2014 22nd International Conference on Pattern Recognition

Head, Eye, and Hand Patterns for Driver Activity Recognition

Eshed Ohn-Bar, Sujitha Martin, Ashish Tawari, and Mohan Trivedi
University of California San Diego
{*eohnbar,scmartin,atawari,mtrivedi*}*@ucsd.edu*

***Abstract*—In this paper, a multiview, multimodal vision framework is proposed in order to characterize driver activity based on head, eye, and hand cues. Leveraging the three types of cues allows for a richer description of the driver's state and for improved activity detection performance. First, regions of interest are extracted from two videos, one observing the driver's hands and one the driver's head. Next, hand location hypotheses are generated and integrated with a head pose and facial landmark module in order to classify driver activity into three states: wheel region interaction with two hands on the wheel, gear region activity, or instrument cluster region activity. The method is evaluated on a video dataset captured in on-road settings.**

I. Introduction

Secondary tasks performed in the vehicle have been shown to increase inattentiveness [1], which, in 2012 was a contributing factor in at least 3092 fatalities and 416,000 injuries [2]. According to a recent survey, 37% of the drivers admit to having sent or received text messages, with 18% doing so regularly while operating a vehicle [3]. Furthermore, 86% of drivers report eating or drinking (57% report doing it sometimes or often), and many reported common GPS system interaction, surfing the internet, watching a video, reading a map, or grooming.

Because of the above issues, on-road analysis of driver activities is becoming an essential component for advanced driver assistance systems. Towards this end, we focus on analyzing where and what 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 [4]. Furthermore, in-vehicle activities involving hand movements often demand coordination with head and eye movements. In fact, human gaze behavior studies involving various natural dynamic activities including driving [5], [6], typing [7], walking ([8]), throwing in basketball [9], batting in cricket ([10]) etc., suggest a common finding that gaze shifts and fixations are controlled pro actively to gather visual information for guiding movements.While specific properties of the spatial and temporal coordination of the eye, head and hand movements are influenced by the particular tasks, there is strong evidence to suggest that the hand usually waits for the eyes either for the target selection or for the visual guidance for the reach, or both [11]. For this, a distributed camera setup is installed to simultaneously observe hand and head movements.

The framework in this work leverages two views for driver activity analysis, a camera looking at the driver's hand and another looking at the head. The multiple views framework provides a more complete semantic description of the driver's activity state [12]. As shown in Fig. 1, these are integrated in order to produce the final activity classification. First, the hand detection technique is discussed, then a detailed description of relevant head and eye cues is given, followed by a description of head, eye and hand cueintegration scheme. Lastly, experimental evaluations is presented on naturalistic driving.

II. Feature Extraction Modules

A. Hand Cues

In the vehicle, hand activities may be characterized by zones or regions of interest. These zones (see Fig. 1) are important for understanding driver activities and secondary tasks. This motivates scene representation in terms of these salient regions. Additionally, structure in the scene can be captured by leveraging information from the multiple salient regions. For instance, during interaction with the instrument cluster, visual information from the gear region can increase the confidence in the current activity recognition, as no hand is found on the gear shift. Such reasoning is particularly useful under occlusion, noise due to illumination variation, and other visually challenging settings [13]. In [14], [15], edge, color, texture, and motion features were studied for the purpose of hand activity recognition. Since we found that edge features were particularly successful, in this work we employ a pyramidal representation for each region using Histogram of Oriented Gradients (HOG) [16], with cell sizes 1 (over the entire region), 4, and 8 for a $8+128+512 = 648$ dimensional feature vector.

B. Head and Eye Cues

Knowing where the driver is looking can provide important cues about any on-going driver activities. While precise gaze information is ideally preferred, its estimation is very challenging, especially when using remote eye tracking systems in a real-world environment such as driving. However, a coarse gaze direction, i.e. gaze zone, is often sufficient in a number of applications, and can be relatively robustly extracted in driving environments [17].

