Understanding Head and Hand Activities and Coordination in Naturalistic Driving Videos

Sujitha Martin[1], Eshed Ohn-Bar[1], Ashish Tawari[1] and Mohan M. Trivedi[1]

***Abstract*— In this work, we propose a vision-based analysis framework for recognizing in-vehicle activities such as interactions with the steering wheel, the instrument cluster and the gear. The framework leverages two views for activity analysis, a camera looking at the driver's hand and another looking at the driver's head. The techniques proposed can be used by researchers in order to extract 'mid-level' information from video, which is information that represents some semantic understanding of the scene but may still require an expert in order to distinguish difficult cases or leverage the cues to perform drive analysis. Unlike such information, 'low-level' video is large in quantity and can't be used unless processed entirely by an expert. This work can apply to minimizing manual labor so that researchers may better benefit from the accessibility of the data and provide them with the ability to perform larger-scaled studies.**

I. Introduction

For the past 50 years, most of the data related to vehicular collisions have come from post-crash analysis. Only recently, naturalistic driving studies (NDS) began providing detailed information about driver behavior, vehicle state, and roadways using video cameras and other types of sensors. Consequently, such data holds the key for the role and effect of cognitive processes, in-vehicle dynamics, and surrounding salient objects on driver behavior [1], [2].

The 100-Car Naturalistic Driving Study is the first instrumented-vehicle study undertaken with the primary purpose of collecting large-scale, naturalistic driving data. A 2006 report on the results of the 100-car field experiment [3] revealed that almost 80 percent of all crashes and 65 percent of all near-crashes involved the driver looking away from the forward roadway just prior to the onset of the conflict. It was also shown that 67% of crashes and 82% of near-crashes occurred when subject vehicle drivers were driving with at least one hand on the wheel. More details about the presence or absence of driver's hands on the wheel and the driver's inattention to forward roadway, for crashes and near-crashes as reported in [3] are shown in Table I and Table II.

Because of the above issues, on-road analysis of driver behavior is becoming an increasingly essential component for future advanced driver assistance system [4]. Towards this end, we focus on analyzing where and what the driver's hands do in the vehicle. Hand positions can provide the level of control drivers exhibit during a maneuver or can even give some information about mental workload [5]. Furthermore, in-vehicle activities involving hand movements often demand coordination with head and eye movements. For this, a distributed camera setup is installed to simultaneously observe hand and head movements. Together, this multiperspective approach allows us to derive a semantic level representation of driver activities, similar to research studies on upper body based gesture analysis for intelligent vehicles [6] and smart environments [7].

Hands on wheel	**Crash** (%)	**Near-Crash** (%)
Left hand only	30.4	31.7
Unknown	29.0	15.1
Both hands	24.6	35.1
Right hand only	11.6	15.5
No hands on wheel	4.4	2.6

TABLE I: Hands on wheel when crash and near-crash occurred from 100-car study [3]

Inattention to forward roadway	**Crash** (%)	**Near-Crash** (%)
Left window	9.7	3.2
Talking/listening	8.3	4.8
Passenger in adjacent seat	6.9	6.1
Center mirror	1.4	1.8
Right window	1.4	1.8
In-vehicle controls - Other	1.4	0.0
Adjust radio	0.0	1.3

TABLE II: Inattention to forward roadway when crash and near-crash occurred from 100-car study [3]

The approach is purely vision-based, with no markers or intrusive devices. There are several challenges that such a system must overcome, both for the robust extraction of head [8] and hand cues [9]. For the head, there are challenges of self-occlusion due to large head motion and of privacy implication for drivers in large scale data. Interestingly, a recent study has focused on the design of deidentification filters to protect the privacy of drivers while preserving driver behavior [10]. For the hand, detection is challenging as the human hand is highly deformable and tends to occlude itself in images. The problem is further complicated by the vehicular requirement for algorithms to be robust to changing

[1]The authors are with the Laboratory of Intelligent and Safe Automobiles at the University of California, San Diego, USA scmartin@ucsd.edu,eohnbar@ucsd.edu, atawari@ucsd.edu, mtrivedi@ucsd.edu

