SCIENCE CHINA
Information Sciences

• HIGHLIGHT •

March 2017, Vol. 60 038101:1–038101:3
doi: 10.1007/s11432-015-0792-1

Global fusion of generalized camera model for efficient large-scale structure from motion

Hainan CUI[1], Shuhan SHEN[1*] & Zhanyi HU[1,2]

[1]*National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing* 100190*, China;*
[2]*University of Chinese Academy of Sciences, Beijing* 100049*, China*

Received March 30, 2016; accepted July 8, 2016; published online November 9, 2016

Citation Cui H N, Shen S H, Hu Z Y. Global fusion of generalized camera model for efficient large-scale structure from motion. Sci China Inf Sci, 2017, 60(3): 038101, doi: 10.1007/s11432-015-0792-1

Recently, interest has grown in building large-scale 3D city models [1] from images captured by multi-camera systems, such as cameras mounted on a car, e.g., Google Street View, or on an unmanned aerial vehicle, e.g., oblique airborne photogrammetry. From such images, structure-from-motion (SfM) techniques can be used to reconstruct the 3D scene. However, as shown in Cui et al. [2], many state-of-the-art SfM methods are incapable of reconstructing the ordered street view images because rigidly mounted cameras are considered separately in the SfM problem solving, i.e., they failed to enforce the inherent rigid transformations of the cameras in the system. Thus, the "generalized camera" model, which is to consider multiple cameras as a single one, is used to solve this problem.

Given accurate transformations among the cameras in a multi-camera system, many state-of-the-art SfM methods fuse the generalized camera model in an incremental way [3–5], which aims at consecutively estimating the relative transformation between two adjacent generalized cameras. However, the scene drift cannot be avoided in incremental method due to errors accumulation, and time-consuming bundle adjustment must be repeatedly activated. To our knowledge, the generalized camera model has never been used in the global SfM approaches.

In this article, we propose a global SfM method under the generalized camera model for the reconstructions of both street view images and oblique airborne images. Contrary to incremental methods, our global method initializes all cameras simultaneously, makes error distribute on the epipolar geometry graph, and has better potential in efficiency and accuracy. In addition, instead of calibrating rigid transformations in advance, we integrate their estimations into our SfM pipeline. Extensive experiments show that our method performs better than two state-of-the-art SfM approaches: Bundler [6] and Cui et al. [2], in terms of scene completeness, reconstruction efficiency and scalability.

Generalized camera model. A generalized camera consists of several rigidly mounted common digital cameras. Let M be the number of the cameras. Figure 1(a) is a graphical representation of our generalized camera model with $M = 4$. For the generalized camera, we choose one camera as the reference camera and assign it with the index '1'. Then, the other cameras are assigned a label from index '2' to 'M'. Let C_{i1} be the reference camera at instance i. The transformations are denoted as

* Corresponding author (email: shshen@nlpr.ia.ac.cn)
The authors declare that they have no conflict of interest.

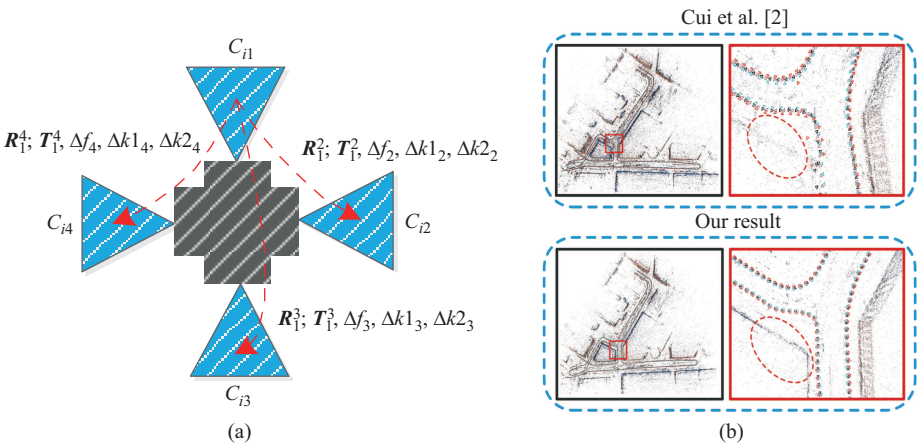

Figure 1 (Color online) (a) An example of the ith generalized camera, which includes four common cameras $C_{i1}, C_{i2}, C_{i3}, C_{i4}$. $\{\boldsymbol{R}_1^j, \boldsymbol{T}_1^j, \Delta f_j, \Delta k1_j, \Delta k2_j, j = 2, 3, 4\}$ denote the camera transformation between cameras C_{i1} and C_{ij}. (b) The reconstruction result comparison between Cui et al. [2] and our method, and cones show the calibrated camera poses.

