In-Vehicle Hand Activity Recognition Using Integration of Regions

Eshed Ohn-Bar and Mohan Trivedi, *Fellow, IEEE*

***Abstract*— In this paper, we focus on the analysis of naturalistic driver behavior using hand activity. To that end, a dataset of color and depth images under varying operating modes and illumination settings was collected. The proposed framework provides a robust solution for localizing the hands by partitioning visible and depth images into disjoint subregions which may be of interest for studying the state of the driver: wheel, lap, hand rest, gear, and infotainment region. Different feature extraction methods are proposed and thoroughly studied in terms of speed and performance for each of the five regions. A model for hand presence is learned for each region separately, and these are integrated using a second-stage classifier. As the appearance of hands varies among regions and the hands can only be found in a subset of the regions chosen, the technique leverages information and confidence from multiple regions to produce hand activity classification.**

I. INTRODUCTION

The study of object detection and tracking, in particular of human hands, has been widely investigated in the computer vision community. This is due to the fact that hands are an important medium for conveying information in human-machine interactivity. Inferring information from hand activity is especially important in the vehicle context, because it may provide vital information about the state of attentiveness of the driver. Together with head pose and other cues, it was previously shown useful for attention monitoring and driver turn intent prediction [1].

Driver distraction is a leading cause of car accidents [2]. There has been extensive psychological research done on quantifying levels of driver distraction in terms of measurable metrics, such as total eyes off the road time and maximum glance duration. Performing secondary tasks in the vehicle has been shown to increase inattentiveness, which, in 2012 was a contributing factor in at least 3092 fatalities and 416,000 injuries [3]. The most known example is of using a cell-phone while driving, which may require visual, manual, and cognitive attention, which significantly hinders driver awareness and reaction capabilities. Other secondary tasks that were also shown to produce increased distraction and are of interest to the scientific community are: reading printed material, eating or drinking, interacting with in-vehicle devices, and grooming [2].

Studying what hands do and where in the vehicle has never been a more pressing matter, as the performance of distracting tasks while driving is widespread. 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 [4]. 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. Naturally, despite of the danger, modern humans seek out such interactions and activities when the primary tasks of driving require a decreased demand.

The authors are with the Department of Electrical and Computer Engineering, University of California San Diego (UCSD), La Jolla, CA, 92092 USA (e-mail: eohnbar@ucsd.edu; mtrivedi@ucsd.edu).

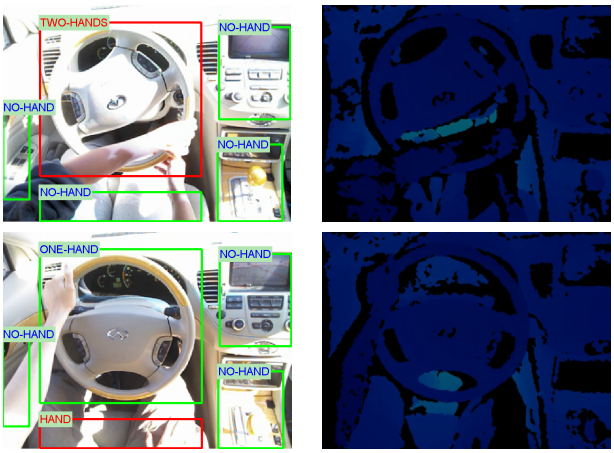


Figure 1: Example images from the dataset. The in-vehicle environment provides a challenge for a vision-based hand localization system. The proposed approach identifies hand activity in chosen sub-regions which may be of interest to researchers studying driver's state. Images were acquired using a Kinect camera. Left: RGB input, right: depth input.

As an initial, yet important step towards hand and hand-object activity recognition we introduce a case study in which we study hand localization in the vehicle. In particular, we are interested to know whether the hand is engaged in a specific region of the vehicle or not, and how many hands are on the wheel. As the hand occludes itself and is subjected to occlusion frequently in our dataset, common trackers (such as recent successful schemes of Tracking-Learning-Detection [5]) and the cascade detectors (based on [6]) performed poorly. Furthermore, the volatile illumination and dense background provide additional challenges for hand detection, under which existing techniques have not been tested. The desired outcome of our algorithm would be to reliably detect hand-related events, which can then be inputted to a semantic model of the scene. Using such model, knowledge of hand activity in the vehicle could result in a suitable assistive technology.

