Large-Scale Structure from Motion with Semantic Constraints of Aerial Images

Yu Chen[1], Yao Wang[1], Peng Lu[2], Yisong Chen[1], and Guoping Wang[1](✉)

[1] GIL, Department of Computer Science and Technology, Peking University, Beijing, China
{1701213988,yaowang95,yisongchen,wgp}@pku.edu.cn
[2] School of Computer Science, Beijing University of Posts and Telecommunications, Beijing, China
lupeng@bupt.edu.cn

Abstract. Structure from Motion (SfM) and semantic segmentation are two branches of computer vision. However, few previous methods integrate the two branches together. SfM is limited by the precision of traditional feature detecting method, especially in complicated scenes. As the research field of semantic segmentation thrives, we could gain semantic information of high confidence in each specific task with little effort. By utilizing semantic segmentation information, our paper presents a new way to boost the accuracy of feature point matching. Besides, with the semantic constraints taken from the result of semantic segmentation, a new bundle adjustment method with equality constraint is proposed. By exploring the sparsity of equality constraint, it indicates that constrained bundle adjustment can be solved by Sequential Quadratic Programming (SQP) efficiently. The proposed approach achieves state of the art accuracy, and, by grouping the descriptors together by their semantic labels, the speed of putative matches is slightly boosted. Moreover, our approach demonstrates a potential of automatic labeling of semantic segmentation. In a nutshell, our work strongly verifies that SfM and semantic segmentation benefit from each other.

Keywords: Structure from Motion · Semantic segmentation
Equality bundle adjustment · Sequential Quadratic Programming

1 Introduction

Structure from Motion (SfM) has been a popular topic in 3D vision in recent two decades. Inspired by the success of Photo Tourism [1] in dealing with a myriad amount of unordered Internet images, respectable methods are proposed to improve the efficiency and robustness of SfM.

Incremental SfM approaches [1–7] start by selecting seed image pairs that satisfy two constraints: wide baseline and sufficient correspondences, then repeatedly register new cameras in an incremental manner until no any camera could

Y. Chen and Y. Wang—Contributed equally.

J.-H. Lai et al. (Eds.): PRCV 2018, LNCS 11256, pp. 347–359, 2018.
https://doi.org/10.1007/978-3-030-03398-9_30

be added in the existing scene structure. This kind of method achieves high accuracy and is robust to bad matches thanks to the using of RANSAC [9] in several steps to filter outliers, but suffers from drift in large-scale scene structures due to the accumulated errors. In addition, incremental SfM is not efficient for the repeated bundle adjustment [10].

Global SfM approaches [11,12] estimate poses of all cameras by rotation averaging and translation averaging and perform bundle adjustment just one time. However, Global SfM approaches are sensitive to outliers thus are not as accurate as incremental approaches.

Far more different from incremental SfM and global SfM approaches, hierarchical SfM methods [13–16] start from two-view reconstructions, and then merge into one by finding similarity transformation in a bottom-up manner.

While a vast of efforts are taken to improve the accuracy of SfM, most SfM approaches are affected greatly by the matching results. The success of incremental SfM is mainly due to the elimination of wrong matches in several steps, such as geometric verification, camera register and repeatedly bundle adjustment. Owing to executing only one bundle adjustment, global SfM is more easily affected by outliers. Thus how to filter outliers out still be a key problem in global SfM.

Recently, more and more works concentrate on semantic reconstruction [17, 18]. They cast semantic SfM as a maximum-likelihood problem, thus geometry and semantic information are simultaneously estimated. So far, semantic 3D reconstruction methods have been limited to small scenes and low resolution, because of their large memory and computational cost requirements. Different from that, our works aim at large scale 3D reconstruction from UAV images.

From our perspective, the state-of-the-art SfM methods still have insufficient geometric/physical constraints. Semantic information is considered as additional constraints for robust SfM process to enhance its accuracy and efficiency. Our contributions are mainly two folds: (1) we propose to fuse the semantic information into feature points by semantic segmentation (2) we formulate the problem of bundle adjustment with equality constraints and solve it efficiently by Sequential Quadratic Programming (SQP).

