Vision-Based Behavior Prediction in Urban Traffic Environments by Scene Categorization

Martin Heracles[12]
heracles@cor-lab.uni-bielefeld.de
Fernando Martinelli[2]
fernando.martinelli@gmail.com
Jannik Fritsch[2]
jannik.fritsch@honda-ri.de

[1] CoR-Lab
Bielefeld University
Germany

[2] Honda Research Institute Europe
Offenbach/Main
Germany

We propose a method for vision-based scene understanding in urban traffic environments that predicts the appropriate behavior of a human driver in a given visual scene. The method relies on a decomposition of the visual scene into its constituent objects by image segmentation and uses segmentation-based features that represent both their identity and spatial properties. We show how the behavior prediction can be formulated as scene categorization problem and how ground truth behavior data for learning a classifier can be automatically generated from any monocular video sequence recorded from a moving vehicle, using structure from motion techniques. We evaluate our method on the recently proposed CamVid dataset [1], predicting the appropriate velocity and yaw rate of the car (and their appropriate change) for both day and dusk sequences. In particular, we investigate the impact of the underlying segmentation and the number of behavior classes on the quality of these predictions.

Given an uncalibrated monocular image sequence $I = (i_1, \ldots, i_n)$ (see Fig. 1a), both the ego-motion of the camera and the 3D structure of the scene can be reconstructed by structure from motion techniques [4]. For image sequences acquired by a car-mounted camera while driving, the ego-motion of the camera largely corresponds to the motion of the car itself, hence it is possible to infer the (normalized) velocity $v \in [-1,1]$ and yaw rate $y \in [-1,1]$ of the car from the camera trajectory $C = (c_1, \ldots, c_n)$. As a result, the camera trajectory C has a corresponding trajectory $B = ((v_1,y_1), \ldots, (v_n,y_n))$ in the 2D space spanned by velocity and yaw rate (Fig. 1b, blue), which largely reflects the behavior of the human driver.

Clearly, the behavior of the driver depends to a great extent on what he or she currently sees, i.e., the pair $(v_k,y_k) \in B$ has a visible correlate in the corresponding image $i_k \in I$. Our goal is to learn such correlations from I and B, which can be seen as examples of the appropriate behavior in different visual scenes provided by a responsible driver, in order to be able to predict the appropriate behavior $(v^*,y^*) \in [-1,1]^2$ for new images $i^* \notin I$. We formulate the behavior prediction as scene categorization problem by considering velocity, yaw rate and their respective changes independently from each other, thus predicting a 1D quantity $q \in [-1,1]$ each, and defining symmetric thresholds $-t_p, \ldots, -t_1, t_1, \ldots, t_p \in [-1,1]$ on the domain of q (Fig. 1b, red). This subdivides B into behavior classes $B_0 = \{(v,y) \in B | -t_1 < q < t_1\}$, $B_{t_1} = \{(v,y) \in B | t_1 \leq q < t_2\}$, $B_{-t_1} = \{(v,y) \in B | -t_2 > q \geq -t_1\}$ etc. Since each $(v_k,y_k) \in B$ has a corresponding image $i_k \in I$, these induce image classes $I_0 = \{i_k \in I | (v_k,y_k) \in B_0\}$, $I_{t_1} = \{i_k \in I | (v_k,y_k) \in B_{t_1}\}$, $I_{-t_1} = \{i_k \in I | (v_k,y_k) \in B_{-t_1}\}$ etc. By classifying a new image $i' \notin I$ as belonging to a certain I_*, based on its similarity to the $i \in I_*$, the behavior represented by B_* is then taken to be the appropriate behavior for the situation depicted in i'.

