Automized Scene Layout Generation

Alexander Hanel
Karl Leiß
alexander.hanel@bit-ts.de
karl.leiss@bit-ts.de
BIT Technology Solutions GmbH
Pfaffing, Germany

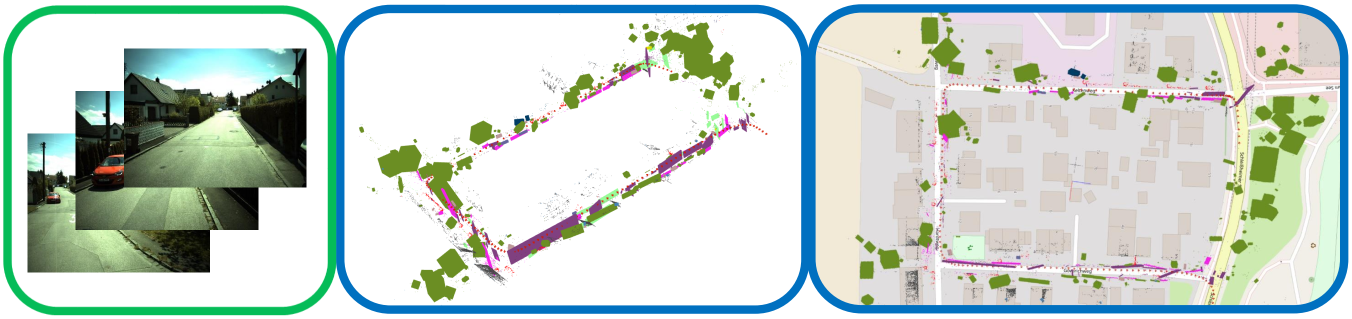

Figure 1: Structure from motion on a RGB image sequence (left) and the corresponding semantic image sequence (center) are used to obtain a semantic 3D point cloud. The scene layout (right; oblique view) is generated from the point cloud and represented by the street network and by oriented bounding boxes for scene objects of different classes (e.g. green for vegetation).

ABSTRACT

Automized scene layout generation gives the opportunity for saving time and scaling up variations in 3D city models. In this contribution, a method for combining structure from motion with semantic segmentation to obtain the street network and bounding boxes of scene objects as parts of the scene layout is proposed. Tests with an image sequence of a suburban scene show the potential to obtain the street network and bounding boxes of multiple objects belonging to different semantic classes.

KEYWORDS

Computer vision, computer graphics, semantic segmentation, structure from motion, 3D city models

ACM Reference Format:
Alexander Hanel and Karl Leiß. 2019. Automized Scene Layout Generation. In *Proceedings of ACM Computer Science in Cars Symposium.* ACM, New York, NY, USA, 3 pages.

1 INTRODUCTION

Synthetic images can be used to train deep networks for computer vision tasks like object detection [6], [9], [8] or semantic segmentation [11], [2], [1]. Especially for automotive applications with hard robustness requirements, synthetic images give the opportunity to obtain data from corner case situations like bad illumination or near accidents [14] and to use them for training. While synthetic images of artificial cities can be derived easily from computer games (e.g. [10], [5]), does creating synthetic images of real cities typically require a vast amount of careful manual modelling of the scene layout and the 3D scene objects like buildings or vegetation to obtain a realistic model of the city for rendering. In contrast to manual city modelling, do methods like structure from motion (e.g. [13]) or visual SLAM (e.g. [7]) allow to obtain a representation of the city as well, for example by a sparse 3D point cloud belonging to scene objects like buildings or vegetation.

Potential for saving time and for increasing the number of variations in training data by using a large number of scene layouts could be exploited by automatizing scene layout generation. One approach to achieve this goal could be combining a 3D point cloud from structure from motion with semantic information obtained from semantic segmentation (Figure 1 center) of an input image sequence (Figure 1 left) as basis for generating the scene layout. While the street network as one part of the scene layout is already available from the camera trajectory estimated by structure form motion, is the process for generating bounding boxes of scene objects as another part of the scene layout (Figure 1 right) more challenging.