Our work expedite the cross field of Structure from Motion and semantic segmentation. Also, to the best of our knowledge, our work achieve state-of-the-art in both efficiency and accuracy.

2 Related Work

2.1 Structure from Motion

With the born of Photo Tourism [1], incremental SfM methods are proposed to deal with large scale scene structures. Though many efforts (Bundler [3], VisualSfM [5], OpenMVG [6], Colmap [7], Theia [8]) are taken, drift and efficiency are still the two main limitations of incremental SfM. Besides, the most 2 time consuming parts of reconstruction are feature matching and repeated bundle adjustment [10].

As mentioned in Multi-View Stereo [19], the integration of semantic information will be a future work for 3D reconstruction. Recently, it appears more and more works about semantic reconstruction. As the first work of semantic SfM is based on geometric constrains [17], the later work [18] takes advantage of both geometric and semantic information. Moreover, they [17,18] deem scene structure as not merely points, but also regions and objects. The camera poses can be estimated more robustly.

Haene et al. [20] propose a mathematical framework to solve the joint segmentation and dense reconstruction problem. In their work, image segmentation and 3D dense reconstruction benefit from each other. The semantic class of the geometry provides information about the likelihood of the surface direction, while the surface direction gives clue to the likelihood of the semantic class. Blaha et al. [21] raise an adaptive multi-resolution approach of dense semantic 3D reconstruction, which mainly focuses on the high requirement of memory and computation resource issue.

2.2 Outdoor Datasets

Street View Dataset. The street view datasets [22,23] are generally captured by cameras fixed on vehicles. The annotations of street views are ample, usually from 12 to 30 classes [22,24]. Since it provides detailed elevation information and lacks roof information, it is essential to fuse it with aerial or satellite datasets in the 3D reconstruction task.

Drone Dataset. The drone datasets [25,26] are mostly annotated for object tracking tasks. There are no public pixel-level annotated datasets.

Remote Sensing Dataset. The remote sensing datasets [27,28], like its name implies, is collected from a far distance, usually by aircraft or satellite. It is so far away from the earth that, the camera view is almost vertical to the ground. It is short of elevation information. In addition, the resolution of the remote sensing image is always unsatisfying.

In a nutshell, constructing a drone dataset with refined semantic annotation is critical to get semantic point cloud for large-scale outdoor scenes.

3 Semantic Structure from Motion

3.1 Semantic Feature Generation

In 3D reconstruction tasks, SIFT [29] is widely adopted to extract feature points. For each feature point, there is a 2-dimensional coordinate representation and a corresponding descriptor. After extracting the feature points and computing the descriptors, exhaustive feature matching is then performed to get putative matches. While the SIFT features are robust to the variation of scale, rotation, and illumination, more robust features are required to produce more accurate

models. The traditional hand-crafted geometric features are limited in complicated aerial scenes. Intuitively, we can take semantic information into consideration to get more robust feature points.

Semantic Label Extraction. Inspired by [30], which deals with the problem of drift of monocular visual simultaneous localization and mapping, uses a CNN to assign each pixel x to a probability vector P_x, and the $(i^t)^h$ components of P_x is the probability that x belongs to class i. By taking the result of semantic segmentation of original images, the process of scene labeling [30] is replaced to avoid a time-consuming prediction. Since we already get its coordinate in the raw image, the semantic label can be easily searched in the corresponding semantic segmentation image. Then each feature point has two main information: 2-dimensional coordinate, and semantic label.

Grouped Feature Matching. Though wrong matches are filtered by geometric verification, some still exist due to the complication of scenes. It suggests that epipolar geometry is not strong enough to provide sufficient constraints. We could apply the semantic label for additional constraints in feature matching. The candidate matches of Brute-Force matching method may not have the same semantic label (a feature point indicates road may match to a building, e.g.). As we annotate the images into three categories, we can simply cluster the feature points into three semantic groups. Performing matches only in each group could eliminate the semantic ambiguity.

