Predicting Driver Maneuvers by Learning Holistic Features

Eshed Ohn-Bar, Ashish Tawari, Sujitha Martin, and Mohan M. Trivedi

***Abstract*— In this work, we propose a framework for the recognition and prediction of driver maneuvers by considering holistic cues. With an array of sensors, driver's head, hand, and foot gestures are being captured in a synchronized manner together with lane, surrounding agents, and vehicle parameters. An emphasis is put on real-time algorithms. The cues are processed and fused using a latent-dynamic discriminative framework. As a case study, driver activity recognition and prediction in overtaking situations is performed using a naturalistic, on-road dataset. A consequence of this work would be in development of more effective driver analysis and assistance systems.**

I. INTRODUCTION

On-road driving behavior consists of intricate, multi-dimensional dynamics. It is the outcome of many variables that interact at a certain place and time. This motivates the study in this paper, which pursues a holistic approach using multiple cues in the scene. When fusing the different modalities in a temporal modeling framework, activity recognition and critical situation prediction can be made more effectively. Furthermore, through studying of temporally discriminative cues, we gain insight into the processes that characterize driver behavior.

One important implication of this work is in the development of driver assistance systems. For instance, human-observing cameras can perceive visual scanning and preparatory movements preceding a maneuver, giving a larger margin of time for prevention of a critical situation. Furthermore, a benefit of integrating driver gestures is in the ability to observe driver *intent* for performing a maneuver. Such knowledge can be incorporated into assitive technologies, which in turn may generate a more effective warning system under unintentional maneuvers (e.g. increased lane deviation), while assessing the need for a warning under intentional maneuvers. In this work, driver, vehicle, and environment cues are modeled to produce prediction of activities. In particular, overtaking event prediction will be studied to show the usefulness in the synergistic approach.

Many current safety systems sense the environment in order to produce a warning to the driver. We motivate the usage of additional cues-in particular, driver-based cues. In 2012 alone, 33,561 people died in motor vehicle traffic collisions in the United States [1]. A majority of such accidents involved an inappropriate maneuver or a distracted driver. Lateral control maneuvers such as overtaking and lane changing represent a significant portion of the total accidents each year. Between 2004-2008, 336,000 such crashes occurred in the US [2]. The major contributing factors are driver related (i.e. due to distraction or inappropriate decision making). Therefore, robust vision systems can be employed to detect driver motion patterns, such as head scanning or pre-control actions with foot or hand, and better mitigate critical situations. This work deals with such analysis in a

E. Ohn-Bar, S. Martin, A. Tawari, and M. Trivedi are with the Laboratory for Intelligent and Safe Automobiles (LISA) at the Department of Electrical and Computer Engineering, University of California, San Diego, La Jolla, CA 92093, USA {eohnbar, scmartin, atawari, mtrivedi}@ucsd.edu

