IEEE 18th International Conference on Intelligent Transportation Systems, 2015 - To Appear

驾驶员手势检测算法的性能评估

On Performance Evaluation of Driver Hand Detection Algorithms: Challenges, Dataset, and Metrics

Nikhil Das, Eshed Ohn-Bar, and Mohan M. Trivedi
Laboratory of Intelligent and Safe Automobiles
University of California, San Diego
La Jolla, California 92093

***Abstract*—Hands are used by drivers to perform primary and secondary tasks in the car. ~~Hence, the study of driver hands has several potential applications, from studying driver behavior and alertness analysis to infotainment and human-machine interaction features~~. The problem is also relevant to other domains of robotics and engineering which involve cooperation with humans. In order to study this challenging computer vision and machine learning task, our paper introduces an extensive, public, naturalistic video-based hand detection dataset in the automotive environment. The dataset highlights the challenges that may be observed in naturalistic driving settings, from different background complexities, illumination settings, users, and viewpoints. In each frame, hand bounding boxes are provided, as well as left/right, driver/passenger, and number of hands on the wheel annotations. Comparison with an existing hand detection datasets highlights the novel characteristics of the proposed dataset.**

I. Introduction

The detection and tracking of human hands has been studied extensively in the vision and learning community. In more recent years, the field has seen growing interest with the introduction of cheaper range sensors [1], [2], egocentric applications [3]–[6], and driver study [7]–[11]. Until recently, the majority of studies have emphasized human-machine interaction (HMI) applications and gesture analysis in relatively constrained settings, as opposed to more naturalistic, out of the lab, social, and "in the wild" settings. Higher level semantic analysis of hand gestures would benefit from better detection and tracking of hands, which is challenging due to the tendency of the hand to deform and occlude itself. The dataset proposed in this paper follows the more recent trends of leaving the constrained, in-front-of-the-sensor lab settings, and provides the full challenge of occlusion, hand-hand and hand-object interaction, illumination variability, and more. Specifically, we strive to create a hand detection dataset that incorporates the conditions encountered in a naturalistic driving setting.

In the domain of driving, several key motivations exist for the vision-based study of human hands. First, in the interest of the safety of a vehicle's occupants and their surroundings, our motivation to pursue the challenge of detecting vehicle occupants' hands is that successful detection will provide a major indication of the driver's level of attentiveness to the road. Drivers who regularly engage in distracting secondary tasks involving hands during vehicle operation, such as text messaging or eating, are reportedly common [12]. Second, driver hands provide a unique modality of understanding driver behavior [13]. When maneuvering on a freeway or turning in an intersection, driver hands provide information of the driver's style and experience level. Third, large scale naturalistic driving studies could immensely benefit from automatic or semi-automatic analysis of driver hands and secondary tasks. Recently, the SHRP 2 Naturalistic Driving Study has been collecting raw data from 3,100 drivers throughout their everyday driving routines, which contain data looking into and out of the vehicle using camera sensors [14]. The purpose of the study is to understand the role of driver behavior in vehicular safety. The study is advantageous because pre-crash conditions and patterns in a driver's behaviors may be examined in detail, which may shine light on the role that driver behavior plays in a crash, demonstrate how drivers use hands to regain control of a vehicle, and provide valuable insight in the design of autonomous driving systems. The SHRP 2 study provides a dashboard view looking into the vehicle, which shows the driver's hands [15], thus demonstrating the direct applicability of the SHRP 2 data to the task of automatic analysis of hand positions and motion patterns in long-term video.

驾驶员的手势会提供一种理解驾驶员行为的独特行为。

This paper presents the following contributions:

Dataset: As a benefit to the research community, we assembled an annotated a video-based dataset for the task of hand detection under challenging naturalistic driving settings. We make this dataset accessible to the community as part of the Vision for Intelligent Vehicles and Applications (VIVA) challenge[1]. We provide a method for participating research groups to publicly compare detection algorithms and results on a readily available online framework.

Analysis: A benchmark algorithm based on boosting decision trees over color and shape descriptors [16] is tuned for the settings of hand detection and is used for experimental analysis. This paper demonstrates how a hand detector can greatly benefit from employing deeper decision trees.