Driver's gaze is inferred using head-pose and eye-state. We use facial features-based geometric approach for head pose estimation. With recent advancements in facial feature tracking methods [18], [19] and two cameras monitoring the driver's head, we can obtain good accuracy and can reliably track the driver's head during spatially large head movements [20]. The tracked facial landmarks can not only be used to estimate head pose, but can also be used to derive other states of the driver, such as the level of eye opening. Head pose alone provides a

1051-4651/14 $31.00 © 2014 IEEE
DOI 10.1109/ICPR.2014.124

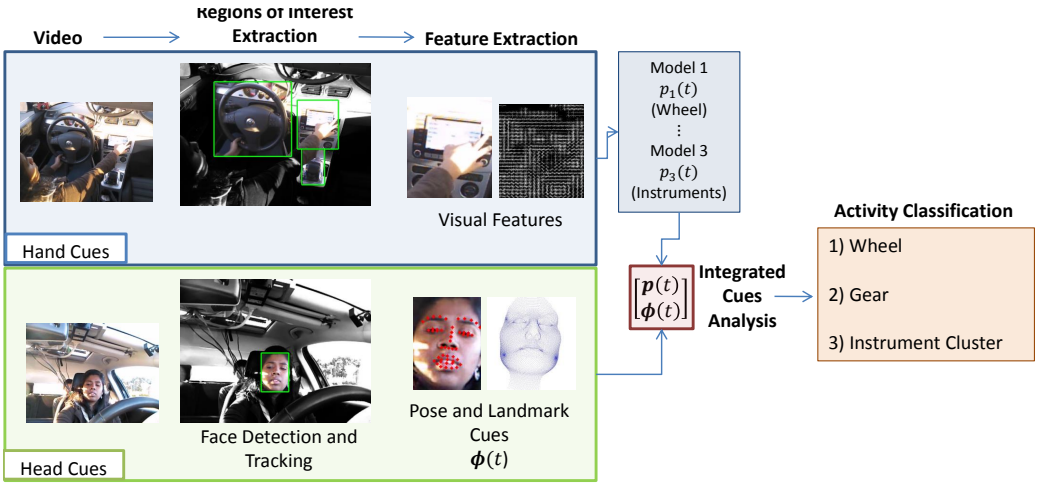


Fig. 1: The proposed approach for driver activity recognition. Head and hand cues are extracted from video in regions of interest. These are fused using a hierarchical Support Vector Machine (SVM) classifier to produce activity classification.

good approximation of gaze zone, but neighboring zones (e.g. instrument cluster region and gear region) are often confused [17]. In such cases, eye-state such as eye-opening can help to disambiguate between confusing zones.

In our implementation, the eye state at time t is estimated using two variables: area of the eye and area of the face. Area of the eye is the area of a polygon whose vertices are the detected facial landmarks around the left or right eye. Similarly, the area of the face is the area of the smallest polygon that encompass all the detected facial landmarks. To compute the level of eye opening, we divide area of the eye by the area of the face at every time t. This normalization will allow the computation of eye opening to be invariable to driver's physical distance to the camera, where closer distances makes the face appear larger in the image plane. Finally, a normalization constant learned for each driver representing his or her normal eye-opening state is used such that after normalization values < 1 represent downward glances and values > 1 represent upward glances (visualized in Fig. 2.

The eye-opening cue in addition to head pose, has potential in differentiating between glances towards the instrument cluster and glances towards the gear, as shown in Fig. 2. Figure 2 shows the mean (solid line) and standard deviation (semi-transparent shades) of two features (i.e. head pose in pitch and eye opening) for three different driver activities, using the collected naturalistic driving dataset. The feature statistics are plotted 6 seconds before and after the start of the driver hand activity, where time of 0 seconds represents the start of the activity. Using the eye opening cues alone, we can observe that when the driver is interacting with the instrument cluster he or she glances towards the IC at the start of the interaction. However, when the driver is interacting with the gear, while there is some indication of a small glance before the start of the activity, there is significant glance engagement with the gear region after the start of the event.