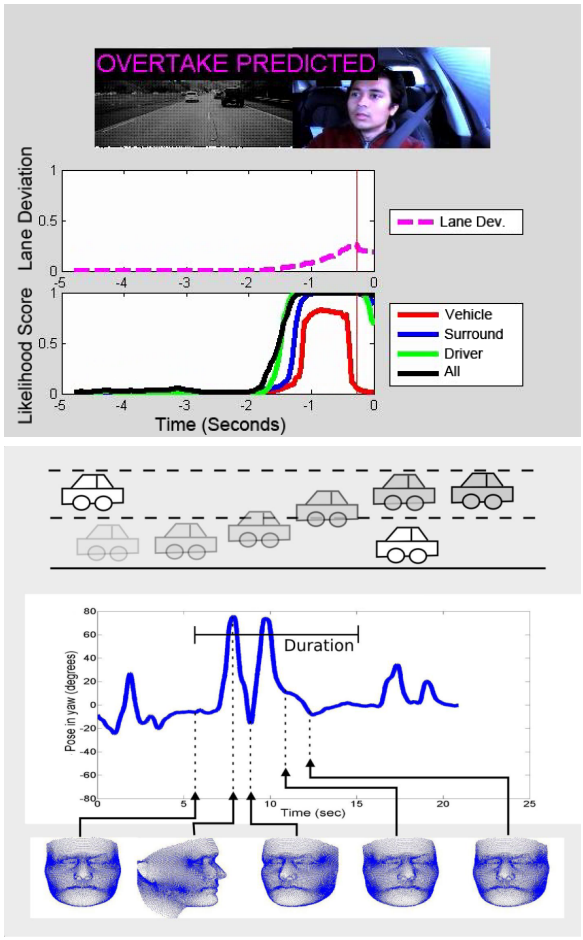


Fig. 1: Different cues are useful for analysis and prediction at different times of the maneuver. Top: An event from our dataset, with lane crossing marked in a vertical red line, and confidence scores from the model. Notice the spike occurring before the event, providing prediction. Bottom: Learning holistic features. For an overtake maneuver, visual scanning and head motion provides predictive value, especially when coupled with surround and vehicle dynamics cues.

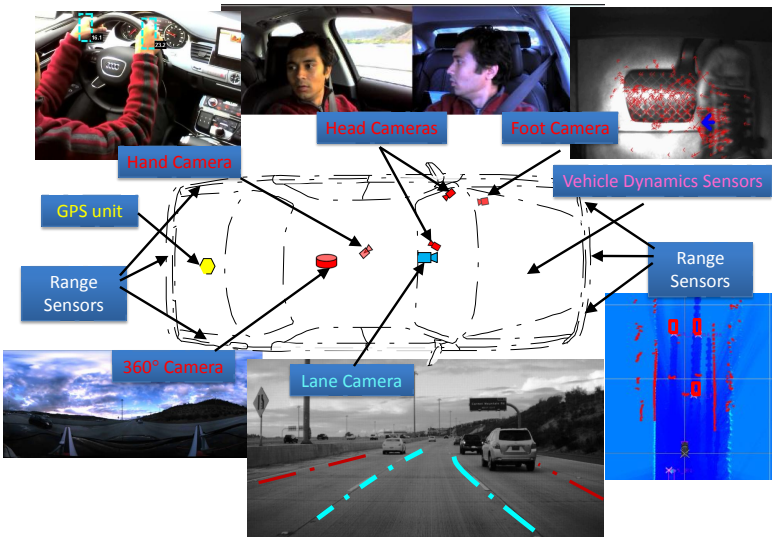


Fig. 2: Safety and driving behavior analysis using a multi-dimensional, holistic framework. The combined system can be used to predict events by a few seconds before they occur, thereby making roads safer and saving lives.

real-time action recognition framework.

First, in Section III we describe the testbed, and Section IV contains the vision algorithms used for signal extraction. The temporal formulation is details in Section V, and experimental analysis on the real-world driving dataset is in Section VII.

II. Related Research

Related research on maneuver recognition and prediction may differ based on cues used, the time-scale of the events of interest, or the type of maneuver studied.

In [3], vehicle dynamics measurements from the CAN-bus are coupled with laser data for preceding vehicles cues to predict driver behavior in a three second window. Yaw and steering-wheel angles were used in [4] for lane departure prediction. In [5] vehicle velocity, steering, and GPS were used to model lane-change behavior and predict collision trajectories. CAN parameters and a Gaussian Mixture Model were used in [6] to study car-following. Vehicle dynamics are used in [7] to identify driver intent at intersections and recognize actions. Vehicle trajectories were used in [8] to model driver intent at intersections using a Probabilistic Finite State Machine. Although methods commonly incorporate ego-vehicle dynamics, others may take on a purely vision approach. In [9], monocular video is used to predict driving behavior in urban environments. A front looking camera was used in [10] to predict future lane change behavior. Recognition of events was done using ego-vehicle dynamics and lane information in [11]. In our work, surrounding agents are modeled by using lidar cues, and a camera is used to extract lane parameters of the scene.

