IEEE Conference on Computer Vision and Pattern Recognition Workshops: Analysis and Modeling of Faces and Gestures, to appear 2013

The Power is in Your Hands: 3D Analysis of Hand Gestures in Naturalistic Video

Eshed Ohn-Bar and Mohan M. Trivedi
Computer Vision and Robotics Research Lab
University of California, San Diego
La Jolla, CA 92093-0434
{eohnbar,mtrivedi}@ucsd.edu

Abstract

We study natural human activity under difficult settings of cluttered background, volatile illumination, and frequent occlusion. To that end, a two-stage method for hand and hand-object interaction detection is developed. First, activity proposals are generated from multiple sub-regions in the scene. Then, these are integrated using a second-stage classifier. We study a set of descriptors for detection and activity recognition in terms of performance and speed. With the overarching goal of reducing 'lab setting bias', a case study is introduced with a publicly available annotated RGB and depth dataset. The dataset was captured using a Kinect under real-world driving settings. The approach is motivated by studying actions-as well as semantic elements in the scene and the driver's interaction with them-which may be used to infer driver inattentiveness. The proposed framework significantly outperforms a state-of-the-art baseline on our dataset for hand detection.

1. Introduction and Motivation

Object detection and tracking, in particular of human hands, has been widely investigated in the research community. Inferring information from hand activity is especially important in the operated vehicle, because it can provide vital information about the state of attentiveness of the driver [12]. The field has been pushed by increasingly difficult datasets comprising of different visual modalities, although the overall majority of these are still captured under controlled environments (some exceptions are the PASCAL VOC challenge [8] and the hand dataset in [15]). The dataset in this work adds upon the aforementioned by incorporating depth images, temporal events, occluding objects, and other appearance artifacts produced in the volatile environment of the vehicle's interior.

In addition to a novel dataset, there are two contributions in this paper. First, motivated by the need for a fast and robust hand localization system, we propose a two-stage method for integrating cues from critical regions of interest in the vehicle. We extend the framework and features proposed in [16, 3] for detecting hand or no hand events from RGB and depth. A hand model is learned for each region using a linear SVM, and the output of each SVM is integrated through a second-stage classifier to produce the final activity classification. The approach is also extended to detect activities of hand-object interaction. The method is experimentally shown to significantly outperform a state-of-the-art sliding window detector for hand detection. It is particularly robust for handling cases of self-occlusion and other occluding objects (see figure 1), as well as in reducing the false positives from appearance artifacts.

The second contribution is in proposing an interactive hand gesture module. To that end, we present a real-time, RGB and depth-based vision system for hand detection and gesture recognition. The method is implemented using a spatio-temporal feature for RGB and depth images based on a modified histogram of oriented gradients (HOG) [4] applied spatially as well as temporally [17].

2. Related Work

Vision-based hand detection is challenging, primarily because of the wide range of configurations and appearances it can assume and its tendency to occlude itself in images. The problem is further complicated by the vehicular requirement for algorithms to be robust to changing illumination. Hand detection was mostly studied in indoor settings, where the hand is segmented in a naive manner. For instance, the hands may be the main salient object in the scene in terms of motion [6], skin-color [10, 21], or it may be segmented using a depth-based threshold [14]. As single cues, such techniques were shown to perform poorly on our dataset. The more reliable schemes were edge-based boosting schemes [13, 19]. Such schemes may incorporate an arm detector as well [11]. A close work to ours is found in [15] where a shape, arm, and skin-based detectors are integrated to achieve state-of-the-art on several benchmarks. Because hand detection of the entire scheme takes over 2

