Vision on Wheels: Looking at Driver, Vehicle, and Surround for On-Road Maneuver Analysis

Eshed Ohn-Bar, Ashish Tawari, Sujitha Martin, and Mohan M. Trivedi
Computer Vision and Robotics Research Laboratory
University of California, San Diego
eohnbar@ucsd.edu, atawari@ucsd.edu, scmartin@ucsd.edu, mtrivedi@ucsd.edu

Abstract

Automotive systems provide a unique opportunity for mobile vision technologies to improve road safety by understanding and monitoring the driver. In this work, we propose a real-time framework for early detection of driver maneuvers. The implications of this study would allow for better behavior prediction, and therefore the development of more efficient advanced driver assistance and warning systems. Cues are extracted from an array of sensors observing the driver (head, hand, and foot), the environment (lane and surrounding vehicles), and the ego-vehicle state (speed, steering angle, etc.). Evaluation is performed on a real-world dataset with overtaking maneuvers, showing promising results. In order to gain better insight into the processes that characterize driver behavior, temporally discriminative cues are studied and visualized.

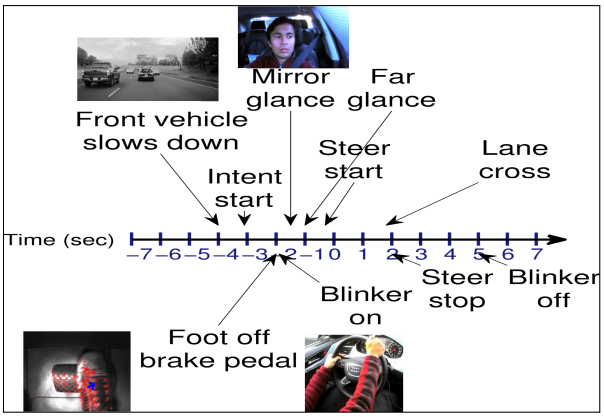


Figure 1. Timeline of an example overtake maneuver. We study the dynamics of several key variables that play a role in holistic understanding of overtake maneuvers. Driver monitoring could allow for more effective warning systems.

1. Introduction

Futuristic 'smart' cars as we envision will be equipped with advanced sensors including GPS (for navigation), cameras (for driver monitoring, lane detection), communications devices (vehicle-to-vehicle, vehicle-to-infrastructure), etc. along with networked mobile computing devices with ever increasing computational power. Automakers have come a long way in improving both safety and comfort of the car users. However, alarming crash statistics have kept safer and intelligent vehicle design an active research area. In 2012 alone, 33,561 people died in motor vehicle traffic crashes in the United States [1]. A majority of such accidents, over 90%, involved human error (i.e. inappropriate maneuver or a distracted driver). Advanced Driver Assistance Systems (ADAS) can mitigate such errors either by alerting the driver or even making autonomous corrections to safely maneuver the vehicle. Computer vision technologies, as non-intrusive means to monitor the driver, play an important role in the design of such systems.

In this work, we propose a holistic framework for real-time, on-road analysis of driver behavior in naturalistic real-world settings. Knowledge of the surround and vehicle dynamics, as well as the driver's state will allow the development of more efficient driver assistance systems. As a case study, we look into overtaking maneuvers in order to evaluate the proposed framework.

Lateral control maneuvers such as overtaking and lane changing contribute to a significant portion of the total accidents each year. Between 2004-2008, 336,000 such crashes occurred in the US [13]. Most of these occurred on a straight road at daylight, and most of the contributing factors were driver related (i.e. due to distraction or inappropriate decision making). This motivates studying a predictive system for such events, one that is capable of fully capturing the dynamics of the scene through an array of sensors. However, the unconstrained settings, the large number of variables, and the need for a low rate of false alarms and further distraction to the driver are challenging.

