Fusion of Sequential Information for Semantic Grid Map Estimation

Frank Bieder[1], Muti Ur Rehman[2], and Christoph Stiller[2]

[1] FZI Forschungszentrum Informatik, Mobile Perception Systems Department,
Haid-und-Neu-Straße 10-14, 76131 Karlsruhe
[2] Karlsruhe Institute of Technology, Measurement and Control Systems,
Engler-Bunte-Ring 21, 76131 Karlsruhe

Abstract In this work, we improve the semantic segmentation of multi-layer top-view grid maps in the context of LiDAR-based perception for autonomous vehicles. To achieve this goal, we fuse sequential information from multiple consecutive lidar measurements with respect to the driven trajectory of an autonomous vehicle. By doing so, we enrich the multi-layer grid maps which are subsequently used as the input of a neural network. Our approach can be used for LiDAR-only 360° surround view semantic scene segmentation while being suitable for real-time critical systems. We evaluate the benefit of fusing sequential information based on a dense ground truth and discuss the effect on different semantic classes.

Keywords Autonomous driving, sensor data fusion, semantic grid map estimation.

1 Introduction

Environmental perception is a crucial task for many applications in robotics and mobile systems. This is particularly true for highly dynamic environments in which human life is at stake, such as urban scenarios. In these situations, autonomous driving systems heavily rely on a robust and accurate environment interpretation and scene understanding. Semantic segmentation plays a key role in efficient,

meaningful and holistic scene representation. With the advent of deep convolutional networks the task has received a lot of attention in the last few years and has shown significant improvements. Many well-developed network architectures are tailored to the image domain due to the data shortage in other domains.

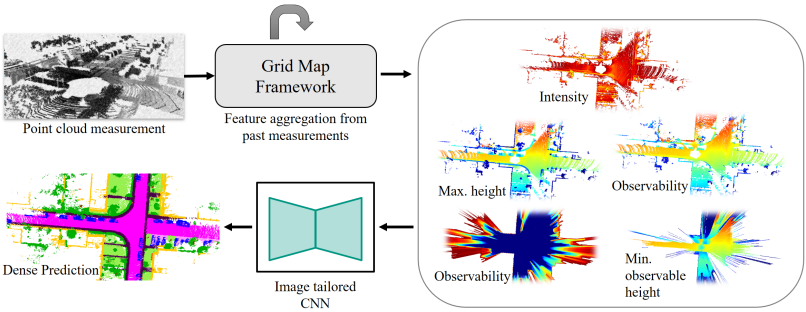

Figure 1.1: System overview including all input and output grid map types. By using our grid map framework we transform lidar measurements into a multi-layer grid map representation. The multi-layer grid maps are processed by an image-tailored CNN to predict semantic grid maps.

Recently, Behley et al. [1] published SemanticKITTI, the first large scale publicly available dataset which provides semantic segmentation for lidar measurements. The publicly available data consists of more than 23.000 single shot lidar measurements with a point-wise annotation distinguishing 28 semantic classes. By doing so the authors also provide information about moving and non-moving objects for classes like vehicle or motorcycle. In a recent work, we [2] consider the transformation of lidar point clouds into a top-view grid map representation to approach an efficient top-view segmentation of lidar measurements. The structured representation of grid maps can be utilized by applying efficient, well-developed CNN architectures from the image domain. In contrast, neural networks which operate on unstructured point clouds often lack real-time capability.

A further advantage of the grid map representation is that it is well-suited for sensor fusion applications. For instance, Nuss et al. [3] fuse radar and laser measurements to estimate the dynamic state of grid cells. Furthermore, Richter et al. [4] used grid maps as

a common fusion structure for semantic information and different range measurements. Besides the information fusion from different sensors, grid maps can also be used to fuse sequential measurement data from one sensor [5]. Another interesting work in this direction was done by Wirges et al. [6] by training a neural network to estimate dense multi-layer grid maps from single shot measurements. The paper shows that this enrichment is improving the performance of object detection algorithms.

This work investigates the fusion of sequential lidar measurements in multi-layer grid maps in the context of top-view semantic grid map segmentation.