Metrics: The paper establishes suitable metrics and evaluation procedures on the dataset. The metrics emphasize overall precision-recall curve as well as performance at low false positives rates.

[1] Dataset publicly available at http://cvrr.ucsd.edu/vivachallenge/

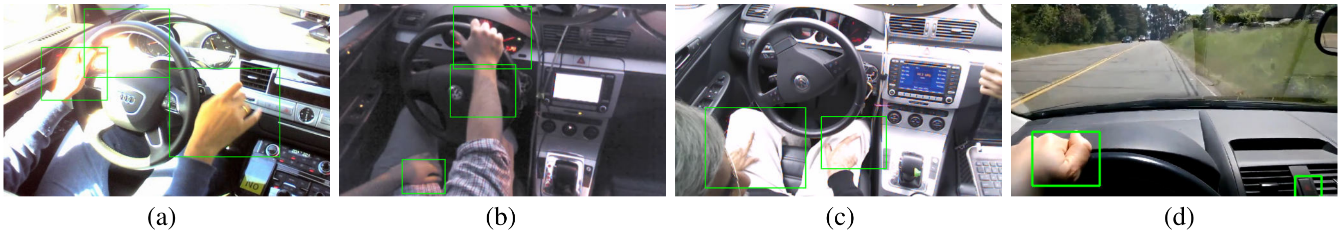

Fig. 1: Challenges in the dataset. (a) Varying illumination conditions may cause false positives and missed detections. Sunlight causes the detector to consider the bright spot on the steering wheel as a hand. Realistic driving scenarios are prone to volatile illumination, and thus the inclusion of severe illumination settings in the hand detection dataset is vital. (b) Skin-colored non-hand objects, such as faces, forearms, and car interiors, may cause false positives in hand detection when the detector relies heavily on color information. Utilizing additional cues, such as context or motion cues, may make the detector more robust against false positives due to skin-colored objects. (c) A detector may miss a hand if it is occluded by another object, self-occluded, or otherwise not completely visible within the frame of the image. In this example, the passenger's left hand is not detected due to being partially out of the frame. (d) Introduction of different viewpoints may cause errors in detection because the perceived size and shape of the hand as well as background variability. This is useful for evaluating detector generalization capacity.

II. Challenges of a Naturalistic Driving Setting

A vision-based hand detection dataset must include the challenges encountered in a naturalistic driving setting in order to be fully representative of hands in a vehicle. Existing hand analysis research often circumvents the issues that are prevalent in realistic driving situations by constraining the hand detection problem such as by limiting the search space [17] or by fixing the hand and background colors [18]. A general hand detection dataset currently exists [19], which occasionally incorporates challenges that overlap with those found in a naturalistic driving setting. However, the occurrences of these challenges are uncommon in the general hand detection dataset as the imagery of said dataset are hand-picked photographs obtained via crowd-sourcing, while imagery found in a naturalistic driving setting and in-vehicle camera system will typically come from videos in a non-selective manner. Thus, when analyzing hands within vehicles, we do not have the ability to control the environment, to enforce an allowable range of clothing colors upon the driver, or to select which images are clear enough to analyze. Instead, the challenges that are often avoided in the field of hand analysis must now be considered in the context of a naturalistic driving setting.

This section outlines some of the challenges that exist in a naturalistic driving setting that we strive to represent within the VIVA hand detection dataset.

Illumination conditions: Varying illumination conditions (Figure 1(a)) and overexposure often cause false positives during detection [15].

Non-hand objects of similar color: Detectors that rely heavily on color features [20] may result in many false positives due to skin-colored non-hand objects, including faces, forearms, clothing, and car interiors (Figure 1(b)). While relying on color for detection may be beneficial in locating potential hand locations, further techniques must be employed to reject non-hand detections, such as a context detector [19].

Occlusion and truncation: Occlusion of hands by other objects and self-occlusion are challenges in the hand detection problem [21]. Figure 1(c) shows a passenger hand on the right that is missed by a detector because the hand is only partially visible. An improved detector must be able to locate hands even when the hands are partly occluded or out of frame. The necessity to detect occluded hands is important because driver hands that are not clearly visible may actually be involved in other activity, which identifies the driver's distracted state.

