ISMCR 2018 - 21th International Symposium on Measurement and Control in Robotics
26-28 September 2018 - International CBRNE Institute, MONS, Belgium
ROBOTICS FOR THE CHANGING WORLD

Semantic Grid Mapping based on Surface Classification with Supervised Learning

Torsten Engler, Felix Ebert and Hans-Joachim Wuensche

***Abstract*— LiDAR-based occupancy grid mapping can lead to overly conservative detection of obstacles in non-urban autonomous driving scenarios, e.g. grass in the middle of the lane is often interpreted as obstacle although it is actually driveable. We therefore aim to augment our current grid-based environment representation with additional information derived from pixel-level semantic segmentation in camera images. We project the resulting segmentation map onto an additional semantic layer in the environment grid representation by utilizing LiDAR data for pixel-to-cell association to improve our driveability analysis.**

We apply supervised machine learning techniques for pixelwise prediction of class labels. Datasets for non-urban environments are rare. Therefore, we created a custom dataset. Due to the huge effort necessary to create such a dataset, its size is relatively small and hence neural networks might not be able to train effectively. Thus, low numbers of training samples require a careful choice of the classifier and/or data augmentation techniques. We therefore compare classification performance of neural networks with random forest classifiers.

I. Introduction

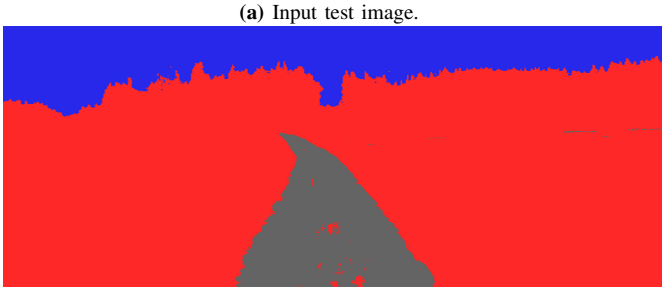
A semantic understanding of the current surroundings is a fundamental ability to improve autonomous driving performance. Thus, we enhance our current environment representation by semantic segmentation of camera images to improve our driveability analysis. We apply the segmentation results to distinguish between areas that are detected as occupied but actually are driveable (e.g. high grass) and those that are correctly detected as non-driveable (e.g. tree, hedge, stones covered with grass).

Our current environment representation is a multi-modal local terrain map obtained from Vision and LiDAR data [1]. It consists of several distinct layers where each layer corresponds to a specific environmental feature such as obstacle probabilities, colors and heights. The sensors used are a Velodyne HDL-64 LiDAR with 360° field of view, a front-facing color camera and a high-precision inertial navigation system. The resulting terrain map is obtained by spatiotemporal fusion of depth measurements, color information and motion estimation. Due to temporal accumulation, outliers are effectively filtered.

While this terrain map provides a conservative basis for obstacle-free path planning, we observe a high false positive rate of cell occupancy classification. Therefore, we propose a classification based approach to semantically distinguish between different surfaces and driveabilities as shown in Figure 1.

All authors are with the Institute for Autonomous Systems Technology (TAS) of the University of the Bundeswehr Munich, Neubiberg, Germany. Contact author email: torsten.engler@unibw.de

(a) Input test image.



(b) Result of the segmentation with the random forest with hand-crafted features.

(c) Segmentation result of the Pyramid Scene Parsing Network.

Fig. 1: Segmentation result for the random forest (b) and the neural network classifier (c). Blue corresponds to label *sky*, red to *vegetation* (undriveable), green to *grass* (driveable vegetation) and gray to *dirt road*.

Acquiring datasets of sufficient size poses one of the main challenges to apply supervised machine learning algorithms. Moreover, most readily available datasets for autonomous driving are designed for urban scenarios and therefore not directly applicable. It is necessary to create a distinct dataset for the situations and vegetations we encounter during offroad driving. Due to the high effort of manually creating large datasets, our dataset is comparably small. While deep neural networks perform very well in pixel-wise semantic classification, their performance degrades when trained with insufficient training data. In order to achieve good classification results, it is necessary to artificially increase the training data size by augmentation and/or transfer learning. In addition to neural networks, we evaluate random forests which don't require

large datasets and compare their classification performance to a neural network classifier.

This document is structured as follows: Section II gives a short overview of related work in the field of semantic segmentation with random forest or deep neural network classifiers. Section III details the dataset deployed for training. After covering our approaches to semantic segmentation in Section IV, we present the fusion approach in Section V and finally provide results and conclusions in Section VI.

