Taking advantage of sensor modality specific properties in Automated Driving

Extended Abstract

Christian Haase-Schuetz
Engineering Cognitive Systems - Automated Driving,
Chassis Systems Control, Robert Bosch GmbH, Institute of
Radio Frequency Engineering and Electronics, Karlsruhe
Institute of Technology
christian.schuetz2@partner.kit.edu

Heinz Hertlein
Engineering Cognitive Systems - Automated Driving,
Chassis Systems Control, Robert Bosch GmbH

ABSTRACT

Deep Learning methods are widely applicable in Robotics and Automated Driving scenarios. The task of perception for Automated Driving in the real world is particularly challenging and requires a sufficient amount of high quality labeled training data for the algorithms to perform well. While the detection of 2D bounding boxes is a well studied problem, only recently researchers are focusing more on 3D bounding box detection. In Automated Driving, a wide variety of sensors are used, such as Radar and Lidar; however, many current approaches focus on image based detections. In this position paper we want to emphasize the importance to study 3D perception with 3D labeling that is not based on image data alone, but ideally uses multi modal data for manual, semi automatic or fully automatic labeling.

CCS CONCEPTS

• **Computing methodologies → 3D imaging**; *Neural networks*;

KEYWORDS

Automated Driving, Multi Modal Datasets, Deep Learning

ACM Reference Format:
Christian Haase-Schuetz and Heinz Hertlein. 2018. Taking advantage of sensor modality specific properties in Automated Driving: Extended Abstract. In *Proceedings of ACM Computer Science in Cars Symposium (CSCS'18).* ACM, New York, NY, USA, 2 pages. https://doi.org/

1 INTRODUCTION

Recent progress in Advanced Driver Assistance Systems (ADAS) and sub-sequentially Automated Driving (AD) enables safer and more comfortable driving. Increasing levels of automation require more reliable and more advanced methods for perceiving information about the surroundings. Deep Learning based methods have shown superior recognition performance in various visual perception tasks in comparison to more traditional techniques based on hand-engineered features and classifiers, often by a large margin. Benchmark datasets such as [8, 11] play an important role in Computer Vision. Similarly in Automated Driving high-quality research datasets like [3] and [5] are well established to evaluate new perception approaches.

2 MULTIMODAL LABELS

Labeled datasets are a key enabler for the success of Deep Learning. In this paper we show emphasize that multi modal labeled datasets can further increase the 3D detection performance.

Current situation Cityscapes [3] and KITTI [5] are well established, high-quality datasets for AD. While Cityscapes is image-based and only 2D information is provided, KITTI contains Lidar measurements as well and 3D bounding boxes are given as labels. However the 3D annotations are given only in the camera frame as this modality was used as support for labeling. While the Lidar used in the measurement setup, as reported, covers 360°, the camera has a narrower field of view (FoV). As a result, only the Lidar point cloud data inside the camera's field of view, i.e. only a subset of the overall Lidar measurement data can be used as a labeled data set for training or evaluation of a Deep Neural Network. This not only reduces the usable size of the data set, but makes the labeled subset less representative, as the Lidar data outside the camera's FoV most likely includes other phenomena with a different distribution in comparison to the data inside the camera's FoV in front of the ego vehicle. Figure 1 illustrates the setup.

Other datasets like [1] provide Lidar measurements but no labels for dynamic objects at all.

Taking advantage of different modalities Different sensors have distinct characteristics, such as a dissimilar FoV, being able to measure certain target attributes but not others or with less accuracy (e.g. Radar measuring velocity, Lidar measuring accurate 3D locations etc.) or a different behavior in varying weather conditions. The power of fusing multiple sensor modalities in order to detect dynamic objects is that each sensor adds additional information to the overall result, based on its characteristics. Recent methods, published on the KITTI leaderboard, fusing information from camera and Lidar [2, 4, 7, 10, 12] outperform vision based approaches, in terms of 3D detection performance. This supports the argument that Lidar adds important information for the 3D detection task. Transferring labels from one modality to another is one possibility to obtain 3D labels. But classical fusion systems try to take advantage of different sensor properties regarding e.g. range or vulnerability to certain environmental conditions. If images are used for labeling the advantage of using different sensor modalities is partially lost. Furthermore in many scenarios it is relevant to detect objects 360° around the ego, which is not always covered by

CSCS'18, September 2018, Munich, Germany
2018.

