Scene Coordinate Regression with Point Clouds for RGB Camera Relocalization

Extended Abstract

Dehui Lin
DFKI, Saarbrücken, Germany
Dehui.Lin@dfki.de

Christian Müller
DFKI, Saarbrücken, Germany
Christian.Mueller@dfki.de

ABSTRACT

Outdoor camera relocalization from a single RGB image remains a challenging task in autonomous driving. State-of-the-art method is based on training a scene coordinate regression (SCoRe) neural network with a 3D mesh model [3]. In this work, we look into training the SCoRe network with 3D point clouds. Our results have shown strong evidence that 3D points optimized under multi-view constraints, such as epipolar constraints, reprojection errors, photometric consistency and global visibility, are effective for training the SCoRe network. This also inspires follow-up research on training with the above constraints, i.e. without explicit 3D models, to achieve a robust and generalized SCoRe network for outdoor relocalization.

1 INTRODUCTION

Camera relocalization is an essential component of visual-SLAM in autonomous driving and augmented reality [18, 22]. Camera relocalization to estimate 6DoF camera poses w.r.t. a known 3D scene is an ongoing research problem. Recent research works exploiting learning-based methods mainly revolve around direct pose regression [4, 13–15] and scene coordinate regression [1, 3, 5, 17, 24].

For outdoor camera relocalization *from a single RGB image*, state-of-the-art relocalization accuracy is achieved based on scene coordinate regression (SCoRe) network trained with a 3D mesh model [3, 5]. Such a 3D model could be time costly and impractical to obtain, especially for outdoor scenes with large scales. Instead of using 3D models, research works [3, 14, 17] have shown evidence of using the deep convolutional neural network (CNN) to learn outdoor scene geometry implicitly with a geometric constraint of single-view reprojection error.

To inspire further research on SCoRe network learning scene geometry implicitly with more constraints, we investigate in this work the performance of SCoRe network trained with point clouds, i.e. SfM point cloud [23, 25, 30] and PMVS point cloud [10]. These point clouds are reconstructed from only RGB images under optimization of constraints such as multi-view epipolar constraint, reprojection error, photometric consistency and global visibility. Our experiments are based on the outdoor scene dataset, i.e. Cambridge Landmarks [14, 15], and the state-of-the-art method for camera relocalization from a single RGB image, i.e. DSAC++ [3].

ACM Computer Science in Cars Symposium, 2019, Kaiserslautern, Germany

2 RELATED WORK

Camera relocalization methods such as keyframe-based [11, 12] or keypoint-based [16, 28] approaches utilize handcrafted image-level or feature-level descriptors, while the learning-based approaches avoid such explicit extraction and matching. With the capability of deep CNN, direct absolute pose regression methods are possible but they achieve coarse accuracy. On the other hand, SCoRe methods, utilizing forests [2, 6, 7, 24] or CNN [1, 3, 5, 9, 17], achieves significantly higher accuracies. Such methods predict dense 2D-3D correspondences that are subsequently fed into a robust RANSAC-based scheme which estimates the final camera pose. Other camera relocalization methods utilizing 3D-structure [8, 21] and other variants of CNN [20, 26, 27] also exist.

3 SCoRe WITH POINT CLOUDS

SfM point cloud and PMVS point cloud for an example scene from the Cambridge Landmarks dataset [14, 15] are shown in Figure 1, where we can see that both point clouds respect the geometric structure of the scene. Quantitatively from Table 1, PMVS point cloud is approximately $10 - 20$ times denser than SfM point clouds.

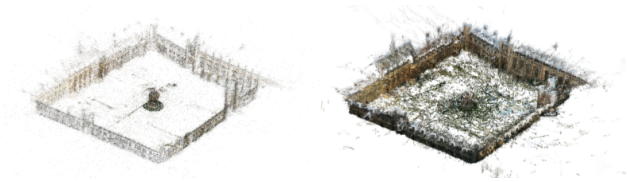

Figure 1: Point Clouds of the *Great Court* scene from Cambridge Landmarks [14, 15] generated with VisualSfM [29, 30]. Left: SfM point cloud. Right: PMVS point cloud.

Table 1: Details of Cambridge Landmarks dataset [14, 15] and the approximate number of reconstructed 3D points.