2 METHODOLOGY

Main contribution of this paper is a method for automized generation of bounding boxes for scene objects by the aforementioned semantic-aided structure from motion. Goal of the proposed method is getting a first impression of the potential to automatize scene layout generation. The method consists of the steps (Figure 2) that are being explained in the following.

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

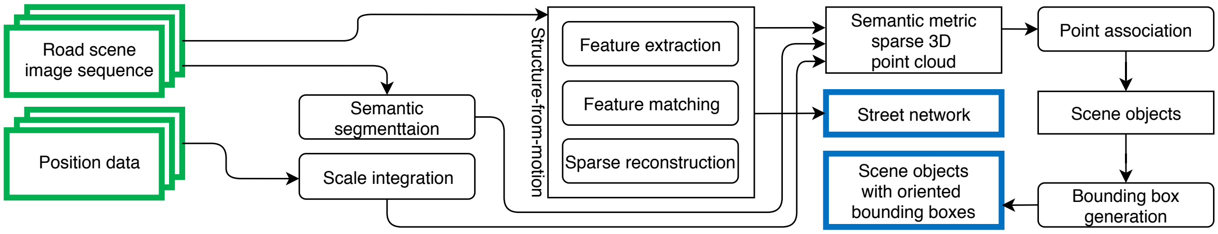

Figure 2: Workflow for automized scene layout generation represented by scene objects and the street network.

2.1 Structure from motion

A sparse point cloud as 3D reconstruction of the scene is obtained by structure from motion using an image sequence recorded with an environment-observing vehicle camera. Thereby, feature correspondences between image pairs are used for triangulating 3D points. An indirect, feature-based, 3D reconstruction method seems favorable over direct methods as the later ones tend to fail in the case of large motion between consecutive images [15] as they might occur on vehicles. Structure from motion allows to obtain a globally consistent point cloud and camera trajectory by more complex optimization [12] and is therefore preferred over visual SLAM or visual odometry. During 3D reconstruction, the same camera model is used for every image to ensure that the point cloud is Euclidean.

2.2 Semantic segmentation

Semantic information is obtained by semantic segmentation of each image of the input image sequence used for 3D reconstruction. Due to the recent success, a deep learning method is used for semantic segmentation. The model is trained on road scene images for typical semantic classes of road scenes, like vegetation, road or building. A semantic class is assigned to each 3D point based on a majority vote of the semantic classes of all related 2D feature points available in the semantic images.

2.3 Scale integration

In addition to an Euclidean coordinate system, metric scale of the scene layout is necessary for consistency when replacing scene object bounding boxes by specific 3D models. Scale integration is done by manually defining 3D coordinates of two camera positions, ideally the most remote camera positions to keep the scale error low. The 3D coordinates can be obtained from GPS data, for instance.

2.4 Point association

The 3D points of the point cloud have to be associated to scene objects for which the bounding boxes should be generated. First, all 3D points belonging to the same semantic class are selected and a seed point is randomly chosen from the selected points. Second, points nearby the seed point are selected based on a given Euclidean distance threshold. Previously selected nearby points are used as seeds for the next iteration of point selection until there are no more points within the distance threshold. Third, all the selected nearby points are assigned to a scene object, assuming that road scene objects can be distinguished by a distance criterion reliably. Forth, the last steps are repeated from step one on, points already assigned to an object are excluded.

2.5 Bounding box generation

An oriented 3D bounding box is generated for each scene object. Its position is defined by the center of gravity of the 3D points of the scene object. The orientation of the box is obtained from the Eigenvectors of a principal component analysis using these 3D points. The extension of the box is defined as smallest box enclosing these 3D points.

3 EXPERIMENTS

The proposed method is tested on an 150 image sequence recorded in a suburban area with an environment-observing forward-looking vehicle camera. COLMAP [13] is used for 3D reconstruction; pairwise extensive matching between all images is applied as a high number of 3D points is desired and no time requirements have to be fulfilled. Semantic segmentation is done with the Deeplabv3+ network [3] using a model trained by the same authors using the Cityscapes road scene image dataset [4]. 3D coordinates of the two most remote camera positions are approximated from aerial imagery and the position data provided by Google Earth.

