 DOI: 10.3217/978-3-85125-663-5-45

Evaluation Study on Semantic Object Labelling in Street Scenes

Andreas Wittmann[1], Margrit Gelautz[2], Florian Seitner[3]

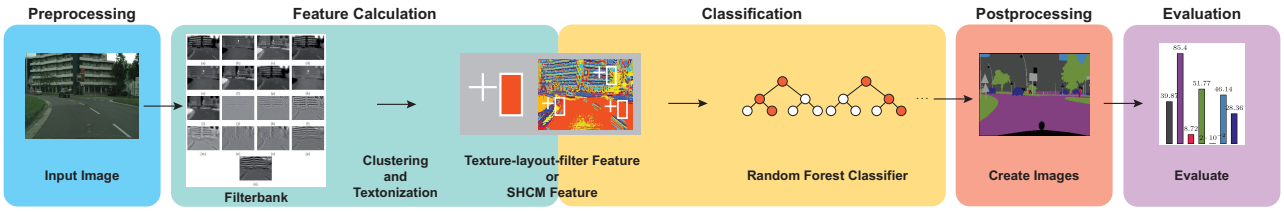

Fig. 1: Overview of our semantic segmentation system.

***Abstract*— We present a processing pipeline for semantic scene labelling that was developed in view of autonomous driving applications. Our study focuses on two different methods for feature selection - Texture-layout-filter (TLF) and Single Histogram Class Models (SHCM) - whose influence on the performance of a random forest classifier is investigated. In tests on the Cityscapes dataset, we assess the effects of parameter variation and observe an improvement of the Intersection over Union score by 44 percent when substituting the TLF by the computationally more demanding SHCM feature.**

I. INTRODUCTION

Advanced driving assistance systems and fully autonomous vehicles require detailed knowledge about vehicle surrounding and real-time interpretation of street scenes. Deep learning approaches are computationally demanding and need powerful and expensive hardware. On the other hand, conventional machine learning methods can be used for real-time semantic scene labelling at a fraction of the computational complexity and on cost-efficient hardware. In view of a subsequent transfer to an embedded platform, we implement a semantic scene labelling pipeline based on a random forest classifier and evaluate its performance regarding the overlap of the predicted semantic segmentation results and the ground-truth labelling by using the Intersection over Union score.

Thoma [8] gives a comprehensive survey of different techniques and features used for pixel-wise semantic segmentation. Several semantic scene labelling approaches rely on contextural features [9], [1], [4], [2], [7]. We employ two features that make use of textural and contextural properties proposed in the literature [7], [6] and analyze and compare their results on the publicly available Cityscapes dataset [3].

*This work was supported by the project CarVisionLight (Project No. 861251), which is funded by the Austrian Research Promotion Agency (FFG) under the programme "ICT of the Future", an initiative of the Federal Ministry of Transport, Innovation and Technology (BMVIT).

[1]Andreas Wittmann is with emotion3D, 1040 Vienna, Austria awi@emotion3d.ai

[2]Margrit Gelautz is with the Institute of Visual Computing and Human-Centered Technology, TU Vienna, 1040 Vienna, Austria margrit.gelautz@tuwien.ac.at

[3]Florian Seitner is with emotion3D, 1040 Vienna, Austria fse@emotion3d.ai

II. SEMANTIC SCENE LABELLING PIPELINE

Figure 1 visualizes the semantic segmentation pipeline realized by our system. Our segmentation pipeline consists of five modules. The *preprocessing module* serves to prepare incoming images and datasets. The *feature module* prepares the extraction of the features by providing a feature factory that can create variations of the feature within the predefined parameter ranges. The *classification module* instantiates features by using the *feature module* and calculates the feature response in real-time. It uses the feature responses in a random forest classifier. The *postprocessing module* generates prediction label images from the output of the random forest. Finally, the *evaluation module* calculates the performance of our algorithm in terms of the per-class prediction accuracy.

The feature module contains the computation of two features: the Texture-layout-filter (TLF) feature and the Single Histogram Class Models (SHCM) feature. The TLF feature combines textural information represented by textons with semantic, spatial and scale context by using information of neighbouring pixels and regions. The computation of the TLF feature is explained in detail by Shotton et al. [7]. The SHCM feature is based on the semantic segmentation approach of Schroff et al. [6]. The authors compare a histogram of a selected region with precalculated SHCM [5] in a random forest classifier. The SHCM feature can be seen as an extension of the TLF feature. While the TLF feature counts the occurrences of one single texton id within a region, the SHCM feature calculates the distribution of all texton ids within a region. Thus, the SHCM feature module extracts a more dense form of information for a specific region.

III. EVALUATION AND RESULTS

The aim of our evaluation study is to gain insight into how well a computationally cheap semantic segmentation system can detect regions and objects in street scenes. Furthermore, we want to explore how much information of a neighbouring region of a pixel is needed to semantically classify the pixel. We use the Intersection over Union score to measure the performance of our system.

We evaluate the performance of our system on the Cityscapes per-category dataset [3]. The data of the dataset is captured in different seasons, under different weather conditions and contains crowded street scenes with a high variety of vegetation, which makes the dataset challenging.

By using the Intersection over Union performance score and by applying our system on the Cityscapes dataset, we make it possible to compare the performance of our system with other approaches published in the Cityscapes benchmark.