To reconstruct the semantic point clouds, 2D semantic labels should be transmitted to 3D points. After performing triangulation, the 2D semantic label is assigned to the triangulated 3D point accordingly.

(a) Auditorium (b) Water (c) Tower (d) Pitch

(e) Road (f) Bungalow (g) Building complex

Fig. 1. Example images from UDD. (a)–(g) are typical scenes in drone images. Best viewed in color.

3.2 Equality Constrained Bundle Adjustment

As mentioned in Sect. 3.1, each 3D feature has a semantic label. Then we seek approaches to optimize the structures and camera poses further.

Review the unconstrained bundle adjustment equation below:

$$min\ \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m}\|x_{ij}-P_i(X_j)\|^2 \tag{1}$$

where n is the number of cameras, m is the number of 3D points, and x_{ij} is the 2D feature points, X_j is the 3D points, P_i is the nonlinear transformations of 3D points.

While Eq. (1) minimizes the re-projection error of 3D points, due to the existence of some bad points, an additional weighting matrix W_e should be introduced. As a result, the selection of W_e affects the accuracy of the final 3D model, and the re-projected 2D points may be located at some wrong places (For example, a 3D building point corresponds to a 2D tree point). Intuitively, we can force the 3D points and the re-projected 2D points satisfy some constraints, that is *Semantic Consistency*, which means the 3D points and re-projected 2D points have the same semantic label.

Different with traditional bundle adjustment, with additional semantic constraints, we modify the bundle adjustment as an equality constrained nonlinear least square problem. Take semantic information from features, we can rewrite Eq. (1) as follows:

$$min\ \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m}\|x_{ij}-P_i(X_j)\|^2, s.t.\ L(x_{ij})=L(P_i(X_j)) \tag{2}$$

where L represents the semantic label of observations.

Then we show how to transform Eq. (2) into a Sequential Quadratic Programming problem. Let $f(x)$ be a nonlinear least square function that need to be optimized, $c(x)=L(x_{ij})-L(P_i(X_j))=0$ be the equality constraints, A be the Jacobian matrix of the constraints, then the Lagrangian function for this problem is $F(x,\lambda)=f(x)-\lambda^T c(x)$. By the first order KKT condition, we can get:

$$\nabla F(x,\lambda)=\begin{bmatrix}\nabla f(x)-A^T\lambda\\ -c(x)\end{bmatrix}=0 \tag{3}$$

Let W denotes the Hessian of $F(x,\lambda)$, we can get:

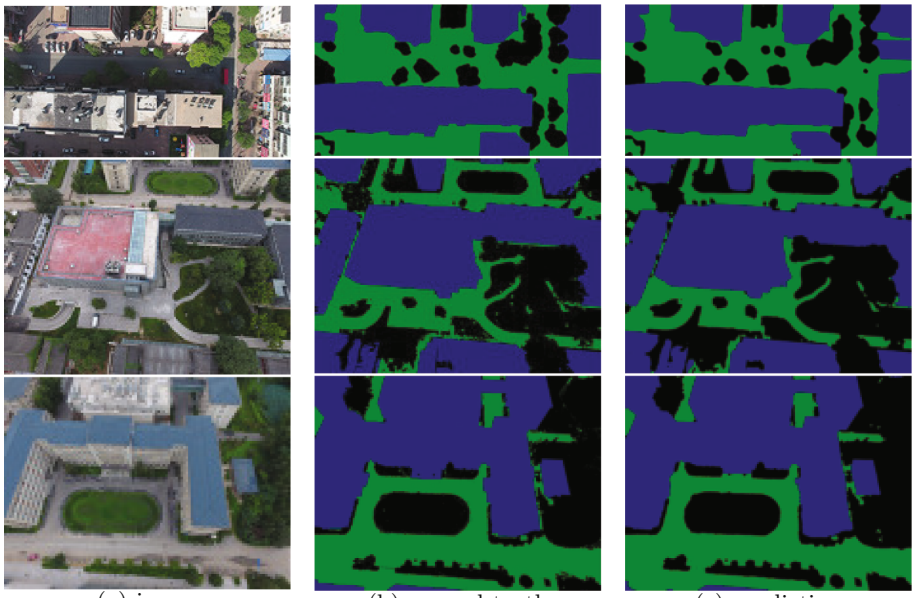
$$\begin{bmatrix}W & -A^T\\ -A & 0\end{bmatrix}\begin{bmatrix}\delta x\\ \lambda_k\end{bmatrix}=\begin{bmatrix}-\nabla f+A^T\lambda_k\\ c\end{bmatrix} \tag{4}$$