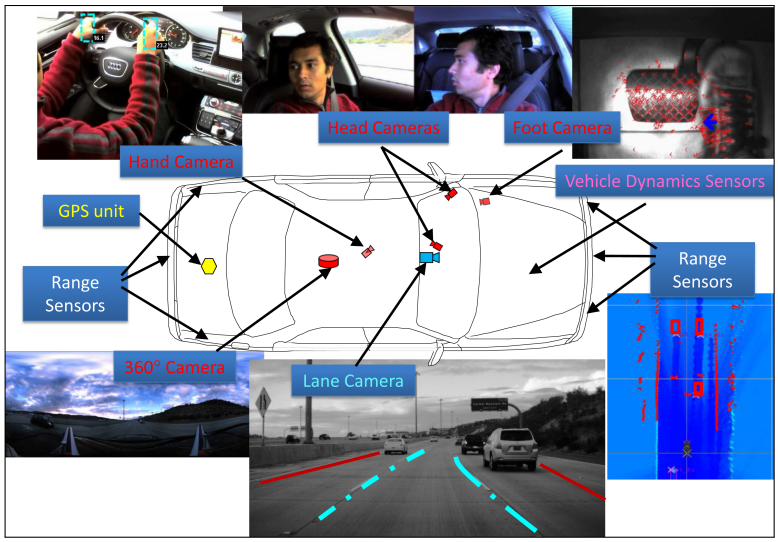


Figure 2. A holistic representation of the scene allows for prediction of driver maneuvers and inferring driver intent. Even a few hundred milliseconds of early identification of a dangerous maneuver could make roads safer and save lives. Best viewed in color.

2. Problem Statement and Motivation

Our goal is defined as follows: The early detection of an intended maneuver using driver, vehicle, and surround information.

As a case study, an on-road, naturalistic dataset of overtake maneuvers was collected. Fig. 1 illustrates the temporal evolution of different events in the course of a typical overtake maneuver, although the order and combination of the shown events may differ among different overtake maneuvers. First, the distance between the front and ego-vehicle may decrease, causing the driver to scan the surround (mirror and far glances). With the awareness that an option for a maneuver is possible, the driver may perform preparatory hand and foot gestures. Steering starts as the driver accelerates to the adjacent lane. The zero on the time axis marks the time of the beginning of the lateral motion. This temporal dissection of the overtake maneuver suggests that a rich set of information lies in the 3 components (i.e. driver, vehicle and surround) and their temporal analysis will benefit towards achieving our goal. The challenges, however, lie in the development of vision algorithms that work with high accuracy for detection of subtle movements, as well as robust to large illumination changes and occlusion.

A distributed camera network, see Fig. 2, is designed for this purpose. The requirement for robustness and real-time performance motivates us to study feature *representation* as well as techniques for *recognition* of key temporal events. The implications of this study are numerous. First, early warning systems could address critical maneuvers better and earlier. Knowledge of the state of the driver allows for customization of the system to the driver's needs, thereby avoiding further distraction caused by the system and easing user acceptance [9, 8]. On the contrary, a system which is not aware of the driver may cause annoyance. Additionally, under a dangerous situation (e.g. overtaking without turning on the blinker), a warning could be conveyed to other approaching vehicles (e.g. turning blinkers on automatically). Finally, in the process of studying the usability and the discriminative power of each of the cues alone and combined, we gain further insight into the underlying processes of driver behavior.

3. Instrumented Mobile Testbed

A uniquely instrumented testbed vehicle was used in order to holistically capture the dynamics of the scene: the vehicle dynamics, a panoramic view of the surround, and the driver. Built on a 2011 Audi A8, the automotive testbed has been outfitted with extensive auxiliary sensing for the research and development of advanced driver assistance technologies. Fig. 2 shows a visualization of the sensor array, consisting of vision, radar, LIDAR, and vehicle (CAN) data. The goal of the testbed buildup is to provide a near-panoramic sensing field of view for experimental data capture. The experimental testbed employs a dedicated PC, which taps all available data from the on-board vehicle systems excluding some of the camera systems which are synchronized using UDP/TCP protocols. On our dataset, the sensors are synchronized on average by 22ms or less.