We would like to study hand and hand-object activities in the car using a Kinect camera. The unsatisfactory results of

available state-of-the-art hand detection and tracking algorithms produced on our collected dataset, both in terms of speed and accuracy, inspired the multi-cue, comprehensive analysis of features derived from both RGB and depth modalities. In addition to such novel experimental analysis, we propose a scheme for further cue integration from multiple regions of interest (ROIs). Due to the difficult problem of hand tracking under volatile illumination changes, we propose a framework of hand activity analysis in which the image is partitioned to five sub-regions. These are strategically located in key interest regions as shown in Fig. 1. Restricting the problem to multiple ROIs provides us with the ability to reason over information and confidence from these smaller regions in order to produce an activity classification with higher confidence. We experimentally show how the scheme benefits activity detection in large and difficult areas, such as the wheel region, as well as significantly suppressing a state-of-the-art hand detector.

In the peripheral, smaller sub-regions around the wheel we will study two classes: hand or no hand. The central region of the wheel will be studied in terms of 'normal' and 'abnormal' events, where two hands in the region is defined as 'normal', and one or no hand is defined as 'abnormal'. Next, different image descriptors will be compared, including skin-based detection, a modified form of the histogram-of-oriented-gradients (HOG) descriptor [7], [8], and other image features such as GIST [9] (Section III). We also propose additional descriptors which were shown to improve the detection performance when used with the aforementioned. Both the individual first-stage classifier in each ROI, and the second-stage classifier will employ a linear SVM.

II. Related Research Studies

Most existing works involve hand detection under indoor or naive settings. Under such constraints, the hand may be the main salient object in the scene or exhibiting the most motion [10], skin-color techniques may be used [11], [12], [13], [14], or a depth-based threshold could provide the main cue [15]. As single cues, such techniques were shown to perform poorly on our dataset. The more reliable schemes were edge-based boosting schemes [16], [17]. A related work to ours is in Mittal *et al.*[18] where a shape, arm, and skin-based detectors are integrated to achieve state-of-the-art on several benchmarks. The method runs at about two minutes per frame, and so we seeked a faster solution. Furthermore the base model of hand shape built using a deformable part model (Felzenszwalb *et al.* [6]) had a significant amount of false positive detections, even in images without volatile illuminations. The hand model (trained on the PASCAL VOC challenge [19] and several other hand datasets [18]) will be used as the baseline for our method, but is still significantly slower in comparison to the proposed approach.

Most of the work published in hand detection and tracking using depth images make use of the depth for segmentation purposes, for instance as the closest object to the camera [15]. In the vehicle, even assuming no light interference with the depth input, such an approach provides poor detection since most of the time the hand lies in the same plane as other objects and can not be easily separated by depth. Nonetheless, depth information provides a distinguishing volumetric representation that can be leveraged to detect hands and objects in the scene in other ways, as will be shown.

Our work is inspired by the robustness of the system proposed in [20], where a HOG and a RBF SVM were used to detect whether there's a driver, passenger, or no hand in monochrome images of the hand rest by the gear shift. We extend the work to infer hand location in multiple regions in the vehicle. The main observation that motivates our approach is that hand presence in a certain small region can be detected, but the difficult settings makes sliding-window detectors over the entire image perform poorly with many false positives. Therefore, we constrain the problem into a number of sub-regions (ROIs) that researchers may be interested in for studying the driver's state.

III. Hand Event Detection from Multiple Cues

A linear SVM model is learned for each region using a different set of descriptors (as the hand appears differently in different regions). Then, the probability output from these SVMs are given to a second-stage linear SVM to perform the final activity classification. The features that will be compared for the purpose of detecting a hand or hands in a sub-region are described below. Some are well known image descriptors, yet they were little tested before on a depth image. Furthermore, their performance in terms of speed, complementary information to other descriptors, and the size of ROI need to be thoroughly studied. For each descriptor we also specify the extraction time for extraction in the largest ROI, the wheel region, using a MATLAB implementation.