Camera viewpoints: Varying camera viewpoints may contribute to both false positives and false negatives due to representations of the hand that are rarely seen from other viewpoints. Changing the viewpoint may drastically change the perceived size of the hand, the orientation of the hand, and the level of occlusion of the hand. Figure 1(d) demonstrates both a false negative and a false positive that occurs in the first-person viewpoint. The hand is incompletely detected, and is thus considered a miss, while the hazard light button is falsely detected as a hand. An improved, generalized hand detector should be able to detect hands regardless of the viewpoint. While the camera viewpoint would typically be known if a hand detection system were built into a vehicle, we create a dataset with varied viewpoints with the intent to encourage the generalizability of detector submissions.

III. Description of the Dataset

In this section, we describe the VIVA hand detection dataset in detail, including the annotation format, sources of imagery, and categorized counts of images.

A. *Annotations*

Placement of bounding boxes: Each hand present in a given image is annotated with an axis-aligned bounding box. ~~Partially occluded hands have a bounding box that encompasses the entire hand including the occluded portions of the hand.~~ When a hand is partially out of frame, a bounding box is drawn only around the portion of the hand within the frame. Completely occluded hands and hands completely out of frame have no bounding box. Each image in the training and test sets has at least one annotated hand belonging to the driver and at most four annotated hands belonging to the driver and a single passenger. Figure 2 exemplifies typical annotated images from the dataset.

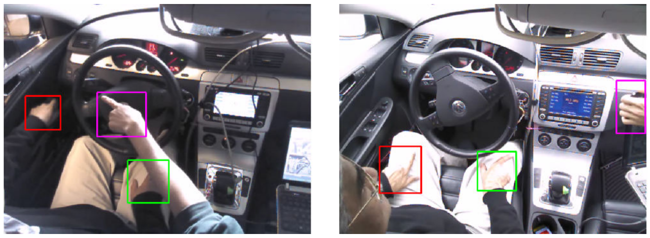

Fig. 2: Visualization of annotations for a given video. Passenger hands are also annotated in the VIVA dataset as they may influence the behavior of the driver or may provide a further challenge in hand detection.

Format of ground truth: The format of the annotations follows the format supported by Piotr's Computer Vision MATLAB Toolbox (PMT) [22]. Each bounding box is specified by its top-left point, width, and height $[x, y, w, h]$. Additionally, each bounding box is assigned to one of four classes depending on whether the hand belongs to the driver or to a passenger and whether the hand is the owner's left or right hand. We note that left-right hand information is useful for many potential in-vehicle applications [23].

B. Sources of Imagery and Camera Positions

We collect and annotate data from various sources and viewpoints with the intent to create a diverse and challenging detection task.

The VIVA detection dataset is comprised of images gathered primarily from videos recorded from our lab. Three lab test-beds were used, labeled as LISA X, LISA Q, and LISA A. The viewpoints in these are either from behind the driver or top down from the rear view mirror. We also include images from YouTube videos of drives to further diversify the VIVA hand detection dataset. The majority of the selected YouTube videos have similar viewpoints as those observed in our testbeds imagery. The remaining YouTube imagery uses unfixed cameras, such as head-mounted or handheld cameras.

Figure 3 shows the possible camera positions from viewpoint 5 (top view). Handheld camera imagery in our dataset is viewed in a position similar to viewpoint 3 or 4, but are not classified as such because the camera position is not fixed in these cases.

C. Temporally Preceding Frames

For each image in the VIVA detection dataset, we make available up to three temporally preceding frames as is provided with the KITTI detection dataset [24]–[26]. The set of temporally previous frames do not have bounding box annotations and serves only to augment the detection data. The temporally previous frames will be useful to detection algorithms that utilize motion cues.

D. Annotation Statistics

In this section we present the counts of each image type and each image source.

Figure 4(a) shows we have over 2000 annotated images from each of our three testbed vehicles. To further diversify the dataset, we also include over 2000 images from YouTube which use imagery in unknown vehicles.