A close effort to our work can be found in [12], where radar, lane, head, and vehicle cues are integrated using a Relevance Vector Machine in order to recognize and predict lane change maneuvers in a two seconds window. A main difference is in the temporal modeling, as well as the cue representation. For instance, we produce a multi-level representation of the driver's state using head, hand, and foot motion cues. These cues provide additional context information for upcoming maneuvers. Furthermore, the temporal segmentation of actions is done automatically using the model.

Due to the holistic framework, it will be shown that prediction can be accomplished much earlier than in the aforementioned works. The event definition will be explored to highlight the earlier prediction. Generally, the aforementioned work predicts a maneuver after it has began using trajectory forecasting approaches. Nonetheless, we point out that the intent to perform the maneuver existed before the trajectory of the vehicle was altered and can be observed earlier. Within such an early time, a different set of cues may be useful for prediction, these are driver-based.

The closest work to ours of Doshi *et al.* [13] defines a lane change at the maximum lane deviation (i.e. when the vehicle crosses the lane marker). Nonetheless, the driver had the intent to change lanes much earlier, even before any lane deviation occurred. We therefore experiment with an alternate definition (used in [14]), which is at the beginning of the lateral movement. Both definitions will be studied in this work.

III. Experimental Setup: Testbed and Dataset

A 2011 Audi A8 was uniquely instrumented in order to capture the scene holistically: the dynamics of the ego-vehicle, surround agents and road geometry, and the driver's state. Fig. 2 shows a visualization of the sensor array, consisting of vision, radar, lidar, and vehicle CAN data. An on-board PC provides the computational resources. Sensor

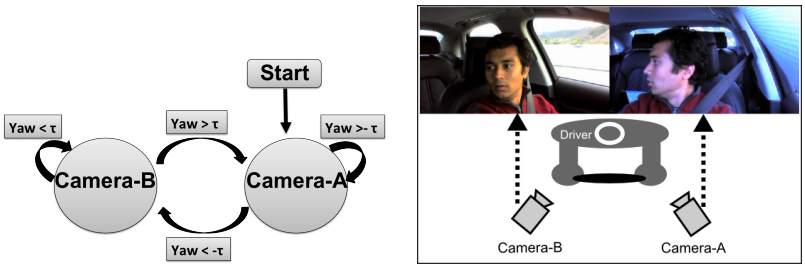


Fig. 3: A two camera system overcomes challenges in head pose estimation and allow for continuous tracking even under large head movements.

data from the radars and lidars are fused to generate a near-panoramic sensing of surrounding objects. These are tracked and associated using a module developed by Audi. UDP/TCP protocols are employed for synchronization of some of the sensors with the main PC. On our dataset, the sensors on average are synchronized up to 22ms.

The array of sensors includes vision-based, non-intrusive driver observing cameras. Two cameras for head pose tracking are used. These are aimed to capture head gestures, such as visual scanning. Two additional cameras provide inputs regarding gestures related to vehicle control. One camera is utilized for hand detection and tracking, and another camera for foot motion analysis.

For sensing the surround, a forward looking camera for lane tracking. 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) on top of the vehicle. Combined, these sensors allow for comprehensive and accurate on-line and off-line analysis and annotation.

The sensors are integrated into the vehicle body or placed in non-distracting regions to ensure minimal distraction while driving. Vehicle parameters are recorded into 13 measurements, such as steering angle, throttle and break, and vehicle's yaw rate.

In order to study the feasibility of the proposed framework, a dataset of 54 minutes of video containing 78,018 video frames was used (at 25 frames per second). All results reported employ 2-fold cross validation, with half of the data used for training and the rest for testing. Overall, we use 1000 normal events driving (each defined in a three second window leading to about 75,000 frames) as well as 13 overtaking instances (975 frames) which occurred throughout the video.