Modified HOG: This is a modified version of the original HOG descriptor [7], inspired by [20]. It has been used before for the purpose of hand detection in a small area in the vehicle. The modified HOG descriptor is created as follows: the gradient image of the image patch is divided into rectangular cells along the x- and y-directions. Unlike in [20], we use a 50% overlap between the cells. 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 2×2 grid of cells with 8 histogram bins on the image results in a 32-D feature vector. The histograms are normalized in each cell according to the *L2-norm* scheme in [7]. **Extraction Time: 10 ms.**

Difference of HOG (DIFFHOG): This feature involves differences of the modified HOG descriptor in sub-patches of a region of interest. It's computed for image I as an absolute value of a difference of HOG features in the left and right part of an image:

$$DIFFHOG(I) = |HOG(I(:, 1 : m/2)) - HOG(I(:, m/2 + 1 : end))| \quad (1)$$

The above is given in MATLAB notation, where the colon operator in the first coordinate gives the range of all of the rows in the image, and the image has m rows. We show that this operation contains complementary information to the modified descriptor, possibly by better capturing symmetry (important for the wheel region). **Extraction Time: 10 ms.**

GIST: Another widely known global image descriptor is the GIST descriptor [9]. Although it is slower to compute, it proved successful in difficult cases of hand detection. For instance, when the hand is interacting with the CD region, part of it or part of the arm may be in the gear region. The GIST significantly outperformed HOG under these settings. **Extraction Time: 370 ms.**

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 composed of the area and area/perimeter ratio of the two largest connected components in the image. **Extraction Time: 10 ms.**

EUC: By applying a distance function directly on the column pixel intensities on an image, this descriptor measures co-variance features of pixels. Given two columns in an image, $I(:,i), I(:,j)$, the Euclidean distance between them is

$$EUC(I(:,i), I(:,j)) = \|I(:,i) - I(:,j)\|_2 \tag{2}$$

Extraction Time: 4 ms.

GLOBAL: Statistical properties in the region provides a rough depth or pixel indicator for an object presence. This feature set (3-dimensional) is composed of

$$GLOBAL(I) = \begin{bmatrix} mean(vec(I)) \\ median(vec(I)) \\ var(vec(I)) \end{bmatrix} \tag{3}$$

where *vec* is the vectorize operation. **Extraction Time: 1 ms.**

A. SVM Classification

We will use a linear kernel SVM both for individual ROI hand or no hand detection, and abnormal wheel region hand activity (one or no hands on the wheel vs. two hands). The suitability of the our scheme on a large and difficult region-such as the wheel region-is not trivial, as the hand can appear in many different parts of the region. We will provide deeper analysis of this region in the evaluation Section IV.

Because the dataset collected is unbalanced-hand events in the peripheral ROIs may be rare-we slightly modify the classical SVM formulation. This is done to preserve all of the samples in training, which is desired because the dataset contains a large amount of intra-class variation in the scene. One possible way to address this is through penalizing parameters in the SVM formulation so that the optimization problem is written as

$$\begin{aligned} \min_{w,b,\xi} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} + C^+\sum_{t_i=1}\xi_i + C^-\sum_{t_i=-1}\xi_i \\ \text{subject to} \quad & t_i(\mathbf{w}^T\phi(x_i) + b) \geq 1 - \xi_i, \\ & \xi_i \geq 0,\ i = 1, \ldots, l. \end{aligned} \tag{4}$$

and LIBSVM [21] can be used in order to solve the max-margin problem.

B. Integration of Multiple Regions

The following assumption is made: a hand can only exist in a subset of the regions. Therefore, a high confidence of the hand in a specific region can be used to infer likelihood of another hand being in a different region. For example, 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 (which has a weaker definition of the classes as it is large and prone to illumination change).