The final prediction result of our semantic scene labelling system is dependent on several parameters whose influence we investigate in more detail. Both features use k-means clustering with κ clusters in their textonization process, a window size ω and a window offset o. The window size and offset to the candidate pixel define the size and position of the region that is used to extract information of neighbouring pixels. We evaluate the performance of our system with parameter values within the following ranges: $\kappa = [50, 8000]$, $\omega = [4, 49]$ and $o = [0, 100]$.

In the first set of experiments, we evaluate the performance of the TLF by varying the number of clusters, and the ranges for the window size and the window offset. We achieved best results with an Intersection over Union score of 26.50 by using the TLF feature with $\kappa = 100$, $\omega = [4, 24]$ and $o = [0, 23]$.

By using the SHCM feature in the second part of our experiments, we could increase the Intersection over Union score to 38.29 with $\kappa = 8000$, $\omega = [4, 17]$ and $o = [0, 23]$.

By looking at example segmentation results in Figure 3, we can observe that the TLF feature (column 3) is not able to distinguish between all object classes. By comparing the results of the TLF feature in Column 3 with the results of the SHCM feature in Column 4 of Figure 3, one can recognize a clear improvement of the semantic segmentation. Especially region borders are refined by using the SHCM feature. However, one can see that the classifier has difficulties to distinguish between the class *vehicle* and the class *human* in Column 4. These missclassifications cannot be observed in Column 3, because the classifier hardly detects any pixels that belong to the class *human* by using the TLF feature.

In particular, we found that the system has a bias towards classes representing big object regions such as the classes *flat (road)*, *nature* and *sky*, which can be detected to a certain extent. We assume that a reason for this behaviour is the context-driven classification approach of both features, which tends to interpret small objects in large neighbouring regions as noisy pixels. This bias gets smaller by using the SHCM feature, because the SHCM feature extracts more detailed information from neighbouring regions.

IV. CONCLUSIONS

Semantic, spatial and scale context as well as textural properties play an important role for semantic scene labelling. The TLF feature is a computationally cheap way to extract textural properties and context of data points. In our experimental evaluation, using the SHCM feature instead of the TLF feature increases the accuracy by around 44%. However, the SHCM feature is computationally more complex than the TLF feature. Both features have a bias towards classes representing big objects when used in a random forest classifier and applied on the Cityscapes dataset.

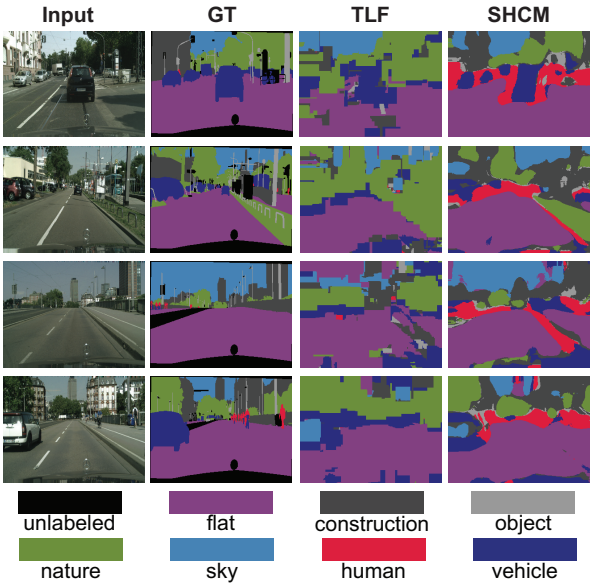

Fig. 3: Example segmentation results. 1st column: input image; 2nd column: ground-truth; 3rd column: results by using the TLF feature; 4th column: results by using the SHCM feature.

REFERENCES

[1] A. Arnab, S. Zheng, S. Jayasumana, B. Romera-Paredes, M. Larsson, A. Kirillov, B. Savchynskyy, C. Rother, F. Kahl, and P. H. Torr, "Conditional Random Fields Meet Deep Neural Networks for Semantic Segmentation: Combining Probabilistic Graphical Models with Deep Learning for Structured Prediction," *IEEE Signal Processing Magazine*, vol. 35, no. 1, pp. 37–52, Jan. 2018.

[2] L. 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*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[3] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The Cityscapes Dataset for Semantic Urban Scene Understanding," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[4] Z. Liu, X. Li, P. Luo, C. C. Loy, and X. Tang, "Deep Learning Markov Random Field for Semantic Segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 8, pp. 1814–1828, Aug. 2018.

[5] F. Schroff, A. Criminisi, and A. Zisserman, "Single-Histogram Class Models for Image Segmentation," in *Computer Vision, Graphics and Image Processing*. Springer Berlin Heidelberg, 2006, pp. 82–93.

[6] ——, "Object Class Segmentation using Random Forests," in *Conference on British Machine Vision (BMVC)*. British Machine Vision Association, 2008, pp. 54.1–54.10.

[7] J. Shotton, J. Winn, C. Rother, and A. Criminisi, "TextonBoost for Image Understanding: Multi-Class Object Recognition and Segmentation by Jointly Modeling Texture, Layout, and Context," *International Journal of Computer Vision (IJCV)*, vol. 81, no. 1, pp. 2–23, Jan. 2009.

[8] M. Thoma, "A Survey of Semantic Segmentation," *arXiv:1602.06541*, 2016.

[9] J. Xie, L. Yu, L. Zhu, and X. Chen, "Semantic Image Segmentation Method with Multiple Adjacency Trees and Multiscale Features," *Cognitive Computation*, vol. 9, no. 2, pp. 168–179, Apr. 2017.