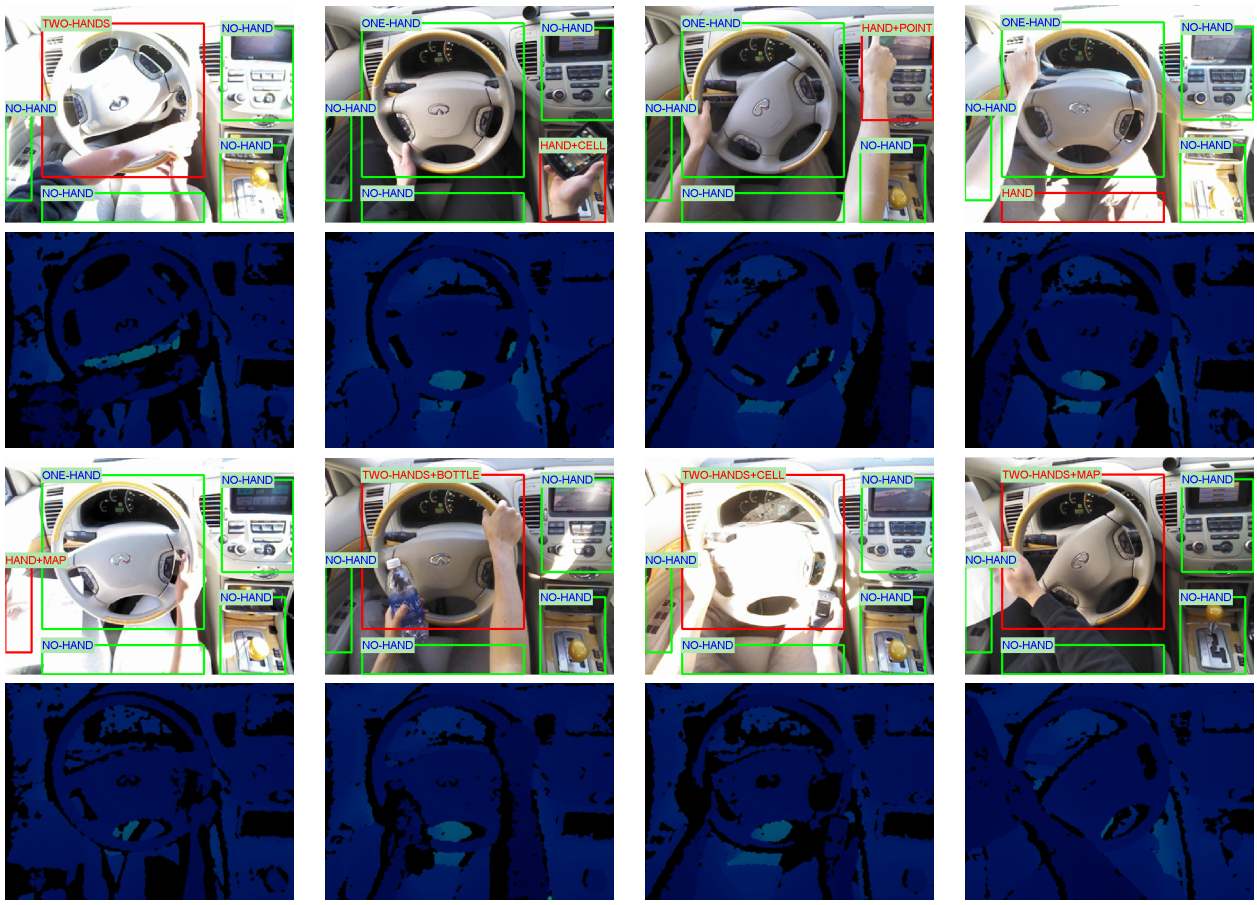


Figure 1: The naturalistic **CVRR-HANDS 3D** dataset, collected while observing the driver of an operating vehicle. One proposed evaluation is in terms of detecting hand presence or hand-object activity in five regions of interest, for which the labels are visualized. The dataset poses many challenges, such as volatile illumination and frequent occlusion of objects and hands. In addition to hand location, six classes of hand-object interaction types and 11 hand swipe gestures for human-machine interaction were annotated.

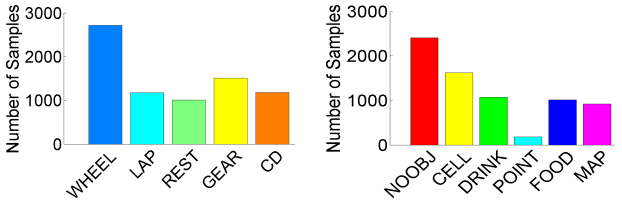


Figure 2: Hand (left) and object (right) occurrence distribution in our dataset. *NOOBJ* are instances in the dataset with hand but no hand-held object.

minutes per image, we only use the hand shape model as a baseline. This model was built using a deformable part model (Felzenszwalb *et al*. [9]) and trained on hand instances from several hand datasets and the PASCAL VOC dataset (see [15]).

We tie hand and object detection with activity recognition by proposing a hand-gesture based module for occupant-vehicle interaction. One reason for this is in order to study RGB and depth-based spatio-temporal descriptors performance under the harsh visual settings, including illumination artifacts and occlusion. A second reason is domain specific: as we are concerned with driver distraction caused by secondary tasks-a contact-less hand gesture-based interface may be intuitive for the driver. To that end, we incorporate the frame-by-frame features extracted for hand detection using a modified HOG algorithm in order to produce a fast to compute, spatio-temporal descriptor for gestures [17]. The method is compared against two other common spatio-temporal feature extraction methods: the Cuboids descriptor proposed by Dollár *et al*. [7] and the motion history image (MHI) [1] coupled with HOG. These are studied both on the RGB and depth images.