As the above cues may occur before or after an associated hand cue (i.e. looking and then reaching to the instrument cluster), the head and eye features are computed over a temporal window. Let $\boldsymbol{h}(t)$ represent the features containing the head pose (in pitch, yaw and roll in degrees) and the level of eye opening (for both left and right eye) at time t and δ be the size of the time window to be used for temporal concatenation. Then, the time series $\boldsymbol{\phi}(t) = [\boldsymbol{h}(t-\delta), \ldots, \boldsymbol{h}(t)]$ is the feature set extracted from the head view at time t to be further used in the integration with hand cues.

III. Activity Recognition Framework

In this section, we detail the learning framework for fusion of the two views and performing activity classification. The classifier used is a linear kernel SVM [21], and fusion is done using a hierarchical SVM which produces the final activity classification.

Because the hand and head cues are different in nature, first a multiclass Support Vector Machine (SVM) [22] is trained to produce activity classification based on the hand view region features only. A weight, $\mathbf{w}_i$ is learned for each class $i \in \{1, \ldots, n\}$ where n is the number of activity classes. In this work, we focus on three activity classes: 1) Wheel region interaction with two hands on the wheel; 2) Gear region interaction; 3) Instrument cluster interaction. The weights for all of the classes are learned jointly, and classification can be performed using

$$i^\star = \underset{i \in \{1,\ldots,n\}}{\arg\max}\ \mathbf{w}_i^T \mathbf{x} \tag{1}$$

where $\mathbf{x}$ is the feature vector from all the regions in the hand view.

In order to measure the effectiveness and complementarity of the hand and head cues, activity recognition will be studied