IV. Holistic Representation of Maneuvers

The features used to represent vehicle, surround, and driver, are detailed below. The output from each vision algorithm provides a time-series which is used in the temporal modeling for activity prediction. In this work, the panoramic 360° camera (composed of 6 individual streams) was used for annotation and offline analysis.

A. Head Pose Signal

Real-time, robust head pose is key to early prediction. The head provides a different set of cues (compared to hand and foot) because it is used by drivers for visual scanning and the retrieval of information from the surround. For instance, head motion may precede an overtaking maneuver in order to scan for an available space in the adjacent lane. On the other hand, controlling-gestures are associated with the foot and hand signals. These occur with the driver intention to operate a controller in the vehicle, such as in turning on a lane change indicator.

For capturing the wide motion of the head under such critical situations, we use a spatially distributed set of two cameras around the driver, as in [15].

First, head pose is estimated independently on each camera perspective from some of the least deformable facial landmarks (i.e. eye corners, nose tip), which are detected using supervised descent method [16], and their corresponding points on a 3D mean face model. The system runs at 50Hz. It is important to note that head pose estimation from each camera perspective is with respective camera coordinates. One-time calibration is performed to transform head pose estimation from respective camera coordinates to a common coordinate where a yaw rotation angle equal to, less than and greater than 0° represent the driver looking forward, rightward and leftward, respectively.

A simple camera selection module is used over the wide operational range in the yaw rotation angle, as shown in Fig. 3. In order to handle camera selection and hand-off, we use the yaw angle of the head.

B. Hand Location Signal

Hand gestures are incorporated in order to study preparatory motions before a maneuver is performed. Below, we specify the hand detection and tracking module. Hand detection is a difficult problem in computer vision, due to the hand tendency to occlude itself, deform, and rotate, producing a large variability in its appearance. In [17], motion and appearance cues were studied for hand activity classification in on-road data. As mentioned in [18], hand detection is particularly difficult in the car due to illumination variation. Therefore, we turned to training a hand detector on data from the same vehicle to maximize robustness to background artifacts. We use the fast to compute integral

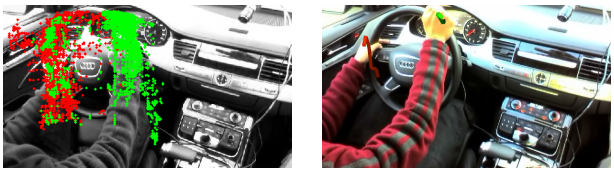


Fig. 4: Scatter plot of left (in red) and right (in green) hand detection for the entire drive. A hand trajectory of reaching towards the signal before an overtake is shown (brighter is later in time).

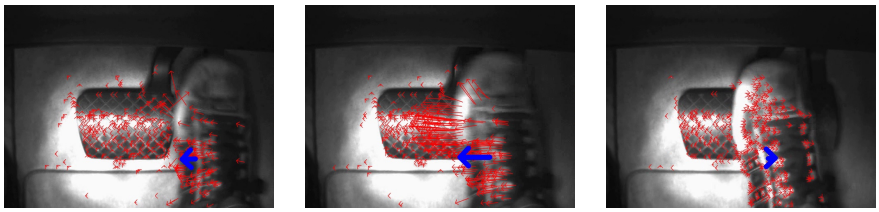


Fig. 5: Foot tracking using iterative pyramidal Lucas-Kanade optical flow. Features of location and velocity are extracted by majority vote.

channel features [19]. 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 and HSV) are extracted. 2438 instances of hands were annotated. An AdaBoost classifier [20] was learned over 2000 level-2 decision trees containing 3 stumps. Bootstrapping was performed, with the first stage sampling 5000 random negative samples, and two additional stages of training using hard negatives.