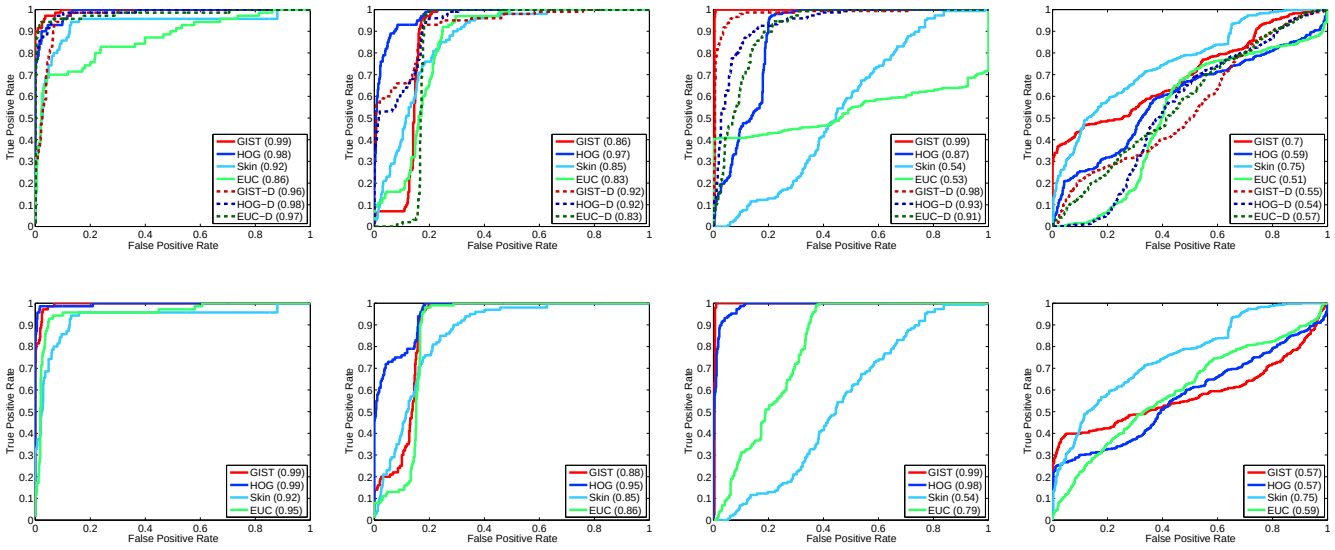
We obtain an SVM model trained on a combination of RGB and/or depth-based descriptors of either: 1) Hand or no hand in the ROI (in the peripheral ROIs) 2) Two hands or one or no hands in the ROI (the center wheel ROI). By training five individual SVMs, a score $\alpha_i, i \in \{1, \ldots, 5\}$ is obtained for each region. The five scores can be combined to form a feature vector, and a linear SVM classifier is learned using the scores from multiple ROIs. This second-stage classifier is used to make the final activity classification.

IV. Experimental Setup in LISA Testbeds

In order to demonstrate the feasibility of the proposed system, four video sequences during a total of 47 minutes of video in different illumination conditions were collected out of which 1923 samples were extracted. The Kinect was mounted behind the driver's head. Five ROIs, seen in Fig. 1, were defined and monitored. In annotation, the region of activity was the one with the most area occupied by the hand out of the five. Because this is not sufficient to define whether a hand is in the lap region (see Fig. 4-top), we require less than 10% overlap between the bottom part of the wheel region and the hand box in order to annotate the hand location ROI as lap. Training and testing was performed using cross-subject testing.

V. Experimental Evaluation and Discussion

Fig. 2 shows the results of the top performing descriptors on the four peripheral ROI: the lap, hand rest, gear, and infotainment area. As each ROI differs in size, area, and exposure to background and illumination changes, different results for each ROI are exhibited. The clearly defined regions enclosing the hand well are the infotainment and gear regions, which show high performance despite being prone to illumination changes. As the hand crosses between the lap and wheel regions, it produces hand event cues in both regions. Because the arm is also present in the lap region as it's defined, this creates a poor separation in the feature space (see Fig. 5 and the confusion matrix in Fig. 7).



(a) Infotainment (b) Gear (c) Side Hand Rest (d) Lap

Figure 2: Linear SVM and top performing descriptors for the periphery ROIs. **Top**: descriptors performance using RGB or depth input ('-D' stands for descriptors derived from the depth image). **Bottom**: the same descriptors but with RGB and depth descriptors concatenated. In parenthesis is the AUC for the given descriptor. Skin cues are only available for color images.