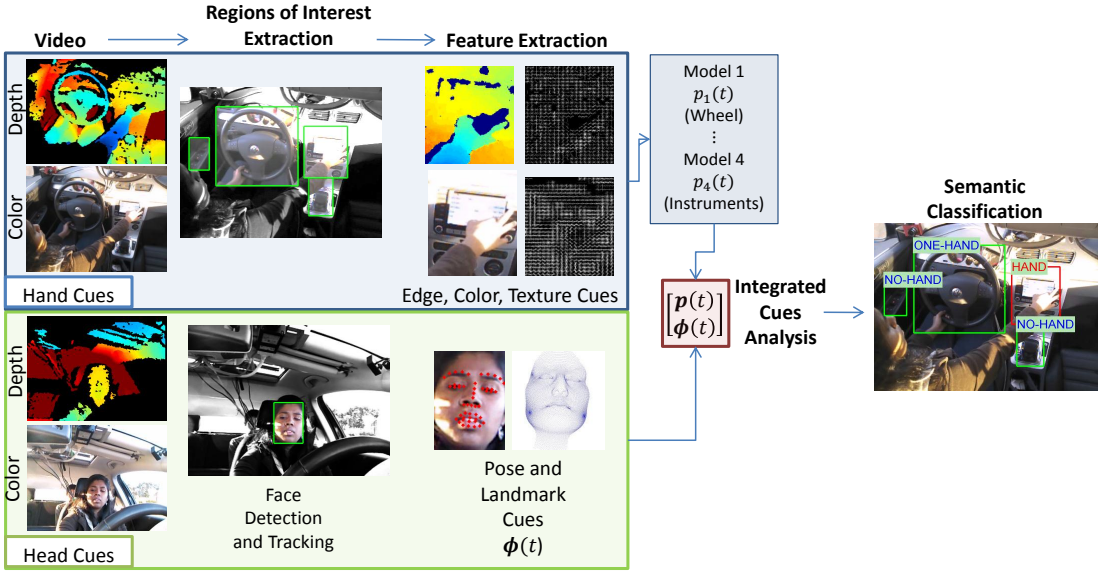


Fig. 1: The proposed approach for driver activity recognition. Head and hand cues are extracted from color and depth video in regions of interest. A classifier provides an integration of the cues, and the final activity classification.

illumination. Therefore, we are interested in incorporating head and eye cues to better represent the driver's interaction with the steering wheel, the instrument cluster and the gear. A more detailed analysis of various feature extraction on individual perspectives and their integration can be found in [11].

II. Activity Analysis Framework

The framework in this work leverages two views for activity analysis, a camera looking at the driver's hand and another looking at the head. As shown in Fig. 1, these are integrated in order to produce the final activity classification.

A. Hand Cues

Localizing the hands in the vehicle with a high degree of accuracy is highly desired. One approach for hand detection relies on a sliding-window. This is a common technique for generic visual object detection, where a model is learned based on positive samples (i.e. hands in different poses) of fixed size and negative samples which don't contain the object of interest. A classifier is then used to learn a classification rule. Such a scheme can be applied on multiple scales of the image in order to detect objects at different sizes. Specifically for hand detection, these techniques are faced with challenges as the hand is highly deformable and tends to occlude itself. Models are often sensitive to even small in-plane rotation [12] and deformation. A more sophisticated set of models (usually referred to in literature as a 'part-based deformable model' [13]) allows for learning a model for different configurations, deformations, and occlusion types. A pre-trained model for hand shape, however, resulted in many false positives on naturalistic driving dataset [14], [15].

Instead of learning a model for hand and searching for it throughout the entire cabin, we constrain the problem to a number of regions of interest which may be useful for studying the driver's state. This provides several benefits:

1) As the variation in hand appearance differs based on the region in which it is in, a *model learned for each region* could potentially better generalize over the variations in that specific region.
2) This phrasing of the problem allows us to study the performance of visual descriptors *for each region*. For instance, some regions are less prone to illumination changes.
3) *Integration*: in the context of our problem, the hand may be commonly found in only parts of the scene. Assuming that the hands must be in one of three regions of interest reduces the complexity of the problem and opens up the door for leveraging cues among the different regions. Integration also provides a model with the opportunity to perform higher-level reasoning of the hands configuration.