Scene	Spatial Extent	# Frames Train/Test	# SfM Points	# PMVS Points
Great Court	$95 \times 80m^2$	1532/760	0.21M	4.0M
King's College	$140 \times 40m^2$	1220/343	0.17M	3.5M
Old Hospital	$50 \times 40m^2$	895/182	0.11M	5.9M
Shop Facade	$35 \times 25m^2$	231/103	0.06M	1.2M
St Mary's Church	$80 \times 60m^2$	1487/530	0.42M	4.3M

*similar to other research work [3, 5], we ignored the *Street* scene

The sparsity of the point clouds is more distinguishable in the rendered depth images shown in Figure 2. Such rendered depth images are obtained by projecting the 3D points with the ground-truth pose and intrinsic parameters. When a pixel has multiple projected depths, we simply assign it with the smallest depth. Pixels without any projected depth are set to zero depth. The depth image from SfM point cloud is particularly sparse, with valid depths only in the strongest structural locations. On the other hand, the depth image from PMVS point cloud is denser, covering almost the whole structure. However, there are noisy or invalid depths in the large homogeneous regions. In general, the rendered depth image in DSAC++ [3] is more complete with a large percentage of valid depths, covering both the structures and the homogeneous areas.

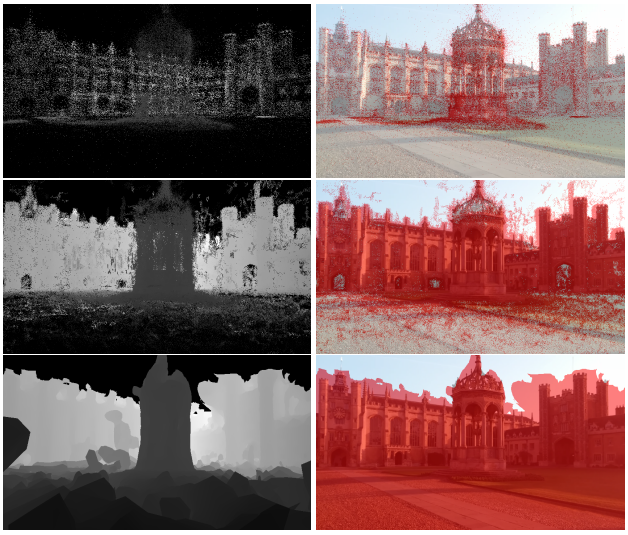

Figure 2: Left column: rendered depth images from the SfM point cloud, PMVS point cloud and 3D mesh model (provided by [3]). Right column: rendered depth images overlayed with RGB images.

4 RESULTS AND DISCUSSIONS

Mesh Model vs. Point Cloud. From Table 2, we can see that training the SCoRe network with point clouds achieves overall better relocalization results than training with a mesh model. Even though the point clouds are more sparse than the mesh model, as observed from the rendered depth images in Figure 2, their 3D points are sufficient and probably more accurate for training the SCoRe network. In other words, accurate ground-truth scene coordinates are more important than completeness for camera relocalization.

Recall that the 3D point clouds are optimized under the multi-view geometric, photoconsistent and visibility constraints. Intuitively, scene coordinate predictions from the SCoRe network could be optimized with such constraints for learning good scene geometry as well. This idea is also in agreement with the work by [17]. Given such positive sign from the results, further research work could be conducted to train the SCoRe network for outdoor relocalization with only constraints, i.e. no 3D models.

Table 2: Median 6D localization errors of camera pose estimation from RGB images using DSAC++ [3] pipeline. Best results are marked in bold.

Scene	Mesh Model [3]	SfM Point Cloud	PMVS Point Cloud
Great Court	0.40m, 0.2°	0.41m, 0.2°	**0.39m, 0.2°**
King's College	0.18m, 0.3°	**0.13m, 0.3°**	0.14m, 0.3°
Old Hospital	0.20m, 0.3°	0.21m, 0.4°	**0.19m, 0.3°**
Shop Facade	**0.06m, 0.3°**	**0.06m, 0.3°**	**0.06m, 0.3°**
St Mary's Church	0.13m, 0.4°	0.19m, 0.6°	**0.11m, 0.3°**

Table 3: Median 6D localization errors of SCoRe network trained with only Scene Coordinate Initialization step.