$\{\boldsymbol{R}_1^j, \boldsymbol{T}_1^j, \Delta f_j, \Delta k1_j, \Delta k2_j, j = 2, 3, 4\}$, where $\boldsymbol{R}_1^j$ denotes the relative camera rotation between camera C_{i1} and C_{ij}, $\boldsymbol{T}_1^j$ denotes the camera C_{ij}'s location in the camera C_{i1}'s coordinate system, Δf_j denotes the difference of focal length between camera C_{ij} and C_{i1}, and $\Delta k1_j, \Delta k2_j$ denote the differences of radial distortions between camera C_{ij} and C_{i1}. Thus, if we get the camera model of C_{i1} and all the relative transformations, all the camera models of C_{i2}, C_{i3}, C_{i4} could be computed.

Our global SfM algorithm. Based on this generalized camera model, the input of our SfM problem consists of: (a) image sets captured by a M-camera system; (b) noisy imaging information for each reference camera, including geotags, compass angle, and focal length. Our goal is to estimate: (1) a 9 degree-of-freedom camera model for each reference camera, including camera rotation matrix $\boldsymbol{R}_{i1}$, camera center $\boldsymbol{T}_{i1}$, and camera intrinsic parameters $f_1, k1_1, k2_1$; (2) the rigid transformations in the generalized camera model; (3) a 3D position for each scene point.

Under the generalized camera model, the number of parameters in the bundle adjustment decreases dramatically. For conventional SfM methods where the cameras are considered separately, the number of parameters in the bundle adjustment is $9NM$. However, for our global SfM method under the generalized camera model, the number of parameters is only $6N + 9M - 6$. Thus, for large-scale scene reconstruction applications, our method has better efficiency and scalability.

Our global SfM method consists of three main steps. The first step is to build an epipolar geometry graph (EG), the second is to perform rotation averaging on the generalized cameras, and the third is for scene reconstruction. SIFT points are extracted from each image, and then matched using cascade hashing strategy. The matching result is represented by an epipolar geometry graph, where vertices denote images and edges link matched pairs. Let $\boldsymbol{R} = \{\boldsymbol{R}_{ij}, i = 1, \ldots, N, j = 1, \ldots, M\}$ be the absolute camera rotations, where $\boldsymbol{R}_{ij}$ denotes the rotation of the jth camera in the ith generalized camera; $\{\boldsymbol{R}_1^j\}$ be the camera rotation transformation between reference camera C_{i1} and camera C_{ij}. For the camera C_{ij}, the corresponding rotation $\boldsymbol{R}_{ij}$ is calculated by $\boldsymbol{R}_{ij} = \boldsymbol{R}_1^j \boldsymbol{R}_{i1}$.

Given pairwise relative rotation estimate $\boldsymbol{R}_{ij}^{pq}$, we perform the rotation averaging in an Iterative Reweighed Least Square manner. In the lth iteration, the goal is to find a set of reference camera rotations and rotation transformations to minimize

$$\sum_{i,p=1}^{N} \sum_{j,q=1}^{M} w_{ij}^{pq(l)} \epsilon_{ij}^{pq(l)}, \tag{1}$$

where $\epsilon_{ij}^{pq(l)} = \|\boldsymbol{R}_{ij}^{pq} - \boldsymbol{R}_1^{q(l)} \boldsymbol{R}_{p1}^{(l)} (\boldsymbol{R}_1^{j(l)} \boldsymbol{R}_{i1}^{(l)})^{\mathrm{T}}\|_F$, and the weighting factor w_{ij}^{pq} is an indicator function. $w_{ij}^{pq(l)} = 0$ when $\epsilon_{ij}^{pq(l)} > \tau_1$, and otherwise set to 1. $\tau_1 = \max\{0.5, 1.2 \times \max_{\mathrm{mst}}\}$, where $\max_{\mathrm{mst}}$ is the largest residual among the edges in the minimum spanning tree (MST) of EG. When the threshold τ_1 is not changed between two consecutive iterations, the rotation averaging terminates. As a result, all the camera rotations $\{\boldsymbol{R}_{ij}\}$ are obtained.