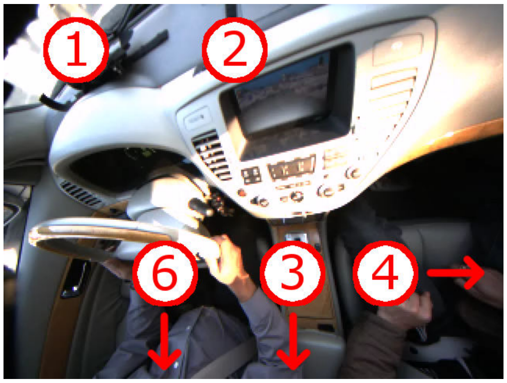

Fig. 3: Camera positions indexed as in the dataset: 0 - handheld (not shown), 1 - front left, 2 - front right, 3 - back, 4 - side, 5 - top (current view), 6 - first-person.

Figure 4(b) presents the number of images provided for each viewpoint. The distribution of imagery by viewpoint was selected based on the availability of imagery and our endeavor to create various levels of difficulty within the dataset. Imagery from the back view is most common in our dataset, and we intend for this viewpoint to be the easier portion of the dataset. A subset of the test data consisting of only back view imagery and larger instances (above 70 pixels in height) constitutes the easier difficulty level in the hand detection challenge which we denote as level-1 (**L1**) evaluation setting. Imagery from other viewpoints and instances greater than 25 pixels in height serve as the more difficult portion of our dataset, and correct hand detection for these viewpoints is reserved for detection algorithms that are capable of hand detection regardless of the camera viewpoint. The level-2 (**L2**) setting includes imagery from all viewpoints (including the images from the L1 setting) and serves as the more difficult evaluation setting.

The majority of the dataset uses imagery in which both of the driver's hands are visible and neither of a passenger's hands are visible. We provide counts of images with a specified total number of visible hands and a specified number of visible driver and passenger hands in Figures 4(c)–(d).

The annotated bounding box dimensions for both the training and test sets are plotted in Figure 5. The majority of the hand sizes are similar between the training and test set, though the amount of overlap decreases as the size of hands increases. The test set uses some YouTube videos that are of higher resolution than other imagery in our dataset, which causes hands to appear larger in terms of pixels.

IV. Experimental Evaluation

We use the Aggregate Channel Features (ACF) object detector [16] from the PMT [22] to test the viability of the VIVA hand detection dataset. This section describes evaluation metrics, the ACF detector, and the results of the detector on the hand detection set when we sweep through basic model parameters. We use the precision-recall (PR) curve and the area under the PR curve (AP) to evaluate how a parameter affects performance. We also publicize the average recall (AR)

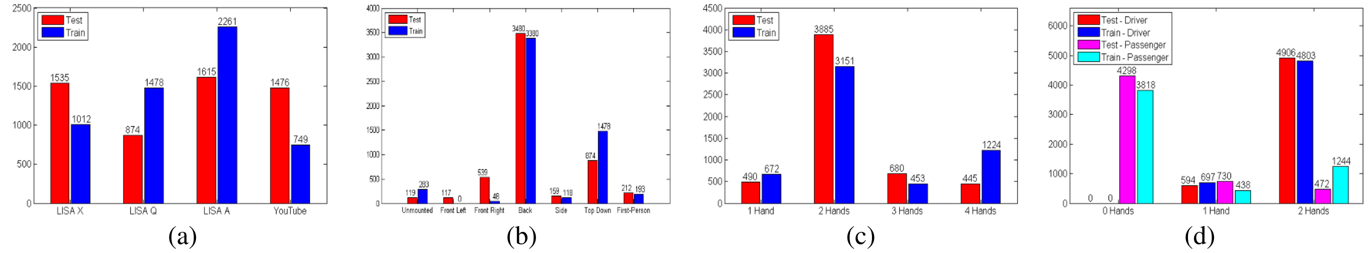

Fig. 4: (a) Counts of images by vehicle type. Our three testbed vehicles are marked separately, but all vehicles from YouTube videos are grouped together. (b) Counts of images by viewpoint. The number of images from the back viewpoint largely dominates over the other viewpoints, and thus we consider the imagery from the back viewpoint as the easier of two levels of difficulty in our dataset. (c) Counts of images by the total number of visible hands. The maximum number of visible hands is 4, and there is always at least 1 hand visible in each image. (d) Counts of images by the number of visible driver and passenger hands. There is always at least 1 driver hand, and there is usually no visible passenger hands.

