Vision-Based Infotainment User Determination by Hand Recognition for Driver Assistance

Shinko Y. Cheng, *Member, IEEE*, and Mohan M. Trivedi, *Fellow, IEEE*

***Abstract*—We present a novel real-time computer-vision system that robustly discriminates which of the front-row seat occupants is accessing the infotainment controls. The knowledge of who is the user—that is, driver, passenger, or no one—can alleviate driver distraction and maximize the passenger infotainment experience. The system captures visible and near-infrared images of the front-row seat area in the vehicle. The algorithm uses a modified histogram-of-oriented-gradients feature descriptor to represent the image area over the infotainment controls and a support vector machine (SVM) and median filtering over time to classify each image to one of the three classes with 97.9% average correct classification rate. This rate was achieved over a wide range of illumination conditions, human subjects, and times of day. With an offset of 5 pixels in any direction, the rate could still be maintained at better than 85%. This approach represents an alternative to detecting and tracking the hand movements and then classifying the hands into the respective classes. This approach demonstrates the ability to achieve good classification rates, despite the presence of vast illumination changes of the vehicle environment.**

***Index Terms*—Computer vision, hand-gesture recognition, infotainment, user determination.**

I. Introduction

A BROAD NEW array of information devices is finding its place in today's vehicles. The infotainment system graduated from a term referring to the radio to a collective term describing the various functionalities regarding navigation, vehicle state, climate control, cell-phone control, MPEG-1 Audio Layer 3 player control, web browsing, and even television entertainment at the front console area of the vehicle [1], [2]. With all of these opportunities for the drivers to be distracted, the solution has been to limit the functionality or the content from these infotainment systems and make them less distracting. Often, the information provided by these devices then becomes oversimplified and less useful for passengers. It is far more desirable to be able to both alleviate driver distraction and provide better information to passengers.

We propose a novel real-time vision-based user determination (VUD) system to determine which front-row seat occupant is accessing the infotainment controls. The controls of the infotainment system consist of buttons and a knob with rotational and directional degrees of freedom. They are centrally located in the aisle area between the driver and the passenger. The VUD is intended to improve vehicle safety by providing information about the type of user (driver, passenger, or no one) so that the infotainment system can provide an adequate level of information. Driver distraction can be mitigated while the passenger can be granted full access to the information device. The infotainment system can also default to an appropriate mode when it senses no one is using the device.

The challenges of developing such a system center on developing a vision-based classification algorithm robust to the various operating modes of the vehicle. The performance should be maintained through changes in the appearance of different people, changes in lighting from different times of the day, and changes in the camera position. Because vehicles are likely to receive maintenance only between several months of operation, if at all, much of the functionality must also require little or no maintenance. These attributes we show can be obtained with appropriate choices of the pattern classifier and system components.

The proposed system takes as input visible and near-infrared images of the front-seat and center-console area illuminated with a bank of near-infrared LEDs. The system then uses these images to determine which front-row occupant is accessing the device, if there is anyone at all. A modified form of the histogram-of-oriented-gradients (HOG) image descriptor was chosen to create the feature vectors [3], [4]. The system then utilizes the support vector machine (SVM) with a radial basis function (RBF) kernel to classify the observed image features into three classes: 1) driver; 2) passenger; or 3) no one. The processing takes 10.7 ms/frame on an Intel Pentium D 3.2-GHz PC.

The evaluation of this approach uses two metrics: 1) the correct classification rates of each class that are part of the confusion matrix and 2) the average correct detection rate over the three classes. In the training process, care was taken to ensure that a representative data set was used to train the pattern classifier. Data were collected at four different times of the day with eight different individuals, for a total of 18 test runs, over 1 h of data at National Television System Commission frame rates (29.97 FPS). We also analyzed the system's invariance to translation in the x- and y-directions of the image patch, where the features are extracted. These qualities influence the flexibility of camera placement during installation and maintenance. The resulting trained system can correctly recognize whether the driver, passenger, or no one has their hand over the infotainment controls with better than 95% average

Manuscript received July 28, 2009; revised April 11, 2010; accepted April 18, 2010. This work was supported by Volkswagen of America, Electronics Research Lab, Palo Alto, CA, and by UC Discovery, Digital Media Innovations program. The Associate Editor for this paper was C. Stiller.

S. Y. Cheng was with the University of California, San Diego, La Jolla, CA 92037 USA. He is now with HRL Laboratories, LLC, Malibu, CA 90265 USA (e-mail: shinko.cheng@gmail.com).

M. M. Trivedi is with the University of California, San Diego, La Jolla, CA 92037 USA (e-mail: mtrivedi@ucsd.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TITS.2010.2049354

correct classification rate and 97.9% with additional filtering in postprocessing.