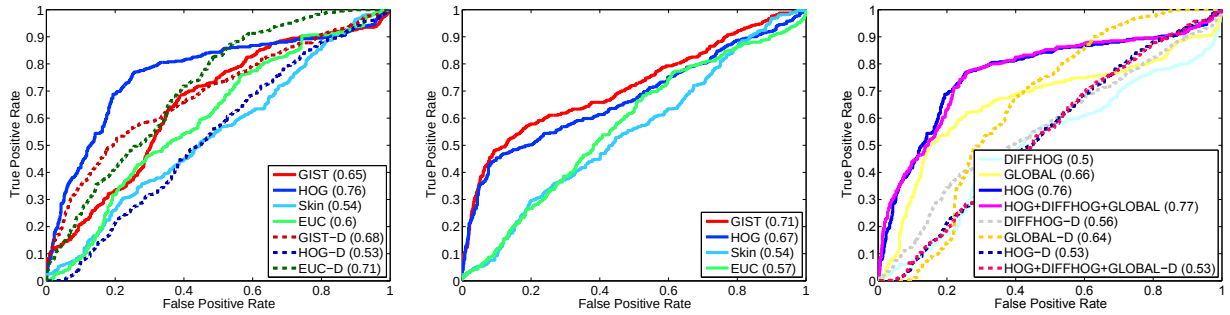


Figure 3: Linear SVM and top performing descriptors for the difficult wheel area. In parenthesis is the AUC for the given descriptor. **Left**: Evaluation with top performing descriptors when used separately. **Middle**: Evaluation of concatenated descriptors from color and depth image. **Right**: Evaluation of different combinations of the descriptors.

Generally, incorporating the depth and color cues together results in a small performance increase (Fig. 2, bottom row). The skin cues are only available for color input, but it's plotted together with the depth-based classifiers for comparison (Fig. 2- bottom). Skin-cues were shown to produce mixed results as there is an ambiguity in the feature space as both the hand and the arm can provide similar cues. Nonetheless, the skin cues are useful in the lap region, where the size of the blobs is generally bigger when an arm is in the scene as opposed to the hand only. Overall, we see that better descriptors that capture the properties of hand vs. arm and background need to be developed.

As previously mentioned, we aim to build on mutual information from the detectors in the ROIs to produce the final classification of whether there are two hands or not on the wheel and raise the overall performance of the hand localization scheme. Fig. 3 shows a similar analysis to the one in Fig. 2, but focuses on the difficult but important ROI-the wheel. This is due to the large area covered by this ROI, as well as volatile background and illumination (see Fig. 4 and Fig. 5 for example images).

We see that the one of the top performing descriptor in the wheel ROI, when used alone with a linear SVM, is a depth derived EUC descriptor (Fig. 3 left), significantly outperforming the other depth-derived descriptors. The modified HOG produced comparable results to the original HOG in all of the regions. We also note that concatenating the RGB and depth-based descriptors doesn't produce improved per-

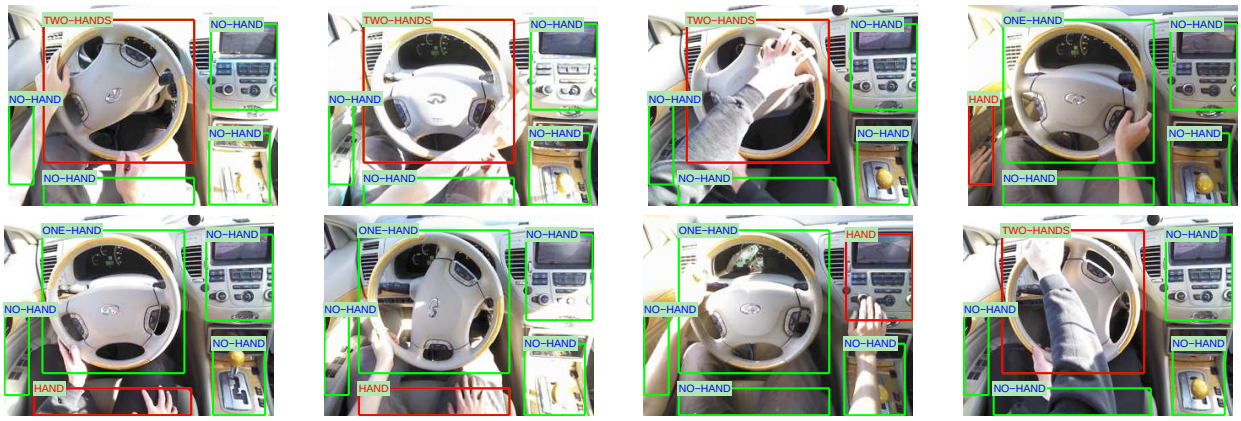
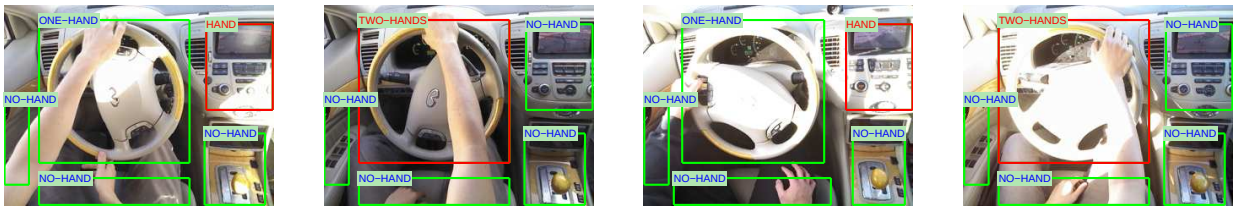