By subtracting $A^T\lambda$ from both side of the first equation in Eq. (4), we then obtain:

$$\begin{bmatrix}W & -A^T\\ -A & 0\end{bmatrix}\begin{bmatrix}\delta x\\ \lambda_{k+1}\end{bmatrix}=\begin{bmatrix}-\nabla f\\ c\end{bmatrix} \tag{5}$$

Equation (5) can be efficiently solved when both W and A are sparse. It is also easy to prove that W and A are all sparse in unconstrained bundle adjustment problem by the Levenburg-Marquart method.

Then the original constrained bundle adjustment problem is formulated to an unconstrained problem, and we seek approaches to solve the linear equation set $Ax = b$. Since A is symmetric indefinite, LDL^T factorization can be used. Besides, to avoid the computation of Hessian, we replace W with reduced Hessian of Lagrangian.



(a) image (b) ground truth (c) prediction

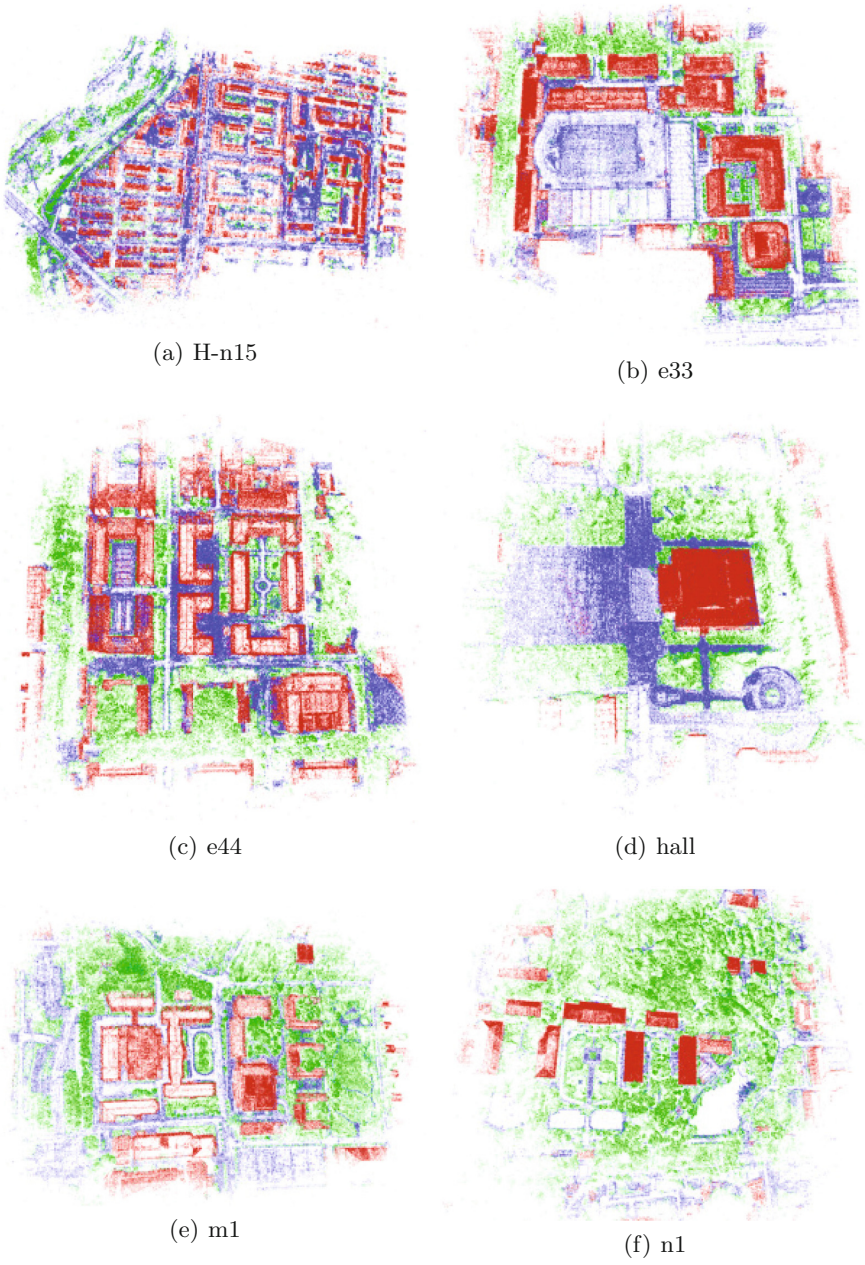
Fig. 2. Visualization of Urban Drone Dataset (UDD) validation set. **Blue**: Building, **Black**: Vegetation, **Green**: Free space. Best viewed in color. (Color figure online)