The hand detector runs at 30Hz on a CPU. We noticed many of the false detections occurring in the proximity of the actual hand (the arm, or multiple detections around the hand), hence we used a non-maximal suppression with a 0.2 threshold. In order to differentiate the left from the right hand, we train a histogram of oriented gradients (HOG) with a support vector machine (SVM) detector. A Kalman filter is used for tracking the hands.

C. Foot Motion Features

Fist, low-level cues of the driver's foot motion patterns are extracted. Such motion cues can provide a few hundred milliseconds in prediction by providing information on before and after pedal presses [21]. These also provide additional context when combined with the other modalities. Therefore, such analysis can 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 30Hz) based motion cues is employed to determine the location and magnitude of relatively significant motions in the pedal region. Motion cues are robust for analyzing foot behavior because of little to no illumination changes and the lack of other moving objects in the region. First, optical flow vectors are computed over sparse interest points, which can be detected using Harris corner detection in the image plane. Second, a majority vote over the computed flow vectors reveals an approximate location and magnitude of the global flow vector.

D. Lidar and Radar Surround Features

The maneuvers we study correlate with surrounding events. Such cues are studied using an array of range sensors that track vehicles in term of their position and relative velocity. The sensor-fusion module, developed by Audi, tracks and re-identifies vehicles across the sensor modalities, allowing for hand-offs between the lidar and radar systems, tracking vehicles 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.

E. Lane Signal

A front-observing gray-scale camera is used for lane marker detection and tracking using a built-in system. The system can detect up to four lane boundaries. This includes the ego-vehicle's lanes and two adjacent lanes to those. The signals we consider are the vehicle's lateral deviation (position within the lane) and lane curvature.

F. Ego-Vehicle Dynamics

The dynamic state of vehicle is measured using a CAN bus, supplying 13 parameters 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.

G. Trajectory Features

We use trajectory features for each of the signals outputted by one of the sensors above at each time, f_t.

First, we only consider a part of the signal in a time window of size L,

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

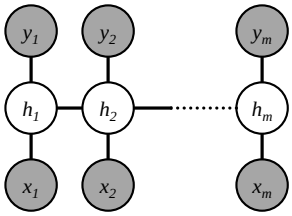


Fig. 6: The Latent-Dynamic Conditional Random Field (LDCRF) model, observations are denoted by x_i, labels as y_i, and h_i is a hidden state coupled with observation x_i.

The time window in our experiments is fixed at three seconds. Each signal is then converted to a histogram descriptor by a splitting of the incoming signal into bins, and consequently quantization of the signal amplitude in each bin. For instance, for 10 bins and 4 sub-segment splits, we get a 40 dimensional histogram representation.

V. Temporal Modeling

For modeling of the temporal dynamics, we use a Latent-Dynamic Conditional Random Field (LDCRF) model [22]. LDCRF is a discriminative approach, learning a latent structure for differentiation of activity classes (unlike a generative approach such as a Hidden Markov Model). The model provides advantages over both CRF and Hidden-state CRFs, as it learns internal sub-structures unique for each maneuver and also addresses the automatic segmentation of the data.

In order to recognize and predict events in the temporal context of driving, a set of labels $\mathbf{y} = \{y_1, y_2, \ldots, y_m\}$ is coupled with an observed sequence of signals, $\mathbf{x} = \{x_1, x_2, \ldots, x_m\}$. Each frame produces an observation, x_i, and is associated with label y_i. The labels are drawn from a pre-defined set, for instance $Y = \{lanekeeping, overtake, \ldots\}$. To capture sub-structure in class sequences, hidden variables $\mathbf{h} = \{h_1, h_2, \ldots, h_m\}$ are introduced, so that the conditional probability is

$$P(\mathbf{y}|\mathbf{x}; \Lambda) = \sum_{\mathbf{h}:\forall h_i \in H_{y_i}} P(\mathbf{h}|\mathbf{x}; \Lambda) \tag{2}$$

where, as in [22], each class label has a disjoint set of associated hidden states by assumption. Λ is the set of weight parameters. Under a simple chain assumption, in the conditional random field this joint distribution over $\mathbf{h}$ has an exponential form,

$$P(\mathbf{h}|\mathbf{x}; \Lambda) = \frac{\exp\left(\sum_k \Lambda_k \cdot \mathbf{F}_k(\mathbf{h}, \mathbf{x})\right)}{\sum_{\mathbf{h}} \exp\left(\sum_k \Lambda_k \cdot \mathbf{F}_k(\mathbf{h}, \mathbf{x})\right)} \tag{3}$$