4 RESULTS AND DISCUSSION

For the given image sequence, several scene objects have been obtained (Figure 1 right) especially for the semantic classes vegetation (green), terrain (light green), sidewalk (pink) and road (purple). Besides, the estimated camera trajectory covers all 150 images. Thereby, it has to be mentioned that currently only a very simple point association algorithm is used and only a single distance threshold was tested, leaving space for further research and optimization. Nevertheless, the aforementioned observations show the potential of automatizing the scene layout generation by means of a semantic-aided structure from motion method. In particular, the large number of vegetation objects that has been generated seems promising, as such objects can not be retrieved from maps in many cases.

5 CONCLUSION

This paper has shown the potential for combining semantic segmentation and structure from motion for automized scene layout generation. Further improvements can be expected by dense instead of sparse point clouds and by integrating additional computer vision tasks like single image depth estimation into the workflow.

REFERENCES

[1] V. Badrinarayanan, A. Kendall, and R. Cipolla. 2017. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39, 12 (Dec. 2017), 2481–2495. https://doi.org/10.1109/TPAMI.2016.2644615

[2] Liang-Chieh Chen, Jonathan T. Barron, George Papandreou, Kevin Murphy, and Alan L. Yuille. 2016. Semantic Image Segmentation with Task-Specific Edge Detection Using CNNs and a Discriminatively Trained Domain Transform. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. 4545–4554. https://doi.org/10.1109/CVPR.2016.492

[3] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. 2018. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In *Computer Vision – ECCV 2018*, Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (Eds.). Springer International Publishing, Cham, 833–851.

[4] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. 2016. The Cityscapes Dataset for Semantic Urban Scene Understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[5] Matthew Johnson-Roberson, Charles Barto, Rounak Mehta, Sharath Nittur Sridhar, Karl Rosaen, and Ram Vasudevan. 2016. Driving in the Matrix: Can virtual worlds replace human-generated annotations for real world tasks? *2017 IEEE International Conference on Robotics and Automation (ICRA)* (2016), 746–753.

[6] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott E. Reed, Cheng-Yang Fu, and Alexander C. Berg. 2016. SSD: Single Shot MultiBox Detector. In *European Conference on Computer Vision (ECCV) (Lecture Notes in Computer Science)*, Vol. 9905. Springer, 21–37.

[7] Raúl 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, 5 (2017), 1255–1262. https://doi.org/10.1109/TRO.2017.2705103

[8] J. Redmon and A. Farhadi. 2017. YOLO9000: Better, Faster, Stronger. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 6517–6525. https://doi.org/10.1109/CVPR.2017.690

[9] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2017. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* 39, 6 (June 2017), 1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031

[10] Stephan R. Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. 2016. Playing for Data: Ground Truth from Computer Games. In *European Conference on Computer Vision (ECCV) (LNCS)*, Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (Eds.), Vol. 9906. Springer International Publishing, 102–118.

[11] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi (Eds.). Springer International Publishing, Cham, 234–241.

[12] D. Scaramuzza and F. Fraundorfer. 2011. Visual Odometry [Tutorial]. *IEEE Robotics & Automation Magazine* 18, 4 (Dec. 2011), 80–92. https://doi.org/10.1109/MRA.2011.943233

[13] J. L. Schönberger and J. Frahm. 2016. Structure-from-Motion Revisited. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4104–4113. https://doi.org/10.1109/CVPR.2016.445

[14] Oliver Wasenmüller, Rene Schuster, Didier Stricker, Karl Leiss, Jürgen Pfister, Oleksandra Ganus, Julian Tatsch, Artem Savkin, and Nikolas Brasch. 2018. Automated Scene Flow Data Generation for Training and Verification. In *ACM Computer Science in Cars Symposium (CSCS). ACM Computer Science in Cars Symposium (CSCS-2018), Munich, Germany*. ACM.

[15] Georges Younes, Daniel Asmar, and John Zelek. 2019. FDMO: Feature Assisted Direct Monocular Odometry. In *14th International Conference on Computer Vision Theory and Applications*.