Figure 4: Correct classification results using the proposed approach. Difficult cases of illumination and occlusion are handled by incorporating information from the five regions.



(a) (b) (c) (d)

Figure 5: Incorrect classification using the proposed system. (a) Although the wheel model outputs a prediction of two hands in the wheel region, so does the infotainment due to an illumination artifact. In this case, the integration produces incorrect results since the model learns to give high confidence to the infotainment score. (b) The lap region produces incorrect classifications due to poor separation in the feature space. (c) and (d): Illumination produces false positives.

formance. Nonetheless combining different features, shown in Fig. 3 (right), produces a slight improvement in the AUC rate. Although feature sets such as DIFFHOG and GLOBAL perform poorly when used alone with a linear SVM, when incorporated with the modified-HOG descriptor, we reach a classifier for the wheel region with an AUC of 0.77. These additional descriptors are very fast to compute. A simple rule-based approach shows the strength of the integration scheme in Fig. 6, where the top performing descriptor combination was used from each region and used to infer where the hand activity in the wheel region as a two-class problem. The basedline is the HOG+DIFFHOG+GLOBAL, which is the best we can do without ROI integration. This ROI integration results in a significant increase to the AUC rate, from 0.77 to 0.92 (Fig. 6).

Finally, Fig. 7 depicts the results of the activity classification as a five-class classification problem with the ROI integration scheme. The basline is the hand shape model from [18], which is a HOG-based part-based deformable model of a mixture over three components [6]. Testing is performed at 36 different rotations of the 240×320 image (performance sharply reduces without this step). This detector runs at about 32 seconds per image on an Intel Core i7 3.02-GHz PC. Although the baseline outperforms at overall correct classification rate of 58.4% vs. our system at 53.4%, it's because of the poor definition of some of the regions. The proposed system suppresses the baseline in every region besides the hand rest. We also notice an incorrect bias learned towards the wheel region activity in the second-stage classification scheme, also possibly due to the regions with high ambiguities in the feature sets. Nonetheless, we can directly compare the performance of our scheme on regions of the CD and gear, where the baseline doesn't perform well. With three regions integration scheme (Fig. 7(c)), we can directly compare the results to the baseline in Fig. 7(a) as it's a sliding window detector. In this case, overall correct classification rate is 79.7% compared to the baseline performing at 67%.

VI. Concluding Remarks and Future Work

In this work, we proposed a system for addressing the difficult problem of hand activity classification in large regions. The feasibility of using a multi-cue integration system from multiple ROIs in order to improve overall event detection rates was experimentally validated. Future work would include extending the activity grammar to include additional hand-object interactions, and more intricate maneuvers and driver gestures. The hand activity model in this work can be used for semantic analysis of the scene, in combination with a head-pose model.

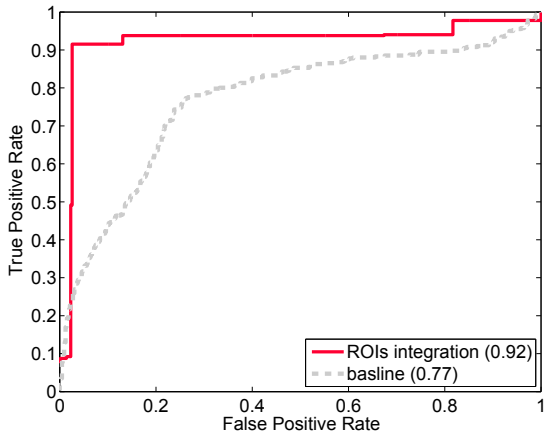
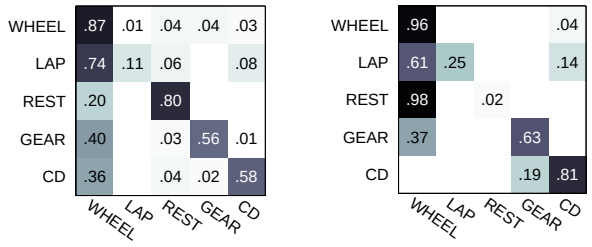
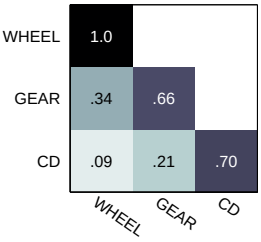