4 Experiments

4.1 Dataset Construction

Our dataset, Urban Drone Dataset (UDD)[1], is collected by a professional-grade UAV (DJI-Phantom 4) at altitudes between 60 and 100 m. It is extracted from 10 video sequences taken in 4 different cities in China. The resolution is either 4k (4096 * 2160) or 12M (4000 * 3000). It contains a variety of urban scenes (see Fig. 1). For most 3d reconstruction tasks, 3 semantic classes are roughly enough [31]: Vegetation, Building, and Free space [32]. The annotation sampling rate is between 1% to 2%. The train set consists of 160 frames, and the validation set consists of 45 images.

[1] https://github.com/MarcWong/UDD.



(a) H-n15 (b) e33

(c) e44 (d) hall

(e) m1 (f) n1

Fig. 3. Semantic reconstruction results with our constrained bundle adjustment. **Red**: Building, **Green**: Vegetation, **Blue**: Free space. Best viewed in color. (Color figure online)

Table 1. Statistics of reconstruction results of original and semantic SfM. **Black**: Original value/unchanged value compared to the original SfM, **Green**: Better than the original SfM, **Red**: Worse than the original SfM.

	Dataset	Images	Poses	Points	Tracks	RMSE	Time
Original SfM	cangzhou	400	400	1,287,539	2,541,961	0.819215	16 h 49 min 23 s
	e33	392	392	559,065	810,390	0.565699	3 h 28 min 43 s
	e44	337	337	468,978	641,171	0.546114	3 h 17 min 16 s
	hall	195	195	476,853	760,769	0.536045	2 h 10 min 39 s
	m1	288	288	422,158	650,072	0.564724	2 h 32 min 10 s
	n1	350	350	479,813	622,243	0.471467	4 h 7 min 21 s
	n15	248	244	484,229	667,029	0.529639	2 h 40 min 07 s
Semantic SfM	cangzhou	400	400	1,326,858	2,660,869	0.719897	14 h 28 min 51 s
	e33	392	392	554,449	803,395	0.561667	3 h 21 min 29 s
	e44	337	337	469,371	635,279	0.538501	3 h 07 min 13 s
	hall	195	195	473,056	745,969	0.531877	2 h 05 min 39 s
	m1	288	288	420,044	644,405	0.560242	2 h 30 min 49 s
	n1	350	350	481,983	617,487	0.466910	4 h 16 min 02 s
	n15	248	248	484,915	647,101	0.520202	2 h 37 min 10 s