We follow [22], where function $\mathbf{F}_k$ is defined as a sum of state (vertex) or transition (edge) feature functions,

$$\mathbf{F}_k(\mathbf{h}, \mathbf{x}) = \sum_{i=1}^{m} l_k(h_{i-1}, h_i, \mathbf{x}, i) \tag{4}$$

The model parameters are learned using log-likelihood maximization with gradient ascent. In testing time, the most probable sequence of labels is the one that maximizes Eqn. 2. This is done using marginal probabilities [22].

VI. Experiment Settings

We perform two experiments to test the proposed framework for recognition and prediction of maneuvers. As mentioned in Section II, we experiment with two definitions of the beginning of an overtake event. An overtake event may be marked when the vehicle crosses the lane marking or when lateral movement began. These are referred to as **overtake-late** and **overtake-early**, respectively. Normal driving is defined as events where the paddle break was not engaged or significant lane deviation occurred, but the driver was simply keeping within the lanes. Furthermore, we do not require a minimum speed for the events. Normal and overtake events may occur without a minimum speed threshold.

We are concerned with how each of the above events is characterized compared to normal driving.

- **Experiment 1**: Overtake-late events (at lane crossing) vs. normal driving events
- **Experiment 2**: Overtake-early events (at initial lateral motion) vs. normal driving events

VII. Experimental Evaluation

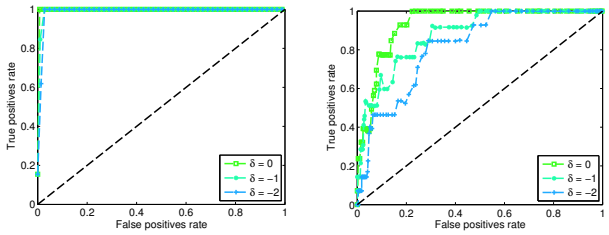
We first verify that training on overtake-late events and testing on overtake-late events results in high recognition rates, as shown in Fig. 7. Furthermore, training and testing on earlier maneuver cues also showS promise, but is a significantly more challenging problem. Prediction of an early overtake maneuver of even a short amount, as shown in the plots, could provide a great advantage towards more effective driver assistance systems. We note that in testing for early-maneuver recognition, the models are re-trained for that specific temporal window to best accommodate the dynamics of that period of time into the maneuver.

In Fig. 8, we plot the advantage of fusion over the three components of driving: the driver, the vehicle, and the surround. Each cue is measured in terms of its predictive value two seconds before an event.

VIII. Concluding Remarks

In this work, a holistic learning framework allowed for the early prediction of driver maneuvers. Components in the framework were studied in term of their predictive power, which is essential for a driver assistance systems where even several hundred milliseconds may be critical.

In order to fully capture the development of complex temporal inter-dependencies in the scene, robust cues at multiple levels of activity were extracted in real-time. Testing was performed on an on-road, naturalistic driving sequence. Having a clear head pose signal with combination of other surround cues proved key to the early observation of behavior and intent. The framework was used to produce predictions earlier than in existing literature. In the future, additional maneuver types will be studied, such as at intersections.



(a) (b)

Fig. 7: Measuring prediction by varying the time in seconds before an event, δ. Results are shown for the fusion of all the cues. (a) Experiment 1: Overtake-late vs. normal (b) Experiment 2: Overtake-early vs. normal. Prediction of overtake-early events, which occur seconds before the beginning of an overtake-late events, is more difficult.