II. Related Work

Semantic segmentation is the process of assigning a label to each pixel of an input image resulting in a segmentation map with same width and height as the input image. This is an essential task for image analysis and scene understanding in autonomous driving tasks. With the emergence of deep neural networks in recent years, significant progress has been made in the field of semantic segmentation, surpassing traditional methods. For a recent overview, the reader is referred to [2]. The Pascal Visual Object Classes [3] was the first challenge to incorporate semantic image segmentation. More recently, the Robust Vision Challenge provides good comparisons [4]. In this work, we compare a traditional segmentation method, the random forest classifier, with a deep neural network classifier.

A. Random Forests

The random forest classifier [5] is an ensemble classifier grown from a number of random binary decision trees. Given a training set $\{X, y\}$ with input X and corresponding class label y, each tree node selects a feature function $h(X)$ and a threshold to split the input set to maximize the information gain. In this way, several trees are learned and together form the random forest. Prediction is achieved by applying the query input x to each tree individually, branching down until a leaf node is reached. A majority vote over all trees gives the final classification result.

Due to randomization, the random forest is able to mitigate the problem of overfitting and lack of generalization connected to single tree classifiers. Randomization is achieved by selecting a random subset of the input data either for each tree or each node and a random subset of the feature functions available for each node splitting.

Due to the general nature of feature functions applicable for this classifier, several approaches have been proposed for semantic image segmentation like structured class-labels [6], semantic textons [7], or a combination of multiple features like textons, color, filterbanks and HOG features [8].

B. Deep Neural Networks for Semantic Image Segmentation

Deep neural networks have their origin in object classification. By introducing masks for convolutions and learning the weights of the masks, the network is able to differentiate image regions based on their context. The task of pixel-level image segmentation is inherently more challenging and can be done in different ways. The information gathered by the network has to be upsampled again to obtain a pixel-wise information. The first possibility are R-CNNs (Regions with CNN Features) [9]. In a first step, the network extracts image regions based on a trained object detection. Afterwards, the classification decision is made based on a majority of all image regions where the pixel is part of.

Another popular approach is the encoder-decoder architecture. The encoder is used to extract discriminative features of the image region and the decoder is used to output dense pixel-wise labels based on the features. The category of these networks is also called fully convolutional networks (FCN) [10]. In contrast to object detection networks, these networks don't have fully connected layers and consist solely of convolutional layers. Therefore, there are no constraints on the size of the input image. However, the output of those networks usually is of low resolution and necessary upsampling leads to fuzzy object boundaries. Different strategies are applied to reduced this issue, e.g. SegNet [11]. Most modern segmentation networks differ in terms of the upsampling part, while using proven techniques for the feature extraction.

III. Dataset

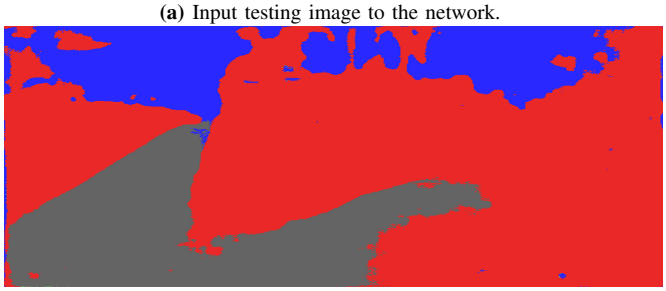
We deploy a handcrafted dataset for training the random forest and the neural network classifier. Data was recorded for five different dirt roads where camera images, LiDAR scans and egomotion data was captured. Special attention was paid to recording data at different times of day and different exposure conditions. Four of these five recordings are used for training and the fifth exclusively for testing. A random set of around 200 images with low correlation between the images was selected for training data generation. Data is split into 80% training data, 20% validation data and 15 images from the test stream that are not used during training. In order to keep inference times small and be able to process images from different sources, the images are downscaled to $512\,\text{px} \times 512\,\text{px}$ which proved sufficient for our classification task. These images were labeled manually pixel-wise with four classes: *dirt road*, *sky*, *vegetation* and *grass*, where *grass* is driveable and *vegetation* is not. The intention to distinguish between driveable and undriveable vegetation is to classify patches of grass in the middle of the lane that might be misclassified as obstacle by the environment mapping module.