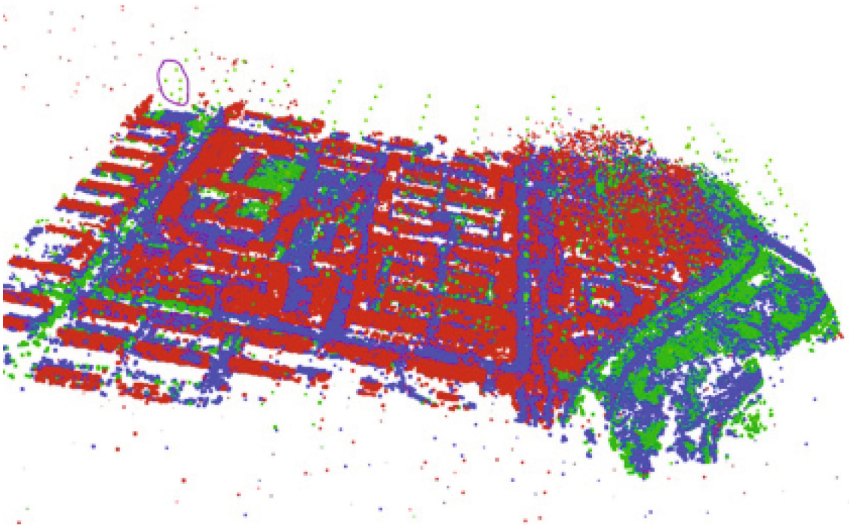
4.2 Experiment Pipeline

For each picture, we predict the semantic labels first. Our backbone network ResNet-101 [33] is pre-trained on ImageNet [34]. We employ the main structure of deeplab v2 [35] and fine-tune it on UDD. The training is conducted on single GPU Titan X Pascal, with tensorflow 1.4. The fine-tuning is 10 epochs in total, with crop size of 513 * 513, and Adam optimizer (momentum 0.99, learning rate 2.5e−4, and weight decay 2e−4). The prediction result is depicted in Fig. 2.

Then, SfM with semantic constraints is performed. For reconstruction experiments that without semantic constraints, we just perform a common incremental pipeline as described in [6], and referred as *original SfM*. Our approach refers to *Semantic SfM* in this article. All the experiments statistics are given in Table 1, and the reconstruction results are depicted in Fig. 3.

4.3 Reconstruction Results

Implementation Details. We adopt SIFT [29] to extract feature points and compute descriptors. After extracting feature points, we predict their semantic label according to views and locations. For feature matching, we use cascade hashing [36] which is faster than FLANN [37]. After triangulation, each semantic label of a 2D feature is assigned to a computed 3D point, and every 3D point has a semantic label. Constrained bundle adjustment is realized by the algorithm given in Sect. 3.2. All of our experiments perform on a single computer and an Intel Core i7 CPU with 12 threads.



(a) semantic reconstruction result of dataset H-n15

(b) original reconstruction result of dataset H-n15

Fig. 4. Results of dataset H-n15. We can see from the left-up corner of (a) and (b), our semantic SfM can recover more camera poses than original SfM. Best viewed in color. (Color figure online)

Efficiency Evaluation. As shown in Table 1, our semantic SfM is slightly faster than original SfM. It's quite important, because as the additional constraints are added, the large-scale SQP problem may not always be solved efficiently in practice. In datasets of e44 and n1, however, the time spent by original SfM is

much higher than expected, it may be caused by other usages of CPU resources when running the program, so we marked it out by red color.

Accuracy Evaluation. For most of the datasets, original SfM and our semantic SfM can recover the same number of camera poses. But in the n15 dataset, our method recovers all of the camera poses while the original SfM misses 4 camera poses. Detailed result is depicted in Fig. 4. As there are more than 200 hundred cameras, we just circled one part for demonstration. Besides, the number of 3D points reconstructed by our semantic SfM reduced slightly in m1, e33 and hall datasets, but in cangzhou, e44, n1 and n15 dataset, the number of points increased. Though the number of tracks decreased in most of our datasets. We use the Root Mean Square Error (RMSE) of reprojection as the evaluation. The RMSE of our semantic SfM is less than the original SfM in all of the datasets. Especially in cangzhou, a much more complicated dataset, the accuracy of RMSE has improved by almost 0.1, which suggests the accuracy of our semantic SfM surpasses original SfM, and our semantic SfM has advantages over the original one in complicated aerial image datasets.