For *sensing inside* the vehicle, two cameras for head pose tracking, one camera for hand detection and tracking, and one camera for foot motion analysis are used. For *sensing the surround* of the vehicle, a forward looking camera for

lane tracking is employed, as well as two LIDAR sensors (one forward and one facing backwards) and two radar sensors on either side of the vehicle. A Ladybug2 360° video camera (composed of an array of 6 individual rectilinear cameras) is mounted on top of the vehicle. Finally, information is captured from the CAN bus providing 13 measurements of the vehicle's dynamic state and controls, such as steering angle, throttle and brake, and vehicle's yaw rate.

4. Feature Extraction

In this section we detail the vision and other modules used in order to extract useful signals for analysis of activities.

4.1. Driver Signals

Head: Head dynamics are an important cue in prediction, as head motion may precede a maneuver in visually scanning for retrieving information about the environment. Unfortunately, many head pose trackers do not provide a large operational range, and may fail when the driver is not looking forward [19]. Therefore, we follow the setup of [19] where a two camera system provides a simple solution to mitigate the problem.

Head pose is estimated independently on each camera perspective from facial landmarks (i.e. eye corners, nose tip), which are detected using the supervised descent method [22], and their corresponding points on a 3D mean face model [19]. The system runs at 50 frames per second (fps). A one-time calibration is performed to transform head pose estimation among the respective camera coordinate system to a common coordinate system.

Hand: The hand signal may provide information on preparatory motions before a maneuver is performed. Hand detection is a difficult problem in computer vision, due to the hand's tendency to occlude itself, deform, and rotate, producing a large variability in its appearance [14, 16]. We use integral channel features [7] which are fast to extract. Specifically, for each patch extracted from a color image, gradient channels (normalized gradient channels at six orientations and three gradient magnitude channels) and color channels (CIE-LUV color channels were experimentally validated to work best compared to RGB or HSV) are extracted. 2438 instances of hands were annotated, and an AdaBoost classifier with decision trees as the weak classifiers is used for learning [23]. The hand detector runs at 30 fps on a CPU. For non-maximal suppression, a 0.2 threshold is used. In order to differentiate the left hand from the right hand and prune false positives, we train a histogram of oriented gradients (HOG) with a support vector machine (SVM) detector for post-processing of the hypothesized hand bounding boxes provided by the hand detector. A Kalman filter is used for tracking.

Foot: One camera is used to observe the driver's foot behavior near the brake and throttle pedal. Due to lack of lighting, an illuminator is used. While embedded pedal sensors already exist to indicate when the driver is engaging any of the pedals, vision-based foot behavior analysis has additional benefits of providing foot movements before and after pedal press. Such analysis can be used to predict a pedal press before it is registered by the pedal sensors.

An optical flow (iterative pyramidal Lucas-Kanade, running at 30 fps) based motion cue is employed to determine the location and magnitude of relatively significant motions in the pedal region. Optical flow is a natural choice for analyzing foot behavior due to little illumination changes and the lack of other moving objects in the region. First, optical flow vectors are computed over sparse interest points, detected using Harris corner detection. Second, a majority vote over the computed flow vectors reveals an approximate location and magnitude of the global flow vector. Optical flow-based foot motion analysis have been used in [21] for prediction of pedal presses.

4.2. Vehicle Signals

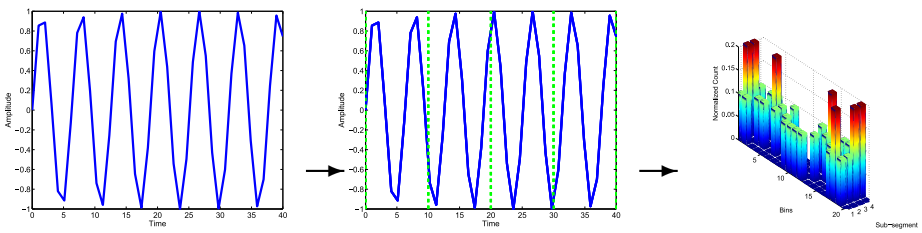
Commonly, analysis of maneuvers is made with trajectory information of the ego-vehicle [4, 10, 11, 2, 3]. In this work, the dynamic state of the vehicle is measured using the CAN bus, which supplies 13 parameters ranging from blinkers to the vehicle's yaw rate. In understanding and predicting the maneuvers in this work, we only use steering wheel angle information (important for analysis of overtake events), vehicle velocity, and brake and throttle paddle information.