Our approach attempts to separate the scene into differently sized regions, and model two classes: no hand and hand presence. To that end, a linear kernel binary support vector machine (SVM) classifier is trained where input features are Histogram of Orientations (HOG) as applied in multiple scales. The linear SVM is used to learn a hand presence model in each of the periphery regions (the side hand rest, gear shift, and instrument cluster) and a 'two hands on the wheel' model for the wheel region. LIBSVM [16] allows for approximating the probability for hand presence in each of the regions at time t,

$$\mathbf{p}(t) = \begin{bmatrix} p_1(t) \\ \vdots \\ p_n(t) \end{bmatrix} \quad (1)$$

Time ⟶

Fig. 2: Hand, head, and eye cues can be used in order to analyze driver activity. Notice the guiding head movements performed in order to gather visual information before and while the hand interaction occurs.

where n is the number of regions considered. For head and hand integration, it will be useful for us to study $n = 3$, where the three regions are the wheel, gear shift, and instrument cluster. These probabilities are a powerful tool for analyzing semantic information in the scene, as they each correspond to our belief of a certain hand configuration. The probability output may be more reliable in certain regions, such as in the gear shift region, or noisier in others, such as in the difficult wheel region which is large and prone to volatile illumination. This motivates their integration, which can be done in multiple ways. A simple way which showed good results and opens up the door for integration with other views and modalities (for instance, head or CAN cues) is by letting a second-stage classifier reason over the probabilities outputted by the regional models. Therefore, a linear SVM is provided with the probability vector, $\mathbf{p}(t)$ to solve the multiclass problem and assign each frame with an activity label, from 1 to n.

B. Head Cues

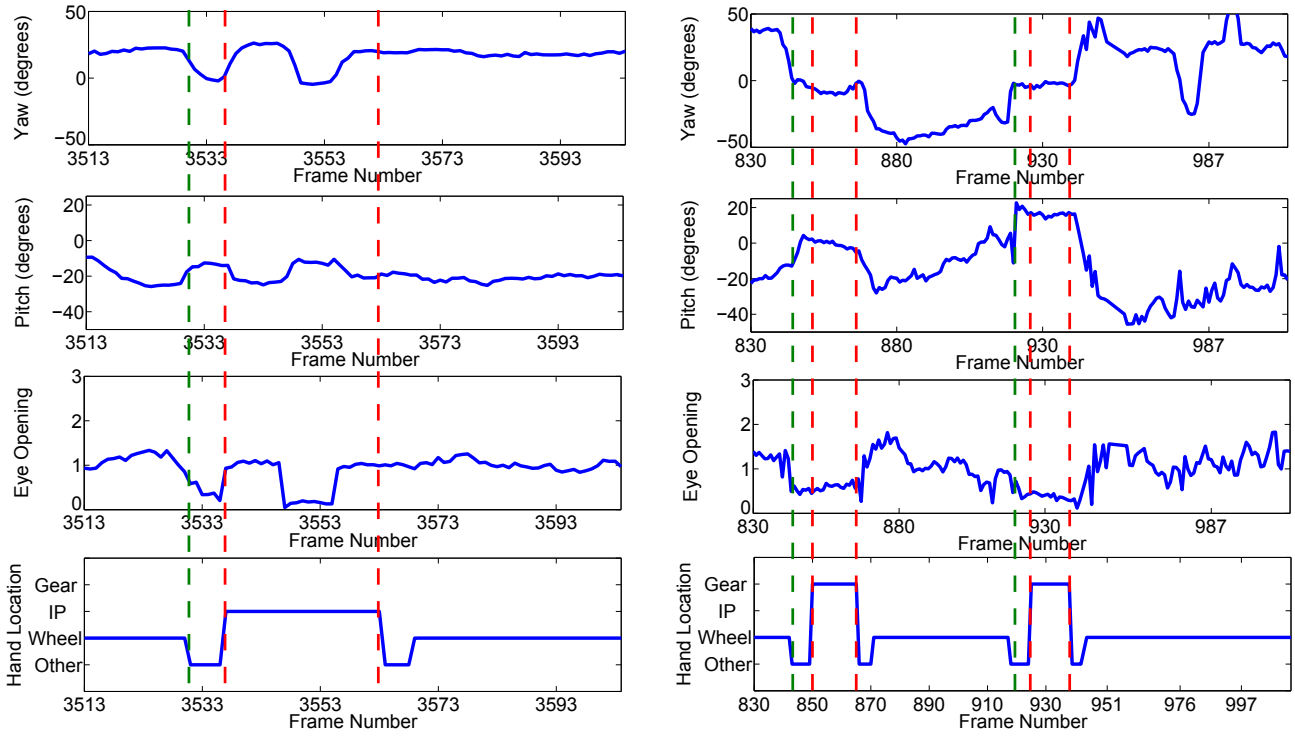
One type of features representative of the driver's head is head pose. Head pose estimator, however, needs to satisfy certain specifications to function robustly in a volatile driving environment. Continuous Head Movement Estimator (CoHMEt) [17] outlines these necessary specifications as: automatic, real-time, wide operational range, lighting invariant, person invariant and occlusion tolerant. Facial features-based approaches for extracting the head pose, such as the mixture of tree structure [18] and supervised descent method for face alignment [19], show promise of meeting many of the requirements. An additional benefit of using facial features for estimating head pose is that it allows for facial landmark analysis, such as level of eye opening. While the percent of eye opening has been vastly studied for detecting driver fatigue, measuring the openness of eyes can benefit in estimating the driver's gaze. For instance, when interacting with the instrument panel, distinctive eye cues arise (see Fig. 3). In this work, we explore the possibility of using head pose and eye opening as features in monitoring the in-vehicle driver activities, summarized in a feature vector we call as $\boldsymbol{\phi}(t)$ at time t.