5 Conclusion

As mentioned above, we propose a new approach for large-scale aerial images reconstruction by adding semantic constraints to Structure from Motion. By assigning each feature point a corresponded semantic label, matching is accelerated and some wrong matches are avoided. Besides, since each 3D point has a semantic constraint, nonlinear least square with equality constraints is used to model the bundle adjustment problem, and our result shows it could achieve the state-of-the-art precision while remaining the same efficiency.

Future Work. Not only should we consider the semantic segmentation as additional constraints in reconstruction, but to seek approaches taken the semantic label as variables to be optimized. What's more, with the rise of deep learning, and some representation works on learning feature [38], we would seek approaches to extract features with semantic information directly. With our approaches proposed in this article, we could further generate a dense reconstruction, which leads to automatic semantic segmentation training data generation.

Acknowledgements. This work is supported by The National Key Technology Research and Development Program of China under Grants 2017YFB1002705 and 2017YFB1002601, National Natural Science Foundation of China (NSFC) under Grants 61472010, 61632003, 61631001, and 61661146002, Equipment Development Project under Grant 315050501, and Science and Technology on Complex Electronic System Simulation Laboratory under Grant DXZT-JC-ZZ-2015-019.

References

1. Seitz, S.M., Szeliski, R., Snavely, N.: Photo tourism: exploring photo collections in 3D. ACM Trans. Graph. **25**(3), 835–846 (2006)
2. Agarwal, S., Snavely, N., Simon, I.: Building Rome in a day. Commun. ACM **54**(10), 105–112 (2011)
3. Snavely, K.N.: Scene Reconstruction and Visualization from Internet Photo Collections. University of Washington (2008)
4. Frahm, J.-M., et al.: Building Rome on a cloudless day. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6314, pp. 368–381. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15561-1_27
5. Wu, C.: Towards linear-time incremental structure from motion. In: International Conference on 3DTV-Conference. IEEE, pp. 127–134 (2013)
6. Moulon, P., Monasse, P., Marlet, R.: Adaptive structure from motion with *a Contrario* model estimation. In: Lee, K.M., Matsushita, Y., Rehg, J.M., Hu, Z. (eds.) ACCV 2012. LNCS, vol. 7727, pp. 257–270. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-37447-0_20
7. Schönberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Computer Vision and Pattern Recognition. IEEE (2016)
8. Sweeney, C., Hollerer, T., Turk, M.: Theia: a fast and scalable structure-from-motion library, pp. 693–696 (2015)
9. Fischler, M.A., Bolles, R.C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Read. Comput. Vis. **24**(6), 726–740 (1987)
10. Triggs, B., McLauchlan, P.F., Hartley, R.I., Fitzgibbon, A.W.: Bundle adjustment — a modern synthesis. In: Triggs, B., Zisserman, A., Szeliski, R. (eds.) IWVA 1999. LNCS, vol. 1883, pp. 298–372. Springer, Heidelberg (2000). https://doi.org/10.1007/3-540-44480-7_21
11. Wilson, K., Snavely, N.: Robust global translations with 1DSfM. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8691, pp. 61–75. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10578-9_5
12. Crandall, D., Owens, A., Snavely, N., et al.: Discrete-continuous optimization for large-scale structure from motion. In: Computer Vision and Pattern Recognition, pp. 3001–3008. IEEE (2011)
13. Farenzena, M., Fusiello, A., Gherardi, R.: Structure-and-motion pipeline on a hierarchical cluster tree. In: IEEE International Conference on Computer Vision Workshops. IEEE, 1489–1496 (2009)
14. Gherardi, R., Farenzena, M., Fusiello, A.: Improving the efficiency of hierarchical structure-and-motion. In: Computer Vision and Pattern Recognition, pp. 1594–1600. IEEE (2010)
15. Toldo, R., Gherardi, R., Farenzena, M., et al.: Hierarchical structure-and-motion recovery from uncalibrated images. Comput. Vis. Image Underst. **140**(C), 27–143 (2015)
16. Chen, Y., Chan, A.B., Lin, Z., et al.: Efficient tree-structured SfM by RANSAC generalized Procrustes analysis. Comput. Vis. Image Underst. **157**(C), 179–189 (2017)
17. Bao, S.Y., Savarese, S.: Semantic structure from motion. In: Computer Vision and Pattern Recognition, pp. 2025–2032. IEEE (2011)
18. Bao, S.Y., Bagra, M., Chao, Y.W.: Semantic structure from motion with points, regions, and objects. IEEE **157**(10), 2703–2710 (2012)