IV. Semantic Image Classification

We apply the Extra-Trees classifier [12] as a representative random forest algorithm. It selects a random subset of the training data and feature functions for each node split. Features are computed as a weighted sum of responses of a rectangular region around the respective pixel location. These include hue, saturation, the green channel, the *ndi* and *exg* vegetation indices [13] as well as parts of the MR8 filter bank [14]. Analysis of feature importances along the trees showed that edge features of the filter bank had no significant impact on the classification result. Therefore, these were left out to reduce computation time of the feature vectors. The final forest consist of 36 random trees.

Correctly implementing and comparing different architectures of deep neural networks still poses a major challenge.

(a) Input testing image to the network.



(b) Segmentation result of the Pyramid Scene Parsing Network.

Fig. 2: Scene of the testing dataset with a forking road. The training data doesn't contain a visually similar scene, yet the network is able to at least partially generalize from the training data.

To alleviate network selection, training and performance comparison of different network architectures, we adopt the Semantic Segmentation Suite [15]. This suite provides methods to train different state-of-the-art network models with any dataset using the deep learning framework *Tensorflow*. After comparing different models trained with our custom dataset, the Pyramid Scene Parsing Network (PSP) [16] turned out to be best performing and hence was selected for further application and comparison with the random forest approach.

Our dataset is comparatively small for training of deep neural networks. Therefore, we applied data augmentation to generate additional training data. During data augmentation it is important to alter the data based on changes that are likely to occur in real-world scenarios. This includes vertical flipping of the image as well as small rotations up to 10°. Other alternations like flipping horizontally, or brightness changes of the images need to be applied with caution. For instance a brightness change of the scene cannot be easily simulated by altering the resulting picture.

V. Fusion Approach

The fusion strategy of the camera and the LiDAR as described in [1] is crucial for all approaches. The fundamentals of the fusion are equal in each tested scenario and only depend on the relative pose of camera and LiDAR and on the egomotion. Each 3D point of the LiDAR is associated with a cell in the occupancy grid and projected back to the camera image. Afterwards the pixel's color is used to color the associated cell. Additionally, each cell has a decay rate to reduce the impact of wrongly accumulated colors. Special consideration is necessary in terms of occlusion. The LiDAR sensor and the camera are mounted on different parts of the vehicle. Thus, points visible by the LiDAR are not necessarily visible in the camera image.

VI. Results and Conclusion

Both approaches, random forests and neural networks, were able to achieve a reasonable classification rate for our dataset. Table I shows a comparison of several performance scores for both approaches. Besides precision and recall, two further performance metrics are given which provide a combined measure. The F1 score, the harmonic mean of precision and recall, and the Intersection over Union (IoU) measure. We achieve average IoU values of 0.933 for the PSP network and 0.920 for the random forest.

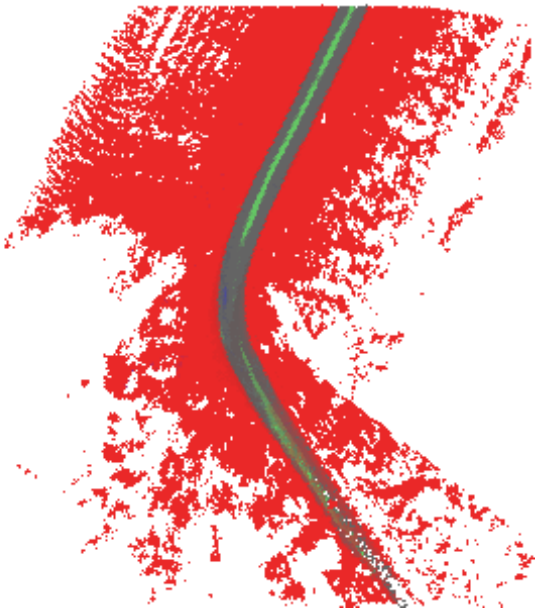
TABLE I: Classification rates and scores of the random forest (RF) and the Pyramid Scene Parsing Network (PSP). The scores are averaged over 15 different images from the validation set (same images for each approach).

Vegetation	PSP	RF
Precision	0.995	0.987
Recall	0.992	0.993
F1 Score	0.993	0.990
IoU	0.987	0.980
Road	PSP	RF
Precision	0.981	0.989
Recall	0.987	0.990
F1 Score	0.984	0.990
IoU	0.968	0.980
Sky	PSP	RF
Precision	0.967	0.937
Recall	0.964	0.919
F1 Score	0.966	0.928
IoU	0.934	0.866
Grass	PSP	RF
Precision	0.883	0.854
Recall	0.948	0.698
F1 Score	0.914	0.768
IoU	0.842	0.624

Note that these values are comparatively high. This is due to the low number of classes. Furthermore we consider relatively simple scenes, where substantial parts of the images consist of sky and vegetation and hence misclassifications in the other two classes have less impact on the overall average values. It can be seen, that the segmentation quality of both approaches is about the same for *vegetation* and *road*. However, the classification scores for *grass* indicate that the neural network is better at distinguishing grass in the middle of the road from other vegetation.