Driver interaction with the infotainment system and the gear show unique pattern combination with head pose, eye opening and hand locations as shown in Fig. 2. Figure 3 shows time synchronized plots of head pose, eye opening, hand activity for two typical events: interacting with IP and interacting with gear. In Fig. 3 head pose in yaw and pitch are measured in degrees, where a decreasing value in yaw represents the driver looking rightward and an increasing value in pitch represents the driver looking downward. In the plot for eye opening, a value of 1 represents the normal size of eyes, values greater than one could represent looking upward, and values less than one could represent looking downward. Hand locations in the image plane are also plotted in a time-synchronized manner, but instead the presences of hands in discrete locations are plotted. The green dotted line indicates the start of supportive head and eye cues to the respective hand activity. The dotted red lines indicates the start and end of the presence of hand in locations respective of its activity. These plots show the presence of hand, head and eye movements while the driver interacts with the infotainment system (Fig. 3(a)) and with the gear (Fig. 3(b)). While the latency of each cue is circumstantial, we experimentally validate the use of head and eye cues to strengthen the detection of hand activity recognition.

C. Integration of Modalities and Perspectives

We obtain an SVM model trained on RGB descriptors of either: 1) Hand or no hand in the ROI (in the peripheral ROIs) 2) Two hands or one or no hands in the ROI (the center wheel ROI). The assumption that the hand can only be found in a subset of the regions of interest allows the second-stage classifier to reason over the likelihood of the driver's two hand configuration. For instance, if the smaller, peripheral regions are known to be more reliable, and all show a 'no hand' event, we would like a model that can reason in such case that both hands are on the wheel.

In addition, the second-stage classifier provides an opportunity for integration with other modalities. Since we observed a correlation between head dynamics and hand activity, we perform a study of head and hand cue integration. Ideally, the second-stage classifier will resolve false positives and increase the likelihood of certain hand configurations by leveraging features extracted from the pose of the head and



(a) Instrument Panel Interaction

(b) Gear Shift Interaction

Fig. 3: Hand, head, and eye cue visualization for (a) an instrument panel activity sequence and (b) gear shift activity sequence. **Green line**: indication of start of head and eye cues (yaw, pitch, and opening) before the hand activity. **Red lines**: start and end of the hand activity. See Section II-B for further detail on the cues.

Loction	Activity Types
Radio	On/Off Radio Change Preset Navigate to Radio Channel Increase/Decrease Volume Seek/Scan for Preferred Channel Insert/Eject CD
	On/Off Hazard Lights
Climate Control	On/Off AC Adjust AC Change Fan Direction
Side Rest	Adjust Mirrors
Gear	Park/Exit Parking

TABLE III: Types of activities in the dataset collected.

eyes. The final feature vector is therefore denoted by

$$\boldsymbol{x}(t) = \begin{bmatrix} \mathbf{p}(t) \\ \boldsymbol{\phi}(t) \end{bmatrix} \tag{2}$$

where $\boldsymbol{\phi}(t)$ is the features extracted from the head view. We compare two possible choices for $\boldsymbol{\phi}(t)$. First, a simple concatenation of the values from pose and landmarks over a time window is used. Second, we use summarizing statistics over the time window, namely the mean, minimum, and maximum for each of the features over the temporal window.

III. Experimental Evaluation and Discussion

Detecting the driver's activity (e.g. adjusting radio, using gear) is an important step towards detecting driver distraction. In this section, we describe the dataset and the results of the proposed framework. By integrating head and hand cues, we show promising results of driver's activity recognition.