2 Contribution

The presented work extends the basic ideas of [2] by making necessary improvements and introducing a fusion concept which replaces the single-shot approach and allows the use of sequential information. The following overview points out the main contributions of the paper:

- We extent our grid mapping framework so that it is capable of combining information from multiple point clouds into one set of grid maps. For each layer we implement a tailored fusion strategy.
- We perform semantic grid map estimation using multi-layer grid maps with accumulated features from the current and past lidar measurements.
- We report the benefit of feature accumulation in multi-layer grid maps for the task of semantic segmentation. By doing so, we evaluate the improvements on a dense semantic ground truth layer.

3 Multi-Layer Grid Maps

This section provides information about the generation and definition of our multi-layer grid maps.

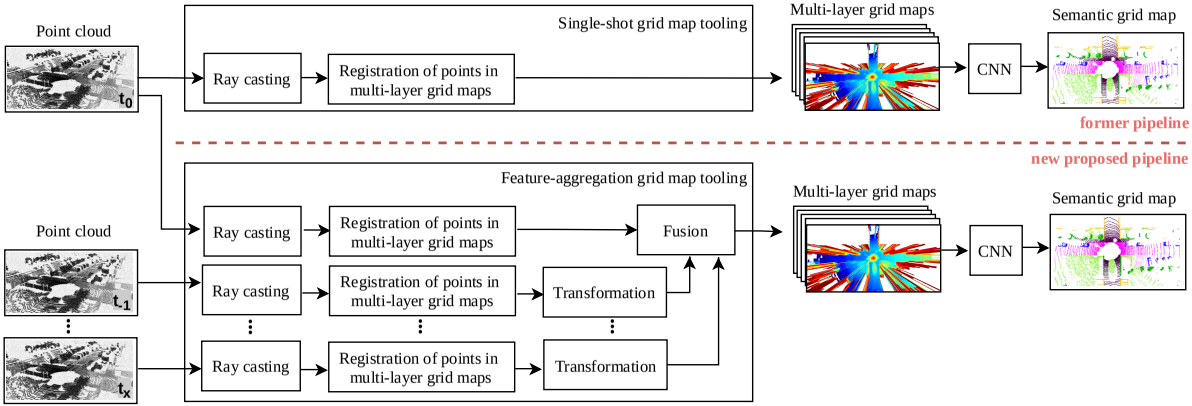

Figure 2.1: Comparison of our proposed feature aggregation pipeline and the initial, single-shot pipeline introduced in [2]. We extended the initial grid map framework so that it is able to fuse point clouds recorded on different time stamps into one grid map representation. As a requirement, we assume that the delta poses between the current pose and past poses are known. By doing so, we enrich the multi-layer grid maps, which are later used as input for a CNN to predict semantics.

Definition of Layers

Our multi-layer grid map input consists of five layers, which store the following features for each grid cell: The mean intensity, the maximum detected height, the minimum detected height,the observability representing the amount of rays through each cell and the minimum observable height with respect to all rays which crossed the cell. The first three layers only carry information in grid cells in which a lidar point is allocated. The information of the last two layers is extracted by casting rays between the sensor origin and the point detections to obtain dense layers in the observable area. In order to facilitate parallel computation and account for geometric sensor characteristics, all layers are first computed in polar coordinates and subsequently remapped into a cartesian coordinate system. An example for each layer can be found in figure 1.1.

Label Set and Data Set Split

We choose the label set and re-mapping strategy according to [2], but further combine the two classes rider and two-wheeler as they are hard to separate in the top view representation. This leads us to the following set of semantic classes: vehicle, person, two-wheel, road, side-walk, other-ground, building, pole/sign, vegetation trunk terrain. The sequences 0-7 and 9-10 of semanticKITTI are used to train the networks and the evaluation is conducted on sequence 8.

Grid Resolution and Sensing Range

The grid cell resolution is set to 10cm $\times$ 10cm. The region registered in one grid map is chosen to be 100m $\times$ 50m with the sensor located in the middle of the gird map. The grid maps are rotated such that the ego vehicles driving direction points to the right of the grid map.