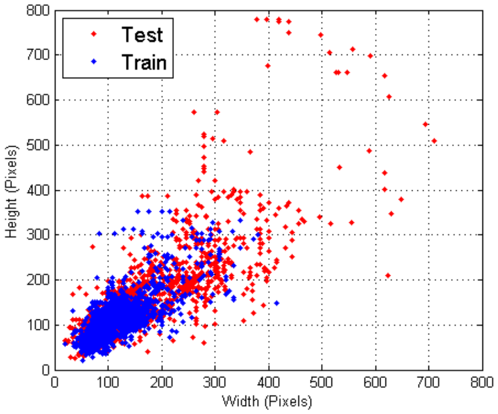

Fig. 5: Annotation bounding box sizes for both the training and test set. The sizes of the hands are largely similar between the training and test set. The test set includes imagery in which the hands appear much larger than the hands in the training set.

metric for each detection submission, computed from the ROC curve over 9 evenly sampled points in log space between 10^{-2} and 10^{0} false positives per image. The AR metric is suitable for summarizing detection performance at lower false positive rates. A detection is considered correct when it satisfies the PASCAL criterion. That is, a detection is correct when the proportion of overlap between the predicted bounding box and the ground truth bounding box is greater than 0.5 [27].

A. ACF Detector Overview

The ACF detector utilizes 10 feature channels, a normalized gradient magnitude channel, 6 gradient orientation channels, and LUV color channels. Features are formed by aggregating and smoothing the channels, and AdaBoost is used to train decision trees based on these features. Object detection is performed using a sliding-window approach. An advantage of the ACF detector is that fast multiscale detection is achieved using feature pyramids which are quickly derived by computing features of octave-spaced scaled images and using approximations for scales between octaves [28]. The output of the ACF detector is a set of axis-aligned bounding boxes along with a score proportional to the confidence of detection for each box [16], [28].

The ACF detector is highly successful in pedestrian detection [28], we thus treat the ACF detector as an effective multiscale object detector to test the viability of the hand detection dataset.

B. ACF Detector Results

To maintain simplicity in training an ACF detector to evaluate the VIVA dataset, we only sweep through parameters that govern the size of the model and the complexity of the weak learners used in AdaBoost. We first select to use boosted trees of depth 2, and we perform a grid search over 6 model heights ranging from 25 to 75 pixels and 5 model aspect ratios from 0.8 to 1.2. The ACF parameters we keep constant are the number of classifiers in each of the four AdaBoost stages ([32, 128, 512, 2048]) and the non-maximal suppression threshold at which lower-scoring bounding boxes are suppressed if they overlap with other bounding boxes (0.2). All other ACF model parameters are left as their default values. We retain the AP obtained by each detector with depth 2 trees. We then repeat this process using detectors with depth 4 trees. Figure 6 shows the AP values obtained in both grid searches.

Using the model dimensions with the highest AP in the depth 4 model size grid search (height of 65 pixels and aspect ratio of 0.9), we sweep the tree depths to ensure that a tree depth of 4 best suits this dataset. Figure 7 shows the PR curves using detectors with tree depths of 2, 3, 4, and 5 on both the L1 and L2 evaluation settings. AP increases as tree depth increases until a depth of 4. The detector with a tree depth of 5 performs worse than the detector with a tree depth of 4, suggesting that a detector with a tree depth higher than 4 suffers from overfitting. Visualizing the models in Figure 8 provides further evidence that the detector with a tree depth of 5 overfits to the training data. In this visualization, warmer colors represent the larger weights assigned to the corresponding locations within each considered window, and the deeper colors in the depth 5 case (far right) suggest that this detector may have overfit to the training data.

Using the detector with model height 65 pixels, aspect ratio 0.9, and tree depth 4, we compute the AP for both the L1 and L2 settings: 70.09% for L1 and 60.06% for L2. We also generate an ROC curve (Figure 9) to better visualize the performance of the detector in terms of its true positive rate and number false positives per image. We also calculate AR for the L1 and L2 settings: 53.84% for L1 and 40.42% for L2.