Figure 6: Evaluation of multiple ROI incorporation for detecting two hands or not in the wheel region. The baseline in this case is the top performing descriptor from Fig. 3 (right) and a first-stage classifier. The AUC rate of each scheme is shown in parenthesis. Detection performance of hand activity as a two class problem-in the periphery ROIs or the central, wheel ROI-is significantly improved.



(a) Baseline. (b) This work.



(c) This work with three regions.

Figure 7: Evaluation of multiple ROI incorporation in a second-stage classifier scheme using a linear SVM as a five-class and three-class problem. The baseline is the part-based deformable hand shape model from [18]. Well-defined regions produces a more reliable cue to the second-stage classifier, as shown in (c). Total correct classification rates for the five-class activity classification are (a) 58.4% and (b) 53.4%. For the three-class activity the baseline performs at 67% and the proposed work at (c) 79.7%.

VII. Acknowledgment

The authors would like to acknowledge support of the UC Discovery Program, associated industry partners, and the reviewers for their constructive comments. We would like to thanks the members of the LISA laboratory for their help in data collection process, and Mr. Sayanan Sivaraman for his helpful advice.

References

[1] S. Y. Cheng and M. M. Trivedi, "Turn-intent analysis using body pose for intelligent driver assistance," *IEEE Pervasive Computing*, vol. 5, no. 4, pp. 28–37, Dec. 2006.

[2] 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.

[3] 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.

[4] 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.

[5] K. Zdenek, J. Matas, and K. Mikolajczyk, "P-N learning: Bootstrapping binary classifiers by structural constraints," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2010.

[6] 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,*, 2010.

[7] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Conf. Computer Vision and Pattern Recognition*, vol. 1, 2005, pp. 886–893.

[8] E. Ohn-Bar, C. Tran, and M. Trivedi, "Hand gesture-based visual user interface for infotainment," in *Conf. Automotive User Interfaces and Interactive Vehicular Applications*, 2012.

[9] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Intl. Journal of Computer Vision*, vol. 42, no. 3, pp. 145–175, May. 2001.

[10] M. V. den Bergh and L. V. Gool, "Combining rgb and tof cameras for real-time 3d hand gesture interaction," in *IEEE Workshop on Applications of Computer Vision*, 2011.

[11] 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.

[12] J. Y. X. Zhu and A. Waibel, "Segmenting hands of arbitrary color," in *Intl. Conf. Autom. Face and Gesture Recog.*, 2012.

[13] C. Tran and M. Trivedi, "Driver assistance for "keeping hands on the wheel and eyes on the road"," in *IEEE Conf. Veh. Electron. Safety*, 2009.

[14] R. Crespo, I. M. D. Diego, C. Conde, and E. Cabello, "Detection and tracking of drivers hands in real time," in *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, 2010.

[15] A. Kurakin, Z. Zhang, and Z. Liu, "A real time system for dynamic hand gesture recognition with a depth sensor," in *European Signal Processing Conf.*, 2012.

[16] M. Kolsch and M. Turk, "Robust hand detection," in *Intl. Conf. Autom. Face and Gesture Recognition*, 2004.

[17] E. Ong and R. Bowden, "A boosted classifier tree for hand shape detection," in *Intl. Conf. Autom. Face and Gesture Recog.*, 2004.

[18] A. Mittal, A. Zisserman, and P. H. S. Torr, "Hand detection using multiple proposals," in *British Machine Vision Conf.*, 2011.

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

[20] S. Y. Cheng and M. M. Trivedi, "Vision-based infotainment user determination by hand recognition for driver assistance," *IEEE Trans. Intell. Transp. Syst.*, vol. 11, no. 3, pp. 759–764, Sep. 2010.

[21] 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, software available at http://www.csie.ntu.edu.tw/˜cjlin/libsvm.