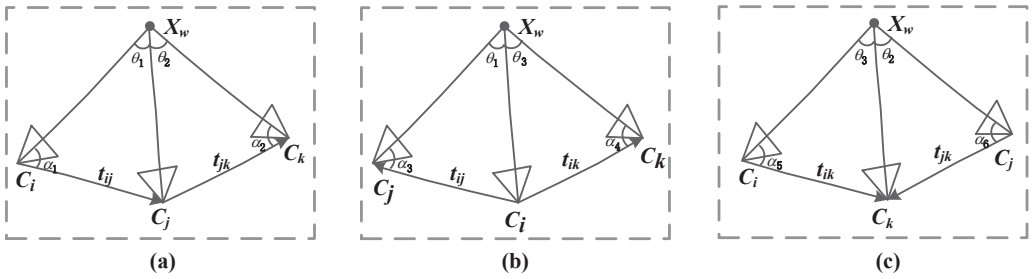

Fig. 2. Different configurations of translations scale-ratio computation.

in [8]. Given the global camera rotations, the translation averaging problem is formulated as finding global camera locations $\mathbf{c}_i$ to satisfy:

$$\begin{aligned} \lambda_{ij}\mathbf{c}_{ij} &= \mathbf{c}_i - \mathbf{c}_j \\ s.t. \quad \mathbf{c}_{ij} &= \mathbf{R}_j^T \mathbf{t}_{ij} \end{aligned} \tag{2}$$

As a result, this formulation mainly relates to the estimated translation $\mathbf{t}_{ij}$ and the scale factor λ_{ij}. Our main **Contribution** is to efficiently and accurately compute the $\mathbf{t}_{ij}$ and λ_{ij}, which is done by the following 3 steps:

- **Translation Refining**: Given global camera rotations, the relative rotation on each epipolar edge could be easily obtained by $\mathbf{R}_{ij} = \mathbf{R}_j\mathbf{R}_i^T$. With known $\mathbf{R}_{ij}$, $\mathbf{t}_{ij}$ is re-calculated by two feature matches with RANSAC technique.
- **Translations scale-ratio Estimation**: For λ_{ij}, we calculate the ratio between two translation scales (scale-ratio) by two adjacent triangles induced by tracks and translations. Since tracks are always redundant, we make a feature tracks selection to accelerate the scale-ratios computation. With scale-ratio estimations, λ_{ij} is solved as a convex L1 optimization problem.

When the $\mathbf{t}_{ij}$ and λ_{ij} are known, then the global camera locations are solved as a convex L1 optimization problem [5].

B. Translation Refining

Given image feature matches, the essential matrix $\mathbf{E}$ could be calculated by the 5-point algorithm, where the normalized feature match inlier $\mathbf{p}, \mathbf{p}'$ should satisfy:

$$\mathbf{p}'^T * \mathbf{E} * \mathbf{p} = \mathbf{p}'^T * [\mathbf{t}_{ij}]_\times * \mathbf{R}_{ij} * \mathbf{p} = 0 \tag{3}$$

Then, with known $\mathbf{R}_{ij}$, let $\mathbf{t}_{ij} = \{t_x, t_y, t_z\}$, $\mathbf{p}' = \{p'_x, p'_y, 1.0\}$, the above equation could be rewritten as:

$$\begin{bmatrix} p'_x & p'_y & 1.0 \end{bmatrix} \begin{bmatrix} 0 & -t_z & t_y \\ t_z & 0 & -t_x \\ -t_y & t_x & 0 \end{bmatrix} \begin{bmatrix} q_x \\ q_y \\ 1.0 \end{bmatrix} = 0 \tag{4}$$

where $\beta * [q_x, q_y, 1.0]^T = \mathbf{R}_{ij} * \mathbf{p}$ and β is a scale factor. Based on this formulation, the translation $\mathbf{t}_{ij}$ computation could be wrote as:

$$\begin{bmatrix} t_x & t_y & t_z \end{bmatrix} \begin{bmatrix} p'_y - q_y \\ q_x - p'_x \\ p'_x * q_y - p'_y * q_x \end{bmatrix} = 0 \tag{5}$$

Since the translation is up to scale, we could assume $t_z = 1$. In this way, $\mathbf{t}_{ij}$ could be determined by only 2 feature matches. Considering the robustness, a RANSAC technique is used here to find the best $\mathbf{t}_{ij}$ which corresponds to the largest number of consistent feature matches.