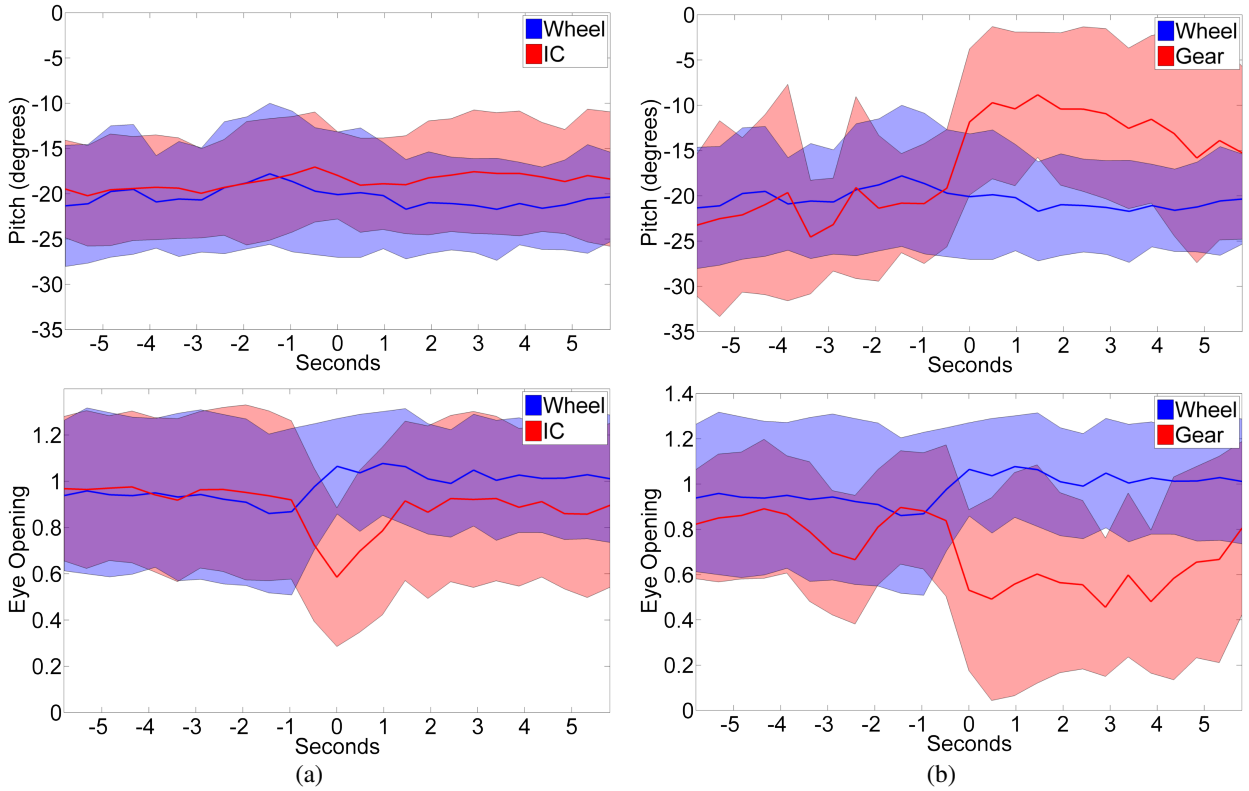


Fig. 2: Head and eye cue statistics visualization for (a) instrument cluster (IC) activity sequences against normal wheel interaction sequences and (b) gear shift activity sequences against normal wheel interaction sequences. Time $t = 0$ represents the start of the respective driver activity. The blue and red line represent the mean statistics of respective cues (i.e. head pose in pitch, eye opening) for 6 seconds before and after the start of the driver hand activity. The lighter shades around the solid line indicate the standard deviation from the respective mean statistics.

using hand-only cues and integrated hand and head cues. Hand cues can be summarized using normalized scores,

$$p(i|\mathbf{x}) = \frac{\exp\left(\mathbf{w}_i^T \mathbf{x}\right)}{\sum_j \exp\left(\mathbf{w}_j^T \mathbf{x}\right)} \tag{2}$$

These posterior probabilities can be calculated at every frame and are abbreviated in Fig. 1 as p_i. For the fusion of the hand and head views, the hand cues are concatenated with the windowed signal of head features to produce the feature set at time t,

$$\mathbf{x}(t) = \begin{pmatrix} p_1(t) \\ \vdots \\ p_n(t) \\ \boldsymbol{\phi}(t) \end{pmatrix}$$

The fused feature vector is given to a hierarchical second-stage multiclass SVM to produce the activity classification.

The classes in our dataset are unbalanced. For instance, one activity class such as wheel region two-hands on the wheel may occur in the majority of the samples. Nonetheless preserving all of the samples for the wheel region in training could be beneficial in producing a robust classifier which can generalize over the large occlusion and illumination challenges occurring in the wheel region. Therefore, we also incorporate a biased-penalties SVM [23], which adjusts the regularization parameter in the classical SVM to be proportional to the class size in training.

IV. Experimental Evaluation and Discussion

The proposed driver hand activity recognition framework is evaluated on naturalistic driving data from multiple drivers. Using hand annotated ground truth data of driver hand activity, we show promising results of integrating head and hand cues.

A. Experimental Setup and Dataset Description

The naturalistic driving dataset is collected using two cameras, one observing the driver's hands and another observing the driver's head. Multiple drivers (three male and one female) of varying ethnicity and varying age from 20 to 30, as well as varying driving experience participated in this study. Before driving, each driver was instructed to perform, at his or her convenience, the following secondary tasks any number of times and in any order of preference:

- *Instrument cluster (IC) region activities:* On/off radio, change preset, navigate to radio channel,

increase/decrease volume, seek/scan for preferred channel, insert/eject a CD, on/off hazard lights, on/off/adjust climate control.

- *Gear region activities:* Observed while parking and exiting parking.
- *Wheel region activities:* Observed under normal driving conditions.