A. Experimental Setup

In order to train and test the activity recognition framework, we collected a dataset using two Kinects, one observing the hands and one observing the head. The dataset was collected while driving, where subjects were asked to perform tasks as listed in Table III. The four subjects (three males and one female) were of various nationalities and ranged from 20 to 30 years of age. The amount of driving experience varied as well, ranging from a few years to more than a decade. The tasks in Table III were first practiced before the drive to ensure the users were familiar with and also comfortable performing the task in the vehicle testbed. For each driver, there are two main consistencies in the process of data collection. First, at the beginning of the drive, the driver was verbally instructed with the list of secondary

Subject	Video Time (min)	# Samples Annotated	Environmen	Time
1	9:08	10115	Sunny	4pm
2	10:05	4491	Sunny	5pm

TABLE IV: Driver activity recognition dataset collected. Training and testing is done using cross-subject cross-validation.

tasks to perform during the drive. Second, the drivers were allowed to drive with control over what secondary task they wanted to perform and when they wanted to perform it. For instance, interaction with the radio was motivated by the driver and not the experiment supervisor. Driving was performed in urban, high-traffic settings.

To ensure generalization of the learned models, all testing is performed by leave-one-subject-out cross validation, where the data from one subject is used for testing and the data from other subjects is used for training. We collected a head and hand dataset with the following statistics: 7429 samples of two hands on the wheel region, 719 samples of hand interacting with the side rest, 679 samples of hand on the gear and 3039 samples of instrument cluster region interaction. Table IV shows the statistics of the entire dataset. As the videos were collected in sunny settings in the afternoon, they contain significant illumination variation that are both global and local (shadows).

B. Evaluating Hand and Head Integration

Although head pose and landmark cues are generated at every frame, they may be delayed in their correlation to the annotated hand activity. Nonetheless, integrating head cues could improve detection in transition among regions as well as to reduce false positives by increasing the likelihood of a hand being present at one of the regions. Two feature sets are compared over a variable sized window of time previous to the current frame. If δ is the size of the time window, we can simply concatenate the time series over $[(t-\delta),\ldots,(t)]$ in order to generate $\boldsymbol{\phi}(t)$ (referred to as **temporal concatenation**) or we may summarize the time window using **global statistics**. In particular, we use the mean, minimum, and maximum values of the window to generate a set of head features. The second approach seems to produce significantly better results as shown in Fig. 4.

For the three region classification problem, head pose and landmark cues exhibit a distinctive pattern over the temporal window. A large window to include the initial glance before reaching to the instrument cluster or the gear shift as well as any head motions during the interaction significantly improves classification as shown in Fig. 5. Both the gear shift and instrument cluster (denoted as **IC**) benefit from the integration.

IV. Conclusion

Automotive systems should be designed to operate quickly and efficiently in order to assist the human driver. To that end,

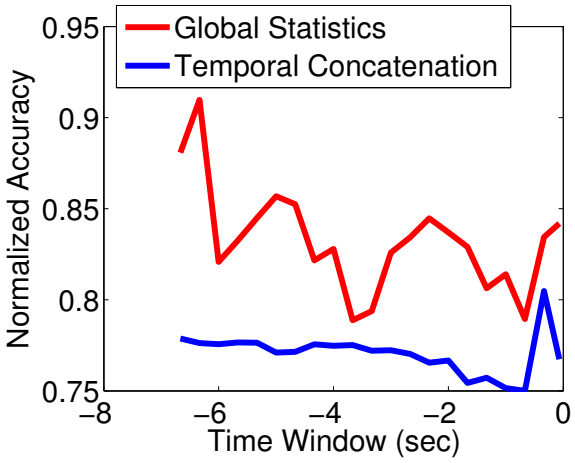
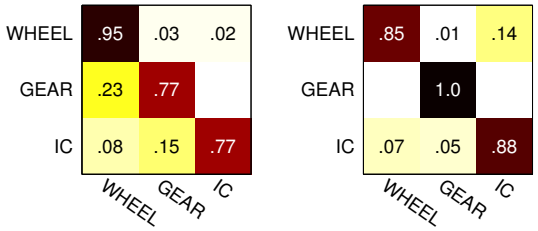


Fig. 4: Integration results for hand and head cues for the three region activity recognition (wheel, gear shift, instrument cluster). The head features are computed over different sized temporal windows (see Section III-B).