It is noted that sometimes the obtained translation may be inverse since t_z is set to positive. Thus, for each estimation, this ambiguity could be solved according to the cheirality [28] of some feature matches triangulation. In each RANSAC iteration, we evaluate the translation estimation result by the number of feature matches inliers. A feature match is considered as an inlier when the distances between the feature points and respectively corresponding epipolar lines are less than a threshold. In our work, the RANSAC is performed 256 times and the inlier threshold is set to 0.4% of the maximum image dimension.

C. Translations scale-ratio Computing

Given the refined EG, tracks are generated along the matched image pairs. Note that a track is defined as a collection of interest points across multiple images which have similar feature descriptors, and each identifies a unique 3D point in the real scene.

A 3D scene point could be viewed as the intersection of multiple rays launched from its visible views. For each ray, it also crosses the image projection of this scene point. Ideally, when three or more views are visible in a track, the rays and the translation estimations between the images could construct many adjacent triangles. Based on these adjacent triangles, the translations scale-ratio could be obtained.

Considering a configuration that a track with three visible images $\{C_i, C_j, C_k; i < j < k\}$, there are three different camera intersection cases between two translations. Note that in the general case, the translation direction of an epipolar edge is from small image index to large image index. Fig. 2

illustrates these three configurations, and the main difference among them is the different shared edge between two adjacent triangles. Since not all of the pairs of visible images in a track have pairwise translation estimations, here we only consider those image pairs that have translation estimations. In fact, for the images in any triplet, if they are connected by two or more epipolar edges, the translations scale-ratio can be obtained.

For the three views C_i, C_j, C_k and a scene point X_w, the projection equation is:

$$\begin{aligned} \mu_{iw}\mathbf{x}_{iw} &= \mathbf{K}_i * \mathbf{R}_i * (\mathbf{X}_w - \mathbf{C}_i) \\ \mu_{jw}\mathbf{x}_{jw} &= \mathbf{K}_j * \mathbf{R}_j * (\mathbf{X}_w - \mathbf{C}_j) \\ \mu_{kw}\mathbf{x}_{kw} &= \mathbf{K}_k * \mathbf{R}_k * (\mathbf{X}_w - \mathbf{C}_k) \end{aligned} \tag{6}$$

Given known global camera rotations and initial camera intrinsic matrices, the directions of rays from the optical center C_i, C_j, C_k to the scene point X_w are computed by ($\simeq$ denotes up to scale):

$$\begin{aligned} \mathbf{X}_w - \mathbf{C}_i &\simeq \mathbf{R}_i^T * \mathbf{K}_i^{-1} * \mathbf{x}_{iw} \\ \mathbf{X}_w - \mathbf{C}_j &\simeq \mathbf{R}_j^T * \mathbf{K}_j^{-1} * \mathbf{x}_{jw} \\ \mathbf{X}_w - \mathbf{C}_k &\simeq \mathbf{R}_k^T * \mathbf{K}_k^{-1} * \mathbf{x}_{kw} \end{aligned} \tag{7}$$

For Fig.2a, we calculate the scale-ratio between translation $\mathbf{t}_{ij}$ and $\mathbf{t}_{jk}$. According to the triangle principle, the ratio between the length of edges is equal to the sine value of its corresponding angle, *i.e.* ,

$$\frac{|\lambda_{ij}|}{|X_w - C_j|} = \frac{\sin(\theta_1)}{\sin(\alpha_1)}, \frac{|\lambda_{jk}|}{|X_w - C_j|} = \frac{\sin(\theta_2)}{\sin(\alpha_2)} \tag{8}$$

Since the triangle $\{C_i - C_j - X_w\}$ shares an edge $\{C_j - X_w\}$ with the triangle $\{C_j - C_k - X_w\}$, the scale-ratio s_{ij}^{jk} could be calculated by:

$$s_{ij}^{jk} = \frac{|\lambda_{ij}|}{|\lambda_{jk}|} = \frac{\sin(\theta_1) * \sin(\alpha_2)}{\sin(\theta_2) * \sin(\alpha_1)} \tag{9}$$