19. Furukawa, Y.: Multi-View Stereo: A Tutorial. Now Publishers Inc., Hanover (2015)
20. Haene, C., Zach, C., Cohen, A.: Dense semantic 3D reconstruction. IEEE Trans. Pattern Anal. Mach. Intell. **39**(9), 1730–1743 (2016)
21. Blaha, M., Vogel, C., Richard, A., et al.: Large-scale semantic 3D reconstruction: an adaptive multi-resolution model for multi-class volumetric labeling. In: Computer Vision and Pattern Recognition, pp. 3176–3184. IEEE (2016)
22. Cordts, M., Omran, M., Ramos, S., et al.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3213–3223 (2016)
23. Sturm, J., Engelhard, N., Endres, F., et al.: A benchmark for the evaluation of RGB-D SLAM systems. In: 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 573–580. IEEE (2012)
24. Brostow, G.J., Shotton, J., Fauqueur, J., Cipolla, R.: Segmentation and recognition using structure from motion point clouds. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008. LNCS, vol. 5302, pp. 44–57. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-88682-2_5
25. Mueller, M., Smith, N., Ghanem, B.: A benchmark and simulator for UAV tracking. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 445–461. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_27
26. Robicquet, A., Alahi, A., Sadeghian, A., et al.: Forecasting social navigation in crowded complex scenes. arXiv preprint arXiv:1601.00998 (2016)
27. Maggiori, E., Tarabalka, Y., Charpiat, G., et al.: Can semantic labeling methods generalize to any city? The inria aerial image labeling benchmark. IEEE International Symposium on Geoscience and Remote Sensing (IGARSS) (2017)
28. Xia, G.S., Bai, X., Ding, J., et al.: DOTA: a large-scale dataset for object detection in aerial images. In: Proceedings of CVPR (2018)
29. Lowe, D.G., Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**(2), 91–110 (2004)
30. Salehi, A., Gay-Bellile, V., Bourgeois, S., Chausse, F.: Improving constrained bundle adjustment through semantic scene labeling. In: Hua, G., Jégou, H. (eds.) ECCV 2016. LNCS, vol. 9915, pp. 133–142. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-49409-8_13
31. Savinov, N., Ladicky, L., Hane, C., et al.: Discrete optimization of ray potentials for semantic 3d reconstruction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5511–5518 (2015)
32. Hne, C., Zach, C., Cohen, A., et al.: Joint 3D scene reconstruction and class segmentation. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 97–104. IEEE (2013)
33. He, K., Zhang, X., Ren, S., et al.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
34. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
35. Chen, L.C., Papandreou, G., Kokkinos, I.: DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFS. IEEE Trans. Pattern Anal. Mach. Intell. **40**(4), 834–848 (2018)
36. Cheng, J., Leng, C., Wu, J., et al.: Fast and accurate image matching with cascade hashing for 3D reconstruction. In: Computer Vision and Pattern Recognition, pp. 1–8. IEEE (2014)

37. Muja, M.: Fast approximate nearest neighbors with automatic algorithm configuration. In: International Conference on Computer Vision Theory and Application VISSAPP, pp. 331–340 (2009)
38. Yi, K.M., Trulls, E., Lepetit, V., Fua, P.: LIFT: Learned Invariant Feature Transform. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9910, pp. 467–483. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46466-4_28