Our initial results are promising, but suffer from false and missed detections. Typical high-scoring false positives are shown in Figure 10. The top row contains hands, but the poor fit of the bounding box prevents these detections from being true positive detections. The false positives in the bottom row suggest that our detector is heavily color-based because faces and red objects are mistakenly detected as hands. Further improvements to our detection system must be able to reject these types of false positives and must form better-fitting bounding boxes for each detection.

C. Cross Dataset Comparison

We performed a cross dataset comparison to assess whether the images provided in the VIVA hand detection dataset may be superseded by images provided in a general hand detection dataset. We selected the diverse hand detection dataset created by Mittal *et al.* [19] which includes annotated photographs in indoor and outdoor settings. Cross dataset training and testing resulted in AP of less than 10% in both cases, showing the difficulty of the hand detection problem and the domain differences among the datasets.

V. Concluding Remarks

Vision-based detection of vehicle occupants' hands may be indicative of the attentiveness and behavior of the driver. This paper introduces the new vision-based detection dataset for hands in a naturalistic driving setting. We assess the feasibility of the VIVA dataset by training and testing object detectors, and we perform a cross dataset comparison using a general hand dataset to illustrate the uniqueness of the hand detection problem in a naturalistic driving setting. Common challenges for hand detection in naturalistic driving settings include volatile illumination conditions, occlusion, non-hand color similarity, and varying viewpoints. The VIVA dataset incorporates instances of these challenges to create a detection task representative of a naturalistic driving setting.

Our initial detector performed adequately on our dataset, indicating that hand detection in our dataset is a possible yet still a challenging task. Future goals include further tuning of the ACF detector using its more complex parameters and designing new detectors that may be better suited for the hand detection task. Additionally, we will continue to update the VIVA hand detection dataset to increase its diversity and update the online evaluation method to better streamline the submission process.

References

[1] J. S. Supančič, G. Rogez, Y. Yang, J. Shotton, and D. Ramanan, "Depth-based hand pose estimation: methods, data, and challenges," in *CVPRW-HANDS*, 2015.

[2] P. Molchanov, S. Gupta, K. Kim, and K. Pulli, "Multi-sensor system for drivers hand-gesture recognition," in *IEEE Conf. Automatic Face and Gesture Recognition*, 2015.

[3] G. Rogez, J. S. Supančič, and D. Ramanan, "First-person pose recognition using egocentric workspaces," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2015.

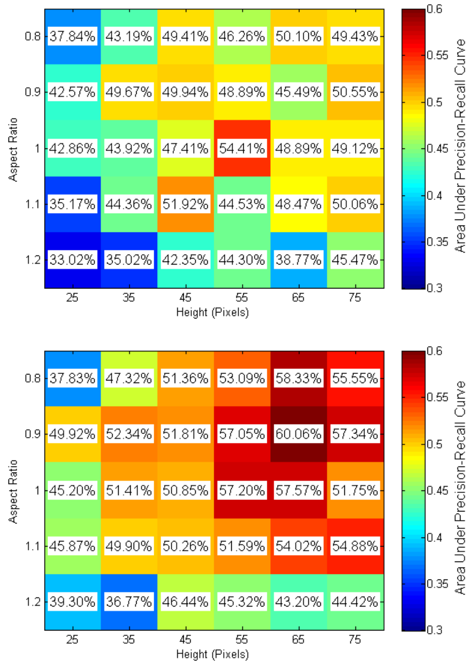

Fig. 6: AP values for a grid search over model heights and aspect ratios with tree depth 2 (top) and tree depth 4 (bottom).

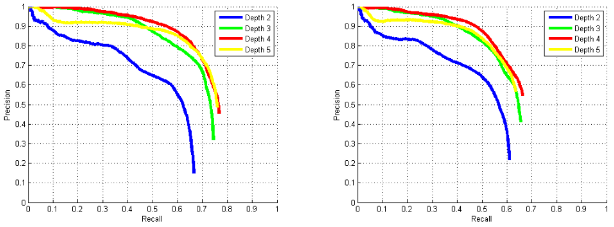

Fig. 7: PR curves using boosted trees of depth 2, 3, 4, and 5 for both the L1 (left) and L2 (right) difficulty levels. The model height is held constant at 65 pixels and aspect ratio 0.9. Increasing the tree depth improves performance in terms of AP until a depth of 4. Further increases to the tree depth decrease performance due to overfitting.