Similarly, the scale-ratio s_{ij}^{ik} between translation $\mathbf{t}_{ij}$ and $\mathbf{t}_{ik}$ (Fig.2b) is computed by:

$$s_{ij}^{ik} = \frac{|\lambda_{ij}|}{|\lambda_{ik}|} = \frac{\sin(\theta_1) * \sin(\alpha_4)}{\sin(\theta_3) * \sin(\alpha_3)} \tag{10}$$

and the scale-ratio s_{ik}^{jk} between translation $\mathbf{t}_{ik}$ and $\mathbf{t}_{jk}$ (Fig.2c) is computed by:

$$s_{ik}^{jk} = \frac{|\lambda_{ik}|}{|\lambda_{jk}|} = \frac{\sin(\theta_3) * \sin(\alpha_6)}{\sin(\theta_2) * \sin(\alpha_5)} \tag{11}$$

For each track with three or more visible views, a group of three views which has two or more epipolar edges inside could construct one of the above functions. As a result, the scale-ratio between two translations may have different values which are generated by different tracks. However, since the track outliers are inevitable, some of the estimations may be erroneous. Considering the robustness, for the scale-ratio estimations corresponding to the same pair of translations, we set the median value of these estimations as the final scale-ratio of these two translations.

Since the tracks are always redundant, we perform a tracks selection for the sake of efficiency. To guarantee that all the images in the epipolar geometry graph (EG) are involved in the solving process, we sort all feature tracks by their lengths in descending order, and then a compact set of tracks is selected to cover γ spanning trees of EG. In our work, γ is set to 30 and the strategy of the construction of a spanning tree is breadth-first with a random initial image vertex.

D. Camera location Estimation

As proposed in [5], given the translation scale factors, the camera locations could be estimated by convex $L1$ optimizations. Given the scale-ratio s_{ij}^{jk} between two epipolar edges e_{ij} and e_{jk}:

$$s_{ij}^{jk} = \frac{|e_{ij}|}{|e_{jk}|} \tag{12}$$

by taking log of both sides, we have

$$log(|e_{ij}|) - log(|e_{jk}|) = log(s_{ij}^{jk}) \tag{13}$$

Collecting this equation from all the scale-ratio estimated in the Sec III-C, we stack them into a linear equation system:

$$\mathbf{A}_s * \mathbf{x}_s = \mathbf{b}_s \tag{14}$$

where $\mathbf{x}_s$ and $\mathbf{b}_s$ are the vectors by concatenating $log(|e_{ij}|)$ and $log(s_{ij}^{jk})$ respectively, and $\mathbf{A}_s$ is a sparse matrix where nonzero values are only 1 and -1. As the translation length estimation is up to scale, to remove the gauge ambiguity, we set the first epipolar edge, for example e_{12}, as unit, *i.e.* , $log(e_{12}) = 0$. Then, the equation system is solved by the following convex $L1$ optimization problem:

$$\arg\min \|\mathbf{A}_s * \mathbf{x}_s - \mathbf{b}_s\|_{L1} \tag{15}$$

After this optimization, the scale factor λ_{ij} for the epipolar edge e_{ij} is obtained. Thus, with the global camera rotations calculated by [8], the right side of the following equation is also known.

$$\mathbf{c}_i - \mathbf{c}_j = \lambda_{ij}\ \mathbf{R}_j^T \mathbf{t}_{ij} \tag{16}$$

Each epipolar edge could construct such an equation. Since the number of epipolar edge is far more than the number of images in the epipolar geometry graph, thus an over-determined equation system is obtained. Similarly, by collecting this equations from all the edges in the epipolar geometry graph, we have the following linear equation system:

$$\mathbf{A}_c * \mathbf{x}_c = \mathbf{b}_c \tag{17}$$

where $\mathbf{x}_c$ and $\mathbf{b}_c$ are the vectors by concatenating $\mathbf{c}_i$ and $\lambda_{ij}\mathbf{R}_j^T\mathbf{t}_{ij}$, Similarly, we set the first camera location, for example $\mathbf{c}_1$, as original, *i.e.* $\mathbf{c}_1 = \mathbf{0}$. Thus, all the camera locations is solved by the following $L1$ optimization problem:

$$\arg\min \|\mathbf{A}_c * \mathbf{x}_c - \mathbf{b}_c\|_{L1} \tag{18}$$

IV. EXPERIMENT

We evaluate the whole global Structure-from-Motion system on various image datasets, including the MVS benchmark datasets in Strecha [29], and some Internet images, as well as UAV images. The Ceres-solver [31] is used for our final bundle adjustment.

TABLE I

CAMERA CENTERS ACCURACY ON BENCHMARK DATASETS [29]. THE RESULTS OF OTHER METHODS ARE TAKEN THE RESULT OF [3] AS A REFERENCE.