3. Dataset - CVRR-HANDS 3D

We introduce a publicly available dataset of synchronized RGB and depth videos collected in an operating vehicle (available at http://cvrr.ucsd.edu/eshed). The experiments in this work are performed in cross-subject testing over a subset of the dataset containing 7207 sample frames, with subjects performing different tasks while driving (figure 1). Our goal is two-fold: study recognition of naturalistic driver gestures that are related to driver attention, and propose an interactive framework for hand gesture-based user interface. The main observation that motivates our approach is that hand presence in a certain region can be detected, but the difficult illumination and quality makes sliding-window detectors over the entire image perform poorly with many false positives. Therefore, we constrain the problem into a number of regions of interest (ROIs) that researchers may be interested in for studying the driver's state. This leads to a better posed problem, where regions can be integrated to produce the final activity classification. Hands locations were annotated, and we evaluate activity in terms of region (figure 2): Wheel, lap, hand rest, gear, and instrument panel). The region of activity is generally defined as the one with the most area occupied by the hand out of the five, unless its the lap region where annotation in lap must involve no interaction with the wheel. Furthermore, presence of objects in the region is annotated as well for six classes (figure 2-right). The classes were chosen to represent common secondary tasks performed in the vehicle and may be related to less attentive driving [20]. The dataset in whole spans more than 80 minutes of video, allowing for future evaluation of more intricate driver states and maneuver gestures.

In a particular ROI, gestures may be performed for occupant-vehicle interaction. Therefore, we collected a dataset of 11 dynamic gestures: *left-right*,*right-left*, *down-up*, *up-down*, *X-swipe*, *Plus-swipe*, *Z-swipe*, *V-swipe*, *N-swipe*, *clockwise-O* and *counter clockwise-O* swipes. There are 807 and 864 instances of gestures performed by the driver or passenger respectively under different times of the day in different vehicles. A total of 7 subjects participated (each subject performed the gestures both as driver and passenger). In the next section, we turn to our developed approaches for evaluating the hand, object, and interactive gestures categories of the CVRR-HANDS 3D dataset.

4. Naturalistic Driver Gestures Recognition

4.1. Hand and Hand+Object Event Detection in Regions

In the context of driving, naturalistic gestures are dictated by the location of the hands. Localizing the hands is therefore the main evaluation on the dataset. This proved quite challenging as partial-occlusion of the hands occurs often. The method in [15] showed best results compared to other detectors and trackers on our dataset, although performance overall was still quite poor. For instance, a multi-object version of the Tracking-Learning-Detection scheme proposed in [22] failed to track hands correctly in the majority of the frames, even under slight deformation and occlusion. Nonetheless, the dataset can be used for evaluating tracking techniques, yet due to poor performance of state-of-the-art methods such analysis is left for future work. Seeking robustness, we were motivated to introduce a ROIs, which system integrates cues from in order to perform the final activity classification.

We expect the hand to vary in appearance among the different regions. Some regions may require finer-detailed descriptors as parts of the hand may be present while interacting in another region (e.g. the hand may be interact with the instrument panel but be present in the gear region, or similarly for the wheel and lap regions). Furthermore, The size and location of each region produce different challenges for a vision-based system. Therefore, we thoroughly study different descriptors in terms of performance and computational complexity.

4.2. Features

We detail the features we found useful for hand and hand+object detection, with the dimensionality and extraction time give for the largest region, the wheel, in table 1.

Modified HOG (MHOG): The algorithm has been previously used for hand detection [16], as well as action recognition [17]. MHOG differs from HOG mainly in the division of the image into subcells. The parameters are the number of cells to divide the image into in the horizontal and vertical directions, where a 50% overlap between the cells was used. Within each cell, an orientation histogram is generated by quantizing the angles of each gradient vector into a pre-defined number of bins. These resulting histograms are concatenated to form the final spatial feature vector. For instance, a 3×3 grid of cells with 9 histogram bins on the image results in a 81D feature vector.

HOF: The IMHwd descriptor from [5] was used, which amounts to applying Haar-like operator on the optical flow image. The optical flow was calculated between current frame and three frames before. The parameters were set such that the cells are approximately the width of an arm, in the hopes of capturing relative displacement of the hand with respect to the background.