4.3. Surround Signals

Lidar/Radar: Prediction of maneuvers can consider the trajectory of other agents in the scene [17]. This is important for our case study, as a driver may choose to overtake a vehicle in its proximity. Such cues are studied using an array of range sensors that track vehicles in terms of their position and relative velocity. A commercial object tracking module [20] tracks and re-identifies vehicles across LIDAR and radar systems providing vehicle position and velocity in a consistent global frame of reference. In this work, we only consider trajectory information (longitudinal and lateral position and velocity) of the forward vehicle.

Lane: Lane marker detection and tracking [18] is performed on a front-observing gray-scale camera (see Fig. 2). The system can detect up to four lane boundaries. This includes the ego-vehicle's lanes and its two adjacent lanes. The signals we consider are the vehicle's lateral deviation (position within the lane) and lane curvature.

A 360° panoramic image collects visual data of the surround. It is the composed view of six cameras, and used for annotation and offline analysis.



Raw Signal | Split into k sub-segments ($k = 4$ in this case) | $k \times (\#bins)$ histogram descriptor

Figure 3. Two features used in this work: raw trajectory features outputted by the detectors and trackers, and histograms of sub-segments of the signal.

4.4. Time-Series Features

We compare two types of temporal features derived from the aforementioned signals. For each of the signals at each time, f_t, we may simply use a concatenation of the signal in a time window of size L,

$$\mathbf{F}_t = (f_{t-L+1}, \ldots, f_t) \tag{1}$$

The time window in our experiments is fixed at three seconds. In the second set of features, the windowed signal $\mathbf{F}_t$ is split into k equal sub-signals first, followed by a construction of a histogram of each of these sub-signals separately (depicted in Fig. 3). Such a partitioning aims to preserve temporal information. We experimented with $k = 1, 2, 4, 8$ and found that using features of up to $k = 4$ (combined splits used are at levels 1, 2, and 4) worked well with no advantage in increasing the number of sub-segments further. Therefore, this partitioning is used in all the experiments.

5. Temporal Modeling

Given a sequence of observations from Eq. 1, $\mathbf{x} = \{\mathbf{F}_t^{(1)}, \ldots, \mathbf{F}_t^{(c)}\}$, where c is the total number of signals, the goal is to learn a mapping to a sequence of labels.

One approach to capturing signal temporal structure involves using a Conditional Random Field (CRF) [12]. CRF has been shown to significantly outperform its generative counterpart, the Hidden Markov Model [12]. Nonetheless, CRF on its own may not capture sub-structure in the temporal data well, which is essential for our purposes. By employing latent variables, the Latent-Dynamic CRF (LDCRF) [12, 15] improves upon the CRF and also provides a segmentation solution for a continuous data stream.

When considering the histogram features studied in this work, we model each bin as a variable in the LDCRF framework. In this case, temporal structure is measured by the evolution of each bin over time (20 bins are used for each histogram). Possibly due to the increase in dimensionality and the already explicit modeling of temporal structure in the model, using raw features was shown to work as good or better than histogram features for the LDCRF model.

A second approach for temporal modeling is motivated by the large number of incoming signals from a variety of modalities. Fusion of the signals can be performed using Multiple Kernel Learning (MKL) [5].

Given a set of training instances and signal channel c_l, a kernel function is calculated for each channel, $\kappa_{c_l}(\mathbf{x}_i, \mathbf{x}_j) : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ (d is the feature dimension and $\mathbf{x}_i, \mathbf{x}_j$ are two data points). Denote $\{\mathbf{K}^{c_l} \in \mathbb{R}^n \times \mathbb{R}^n, l = 1, \ldots, s\}$ as the collection of s kernel matrices for the data points in the training set, so that $K_{ij}^{c_l} = \kappa_{c_l}(\mathbf{x}_i, \mathbf{x}_j)$. In our implementation, Radial Basis Function (RBF) kernels are derived from each signal, $\kappa(\mathbf{x}_i, \mathbf{x}_j) = \exp(-||\mathbf{x}_i - \mathbf{x}_j||/\gamma)$. The cost and spread parameters are found for each signal using grid search.