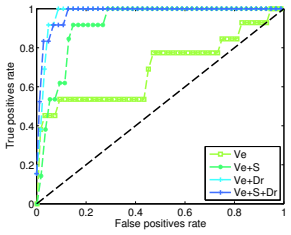


Fig. 8: For a fixed prediction time, $\delta = -2$ seconds, we show the effects of appending cues to the vehicle (**Ve**) dynamics under overtake-late/normal. **S** stands for surround (i.e. lidar and lane). **Dr** stands for driver (hand, head, and foot).

IX. Acknowledgments

The authors would like to thank the support of associated industry partners, the reviewers, and the members of the Laboratory for Intelligent and Safe Automobiles (LISA) for their assistance.

References

[1] "2012 motor vehicle crashes: overview," National Highway Traffic Safety Administration, Washington, D.C., Tech. Rep. DOT HS 811 856, 2013.

[2] 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," National Highway Traffic Safety Administration, Washington, D.C., Tech. Rep. DOT HS 811 731, 2013.

[3] M. Ortiz, J. Schmudderich, F. Kummert, and A. Gepperth, "Situation-specific learning for ego-vehicle behavior prediction systems," in *IEEE Conf. Intelligent Transportation Systems*, 2011.

[4] P. Angkititrakul, R. Terashima, and T. Wakita, "On the use of stochastic driver behavior model in lane departure warning," *IEEE Trans. Intelligent Transportation Systems*, vol. 12, no. 1, pp. 174–183, March 2011.

[5] G. Xu, L. Liu, Y. Ou, and Z. Song, "Dynamic modeling of driver control strategy of lane-change behavior and trajectory planning for collision prediction," *IEEE Trans. Intelligent Transportation Systems*, vol. 13, no. 3, pp. 1138–1155, Sept 2012.

[6] P. Angkititrakul, C. Miyajima, and K. Takeda, "Modeling and adaptation of stochastic driver-behavior model with application to car following," in *IEEE Conf. Intelligent Vehicles*, June 2011, pp. 814–819.

[7] M. Liebner, M. Baumann, F. Klanner, and C. Stiller, "Driver intent inference at urban intersections using the intelligent driver model," in *IEEE Conf. Intelligent Vehicles*, 2012.

[8] A. Kurt, J. L. Yester, Y. Mochizuki, and U. Özgüner, "Hybrid-state driver/vehicle modelling, estimation and prediction," in *IEEE Conf. Intelligent Transportation Systems*, 2010.

[9] M. Heracles, F. Martinelli, and J. Fritsch, "Vision-based behavior prediction in urban traffic environments by scene categorization," in *British Machine Vision Conference*, 2010.

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

[11] R. K. Satzoda, S. Martin, M. V. Ly, P. Gunaratne, and M. M. Trivedi, "Towards automated drive analysis: A multimodal synergistic approach," in *IEEE Conf. Intelligent Transportation Systems*, 2013.

[12] B. Morris, A. Doshi, and M. Trivedi, "Lane change intent prediction for driver assistance: On-road design and evaluation," in *IEEE Conf. Intelligent Vehicles*, 2011.

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

[14] "A comprehensive examination of naturalistic lane-changes," National Highway Traffic Safety Administration, Washington, D.C., Tech. Rep. DOT HS 809 702, 2004.

[15] 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*, vol. 15, no. 3, pp. 818–830, April 2014.

[16] 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.

[17] E. Ohn-Bar, S. Martin, and M. M. Trivedi, "Driver hand activity analysis in naturalistic driving studies: Issues, algorithms and experimental studies," *Journal of Electronic Imaging*, vol. 22, no. 4, 2013.

[18] E. Ohn-Bar and M. M. Trivedi, "In-vehicle hand activity recognition using integration of regions," in *IEEE Conf. Intelligent Vehicles*, 2013.

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

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

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

[22] 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.