Difference of HOG (DIFFHOG): Haar-like operator on the HOG descriptor produces this descriptors.

GIST: Another widely known image descriptor proposed in [18]. Although it is slower to compute, it proved successful in difficult cases of hand detection. For instance, when the hand is interacting with the instrument panel region, part of it or part of the arm may be in the gear region.

Descriptor	Extraction Time (ms)	Descriptor Size
HOG99	6	11780D
MHOG11	10	9D
HOF88	13	1155D
DIFFHOG	10	9D
GIST8	370	2048D
Skin	10	4D
EUC	4	14535D
GLOBAL	1	3D

Table 1: Analysis in terms of speed and dimensionality for each descriptor. The original HOG (HOG99) is the only descriptor with a fast implementation, the rest are in MATLAB, and are likely to be faster than the original HOG extraction once implemented efficiently. The approximate times and sizes are given for the largest ROI, the wheel region. The parameters are followed after the name, where HOG99 has cell size 9, MHOG11 is a 1×1 split to cells, and both are fixed at 9 orientation bins. GIST8 means the free parameters are set to 8.

The GIST significantly outperformed HOG under these settings.

Skin: In order to obtain a skin segmentation model specific to the user, the user's skin color is obtained by an initialization where the driver was asked to maintain the hands over the wheel and in front of the sensor. The hands are segmented using the depth values, and a color likelihood classifier is then constructed in the L*a*b color space. The final descriptor is 4D: the area and area/perimeter ratio of the two largest connected components in the image.

EUC: By applying a distance function between column pixel intensities of an image, such as the Euclidean distance, this feature is produced. In [16] it was shown to perform well on some of the regions, in particularly on the depth image.

GLOBAL: The median, mean, and variance of the intensities in the image.

4.3. Learning from Sparse Exemplars

A linear SVM is learned for each of the regions with a different set of color and depth-based features. Datasets such as the one in this work usually contain sets of unbalanced data-presence of hand in a certain region may be significantly more rare than in other regions (such as the wheel region). Nonetheless, we would like to preserve the large spectrum of training samples to fully capture the intra-class variations in appearance.

We address this through penalizing parameters in a max-margin linear SVM formulation. Given training vectors $x_i \in \mathbb{R}^n$ and $y_i \in \{-1, 1\}$, we use LIBSVM [2] to solve the following optimization problem:

$$\begin{aligned} \min_{w,b,\xi} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} + C^+ \sum_{t_i=1} \xi_i + C^- \sum_{t_i=-1} \xi_i \\ \text{subject to} \quad & t_i(\mathbf{w}^T\phi(x_i) + b) \geq 1 - \xi_i, \\ & \xi_i \geq 0,\ i = 1, \ldots, l. \end{aligned} \tag{1}$$

4.4. Hypothesis Integration from ROIs

By training five individual SVMs, a score is obtained for each region in testing. The five scores can be combined to form a feature vector, and a second-stage linear SVM is learned using the scores. Learning a confidence for each region leverages information from different regions under difficult settings. This is useful because smaller regions are expected to have higher accuracy of hand or object recognition. Other regions might be less prone to illumination changes. Furthermore, it allows for learning patterns of activities with multi-region cues better. For example, it alleviates false positives in situations where the arm in one region and the hand in another. Finally, it provides improved recognition under occlusion, where one hand may not be visible but both are on the wheel (there is no hand in the peripheral regions).

4.5. Evaluation Criteria

Due to unbalanced number of class instances, performance is measured in terms of normalized accuracy (average correct classification rate-CCR)