Feature Aggregation

For the fusion process we collect point clouds from past time stamps, cast them to the grid map representation and transform them in the coordinate system of the current vehicle pose. We only choose past

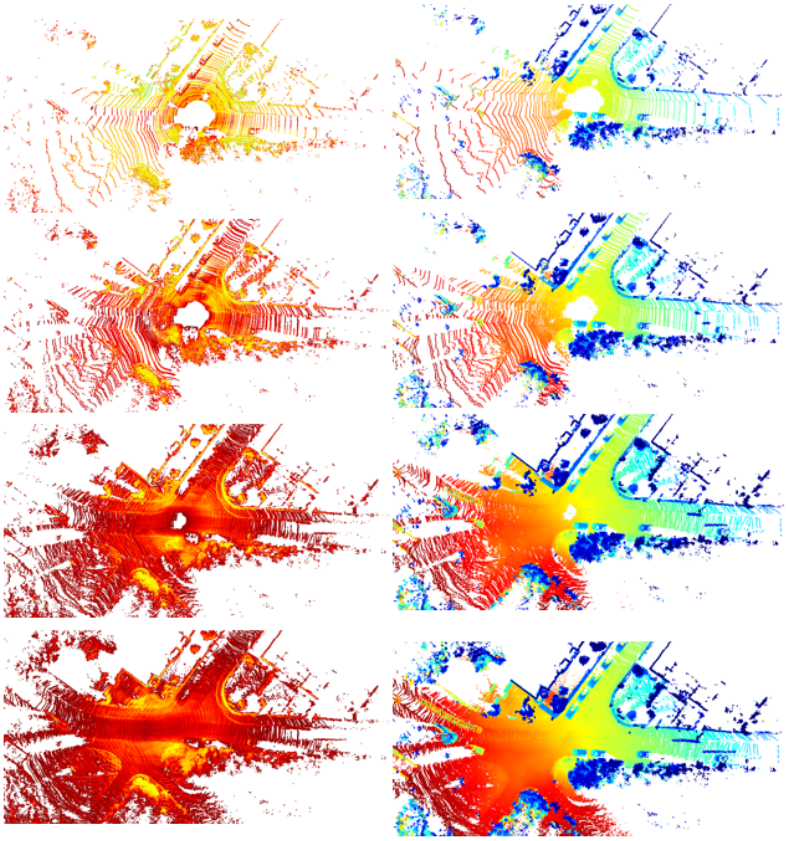

Figure 3.1: Example for feature aggregation for the layer intensity (left) and maximum detected height (right). The first row shows a single shot example, the second row 3 fused frames, the third row 10 fused frames and the last row 20 fused frames.

time stamps to have a causal system which could be applied in a similar fashion on a real-world system like an autonomous car. In order to be able to transform the past measurements into the coordinate system of the current grid maps highly accurate vehicle poses are required. We experienced that the poses of SemanticKITTI are superior of the original KITTI poses [7] and hence, use the former.

A unique fusion strategy is implemented for each layer. Regarding the intensity we calculated the average value for each grid cell considering all available measurements. In contrast we calculate the maximum value for the layer maximum detected height and the minimum value for the layers minimum detected height and minimum observable height. For the observability layer we accumulated the number of rays from each available measurement.

As the computation time for the grid mapping increases with an increasing batch size of point clouds, the number of fused measurements has to be well considered. Hence, we conduct and compare experiments with different point cloud batch sizes. An advantage of this approach is that the computational effort of the neural network does not increase by the accumulation of multiple measurements in the input grid maps.

Semantic Ground Truth

We create a dense semantic ground truth as it is described in [2]. After accumulating the semantic information of all surround poses we register the most likely pose within each grid cell. Here, we do not limit the amount of measurement but select all poses within a given radius for the fusion of semantic information.

4 Experiments

For each experiment we used all five grid map layers and optimized the network using the densely generated ground truth.

We conduct experiments comparing different state-of-the-art deep learning architectures, tailored for image processing. In this paper, all reported experiments are conducted using one architecture: the Deeplab framework with the Xception backbone [8]. We train the

networks using the full image resolution, a batch size of 2 and about 300.000 traing iterations. Besides the single shot experiments we present results for 3, 5, 10 and 20 accumulated frames.

5 Evaluation

We evaluate our experiments using the novel SemanticKITTI data set. Our models are trained to predict 11 classes which are particularly relevant for urban scene understanding. In this paper we choose a dense ground truth which also takes the network's prediction for cells without a detection into account.