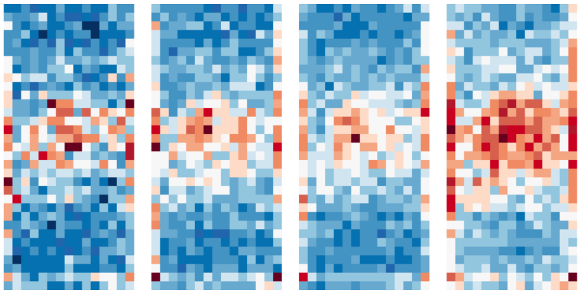

Fig. 8: Model visualizations for detectors with tree depths of 2, 3, 4, and 5 (left to right). Model height and aspect ratio are held constant at 65 pixels and 0.9, respectively. Warmer colors represent the larger weights assigned to the corresponding locations within each considered window. The deeper colors in the visualization for the detector with a tree depth of 5 suggests that this detector may have overfit to the training data.

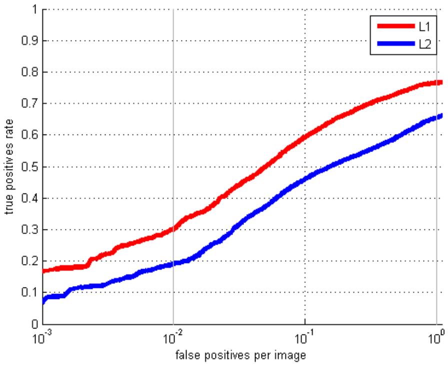

Fig. 9: ROC curves for the detector with height 65 pixels, aspect ratio 0.9, and tree depth 4 on both the L1 and L2 difficulty levels. The incorporation of all viewpoints (L2) provides more challenging settings.

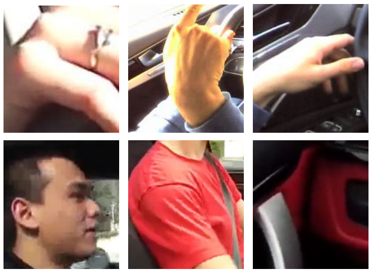

Fig. 10: Typical high-scoring false positives from our trained ACF detector. The bounding boxes for the hands in the top row are poorly fit, thus causing such instances to be marked as false positives. The false positives in the bottom row suggest that our detector is heavily color-based, as skin-colored objects such as faces or objects with a red hue are detected as a hand.

[4] S. Wan and J. K. Aggarwal, "Mining discriminative states of hands and objects to recognize egocentric actions with a wearable rgbd camera," in *CVPRW-HANDS*, 2015.

[5] S. Lee, S. Bambach, D. Crandall, J. Franchak, and C. Yu, "This hand is my hand: A probabilistic approach to hand disambiguation in egocentric video," in *CVPRW-Egocentric Vision*, 2014.

[6] S. Bambach, "A survey on recent advances of computer vision algorithms for egocentric video," *arXiv preprint arXiv:1501.02825*, 2013.

[7] P. Molchanov, S. Gupta, K. Kim, and J. Kautz, "Hand gesture recognition with 3D convolutional neural networks," in *CVPRW-HANDS*, 2015.

[8] M. B. Holte, C. Tran, M. M. Trivedi, and T. B. Moeslund, "Human action recognition using multiple views: A comparative perspective on recent developments," in *Proceedings of the 2011 Joint ACM Workshop on Human Gesture and Behavior Understanding*, 2011, pp. 47–52.

[9] C. Tran and M. Trivedi, "3-D Posture and Gesture Recognition for Interactivity in Smart Spaces," *IEEE Trans. Industrial Informatics*, vol. 8, no. 1, pp. 178–187, Feb 2012.

[10] 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*, vol. 15, no. 6, pp. 2368–2377, Dec 2014.

[11] F. Parada-Loira, E. González-Agulla, and J. L. Alba-Castro, "Hand gestures to control infotainment equipment in cars," in *IEEE Intelligent Vehicles Symposium*, 2014.

