Segmentation-Based Urban Traffic Scene Understanding

Andreas Ess[1]
aess@vision.ee.ethz.ch
Tobias Müller[1]
muelleto@bluewin.ch
Helmut Grabner[1]
grabner@vision.ee.ethz.ch
Luc van Gool[1,2]
vangool@vision.ee.ethz.ch

[1] Computer Vision Laboratory
ETH Zürich
Switzerland
[2] ESAT-PSI / IBBT
K.U. Leuven
Belgium

Recognizing the traffic scene in front of a car is an important asset for autonomous driving, as well as for safety systems. While GPS-based maps abound and have reached an incredible level of accuracy, they can still profit from additional, image-based information. Especially in urban scenarios, GPS reception can be shaky, or the map might not contain the latest detours due to constructions, demonstrations, etc. Furthermore, such maps are static and cannot account for other dynamic traffic agents, such as cars or pedestrians. In this paper, we therefore propose an image-based system that is able to recognize both the road type (straight, left/right curve, crossing, ...) as well as a set of often encountered objects (car, pedestrian, pedestrian crossing). The obtained information could then be fused with existing maps and either assist the driver directly (*e.g.*, a pedestrian crossing is ahead: slow down) or help in improving object tracking (*e.g.*, where are possible entrance points for pedestrians or cars?).

Starting from a video sequence obtained from a car driving through urban areas, we employ a two-stage architecture termed *Segmentation-Based Urban Traffic Scene Understanding (SUTSU)* that first builds an intermediate representation of the image based on a patch-wise image classification. The patch-wise segmentation is inspired by recent work [3, 4, 5] and assigns class probabilities to every 8×8 image patch. As a feature set, we use the coefficients of the Walsh-Hadamard transform (a decomposition of the image into square waves), and, if available, additional information from the depth map. These are then used in a one-versus-all training using AdaBoost for feature selection, where we choose 13 texture classes that we found to be representative of typical urban scenes. This yields a meta representation of the scene that is more suitable for further processing, Fig. 1 (b,c). In recent publications, such a segmentation was used for a variety of purposes, such as improvement of object detection [1, 5], analysis of occlusion boundaries, or 3D reconstruction.

In this paper, we will investigate the use of a segmentation for urban scene analysis. We infer another set of features from the segmentation's probability maps, analyzing repetitivity, curvature, and rough structure. This set is then again used with a one-versus-all training to infer both the type of road segment ahead, as well the additional presence of pedestrians, cars, or pedestrian crossing. A Hidden Markov Model is used for temporally smoothing the result.

SUTSU is tested on two challenging sequences, spanning over 50 minutes video of driving through Zurich. The experiments show that while a state-of-the-art scene classifier [2] can keep global classes such as road types, similarly well apart, a manually crafted feature set based on a segmentation clearly outperforms it on object classes. Example images are shown in Fig. 2.

The main contribution of this paper is the application of recent research efforts in scene categorization research to do vision "in the wild", driving through urban scenarios. We furthermore show the advantage of a segmentation-based approach over a global descriptor, as the intermediate representation can easily be adapted to other underlying image data (*e.g.* dusk, rain, ...), without having to change the high-level classifier.

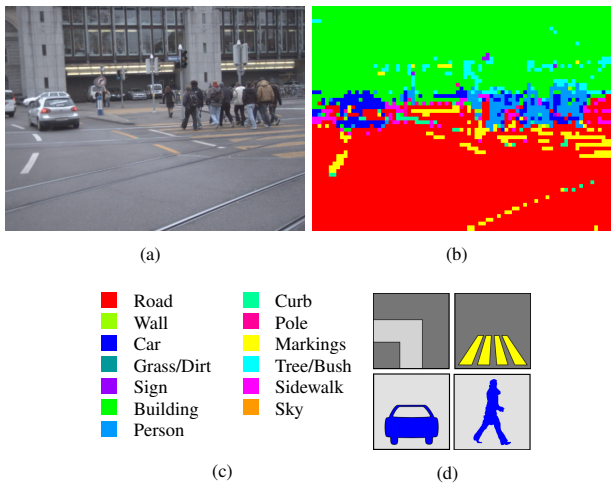

Figure 1: Given an input image (a), we calculate a meta representation based on a patch-wise scene classification (b) into a set of urban texture classes (c), which is then used to classify the scene both with respect to road type, as well as to detect the presence of certain objects (d).

Figure 2: Example images. For each image, the bottom left shows the patch classification output, as well as the scene classification (road type, present objects). Figure is best viewed in color.

[1] D. Hoiem, A. A. Efros, and M. Hebert. Putting objects in perspective. In *CVPR*, 2006.

[2] A. Oliva and A. Torralba. Modeling the shape of the scene: a holistic representation of the spatial envelope. *IJCV*, 42(3):145–175, 2001.

[3] F. Schroff, A. Criminisi, and A. Zisserman. Object class segmentation using random forests. In *BMVC*, 2008.

[4] J. Shotton, M. Johnson, and R. Cipolla. Semantic texton forests for image categorization and segmentation. In *CVPR*, 2008.

[5] C. Wojek and B. Schiele. A dynamic CRF model for joint labeling of object and scene classes. In *ECCV*, 2008.