For combining the kernels, the goal is to learn a probability distribution $\mathbf{p} = (p^1, \ldots, p^s)$, with $p \in \mathbb{R}_+$ and $\mathbf{p}^T\mathbf{1} = 1$, for finding an optimal combination of kernel matrices,

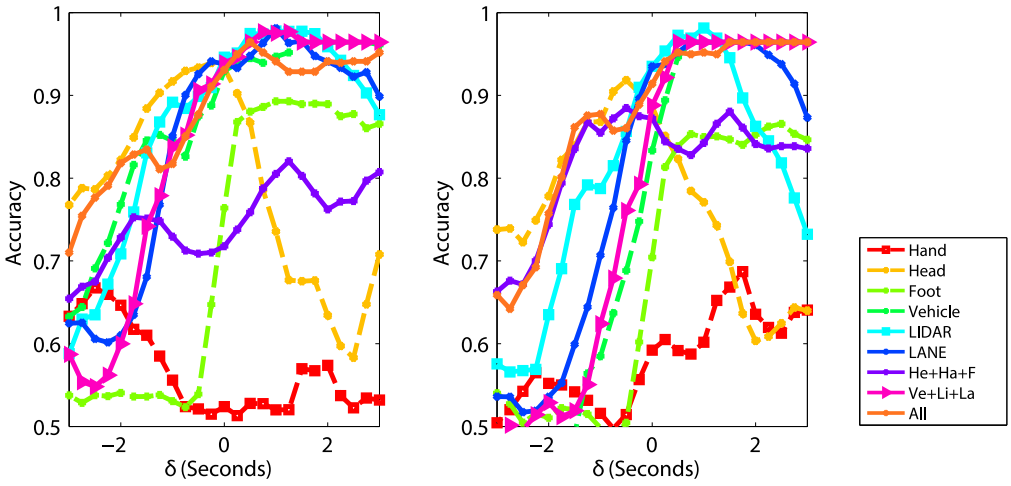
$$\mathbf{K}(\mathbf{p}) = \sum_{l=1}^{s} p^l \mathbf{K}^{c_l} \tag{2}$$

Stochastic approximation is used to learn the weights $\mathbf{p}$ as in [5] with LIBSVM [6].

The histogram features were shown to work well with the MKL features, performing better than simply using the raw features.

6. Experimental Evaluation

Experimental settings: As a case study of the proposed approach for maneuver analysis and prediction, 54 minutes of video containing 78,018 video frames was used (at 25 frames per second). 1000 events of normal driving (each defined in a three second window leading to about 75,000 frames total) were chosen randomly, and 13 with overtaking instances were annotated (a total of 975 frames). Training and testing is done using a 2-fold cross validation. Overtake events were annotated when the lane crossing occurred.



(a) LDCRF (b) MKL

Figure 4. Classification and prediction of overtake/no-overtake maneuvers using LDCRF (raw trajectory features) and MKL (histogram features). He+Ha+F stands for the driver observing cues head, hand, and foot. Ve+Li+La is vehicle, LIDAR, and lane. 'all' comprises of all of the individual cues.

Temporal Modeling: The comparison between the two techniques studied in this paper is shown in Fig. 4. As mentioned in Section 5, LDCRF benefits from the raw signal input, as opposed to treating each bin in the histogram features as a variable. On the contrary, MKL significantly benefits from the histogram features as it lacks a state model and the histogram level pyramid provides distinct temporal structure patterns. In order to visualize the discriminative effect of each cue, a model is learned for each specific cues and then for different combinations. Generally, we notice how the vehicle and surround cues tend to spike later into the maneuver. This can be seen by comparing the 'Ve+Li+La' (vehicle, LIDAR, and lane) curve with the 'He+Ha+F' (driver observing cues, head, hand, and foot). An important observation is that although the trends appear similar in the two temporal modeling techniques, the fusion results differ significantly. For instance, using all the features results in a significantly higher prediction at $\delta = -1$ in MKL when compared to LDCRF. Nonetheless, LDCRF appears to be better at capturing dynamics for individual cues.