Despite having similar IoU values on our test dataset, the random forest approach shows some disadvantages. The size and thus the inference time grows with the size of the training set as well as with the number of classes. Hence, having more data leads to larger trees with larger memory footprint and inference times, whereas these values are constant for the deep learning approach. This is the limiting factor for generalization to other road types or more classes. Furthermore, due to the

(a) Grid colored with the color camera image as described in Section V.



(b) Grid colored with the resulting segmentation map of the Pyramid Scene Parsing Network.

Fig. 3: Accumulation of colors in the grid surrounding the ego vehicle.

current feature selection, the forest approach is not able to incorporate spatial context into the inference. The low recall value for grass in Table I indicates that the amount of false negatives is high which coincides with the observation in Figure 1b that the vegetation at the center of the lane is misclassified as *road* or *vegetation*. The main advantage of random forests is that for a very small training set of around 20 images, it already leads to reasonable classification results.

Despite a small data set, the network is able to learn a good representation of the data. It seems to be helpful to apply data augmentation to artificially increase the training data size. In contrast to the random forest, the segmentation of the network takes spatial information into account. Grass in the middle of the lane is classified as driveable even though it looks visually similar to vegetation in other parts of the image. More importantly, the network is able to generalize very well. In the scene with a parting road (see Figure 2) the segmentation is good even though the training data did not contain a similar case.

In terms of segmentation quality the network outperforms the random forest approach. Therefore, the segmentation results of the CNN are used subsequently to color our environment representation as seen in Figure 3. The segmented grid around the vehicle can be used to mark obstacle cells due to grass in the middle of the road as driveable. In the future, our results can be used to assist road network tracking and crossroad detection.

REFERENCES

[1] H. Jaspers, M. Himmelsbach, and H.-J. Wuensche, "Multi-modal Local Terrain Maps from Vision and LiDAR," in *Proceedings of IEEE Intelligent Vehicles Symposium (IV)*, Redondo Beach, CA, USA, Jun. 2017.

[2] X. Liu, Z. Deng, and Y. Yang, "Recent Progress in Semantic Image Segmentation," *Artificial Intelligence Review*, 2018.

[3] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal Visual Object Classes (VOC) Challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, Jun. 2010.

[4] "Robust Vision Challenge," http://www.robustvision.net/, 2018.

[5] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[6] P. Kontschieder, S. R. Bulò, H. Bischof, and M. Pelillo, "Structured Class-Labels in Random Forests for Semantic Image Labelling," *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, pp. 2190–2197, 2011.

[7] J. Shotton, M. Johnson, and R. Cipolla, "Semantic Texton Forests for Image Categorization and Segmentation," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2008, pp. 1–8.

[8] F. Schroff, A. Criminisi, and A. Zisserman, "Object Class Segmentation using Random Forests," *Proceedings of the British Machine Vision Conference*, pp. 54.1–54.10, 2008.

[9] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," *CoRR*, vol. abs/1311.2524, 2013. [Online]. Available: http://arxiv.org/abs/1311.2524

[10] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 3431–3440.

[11] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *CoRR*, vol. abs/1511.00561, 2015. [Online]. Available: http://arxiv.org/abs/1511.00561

[12] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely Randomized Trees," *Machine Learning*, vol. 63, no. 1, pp. 3–42, 2006.

[13] E. Hamuda, M. Glavin, and E. Jones, "A Survey of Image Processing Techniques for Plant Extraction and Segmentation in the Field," *Computers and Electronics in Agriculture*, vol. 125, pp. 184–199, 2016.

[14] M. Varma and A. Zisserman, "Classifying Images of Materials: Achieving Viewpoint and Illumination Independence," in *Computer Vision — ECCV 2002*, A. Heyden, G. Sparr, M. Nielsen, and P. Johansen, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2002, pp. 255–271.

[15] "Semantic Segmentation Suite," https://github.com/GeorgeSeif/Semantic-Segmentation-Suite, 2018.

[16] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6230–6239, 2017.