The results of this paper reveal two main conclusions: 1) It is possible to achieve extremely reliable pattern recognition using a visible-wavelength imaging sensor in the volatile environment of the vehicle interior, where illumination is changing constantly and high in dynamic range. 2) The proposed system introduces a novel method of extracting information that is discriminant for determining when to provide more informative or less distracting driving assistance.

The organization of this paper is as follows: We discuss related research in Section II. We then present the VUD system and experimental evaluation in Sections III and IV. Finally, we conclude in Section V.

II. Related Research

The idea of tailoring vehicle information system functions, input and output devices, and user interfaces depending on whether the user is the driver or passenger was first proposed by Chou *et al.* [5]. They proposed the use of weight sensors to determine the presence of a passenger before enabling either full or partial functionality of an infotainment system. Harter *et al.* [6] also proposed to switch between "enhanced" and "base" functionality of the information system by determining the presence of a passenger but by using proximity sensors instead. Harter *et al.* take a step further to determine when to engage "base functionality" by determining whether the driver has gazed into the infotainment monitor for more than 2 s (considered too long) using a vision-based eye-gaze tracking system. We propose to use the same vision modality but analyze the hands of the occupants rather than the driver's head to determine when to switch between functionality modes. Our proposed solution can replace or complement these other systems by providing the following advantages.

1) The proposed system is arguably simpler to implement and maintain than an eye-gaze tracking system. The proposed solution requires neither camera geometry information nor person calibration.
2) The proposed system senses when the hand gets near the controls, therefore immediately detecting when a driver's hands are not on the wheel.
3) The proposed system actively detects the case of when no one is accessing the information system. This case can be used to automatically show and hide access controls in the display or interpreted as having a more attentive driver, thus requiring less driver assistance. Weight-sensor-based systems can detect if no one is present but cannot detect if someone is there but not accessing the system.

The problem of hand-based user determination using vision can be approached in two ways: 1) active tracking of occupant hands as the hand passes into and out of the area over the infotainment controls [or region of interest (ROI)] to detect intent to access, and 2) learning of the appearance of the driver's hand, the passenger's hand, or no one's hand over the ROI. A number of efforts have addressed the first approach.

The first type consists of a detector that locates the hand in the images and then tracks the hand. Tracking is associating one hand detection with the next over time. The challenge of this approach is in obtaining a good description of the appearance of the hand in its various poses and a way to efficiently check all areas of the image for the existence of a hand. The characteristics of a good descriptor are one that would correctly associate two hand detections of the same hand in different positions and poses.

One such hand detection algorithm devised by Kölsch and Turk [7] employs a cascade of boosted classifiers using Haar-wavelet-like image features and their extensions to determine whether an image patch, among all possible patches in an image, consists of a hand or not. The rates reported for real-time operation were very good (92% detection rate with a false positive rate of 1e-8 per window), but the approach detected hands in a standard canonical position: fingers up and thumb to the left and seven other similar forms. Kölsch *et al.* proposed using a flock-of-features approach to track the positions of the hands after initial detection, thus addressing the problem of maintaining track of hands through its many poses.

We employed a similar detection technique with long-wavelength thermal infrared images of hands [8]. The thermal infrared modality is particularly appropriate in the vehicular domain because image appearance is not at all affected by changing visible illumination conditions and the pixel intensity of the skin stays relatively constant. Because of the special quality of hands in thermal images, this detector was also effective in detecting hands in various poses. The detector is applied on each incoming frame, and the multiple hands are tracked using the Kalman filter and the probabilistic data association filter to disambiguate one track from another. This approach, however, requires the use of a thermal camera, which are still expensive compared with the visible-wavelength camera.

Yuan *et al.* [9] proposed a hand tracking system that utilizes color and motion information to detect and the Viterbi algorithm to track the hands. Because of the vehicle's high and low light operating conditions and the need to discreetly illuminate the vehicle interior, color cameras—and therefore skin-color-based algorithms—were not an option.

There has also been a significant amount of work on detecting the articulated body pose of the hand. From pose, a slew of other hand gestures can be detected. A recent extensive survey of hand-gesture recognition and pose-estimation systems can be found by Pavlovic *et al.* [10].

We address the user determination problem using the second approach, which is the direct classification of images of the infotainment control region. No tracking is required, although some basic filtering of the classification responses will further increase the correct classification rate. This approach takes advantage of the fact that the ROI has a stable background, which is the vehicle interior. To the best of our knowledge, no other work addresses the user determination problem in a similar way.