Let $\boldsymbol{T} = \{\boldsymbol{T}_{ij}, i = 1, \ldots, N, j = 1, \ldots, M\}$ be the absolute camera centers, and $\boldsymbol{T}_1^j$ be the centers transformation in the generalized camera. Given the reference camera centers $\boldsymbol{T}_{i1}$ (initialized by geotags) and rough centers transforma-

tion $\boldsymbol{T}_1^j$, the locations of cameras C_{ij} is computed by $\boldsymbol{T}_{ij} = \boldsymbol{R}_{i1}^T \boldsymbol{T}_1^j + \boldsymbol{T}_{i1}$. Given the intrinsic parameters of the reference camera, including focal length f_1, and two radial distortion parameters $k1_1, k2_1$, the intrinsic parameters of the other cameras are computed as $f_j = f_1 + \Delta f_j; k1_j = k1_1 + \Delta k1_j; k2_j = k2_1 + \Delta k2_j$. Thus, given the camera rotations $\boldsymbol{R}$ and initial intrinsic parameters, we get the initial projection matrix of each common camera $\boldsymbol{P} = \{\boldsymbol{P}_{ij}\}$.

Then, the triangulation is performed and gross outliers are initially ignored. Let $\boldsymbol{X} = \{\boldsymbol{X}_h, h = 1, \ldots, H\}$ be the predicted 3D points, $\boldsymbol{x} = \{\boldsymbol{x}_{ijh}\}$ be the measured 2D image point locations. Our bundle adjustment is formulated as

$$\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{h=1}^{H} \delta_{ijh} \times \left(\gamma\left(\boldsymbol{P}_{ij}, \boldsymbol{X}_h\right) - \boldsymbol{x}_{ijh}\right)_{\text{huber}}. \quad (2)$$

$\gamma\left(\boldsymbol{P}_{ij}, \boldsymbol{X}_h\right)$ is the reprojection function which projects the predicted 3D point to its visible images; $\delta_{ijh} = 1$ if $\boldsymbol{X}_h$ is visible to the jth common camera in the ith generalized camera, otherwise set to 0. In this work, since the initial camera models are noisy, the triangulation and bundle adjustment are carried out iteratively to make more real track inliers into the bundle adjustment. For the efficiency concern, the bundle adjustment is terminated when the number of track inliers between two consecutive iterations is unchanged. We find that iteration times is always less than 5, hence the time-cost of bundle adjustment is acceptable for our method.

Experiments. We perform our SfM method on two typical kinds of images: (1) street view images, including datasets SV1(1504 images), SV2(3270 images) and SV3(2468 images); (2) oblique airborne images OBL(3720 images), which has ground-truth camera centers transformations. We compare our method with both a state-of-the-art incremental SfM approach, Bundler [6], and a recent representative global SfM approach, Cui et al. [2]. For our datasets, the ratio of the number of the parameters in our method to those in the other two methods is respectively 16.9%, 11.3%, 16.8% and 13.5%. As a result, our global method has a better potential in efficiency and scalability.

For our datasets, we find that there are always some uncalibrated images left by Bundler [6], and our method is about three times faster than the global SfM method Cui et al. [2] on SV1 and SV3, while about four times faster on SV2 and OBL. For the calibration accuracy, our result on OBL has a median error of 0.061 m, which is much smaller than 5.20 m in Bundler [6] and 4.16 m in Cui et al. [2].

Figure 1(b) shows an example of reconstructed results on SV3 produced by Cui et al. [2] and our method. For the area marked by squares, our calibrated camera poses and reconstructed scene are apparently more reasonable. More scene reconstruction and comparison results on SV1, SV2 and OBL are showed in the supplementary file.

Conclusion. In this article, we fuse the generalized camera model into a global calibration pipeline. In particular, the global rotation averaging problem for the generalized camera is solved in an iterative way, and the scene reconstruction problem for the generalized camera is tackled by a few alternations of triangulation and bundle adjustment. Extensive experiments demonstrated our SfM system are more efficient, accurate and scalable than the two state-of-the-art SfM approaches, especially in the large-scale scene reconstruction applications.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant Nos. 61333015, 61473292).

Supporting information The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Yin C T, Zhang X, Hui C, et al. A literature survey on smart cities. Sci China Inf Sci, 2015, 58: 100102
2 Cui H N, Shen S H, Gao W, et al. Efficient large-scale structure from motion by fusing auxiliary imaging information. IEEE Trans Image Process, 2015, 22: 3561–3573
3 Klingner B, Martin D, Roseborough J. Street view motion–from–structure–from–motion. In: Proceedings of IEEE International Conference on Computer Vision (ICCV), Sydeney, 2013. 953–960
4 Sweeney C, Fragoso V, Höllerer T, et al. gDLS: a scalable solution to the generalized pose and scale problem. In: Proceedings of European Conference on Computer Vision (ECCV). Berlin: Springer, 2014. 16–31
5 Torii A, Havlena M, Pajdla T. From google street view to 3D city models. In: Proceedings of IEEE 12th International Conference on Computer Vision Workshops (ICCV Workshops), 2009. 2188–2195
6 Noah S, Steven S M, Richard S. Modeling the world from Internet photo collections. Int J Comput Vision, 2008, 80: 189–210