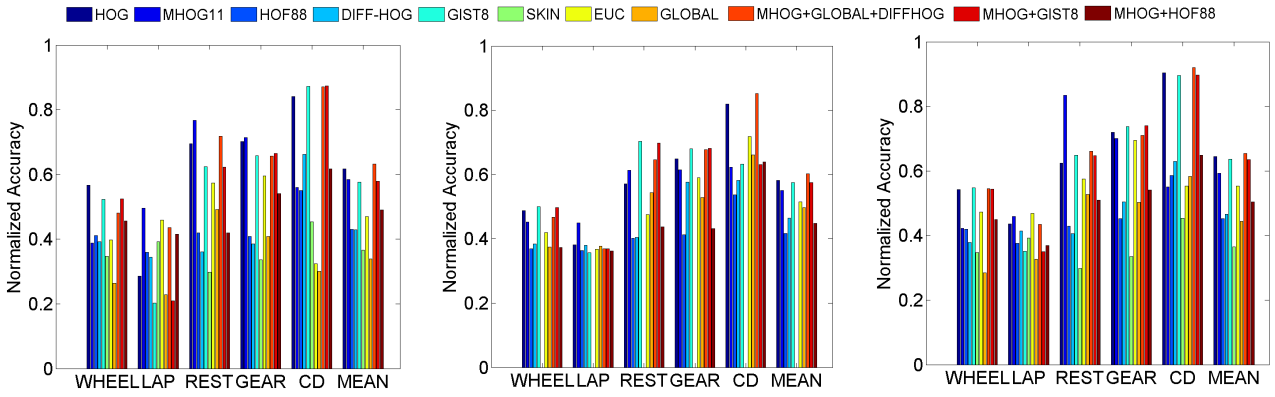
$$CCR = \frac{1}{K} \sum_{c=1:K} p_c \tag{2}$$

where K is the total number of classes, and p_c denotes the percentage of correctly matched instances for class c.

5. Experimental Evaluation - No Hand, Hand, and Hand+Object Events

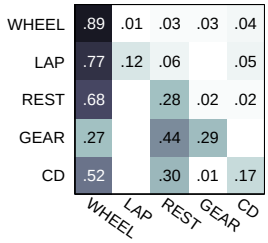
We show top performing descriptors and combinations of descriptors in figure 4 for the three class activity recognition problem of no hand, hand or hand+object detection in each region. Classification of the type of object that is being used out of the six object classes is left for future work.

The top performing descriptors varied for the different regions (figure 4)-mostly alternating between GIST, HOG, and modified HOG. Throughout the two modalities and their combination, the modified HOG together with GLOBAL and DIFFHOG produced the highest results on average. Nonetheless, top performance varied significantly among regions-mostly alternating between GIST, HOG, and MHOG. The EUC descriptor of Euclidean distance among column pixel intensities was useful for analyzing activity in the lap region. Motion (HOF) wasn't shown



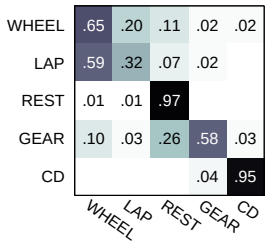
(a) RGB (b) Depth (c) RGB+Depth

Figure 4: Results of normalized accuracy for the three class problem of no-hand, hand, and hand+object detection in each region using top performing descriptors and combinations of descriptors. The analysis is in terms of modality: RGB, Depth, and combined RGB and Depth (concatenated descriptors).



(a) Baseline: **35.2**%.

(b) This work RGB only: **52.1%**.



(c) This work - RGB and Depth: **69.4%**.

Figure 3: Activity recognition as a five class problem of hand localization in five regions. (a) The baseline hand model in [15], a mixture model over three components based on Felzenszwalb *et al*. [9] and trained on hand instances from several hand datasets and the PASCAL VOC dataset. Testing is done at 36 different rotations of the image-10° intervals. (b) and (c) is our ROI integration framework.

to provide performance improvement, and motion cue extraction needs to be further studied.

We choose a collection of the top performing descriptors for each region, and use it to analyze the five class activity classification problem within the five regions (figure 3). All five and final SVM learned use a linear kernel and a one-vs-one multiclass classification. In this evaluation, we are only concerned with whether there is a hand in one of the four peripheral regions or whether both of the hands are in the wheel region. We use the entire dataset containing both instances of hand in the different regions as well as hand holding object. As a baseline, we use the hand shape model from [15] built using a deformable part model (Felzenszwalb *et al*. [9]) and trained on hand instances from several hand datasets and the PASCAL VOC dataset (see [15]). Testing is done at 36 rotations. The technique is significantly slower than ours, and we reach close to or real time (depending on the descriptor used). Secondly, by learning multi-region cues for activity recognition, difficult cases of occlusion are better handled. We notice a significant increase in performance, especially in the rest, gear, or CD regions. The baseline suffers from large amounts of false-positives in the wheel region. Taking four of the top scored detection windows from the baseline (as opposed to just the top two) and checking for activity other than in the wheel region by taking the maximum activity index from one to five (ordered as in figure 3) leads to an improvement overall performance (up to 41.7%, but the wheel region accuracy goes down from 89% to 63%).