Similarity between images is judged using the features proposed by Ess et al. [2]. We did not include the periodicity features since these mainly serve to detect repetitive objects, which is not our focus. Given a segmentation s of image $i \in I$ that assigns an object class label $s(u,v) \in \{l_1, \ldots, l_n\}$ to each pixel (Fig. 1c), we compute n binary maps $s_{l_1}, \ldots, s_{l_n}$ such that $s_{l_*}(u,v) = 1$ iff $s(u,v) = l_*$ (Fig. 1d, top left). Each of the s_{l_*} is downsampled into an 8x8, 4x4, and 2x2 feature map $\overline{s_{l_*}}$ by subdividing s_{l_*} into the corresponding number of blocks and computing the average pixel value per block (Fig. 1d, bottom right). Similarly, we downsample each of the s_{l_*} into an 80x60 map from which we compute a row feature vector $\overline{r_{l_*}}$ and a column feature vector $\overline{c_{l_*}}$ by averaging over its rows and columns, respectively (Fig. 1d, top right and bottom left). Finally, we combine the two binary maps corresponding to lane markings and curbs by computing their pixel-wise maximum and apply a gradient operator (Sobel filter). The resulting gradient image is subdivided into 4x4 blocks and we compute an 18-bin edge orientation histogram per block. In the end, the $\overline{s_{l_*}}$, $\overline{r_{l_*}}$, $\overline{c_{l_*}}$ and histograms are all serialized into one large feature vector of size $n(64+16+4)+n(60+80)+16 \cdot 18$ that represents i.

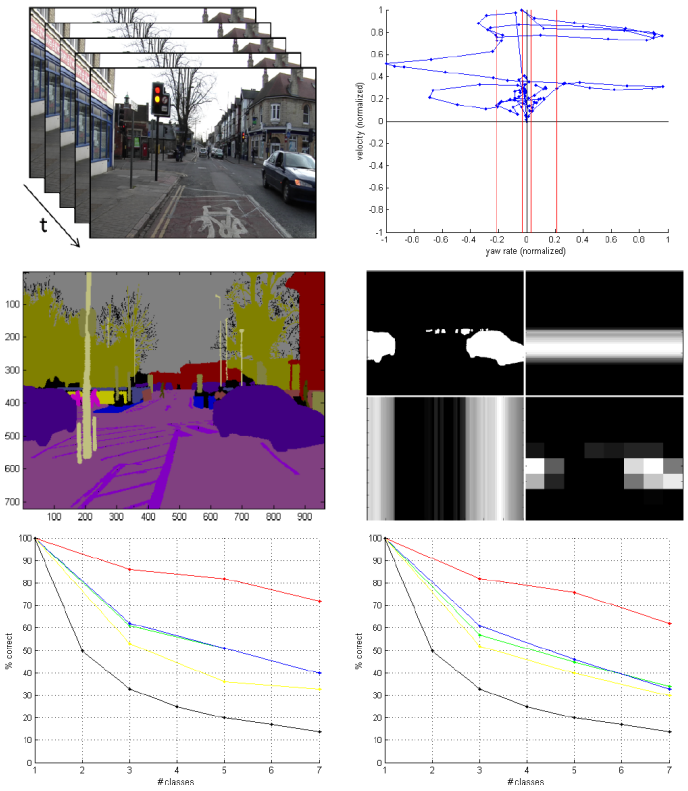

Figure 1: From an image sequence acquired by a car-mounted camera and the corresponding behavior of the driver (top row), our method exploits segmentation-based features (middle row) in order to learn to predict the appropriate velocity and yaw rate of the car for new images (bottom row).

Our evaluation on the CamVid dataset [1] considers 12 different object classes, matching those of [2], and consists of two steps. First, we use the CamVid ground truth segmentation as a basis for our behavior prediction (Fig. 1e). Second, we replace the ground truth segmentation by a realistic segmentation (Fig. 1f), using a Conditional Random Field with unary and pairwise potentials that incorporates color, edge, location and texton features. In both cases, we predict the velocity, yaw rate, velocity change and yaw rate change, defining a 3-, 5-, and 7-class scene categorization problem for each of these quantities by setting the appropriate thresholds (Fig. 1b). We learn a 3-, 5-, and 7-class GentleBoost classifier [3] with decision stumps on the feature vectors, built from one-versus-all classifiers, and determine the optimal number of decision stumps by 10-fold cross-validation on the individual CamVid sequences.

Our results indicate that the segmentation-based features [2] can be used to directly predict the appropriate driving behavior and that performance depends more on the features than on the segmentation quality.

[1] G. J. Brostow, J. Fauqueur, and R. Cipolla. Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters*, 30(2):88–97, 2009.

[2] A. Ess, T. Muller, H. Grabner, and L. van Gool. Segmentation-based urban traffic scene understanding. In *Proc. BMVC*, 2009.

[3] J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: A statistical view of boosting. *Annals of Statistics*, 28:337–407, 2000.

[4] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. 2003.