Features: Fig. 5 depicts the temporal evolution of cue importance using the weight outputs from the MKL framework. Successful cues will correspond to a heavier weight, and cues with little discriminative value will be reduced in weight. To produce this plot, we learn a model using the specific set of cues (driver, vehicle, or surround cues) for each δ time before the maneuver. This provides the kernel weights which are plotted. We observe how driver-related cues are strongest around the time that the lateral motion begins (t=0). After the steering began, there is a shift to the surround cues, such as lane deviation. The results affirm the approach for describing a maneuver using a set of holistic features.

7. Concluding Remarks

Modern automotive systems provide a novel platform for mobile vision application with unique challenges and constraints. In particular, driver assistance systems must perform under time-critical constraints, where even a few hundred milliseconds are essential. A holistic and comprehensive understanding of the driver's intentions can help in gaining crucial time and in saving lives. This shifts the focus towards studying maneuver dynamics as they evolve over longer periods of time. Prediction of overtake maneuvers was studied using information fusion from an array of sensors, required to fully capture the development of complex temporal inter-dependencies in the scene. Evaluation was performed on naturalistic driving showing promising results for prediction of overtaking maneuvers. Having an accurate head pose signal with the combination of other surround cues proved key to early detection.

Acknowledgments

The authors would like to thank UC Discovery program and industry partners, especially Audi and VW Electronic Research Laboratory for their support and collaboration. We also thank the colleagues at the Laboratory for Intelligent and Safe Automobiles for their valuable assistance.

References

[1] 2012 motor vehicle crashes: overview. Technical Report DOT HS 811 856, National Highway Traffic Safety Admin-

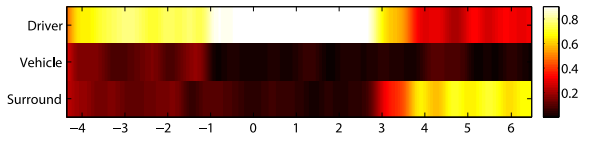


Figure 5. Kernel weight associated with each cue learned from the dataset with MKL (each column sums up to one). Characterizing a maneuver requires cues from the driver (hand, head, and foot), vehicle (CAN), and the surround (LIDAR, lane, visual-color changes). Here, in order to fully capture each maneuver, time 0 for overtake is defined the beginning of the lateral motion, and not at the crossing of the lane marker.

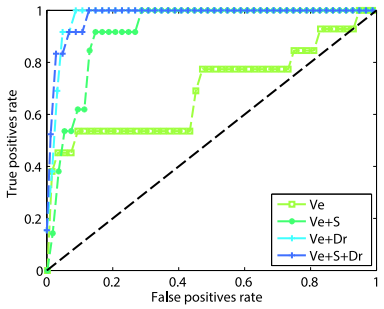


Figure 6. For a fixed prediction time, $\delta = -2$ seconds, we show the effects of appending cues to the vehicle (**Ve**) dynamics. **S** stands for surround (LIDAR and lane). **Dr** stands for driver (hand, head, and foot).

istration, Washington, D.C., 2013.

[2] A. Armand, D. Filliat, and J. Ibañez-Guzmàn. Modelling stop intersection approaches using gaussian processes. In *IEEE Conf. Intelligent Transportation Systems*, 2013.

[3] T. Bando, K. Takenaka, S. Nagasaka, and T. Taniguchi. Automatic drive annotation via multimodal latent topic model. In *IEEE Conf. Intelligent Robots and Systems*, 2013.