6. Experimental Evaluation - Interactive Hand Gestures

The gesture dataset in this work is unique compared to existing datasets as the hand is facing away from the sensor (leading to more self-occlusion) and data was captured in naturalistic driving settings. Furthermore, gestures were performed by the passenger and driver leading to variations in the performance of the gestures. In the evaluation, both

Method	RGB	Depth
Cuboids	20.62%	15.55%
MHI	20.97%	15.21%
HOG^2	**46.65%**	**35.27%**

Table 2: Comparison of gesture classification using three different spatial-temporal feature extraction methods: 1) cuboids [7] with a flattened gradient descriptor 2) HOG applied to the motion history image (MHI) [1] 3) HOG applied to the collection of spatial HOG descriptors over time, HOG^2 [17]. Average correct classification rate is reported using cross-subject cross-validation.

the MHI scheme and the HOG^2 descriptor are inputted to a linear SVM. The parameters for the Cuboids descriptor (see [7]) were grid optimized. Out of the three techniques, the HOG^2 gives the best classification on the dataset for each modality and user (table 2). The depth images are sensitive to illuminations and reflective screens in the car, hence the lower performance. The training and testing for HOG^2 is also significantly faster than the Cuboids descriptor. It's computationally efficient due to the use of the feature set extracted from the initial hand detection step of the system. The implemented system is lightweight, with about 10 ms/frame for spatial feature extraction at every frame and 10 ms for the re-application of the modified HOG on the 2D array of collected histogram descriptors over time. On average, spatio-temporal extraction and gesture classification can be done at about 14 ms/frame.

7. Concluding Remarks

We presented a rich activity recognition dataset in order to contribute to the development of algorithms that work well in naturalistic settings of human activity. A hand localization framework was introduced, with an analysis of different RGB and depth image descriptors for object detection and activity recognition. Interactive gesture recognition was done using a fast bag-of-words approach leading to a real-time interaction module.

8. Acknowledgments

This research was performed at the UCSD Computer Vision and Robotics Research and LISA: Laboratory for Intelligent and Safe Automobiles. We are grateful for the sponsors supporting the lab and to our colleagues.

References

[1] A. Bobick and J. Davis. The recognition of human movement using temporal templates. *PAMI*, 2001.

[2] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Intell. Sys. and Tech.*, 2011.

[3] S. Y. Cheng and M. M. Trivedi. Vision-based infotainment user determination by hand recognition for driver assistance. *IEEE Trans. Intell. Transp. Syst.*, 11(3):759–764, Sep. 2010.

[4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[5] N. Dalal, B. Triggs, and C. Schmid. Human detection using oriented histograms of flow and appearance. In *ECCV*, 2006.

[6] M. V. den Bergh and L. V. Gool. Combining rgb and tof cameras for real-time 3d hand gesture interaction. In *WACV*, 2011.

[7] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *VS-PETS*, October 2005.

[8] M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge 2012 (VOC2012) results.

[9] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *PAMI*, 2010.

[10] V. Harini, S. Atev, N. Bird, P. Schrater, and N. Papanikolopoulos. Driver activity monitoring through supervised and unsupervised learning. *IEEE Trans. Intell. Transp. Syst.*, 2005.

[11] L. Karlinsky, M. Dinerstein, D. Harari, and S. Ullman. The chains model for detecting parts by their context. In *CVPR*, 2010.

[12] 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. Technical Report DOT HS 811 334, National Highway Traffic Safety Administration, Washington, D.C., 2010.

[13] M. Kolsch and M. Turk. Robust hand detection. In *Int. Conf. Autom. Face and Gesture Recog.*, 2004.

[14] A. Kurakin, Z. Zhang, and Z. Liu. A real time system for dynamic hand gesture recognition with a depth sensor. In *EUSIPCO*, 2012.

[15] A. Mittal, A. Zisserman, and P. H. S. Torr. Hand detection using multiple proposals. In *BMVC*, 2011.

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

[17] E. Ohn-Bar and M. M. Trivedi. Joint angles similarities and HOG^2 for action recognition. In *CVPRW*, 2013.

[18] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, 2001.

[19] E. Ong and R. Bowden. A boosted classifier tree for hand shape detection. In *Int. Conf. Autom. Face and Gesture Recog.*, 2004.

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

[21] J. Y. X. Zhu and A. Waibel. Segmenting hands of arbitrary color. In *Int. Conf. Autom. Face and Gesture Recog.*, 2012.

[22] K. Zdenek, J. Matas, and K. Mikolajczyk. Pn learning: Bootstrapping binary classifiers by structural constraints. In *CVPR*, 2010.