The drivers practiced the aforementioned activities before driving in order to get accustomed to the vehicle. In addition, instructors also prompted the drivers to instigate these activities randomly but cautiously. Driving was performed in urban, high-traffic settings.

Ground truth for evaluation of our framework is obtained from manual annotation of the location of driver's hands. A total of $11,147$ frames from many number of driver activities during the drives were annotated: 7429 frames of two hands in the wheel region for wheel region activity, 679 frames of hands on the gear, and 3039 frames of interaction in the instrument cluster region. As the videos were collected in sunny settings at noon or the afternoon, they contain significant illumination variation that is both global and local (shadows). With this dataset, all testing is performed by cross subject test settings, where the data from one subject is used for testing and the rest for training. This ensures generalization of the learned models.

B. Evaluating of Hand and Head Integration

Capturing the temporal dynamics of head and hand cues is evaluated in terms of activity classification out of a three class problem: 1) Wheel region interaction with two hands on the wheel; 2) Gear region interaction; 3) Instrument cluster interaction. Hand cues may be used alone, with results shown in Fig. 4(a). The results are promising, but instrument cluster and gear classification are sometimes confused due to the arm presence in the gear region while interaction occurs with the instrument cluster. Furthermore, under volatile illumination changes the method may also fail.

Incorporating head cues is shown to resolve some of the challenges, as depicted in Fig. 4(b). In order to capture head and hand cue dynamics, head and eye cues are calculated over a temporal window in order to generate $\boldsymbol{\phi}(t)$, the final head and eye feature vector at time t. The effect of changing the time window are shown in Fig. 3. We notice how increasing the window size of up to two seconds improves performance, after which results decline. With a large temporal window, the cue becomes less discriminative and also higher in dimensionality, which explains the decline. Nonetheless, we expect a peak in results for a window size larger than one entry, as head and hand cues may be temporally delayed. For example, a driver may look first and then reach towards the instrument cluster or gear shift.

Fig. 5 visualizes some example cases where hand cues provide ambiguous activity classification due to visually challenging settings, yet these are resolved after the predictions are rescored with the second stage hierarchical SVM and head and eye cues. For each of the depicted scenarios, the hand view, head view, and the fitted head models are shown. Using the hand cue prediction (shown in the purple probabilities) would have resulted in an incorrect activity classification. For

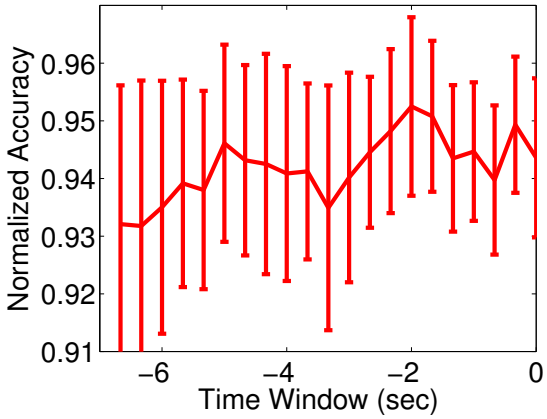
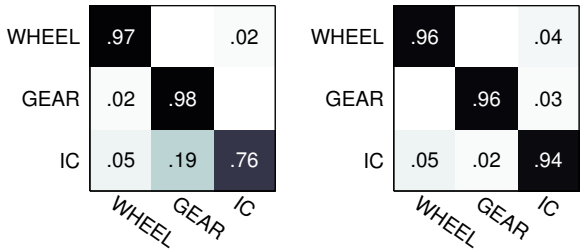


Fig. 3: Effect of varying the time window before an event definition for the head cues. Normalized accuracy (average of the diagonal of the confusion matrix) and standard deviation for activity classification is reported after integration with hand cues.