(a) Hand Only **(83%)** (b) Hand+Head **(91%)**

Fig. 5: Activity recognition based on (a) hand only cues and (b) hand+head cue integration for the three region activity classification. As head cues are common with instrument cluster and gear shift interaction, a significant improvement in results is shown. **IC** stands for instrument cluster.

we investigated leveraging a multiprespective, multimodal approach for semantic understanding of the driver's state. A set of in-vehicle secondary tasks performed during on-road driving was utilized to demonstrate the benefit of such an approach. The cues from two views, of the hand and of the head, were integrated in order to produce a more robust activity classification. The analysis shows promise in temporal modeling of head and hand events. Future work would extend the activity grammar to include additional activities of more intricate maneuvers and driver gestures. Combining the head pose with the hand configuration to produce semantic activities can be pursued using temporal state models, as in [20].

V. Acknowledgment

We acknowledge support of the UC Discovery Program and associated industry partners. We also thank our UCSD LISA colleagues who helped in a variety of important ways in our research studies. Finally, we thank the reviewers for their constructive comments.

REFERENCES

[1] R. Satzoda and M. M. Trivedi, "Automated drive analysis with forward looking video and vehicle sensors," *IEEE Trans. Intelligent Transportation Systems*, to appear 2014.

[2] B. T. Morris and M. M. Trivedi, "Trajectory learning for activity understanding: Unsupervised, multilevel, and long-term adaptive approach," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no. 11, pp. 2287–2301, 2011.

[3] T. A. Dingus, S. Klauer, V. Neale, A. Petersen, S. Lee, J. Sudweeks, M. Perez, J. Hankey, D. Ramsey, S. Gupta *et al.*, "The 100-car naturalistic driving study, phase ii-results of the 100-car field experiment," Tech. Rep., 2006.

[4] A. Doshi, B. Morris, and M. M. Trivedi, "On-road prediction of driver's intent with multimodal sensory cues," *IEEE Pervasive Computing*, vol. 10, no. 3, pp. 22–34, 2011.

[5] D. D. Waard, T. G. V. den Bold, and B. Lewis-Evans, "Driver hand position on the steering wheel while merging into motorway traffic," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 13, no. 2, pp. 129 – 140, 2010.

[6] S. Y. Cheng, S. Park, and M. M. Trivedi, "Multi-spectral and multi-perspective video arrays for driver body tracking and activity analysis," *Computer Vision and Image Understanding*, vol. 106, no. 2, pp. 245–257, 2007.

[7] C. Tran and M. M. Trivedi, "3-d posture and gesture recognition for interactivity in smart spaces," *Industrial Informatics, IEEE Transactions on*, vol. 8, no. 1, pp. 178–187, 2012.

[8] S. Martin, A. Tawari, E. Murphy-Chutorian, S. Y. Cheng, and M. Trivedi, "On the design and evaluation of robust head pose for visual user interfaces: Algorithms, databases, and comparisons," in *ACM Conf. Automotive User Interfaces and Interactive Vehicular Applications*, 2012.

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

[10] S. Martin, A. Tawari, and M. M. Trivedi, "Towards privacy protecting safety systems for naturalistic driving videos," *IEEE Trans. Intelligent Transportation Systems*, 2014.

[11] E. Ohn-Bar, S. Martin, A. Tawari, and M. M. Trivedi, "Towards understanding driver activities from head and hand coordinated movements," in *Pattern Recognition (ICPR), 2012 21st International Conference on*. IEEE, 2012, pp. 605–608.

[12] K. Mathias and M. Turk, "Analysis of rotational robustness of hand detection with a viola-jones detector," in *Intl. Conf. on Pattern Recognition*, 2004.

[13] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Analysis and Machine Intelligence,*, vol. 32, no. 9, pp. 1627–1645, 2010.

[14] E. Ohn-Bar and M. M. Trivedi, "In-vehicle hand activity recognition using integration of regions," in *IEEE Conf. Intell. Veh. Symp.*, 2013.

[15] ——, "The power is in your hands: 3D analysis of hand gestures in naturalistic video," in *IEEE Conf. Computer Vision and Pattern Recognition Workshops*, 2013.

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

[17] 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*, 2014.

[18] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2012.

[19] X. Xiong and F. D. la Torre, "Supervised descent method and its applications to face alignment," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2013.

[20] Y. Song, L. P. Morency, and R. Davis, "Multi-view latent variable discriminative models for action recognition," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2012.