Car Detection Using RGB Image Geometry and Semantic Estimations

Yuanfang Wang, Yinghao Xu, Julian Gao

October 21, 2016

1 Introduction

Car detection has long been a popular topic in computer vision field. With the rise of industrial attention in autonomous driving and research focus on convolutional neural networks (CNN), car detection has seen rapid development recently. Early car detection in autonomous driving relies heavily on expensive devices, such as LIDAR, to sample depth and norm information. Recent works have tried to perform car detection based simply on camera captured images, and have reached considerably high accuracy on specialized datasets such as KITTI. Famous ones include fast R-CNN[], RPN[], etc. However those methods only make use of image data, and subject to problems such as scale variation, occlusion, and truncation[]. To overcome these deficiencies and achieve better accuracy, we propose a new method to incorporate LIDAR into on-board detection system, which shifts the costly part to offline. We want to use CNN to train an image-LIDAR model, that takes in an RGB image and outputs depth, norm, and semantic segmentation obtained from its LIDAR map. We then perform car detection as well as simple 3D reconstruction on these outputs.

2 Motivations

In recent publications, CNN has shown strong ability in both geometry and semantic scene understanding. A lot of work have been done in indoor scene depth and normal estimation using single RGB image and got pretty good results[1][2][3][4]. But research of outdoor scene is hindered by the getting reliable ground truth for training. The depth ground truth of outdoor scene could not be acquired by Kinect, which is able to obtain depth maps with identical sampling rate to those of color images. Two generally used depth measurements in outdoor environment is LiDAR and stereo. Stereo equipment is hard to get perfect calibration between multi cameras and the process of computing depth from raw input is really computational expensive. Also the accuracy range of stereo is limited compared with LiDAR. On the other hand, although LiDAR has farther measuring distance, its sampling rate is extremely sparse.Considering the strong need of understanding geometry information in the outdoor road scene for autonomous driving, we decide to do single image depth estimation in outdoor road environment.

3 Potential Problems & Approaches

We plan to use on-the-shelf single indoor image depth estimation neural network structure from [4], use both stereo and LiDAR captured depth information to generate ground truth for training. We expect to overcome the unsatisfactory depth ground truth problem and the algorithm could predict depth from single outdoor road scene color image.

Besides the geometry estimation, since the geometry information has many applications. If the depth estimation task could be solved on schedule, we also want to applicate one of the applications, car detection using geometry estimation. First reconstruct the 3D scene from our depth prediction, then compute the probability map in the 3D scene showing where cars might be, finally using some on-the-shelf methods to detect cars in our probability map.

4 Dataset

In this project, we choose to use the KITTI Benchmark [5][6][7]. It contains a large number of images with stereo and LiDAR sensing data which can be used in the depth estimation task. For the detection problem, KITTI also has specific problem section for training and testing.

References

[1] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, "Deeper depth prediction with fully convolutional residual networks," *arXiv preprint arXiv:1606.00373*, 2016.

[2] B. Li, C. Shen, Y. Dai, A. van den Hengel, and M. He, "Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1119–1127, 2015.

[3] C. Hane, L. Ladicky, and M. Pollefeys, "Direction matters: Depth estimation with a surface normal classifier," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 381–389, 2015.

[4] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2650–2658, 2015.

[5] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *International Journal of Robotics Research (IJRR)*, 2013.

[6] J. Fritsch, T. Kuehnl, and A. Geiger, "A new performance measure and evaluation benchmark for road detection algorithms," in *International Conference on Intelligent Transportation Systems (ITSC)*, 2013.

[7] M. Menze and A. Geiger, "Object scene flow for autonomous vehicles," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.