Method	Accuracy (mm)						
	VSFM [2]	Jiang [11]	Olsson [30]	Cui *et al.* [15]	Moulon [12]	Sweeney [3]	Ours
FountainP11	7.6	14	2.2	2.5	2.5	2.4	**1.7**
EntryP10	63.0	–	6.9	–	5.9	5.7	**5.4**
HerzJesuP8	19.3	–	3.9	–	**3.5**	**3.5**	3.8
HerzJesuP25	22.4	64	5.7	5.0	5.3	5.3	**4.7**
CastleP19	258	–	76.2	–	25.6	38.2	**12.3**
CastleP30	522	235	66.8	21.2	21.9	32.4	**18.6**

TABLE II

ACCURACY AND TIME-COST COMPARISON. 'NI' AND 'NC' RESPECTIVELY DENOTES THE NUMBER OF INPUT AND CALIBRATED IMAGES, 'X' DENOTES MEDIAN ERROR(METERS), 'T' DENOTES TIME-COST(SECONDS). THE RESULTS OF OTHER METHODS ARE TAKEN THE RESULT OF [5] AS A REFERENCE.

Data		1DSfM [13]			LUD [10]			Jiang *et al.* [11]			Cui *et al.* [5]			Ours		
Name	Ni	Nc	x	T	Nc	x	T	Nc	x	T	Nc	x	T	Nc	x	T
Alamo	627	529	1.1	910	547	**0.4**	750	478	0.6	191	574	0.5	578	**577**	**0.4**	530
Ellis Island	247	214	3.7	171	–	–	–	205	3.2	621	223	0.7	208	**227**	**0.6**	311
Notre Dame	552	507	10.0	1599	536	0.3	1047	518	0.4	1351	549	**0.2**	552	**552**	**0.2**	706

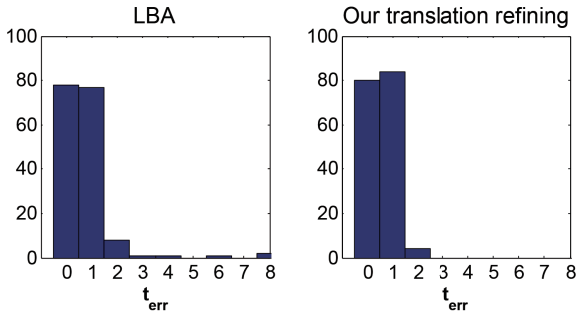

Fig. 3. This figure shows the translation accuracy comparison between local bundle adjustment and our refining method.

A. Evaluation on Benchmark Datasets

We compare our method with a typical incremental method VSFM [2], and several alternative global SfM pipelines. Table I shows the accuracy comparison results on the Strecha MVS benchmark datasets [29], from which we could see that our method performs similarly or better than the other methods. In particular, for the datasets "CastleP19" and "CastleP30", where the tracks and epipolar geometry graph are noisier than the other datasets, our result has a significant improvement on the camera locations accuracy, which mainly owes to the effectiveness of our translation refining module.

To demonstrate the effectiveness of translation refining module, we compare our method with LBA (Local Bundle Adjustment [5]) on the dataset "CastleP30". The translation error is the angular distance(in degrees) between the unit-norm vectors $\mathbf{t}_{ij}$ with ground-truth $\mathbf{t}_{ij}^{gt}$, *i.e.* $t_{err} = acos(\mathbf{t}_{ij}{}^T\mathbf{t}_{ij}^{gt})$. The accuracy comparison is shown in Fig. 3, where 'x' axis denotes the angular distance and 'y' axis denotes the number of epipolar edges in the corresponding bin, from which we could see our method outperforms the LBA, indicating our method is more robust to feature match outliers.

B. Evaluation on unordered and ordered images

To further evaluate our translation averaging module, we perform it on three public unordered Internet datasets "Alamo", "Ellis Island" and "Notre Dame" [13].

Table II shows the comparison of camera centers accuracy and time-cost among five global SfM approaches. As the epipolar edge filtering step may weaken the parallel rigidity of EG, we can see that many of images in the methods [11], [13] are left uncalibrated. Compared to the methods [5], [10], our method can not only calibrate the most images but have the best accuracy on the dataset "Alamo" and "Ellis Island". In addition, for the dataset "Notre Dame", the accuracy and time-cost among these methods [5], [10] are basically comparable, while the number of the calibrated images of our method is larger than them. The reason is two-fold: on one hand, the refined epipolar graph make the tracks more clean; on the other hand, due to the median filtering strategy in the scale-ratio estimations, our method is more robust to the feature track outliers.