Scene	SfM Point Cloud	PMVS Point Cloud
Great Court	0.58m, 0.2°	0.66m, 0.3°
King's College	0.15m, 0.2°	0.15m, 0.3°
Old Hospital	0.21m, 0.4°	0.25m, 0.5°
Shop Facade	0.07m, 0.4°	0.06m, 0.3°
St Mary's Church	0.64m, 1.6°	0.17m, 0.5°

SfM vs. PMVS Point Cloud. From Table 3, training with PMVS point cloud achieves in general the best results with the lowest relocalization error. Even though there are noisy points in PMVS point clouds resulting in some noisy scene coordinate predictions from SCoRe network, the RANSAC-based scheme is able to reject such outliers effectively and estimates good camera poses. In addition, more data points (i.e. higher density in the rendered depth images) from the PMVS point cloud are beneficial to training the SCoRe network.

The relocalization accuracy of SfM point clouds in Table 2 are also remarkable as it is on par with the results of the PMVS point clouds, given the extent of sparsity of the SfM point cloud shown in Table 1. This indicates the possibility of training the SCoRe network with the sparse point clouds generated by visual-SLAM systems, such as ORB-SLAM2 [19]. In addition, this could imply that SCoRe network optimized under the multi-view epipolar constraints and reprojection errors, i.e. without photoconsistent constraint, could recover sufficiently good scene geometry for camera relocalization. This is further verified with the reasonable accuracies of SfM point cloud in Table 3, when we train the SCoRe network with only the initializtion step, i.e. no optimization of single-view reprojection and end-to-end optimization with ground-truth poses.

5 CONCLUSION

We have shown in this work the outstanding performance of using point clouds which are optimized under multi-view geometric, photoconsistent and visibility constraints to train the SCoRe network for outdoor relocalization from a single RGB image. In the next step of research, we would like to adapt such constraints to the optimization of scene coordinate predictions in SCoRe network so as to remove the need of explicit 3D models during training and

achieve a more robust and generalized SCoRe network for outdoor RGB camera relocalization.

REFERENCES

[1] Eric Brachmann, Alexander Krull, Sebastian Nowozin, Jamie Shotton, Frank Michel, Stefan Gumhold, and Carsten Rother. 2016. DSAC âĂŤ Differentiable RANSAC for Camera Localization. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), 2492–2500.

[2] Eric Brachmann, Frank Michel, Alexander Krull, Michael Ying Yang, Stefan Gumhold, and Carsten Rother. 2016. Uncertainty-Driven 6D Pose Estimation of Objects and Scenes from a Single RGB Image. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), 3364–3372.

[3] Eric Brachmann and Carsten Rother. 2017. Learning Less is More - 6D Camera Localization via 3D Surface Regression. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2017), 4654–4662.

[4] Samarth Brahmbhatt, Jinwei Gu, Kihwan Kim, James Hays, and Jan Kautz. 2017. Geometry-Aware Learning of Maps for Camera Localization. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2017), 2616–2625.

[5] Tommaso Cavallari, Luca Bertinetto, Jishnu Mukhoti, Philip H. S. Torr, and Stuart Golodetz. 2019. Let's Take This Online: Adapting Scene Coordinate Regression Network Predictions for Online RGB-D Camera Relocalisation. *ArXiv* abs/1906.08744 (2019).

[6] Tommaso Cavallari, Stuart Golodetz, Nicholas A. Lord, Julien P. C. Valentin, Luigi di Stefano, and Philip H. S. Torr. 2017. On-the-Fly Adaptation of Regression Forests for Online Camera Relocalisation. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), 218–227.

[7] Tommaso Cavallari, Stuart Golodetz, Nicholas A. Lord, Julien P. C. Valentin, Victor Adrian Prisacariu, Luigi di Stefano, and Philip H. S. Torr. 2018. Real-Time RGB-D Camera Pose Estimation in Novel Scenes using a Relocalisation Cascade. *IEEE transactions on pattern analysis and machine intelligence* (2018).

[8] Michael Donoser and Dieter Schmalstieg. 2014. Discriminative Feature-to-Point Matching in Image-Based Localization. *2014 IEEE Conference on Computer Vision and Pattern Recognition* (2014), 516–523.

[9] Nam-Duong Duong, Amine Kacete, Catherine Sodalie, Pierre-Yves Richard, and Jérôme Royan. 2018. xyzNet: Towards Machine Learning Camera Relocalization by Using a Scene Coordinate Prediction Network. *2018 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)* (2018), 258–263.

[10] Yasutaka Furukawa and Jean Ponce. 2007. Accurate, Dense, and Robust Multiview Stereopsis. *2007 IEEE Conference on Computer Vision and Pattern Recognition* (2007), 1–8.