(a) Hand Only (**90%**) (b) Hand+Head (**94%**)

Fig. 4: Activity recognition based on hand only cues and hand+head cue integration for three region activity classification. **IC** stands for instrument cluster.

instance, some of the hand enters the gear shift while still interacting with the instrument cluster in the top figure. This leads to a wrong prediction using hand cues, but pitch and head information rescore the probabilities and correctly classify the activity (final classification after integration is visualized with a red transparent patch). Illumination variation may also cause incorrect activity classification based on hand cues alone, as shown in Fig. 5.

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. 4. Mainly, the gear shift and instrument cluster benefit from the integration.

V. Conclusion

In this work, we proposed a framework for leveraging both a hand and head view in order to provide activity recognition in a car. Integration provided improved activity recognition results and allows for a more complete semantic description of the driver's activity state. A set of in-vehicle secondary tasks

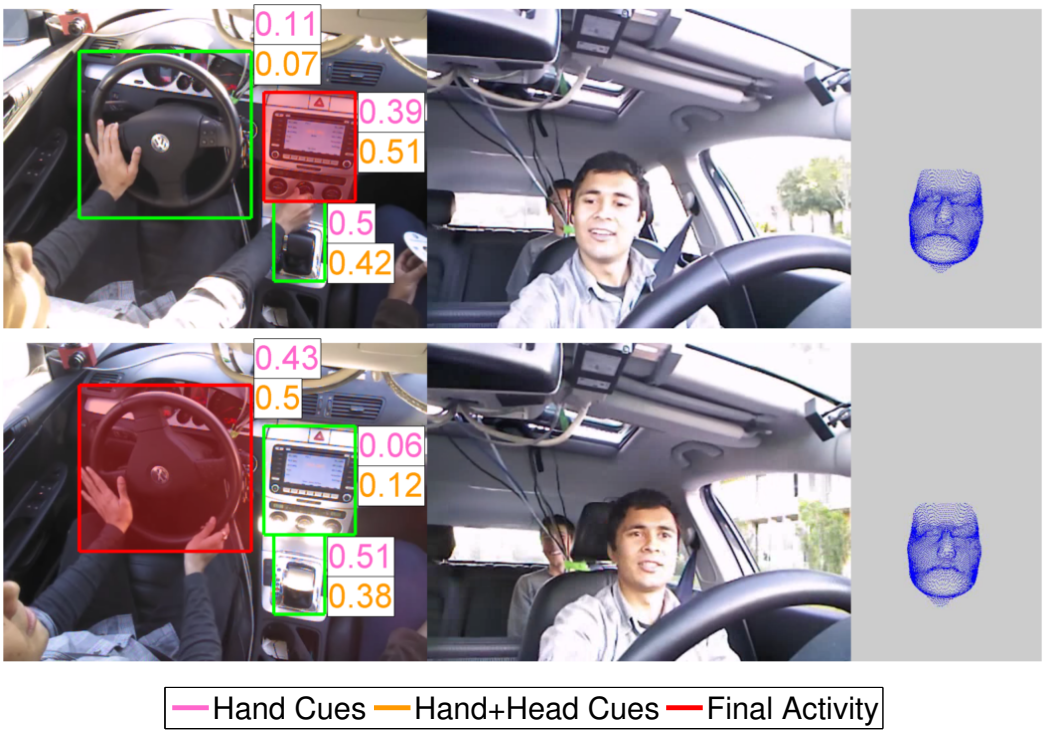


Fig. 5: Visualization of the advantage in integrating head, eye, and hand cues for driver activity recognition. We show the hand view, head view, and the fitted head model. In purple are the probabilities of the activity based on hand cues alone. In orange are the rescored values using a hierarchical SVM and head and eye cues. Note how in the above scenarios, the incorrect hand-based predictions were corrected by the rescoring based on head and eye cues.

performed during on-road driving was utilized to demonstrate the benefit for such an approach, with promising results. Future work would extend the activity grammar to include additional activities of more intricate maneuvers and driver gestures, as in [24], [25]. Combining the head pose with the hand configuration to produce semantic activities can be pursued using temporal states models, as in [26]. Finally, the usefulness of depth data will be studied in the future as well [27].