Besides, our method is also performed on several large ordered image datasets, including a dataset with 1347 UAV images. Some of the detailed reconstruction results are shown in Fig. 4.

V. CONCLUSION

In this paper, we introduce a robust global translation averaging method. Given global camera rotations, the local translations are first computed by a "2-point" method. Then, the ratio of translation scales are determined by utilizing both tracks and refined translations. With accurate translations and their scale-ratios, our translation averaging problem is effectively solved by optimizing convex L1 problems. Extensive experiments show the global SfM approach with our translation averaging module performs similarly or better than the state-of-art SfM approaches in terms of reconstruction accuracy, completeness and efficiency.

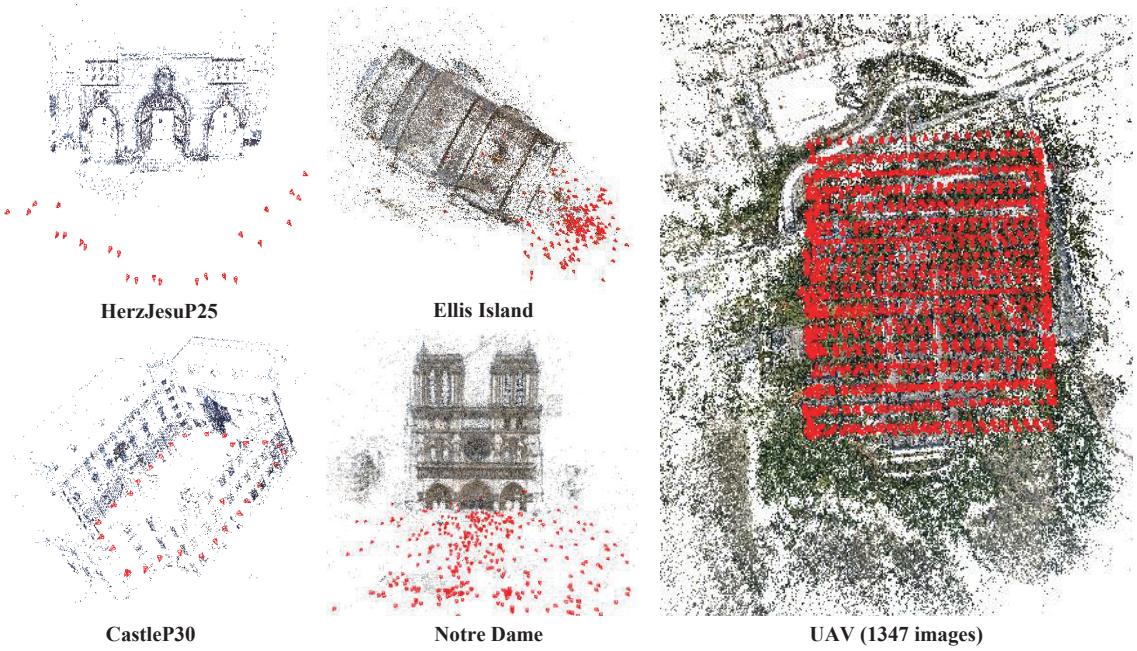

Fig. 4. This figure shows the detailed reconstruction results, where red cones denote the calibrated camera poses.

REFERENCES

[1] N. Snavely, S. M. Seitz, and R. Szeliski, "Modeling the world from internet photo collections," *IJCV*, vol. 80, no. 2, pp. 189–210, 2008.

[2] C. Wu, "Towards linear-time incremental structure from motion," in *IEEE 3DTV 2013*, 2013, pp. 127–134.

[3] C. Sweeney, T. Sattler, T. Hollerer, M. Turk, and M. Pollefeys, "Optimizing the viewing graph for structure-from-motion," in *IEEE ICCV*, 2015, pp. 801–809.

[4] H. Cui, S. Shen, Z. Hu *et al.*, "Efficient large-scale structure from motion by fusing auxiliary imaging information," *IEEE TIP*, vol. 22, pp. 3561–3573, 2015.

[5] Z. Cui and P. Tan, "Global structure-from-motion by similarity averaging," in *IEEE ICCV*, 2015, pp. 864–872.