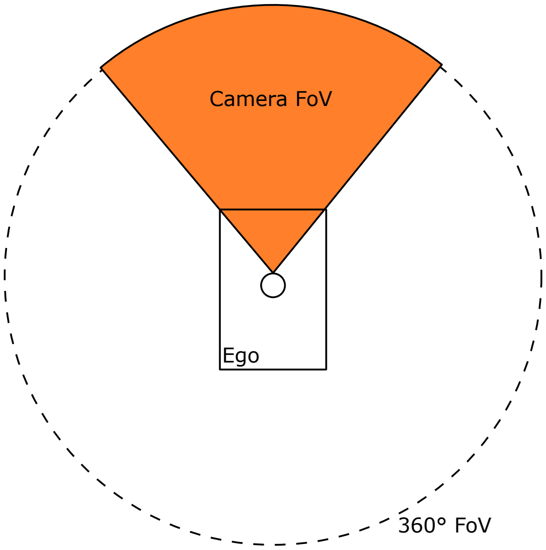

Figure 1: Camera and Lidar often have different FoV. Lidars commonly used, such as the one used in [5] cover 360°**. For cameras this is not the standard case. If Labels are generated using images only a subset of Lidar points will have appropriate labels.**

the FoV of the camera.
Covering 360° around the ego with cameras can be easily achieved by adding more cameras. Calibration of the different sensors needs to be ensured for this approach to work properly. E.g. [9] propose to use trained image segmentation DNNs and project the Lidar pointcloud to the image frame to automatically annotate the pointcloud. While [9] mentions the problem of different FoVs, different sensor range resolutions and varying recording conditions are not taken into account.
In order to have highly reliable and well performing Automated Driving functions it would be required to have per modality labels and algorithms taking advantage of the respective properties. This effect could be fostered by having multi modal labels, i.e. labels generated using multiple modalities, already containing diverse information from the different sensors.

Open challenges Manual labeling of measurements other than images is challenging, as humans are not as used to other modalities employed in AD as they are to vision. Further complexity arises from the presence of having an additional dimension for 3D pointcloud data and the sparse nature of such data.
Suitable tooling and visualization techniques e.g. for 2D/3D point cloud data are necessary to mitigate this difficulty and allow efficient handling of the sensor data. To aid human annotators camera images covering the FoV of the other sensor modalities should be given. As labeling sparse data is tedious and time consuming automated or semi-automated approaches are desirable to facilitate this process.
Recently [6] announced a competitive 3D dataset including high-quality Lidar data, it will be interesting to conduct investigations with this new data when it is finally published.

3 CONCLUSION

While current academic datasets already play a key role for the development of new algorithms for environment perception for AD even without taking full advantage of different sensor properties, we are convinced that making use of sensor specific information through manual, semi-automatic or fully automatic labeling of data sets will be crucial to further accelerate the progress in the field. Recent perception approaches evaluated on benchmark datasets suggest high potential when using multiple modalities. Especially multi modal techniques for labeling are expected to support the creation of labels in the quantity and with the accuracy that is required for deep learning. The resulting labeled multi modal data set can be utilized for training and evaluation of machine learning techniques for data from any single sensor modality contained in the data set as well as for multi modal fusion approaches. New algorithms taking advantage of the strengths of different sensors would bring the goal of highly reliable AD in real world situations closer.

REFERENCES

[1] José-Luis Blanco-Claraco, Francisco-Ángel Moreno-Dueñas, and Javier González-Jiménez. 2014. The Málaga urban dataset: High-rate stereo and LiDAR in a realistic urban scenario. *The International Journal of Robotics Research* 33, 2 (2014), 207–214.

[2] Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. 2017. Multi-view 3d object detection network for autonomous driving. In *IEEE CVPR*, Vol. 1. 3.

[3] 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)*.

[4] Xinxin Du, Marcelo H Ang Jr, Sertac Karaman, and Daniela Rus. 2018. A general pipeline for 3d detection of vehicles. *arXiv preprint arXiv:1803.00387* (2018).

[5] Andreas Geiger, Philip Lenz, and Raquel Urtasun. 2012. Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

[6] Xinyu Huang, Xinjing Cheng, Qichuan Geng, Binbin Cao, Dingfu Zhou, Peng Wang, Yuanqing Lin, and Ruigang Yang. 2018. The ApolloScape Dataset for Autonomous Driving. *arXiv: 1803.06184* (2018).

[7] Jason Ku, Melissa Mozifian, Jungwook Lee, Ali Harakeh, and Steven Waslander. 2017. Joint 3d proposal generation and object detection from view aggregation. *arXiv preprint arXiv:1712.02294* (2017).

[8] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.

[9] Florian Piewak, Peter Pinggera, Manuel Schäfer, David Peter, Beate Schwarz, Nick Schneider, David Pfeiffer, Markus Enzweiler, and Marius Zöllner. 2018. Boosting LiDAR-based Semantic Labeling by Cross-Modal Training Data Generation. *arXiv preprint arXiv:1804.09915* (2018).

[10] Charles R Qi, Wei Liu, Chenxia Wu, Hao Su, and Leonidas J Guibas. 2017. Frustum pointnets for 3d object detection from rgb-d data. *arXiv preprint arXiv:1711.08488* (2017).

[11] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)* 115, 3 (2015), 211–252. https://doi.org/10.1007/s11263-015-0816-y

[12] Danfei Xu, Dragomir Anguelov, and Ashesh Jain. 2017. Pointfusion: Deep sensor fusion for 3d bounding box estimation. *arXiv preprint arXiv:1711.10871* (2017).