[12] T. H. Poll, "Most U.S. Drivers Engage in 'Distracting' Behaviors: Poll," no. FMCSA-RRR-09-042, 2011.

[13] E. Ohn-Bar, A. Tawari, S. Martin, and M. M. Trivedi, "On surveillance for safety critical events: In-vehicle video networks for predictive driver assistance systems," *Computer Vision and Image Understanding*, vol. 134, no. 0, pp. 130–140, 2015.

[14] K. L. Campbell, "The SHRP2 Naturalistic Driving Study: Addressing Driver Performance and Behavior in Traffic Safety," *TR News 282*, 2012.

[15] E. Ohn-Bar, S. Martin, and M. M. Trivedi, "Driver Hand Activity Analysis in Naturalistic Driving Studies: Issues, Algorithms and Experimental Studies," *Journal of Electronic Imaging*, vol. 22, pp. 041 119:1–041 119:10, 2013.

[16] P. Dollár, S. Belongie, and P. Perona, "The Fastest Pedestrian Detector in the West," *British Machine Vision Conf.*, 2010.

[17] R. Lockton and A. Fitzgibbon, "Real-time gesture recognition using deterministic boosting," in *British Machine Vision Conf.*, 2002.

[18] R. Wang and J. Popovic, "Real-time Hand-tracking with a Color Glove," *ACM Trans. Graph.*, vol. 28, no. 3, pp. 63:1–63:8, Jul 2009.

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

[20] E.-J. Ong and R. Bowden, "A boosted classifier tree for hand shape detection," in *IEEE Conf. Automatic Face and Gesture Recognition*, May 2004, pp. 889–894.

[21] E. Ohn-Bar and M. M. Trivedi, "Beyond just keeping hands on the wheel: Towards visual interpretation of driver hand motion patterns," in *IEEE Conf. Intelligent Transportation Systems*, 2014.

[22] P. Dollár, "Piotr's Computer Vision Matlab Toolbox (PMT)," http://vision.ucsd.edu/ pdollar/toolbox/doc/index.html.

[23] S. Y. Cheng and M. M. Trivedi, "Vision-based Infotainment User Determination by Hand Recognition for Driver Assistance," *IEEE. Trans. Intell. Transport. Sys.*, vol. 11, no. 3, pp. 759–764, Sep. 2010.

[24] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite," in *CVPR*, 2012.

[25] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets Robotics: The KITTI Dataset," *International Journal of Robotics Research*, 2013.

[26] J. Fritsch, T. Kuehnl, and A. Geiger, "A new performance measure and evaluation benchmark for road detection algorithms," in *IEEE Conf. Intelligent Transportation Systems*, 2013.

[27] M. Everingham, L. Van Gool, C.K.I Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes (VOC) Challenge," *International Journal of Computer Vision*, vol. 88, pp. 303–338, June 2010.

[28] P. Dollár, R. Appel, S. Belongie, and P. Perona, "Fast Feature Pyramids for Object Detection," *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2014.

[29] M. Trivedi, C. Harlow, R. Conners, S. Goh, "Object Detection Based on Gray Level Cooccurrence," *Computer Vision, Graphics, and Image Processing*, 1984.

[30] E. Ohn-Bar and M. M. Trivedi, "The Power is in Your Hands: 3D Analysis of Hand Gestures in Naturalistic Video," in *CVPRW-AMFG*, 2013.

[31] S. Shivappa, M. Trivedi, and B. Rao, "Audiovisual Information Fusion in Human-Computer Interfaces and Intelligent Environments: A Survey," *Proceedings of the IEEE*, vol. 98, no. 10, pp. 1692–1715, Oct 2010.

[32] M. Holte, C. Tran, M. Trivedi, and T. Moeslund, "Human Pose Estimation and Activity Recognition From Multi-View Videos: Comparative Explorations of Recent Developments," *IEEE Trans. Selected Topics in Signal Processing*, vol. 6, no. 5, pp. 538–552, Sept 2012.

[33] E. Murphy-Chutorian and M. Trivedi, "HyHOPE: Hybrid Head Orientation and Position Estimation for vision-based driver head tracking," in *IEEE Intelligent Vehicles Symposium*, June 2008, pp. 512–517.