[11] Andrew P. Gee and Walterio W. Mayol-Cuevas. 2012. 6D Relocalisation for RGBD Cameras Using Synthetic View Regression. In *BMVC*.

[12] Ben Glocker, Shahram Izadi, Jamie Shotton, and Antonio Criminisi. 2013. Real-time RGB-D camera relocalization. *2013 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)* (2013), 173–179.

[13] Alex Kendall and Roberto Cipolla. 2015. Modelling uncertainty in deep learning for camera relocalization. *2016 IEEE International Conference on Robotics and Automation (ICRA)* (2015), 4762–4769.

[14] Alex Kendall and Roberto Cipolla. 2017. Geometric Loss Functions for Camera Pose Regression with Deep Learning. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), 6555–6564.

[15] Alex Kendall, Matthew Koichi Grimes, and Roberto Cipolla. 2015. PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization. *2015 IEEE International Conference on Computer Vision (ICCV)* (2015), 2938–2946.

[16] Shuda Li and Andrew Calway. 2015. RGBD relocalisation using pairwise geometry and concise key point sets. *2015 IEEE International Conference on Robotics and Automation (ICRA)* (2015), 6374–6379.

[17] Xiaotian Li, Juha Ylioinas, Jakob Verbeek, and Juho Kannala. 2018. Scene Coordinate Regression with Angle-Based Reprojection Loss for Camera Relocalization. In *ECCV Workshops*.

[18] Stefan Milz, Georg Arbeiter, Christian Witt, Bassam Abdallah, and Senthil Yogamani. 2018. Visual SLAM for Automated Driving: Exploring the Applications of Deep Learning. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (2018), 360–370.

[19] Raul Mur-Artal and Juan D. Tardós. 2017. ORB-SLAM2: An Open-Source SLAM System for Monocular, Stereo, and RGB-D Cameras. *IEEE Transactions on Robotics* 33 (2017), 1255–1262.

[20] Noha Radwan, Abhinav Valada, and Wolfram Burgard. 2018. VLocNet++: Deep Multitask Learning for Semantic Visual Localization and Odometry. *IEEE Robotics and Automation Letters* 3 (2018), 4407–4414.

[21] Torsten Sattler, Bastian Leibe, and Leif Kobbelt. 2017. Efficient Effective Prioritized Matching for Large-Scale Image-Based Localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39 (2017), 1744–1756.

[22] Torsten Sattler, Will Maddern, Carl Toft, Akihiko Torii, Lars Hammarstrand, Erik Stenborg, Daniel Safari, Masatoshi Okutomi, Marc Pollefeys, Josef Sivic, Fredrik Kahl, and Tomás Pajdla. 2017. Benchmarking 6DOF Outdoor Visual Localization in Changing Conditions. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2017), 8601–8610.

[23] Johannes L. Schönberger and Jan-Michael Frahm. 2016. Structure-from-Motion Revisited. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), 4104–4113.

[24] Jamie Shotton, Ben Glocker, Christopher Zach, Shahram Izadi, Antonio Criminisi, and Andrew W. Fitzgibbon. 2013. Scene Coordinate Regression Forests for Camera Relocalization in RGB-D Images. *2013 IEEE Conference on Computer Vision and Pattern Recognition* (2013), 2930–2937.

[25] Noah Snavely, Steven M. Seitz, and Richard Szeliski. 2007. Modeling the World from Internet Photo Collections. *International Journal of Computer Vision* 80 (2007), 189–210.

[26] Abhinav Valada, Noha Radwan, and Wolfram Burgard. 2018. Deep Auxiliary Learning for Visual Localization and Odometry. *2018 IEEE International Conference on Robotics and Automation (ICRA)* (2018), 6939–6946.

[27] Florian Walch, Caner Hazirbas, Laura Leal-Taixé, Torsten Sattler, Sebastian Hilsenbeck, and Daniel Cremers. 2016. Image-based Localization with Spatial LSTMs. *ArXiv* abs/1611.07890 (2016).

[28] Brian Patrick Williams, Georg Klein, and Ian D. Reid. 2011. Automatic Relocalization and Loop Closing for Real-Time Monocular SLAM. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33 (2011), 1699–1712.

[29] Changchang Wu. 2011. VisualSFM: A Visual Structure from Motion System. http://ccwu.me/vsfm/

[30] Changchang Wu. 2013. Towards Linear-Time Incremental Structure from Motion. *2013 International Conference on 3D Vision* (2013), 127–134.