[4] S. Bonnin, T. H. Weisswange, F. Kummert, and J. Schm´udderich. Accurate behavior prediction on highways based on a systematic combination of classifiers. In *IEEE Intelligent Vehicles Symposium*, 2013.

[5] S. Bucak, R. Jin, and A. K. Jain. Multi-label multiple kernel learning by stochastic approximation: Application to visual object recognition. In *Advances in Neural Information Processing Systems*, 2010.

[6] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:1–27, 2011.

[7] P. Dollár, R. Appel, S. Belongie, and P. Perona. Fast feature pyramids for object detection. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2014.

[8] A. Doshi, B. T. Morris, and M. M. Trivedi. On-road prediction of driver's intent with multimodal sensory cues. *IEEE Pervasive Computing*, 10:22–34, 2011.

[9] A. Doshi and M. M. Trivedi. Tactical driver behavior prediction and intent inference: A review. In *IEEE Conf. Intelligent Transportation Systems*, 2011.

[10] S. Lefèvre, C. Laugier, and J. Ibañez-Guzmán. Exploiting map information for driver intention estimation at road intersections. In *IEEE Intelligent Vehicles Symposium*, 2011.

[11] M. Liebner, F. Klanner, M. Baumann, C. Ruhhammer, and C. Stiller. Velocity-based driver intent inference at urban intersections in the presence of preceding vehicles. *IEEE Intelligent Transportation Systems Magazine*, 2013.

[12] L. P. Morency, A. Quattoni, and T. Darrell. Latent-dynamic discriminative models for continuous gesture recognition. In *IEEE Conf. Computer Vision and Pattern Recognition*, 2007.

[13] W. G. Najm, R. Ranganathan, G. Srinivasan, J. D. Smith, S. Toma, E. Swanson, and A. Burgett. Description of light-vehicle pre-crash scenarios for safety applications based on vehicle-to-vehicle communications. Technical Report DOT HS 811 731, National Highway Traffic Safety Administration, Washington, D.C., 2013.

[14] E. Ohn-Bar, S. Martin, and M. M. Trivedi. Driver hand activity analysis in naturalistic driving studies: Issues, algorithms and experimental studies. 22:1–10, 2013.

[15] E. Ohn-Bar, A. Tawari, S. Martin, and M. M. Trivedi. Predicting driver maneuvers by learning holistic features. In *IEEE Intelligent Vehicles Symposium*, 2014.

[16] E. Ohn-Bar and M. M. Trivedi. The power is in your hands: 3D analysis of hand gestures in naturalistic video. In *IEEE Conf. Computer Vision and Pattern Recognition Workshops*, 2013.

[17] M. Ortiz, F. Kummert, and J. Schmudderich. Prediction of driver behavior on a limited sensory setting. In *IEEE Conf. Intelligent Transportation Systems*, 2012.

[18] S. Sivaraman and M. M. Trivedi. Integrated lane and vehicle detection, localization, and tracking: A synergistic approach. *IEEE Trans. Intelligent Transportation Systems*, 14:906–917, 2013.

[19] A. Tawari, S. Martin, and M. M. Trivedi. Continuous head movement estimator (CoHMET) for driver assistance: Issues, algorithms and on-road evaluations. *IEEE Trans. Intelligent Transportation Systems*, 15:818–830, 2014.

[20] A. Tawari, S. Sivaraman, M. M. Trivedi, T. Shannon, and M. Tippelhofer. Looking-in and looking-out vision for urban intelligent assistance: Estimation of driver attentive state and dynamic surround for safe merging and braking. In *IEEE Intelligent Vehicles Symposium*, 2014.

[21] C. Tran, A. Doshi, and M. M. Trivedi. Modeling and prediction of driver behavior by foot gesture analysis. *Computer Vision and Image Understanding*, 116:435–445, 2012.

[22] X. Xiong and F. D. la Torre. Supervised descent method and its application to face alignment. In *IEEE Conf. Computer Vision and Pattern Recognition*, 2013.

[23] C. Zhang and P. A. Viola. Multiple-instance pruning for learning efficient cascade detectors. In *Advances in Neural Information Processing Systems*, 2007.