Table 1: Class-wise evaluation using a dense semantic top view ground truth based on the 8 sequence of the semanticKITTI data set

frames	vehicle	two-wheel	pedestrian	road	sidewalk	parking	building	pole/sign	vegetation	trunk	terrain	**overall**
1	0.364	0.000	0.000	0.826	0.461	0.004	0.574	0.093	0.525	0.053	0.583	0.321
3	0.366	0.000	0.000	0.826	0.470	0.105	0.555	0.113	0.579	0.051	0.591	0.332
5	**0.392**	0.000	0.000	0.820	**0.487**	0.089	0.580	**0.138**	0.611	**0.064**	**0.647**	**0.348**
10	0.389	0.000	0.000	0.827	0.480	**0.128**	0.581	0.120	0.622	0.049	0.629	**0.348**
20	0.377	0.000	0.000	**0.831**	0.472	0.119	**0.583**	0.124	**0.631**	0.060	0.626	**0.348**

The quantitative evaluation is based on the *Intersection over Union* (IoU) [9]. The *mean Intersection over Union*, mIoU, is determined by

$$\text{mIoU} = \frac{1}{|K|} \sum_{k \in K} \text{IoU}_k \tag{5.1}$$

where $|K|$ is the the labelset's cardinality and the per-class IoU_k is calculated by

$$\text{IoU}_k = \frac{T_{P_k}}{T_{P_k} + F_{P_k} + F_{N_k}}, \tag{5.2}$$

with k being one of 11 classes. The quantitative results are shown in Table 1. In figure 5.1 some qualitative results are displayed.

Figure 5.1: Comparison of the qualitative results of three different scenes. The first column shows the inference based on single shot grid maps, the second column with 3 frames fused, third with 10 times fused and the last column is showing the ground truth.

6 Discussion

The experiments show that improvements can be achieved by the aggregation of past measurements. The greatest benefit can be obtained for the classes terrain, trunk and vegetation, parking and for pole/sign. However the improvements of additional feature aggregation seem to stagnate if more than 5 measurements are fused. We can also review that even with the feature aggregation the classes pedestrians and two-wheel can not be semantic segmented using the multi-layer grid maps. Here we have no improvement compared to the original paper.

7 Conclusion

We propose a framework to fuse information from sequential lidar measurements in a multi-layer grid map representation. Our experimental evaluations show the benefit of our approach in comparison to a formerly introduced single-shot method. While we review that an aggregation of past measurements brings a benefit, we also show that adding more past measurements only improves the performance to a certain extent.

References

1. J. Behley, M. Garbade, A. Milioto, J. Quenzel, S. Behnke, C. Stachniss, and J. Gall, "SemanticKITTI: A Dataset for Semantic Scene Understanding of LiDAR Sequences," in *Proc. of the IEEE/CVF International Conf. on Computer Vision (ICCV)*, 2019.
2. F. Bieder, S. Wirges, J. Janosovits, S. Richter, Z. Wang, and C. Stiller, "Exploiting Multi-Layer Grid Maps for Surround-View Semantic Segmentation of Sparse LiDAR Data," in *IEEE Intelligent Vehicles Symposium, Proceedings*, 2020.
3. D. Nuss, T. Yuan, G. Krehl, M. Stuebler, S. Reuter, and K. Dietmayer, "Fusion of laser and radar sensor data with a sequential Monte Carlo Bayesian occupancy filter," in *IEEE Intelligent Vehicles Symposium, Proceedings*, 2015.

4. S. Richter, S. Wirges, H. Königshof, and C. Stiller, "Fusion of range measurements and semantic estimates in an evidential framework," *tm - Technisches Messen*, 2019.
5. S. Wirges, C. Stiller, and F. Hartenbach, "Evidential Occupancy Grid Map Augmentation using Deep Learning," in *IEEE Intelligent Vehicles Symposium, Proceedings*, 2018.
6. S. Wirges, Y. Yang, S. Richter, H. Hu, and C. Stiller, "Learned enrichment of top-view grid maps improves object detection," in *IEEE Conference on Intelligent Transportation Systems (ITSC), Proceedings*, 2020.
7. A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.
8. L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.
9. M. Everingham, S. M. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The Pascal Visual Object Classes Challenge: A Retrospective," *International Journal of Computer Vision*, 2014.