REFERENCES

[1] S. Klauer, F. Guo, J. Sudweeks, and T. Dingus, "An analysis of driver inattention using a case-crossover approach on 100-car data: Final report," National Highway Traffic Safety Administration, Washington, D.C., Tech. Rep. DOT HS 811 334, 2010.

[2] J. Tison, N. Chaudhary, and L. Cosgrove, "National phone survey on distracted driving attitudes and behaviors," National Highway Traffic Safety Administration, Washington, D.C., Tech. Rep. DOT HS 811 555, Dec. 2011.

[3] T. H. Poll, "Most U.S. drivers engage in 'distracting' behaviors: Poll," Insurance Institute for Highway Safety, Arlington, Va., Tech. Rep. FMCSA-RRR-09-042, Nov. 2011.

[4] 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.

[5] M. F. Land and D. N. Lee, "Where we look when we steer." *Nature*, vol. 369, no. 6483, pp. 742–744, 1994.

[6] A. Doshi and M. M. Trivedi, "Head and eye gaze dynamics during visual attention shifts in complex environments," *Journal of Vision*, vol. 12, no. 2, 2012.

[7] A. Inhoff and J. Wang, "Encoding of text, manual movement planning, and eye-hand coordination during copy-typing," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 18, pp. 437–448, 1992.

[8] A. E. Patla and J. Vickers, "Where and when do we look as we approach and step over an obstacle in the travel path?" *Neuroreport*, vol. 8, no. 17, pp. 3661–3665, 1997.

[9] J. Vickers, "Encoding of text, manual movement planning, and eye-hand coordination during copy-typing," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 22, pp. 342–354, 1996.

[10] M. F. Land and P. McLeod, "From eye movements to actions: How batsmen hit the ball." *Nature Neuroscience*, vol. 3, pp. 1340–1345, 2000.

[11] J. Pelz, M. Hayhoe, and R. Loeber, "The coordination of eye, head, and hand movements in a natural task," *Experimental Brain Research*, vol. 139, no. 3, pp. 266–277, 2001.

[12] C. Tran and M. M. Trivedi, "Human pose estimation and activity recognition from multi-view videos: Comparative explorations of recent developments," *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, no. 5, pp. 538–552, Sep. 2012.

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

[14] 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.

[15] 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.

[16] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2005.

[17] A. Tawari and M. M. Trivedi, "Dynamic analysis of multiple face videos for robust and continuous estimation of driver gaze zone," in *IEEE Intelligent Vehicles Symposium*, 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] 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. 2, pp. 818–830, 2014.

[21] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A library for large linear classification," *Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.

[22] K. Crammer and Y. Singer, "On the algorithmic implementation of multiclass kernel-based vector machines," *Journal of Machine Learning Research*, vol. 2, pp. 265–292, 2001.

[23] F. R. Bach, D. Heckerman, and E. Horvitz, "Considering cost asymmetry in learning classifiers," *The Journal of Machine Learning Research*, vol. 7, pp. 1713–1741, 2006.

[24] E. Ohn-Bar, A. Tawari, S. Martin, and M. M. Trivedi, "Vision on wheels: Looking at driver, vehicle, and surround for on-road maneuver analysis," in *IEEE Conf. Computer Vision and Pattern Recognition Workshops*, 2014.

[25] ——, "Predicting driver maneuvers by learning holistic features," in *IEEE Intelligent Vehicles Symposium*, 2014.

[26] 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.

[27] E. Ohn-Bar and M. M. Trivedi, "Hand gesture recognition in real-time for automotive interfaces: A multimodal vision-based approach and evaluations," *IEEE Trans. Intelligent Transportation Systems*, 2014.