[6] D. Martinec and T. Pajdla, "Robust rotation and translation estimation in multiview reconstruction," in *IEEE CVPR*, 2007, pp. 1–8.

[7] R. Hartley, J. Trumpf, Y. Dai, and H. Li, "Rotation averaging," *IJCV*, vol. 103, no. 3, pp. 267–305, 2013.

[8] A. Chatterjee and V. M. Govindu, "Efficient and robust large-scale rotation averaging," in *IEEE ICCV*, 2013, pp. 521–528.

[9] M. Arie-Nachimson, S. Z. Kovalsky, I. Kemelmacher-Shlizerman, A. Singer, and R. Basri, "Global motion estimation from point matches," in *IEEE 3DIMPVT*, 2012, pp. 81–88.

[10] O. Ozyesil and A. Singer, "Robust camera location estimation by convex programming," in *IEEE CVPR*, 2015, pp. 2674–2683.

[11] N. Jiang, Z. Cui, and P. Tan, "A global linear method for camera pose registration," in *IEEE ICCV*, 2013, pp. 481–488.

[12] P. Moulon, P. Monasse, and R. Marlet, "Global fusion of relative motions for robust, accurate and scalable structure from motion," in *ICCV 2013*. IEEE, 2013, pp. 3248–3255.

[13] K. Wilson and N. Snavely, "Robust global translations with 1dsfm," in *ECCV*. Springer, 2014, pp. 61–75.

[14] V. M. Govindu, "Lie-algebraic averaging for globally consistent motion estimation," in *IEEE CVPR*, vol. 1, 2004, pp. I–684.

[15] Z. Cui, N. Jiang, and P. Tan, "Linear global translation estimation from feature tracks," *arXiv preprint arXiv:1503.01832*, 2015.

[16] V. M. Govindu, "Combining two-view constraints for motion estimation," in *IEEE CVPR*, vol. 2, 2001, pp. II–218.

[17] C. Rother, "Linear multiview reconstruction of points, lines, planes and cameras using a reference plane," in *IEEE ICCV*, 2003, pp. 1210–1217.

[18] M. Havlena, A. Torii, and T. Pajdla, "Efficient structure from motion by graph optimization," in *ECCV*. Springer, 2010, pp. 100–113.

[19] M. Klopschitz, A. Irschara, G. Reitmayr, and D. Schmalstieg, "Robust incremental structure from motion," in *Proc. 3DPVT*, vol. 2, 2010.

[20] C. Zach, A. Irschara, and H. Bischof, "What can missing correspondences tell us about 3d structure and motion?" in *IEEE CVPR*, 2008, pp. 1–8.

[21] C. Zach, M. Klopschitz, and M. Pollefeys, "Disambiguating visual relations using loop constraints," in *IEEE CVPR*, 2010.

[22] N. Jiang, P. Tan, and L.-F. Cheong, "Seeing double without confusion: Structure-from-motion in highly ambiguous scenes," in *IEEE CVPR*, 2012, pp. 1458–1465.

[23] D. J. Crandall, A. Owens, N. Snavely, and D. P. Huttenlocher, "Sfm with mrfs: Discrete-continuous optimization for large-scale structure from motion," *IEEE TPAMI*, vol. 35, no. 12, pp. 2841–2853, 2013.

[24] M. Pollefeys, D. Nistér, J.-M. Frahm, A. Akbarzadeh, P. Mordohai, B. Clipp, C. Engels, D. Gallup, S.-J. Kim, P. Merrell *et al.*, "Detailed real-time urban 3d reconstruction from video," *IJCV*, vol. 78, no. 2-3, pp. 143–167, 2008.

[25] C. Arth, G. Reitmayr, and D. Schmalstieg, "Full 6dof pose estimation from geo-located images," in *ACCV 2012*. Springer, 2013, pp. 705–717.

[26] E. Candes and J. Romberg, "l1-magic: Recovery of sparse signals via convex programming," *URL: www. acm. caltech. edu/l1magic/downloads/l1magic. pdf*, vol. 4, p. 46, 2005.

[27] D. Nistér, "An efficient solution to the five-point relative pose problem," *IEEE TPAMI*, vol. 26, no. 6, pp. 756–770, 2004.

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

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

[30] O. E. Carl Olsson, "Stable structure from motion for unordered image collections," *Image Analysis*, vol. 6688, pp. 524–535, 2011.

[31] C. http://ceres solver.org/.