III. Vision-Based User Determination System

The VUD system determines the user of the individual whose hand is accessing the infotainment device by classifying patches of captured images of the front-row seat area in a passenger vehicle. *User* is defined as one of three categories:

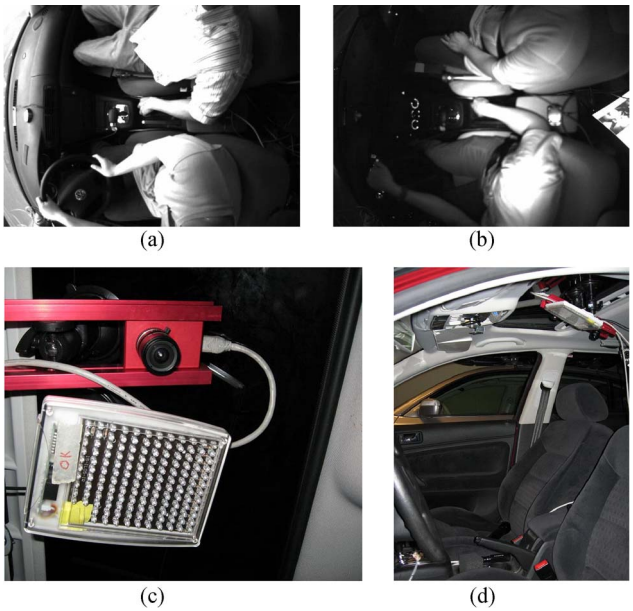

Fig. 1. Example images captured during the day and at night, and the positions of the camera and illuminator in the LISA-P test bed for the VUD system. VidereDesign STH-MDCS2-VAR camera and SUPERCIRCUITS IR14 140 LED IR Illuminator were used for the VUD system. (a) Example image in daylight. (b) Example image at night. (c) Camera and illuminator. (d) Camera and illuminator setup.

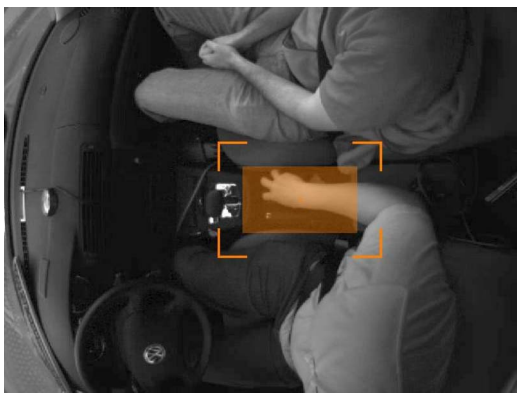

Fig. 2. Image ROI used to determine the user in the VUD system. Patch as depicted is used for the final VUD system.

1) driver; 2) passenger; and 3) no one. The infotainment controls are assumed to be positioned just aft of the gear shift, forward of the hand rest, and beside the hand brake.

The visible and near-infrared spectrum imager was chosen primarily for the passive nature of the camera; at night, the front-row seat area can still be seen by illuminating the area with near-infrared illuminators without distracting the occupants. Example images are shown in Fig. 1(a) and (b). The images were 640×480 pixels in resolution captured at 30 fps. Hardware arrangement is illustrated in Fig. 1(c) and (d).

The system first captures monochrome images, and a rectangular image patch that spans between the edges of the driver's and passenger's seat and the length between the gear shifter and the hand rest is extracted. An example image captured from the front seat area and the image patch is shown in Fig. 2. The modified HOG description of the image patch is calculated and then presented to the multiclass SVM classifier with an RBF kernel to determine which of the three events occurred: 1) the driver's hand; 2) the passenger's hand; or 3) no one's hand accessed the infotainment controls.

A. Modified HOG Feature Extraction

The modified HOG descriptor for an image patch is created by first taking the gradient of the patch. The resulting gradient image is then divided into smaller rectangular patches of pixels specified by the number of cells along the x- and y-directions. Within each cell of pixels, a histogram is collected of the angles (orientations) of the gradient vectors of each pixel. In generating this orientation histogram, the number of bins that span gradient orientations of 0° and 360° is the length of each histogram for each cell. All the orientation histograms are concatenated to form the final feature vector $\mathbf{x}$. Altogether, three parameters determine the dimensions of the final feature vector: 2×2 grid of bins with eight slices in the range of possible gradient orientations results in a 32-D feature vector $(2 \times 2 \times 8)$ for each image patch.

The original HOG [3], as well as the shift-invariant feature transform (SIFT) descriptor portion of the SIFT image feature [11], and other gradient orientation histogram-type descriptors were shown to be very effective in characterizing the appearance and shape of objects in images. The histogram of the gradient orientation component allows a family of local descriptors to be among the most successful in the wide-baseline image matching problem [12] and robust to moderate changes in perspective, level of illumination, and focus. There also exist biological underpinnings arguing for the use of similar features [13].

IV. Experimental Evaluation

A. Performance Metric

The evaluation of the VUD system utilizes two metrics: 1) the confusion matrix summarizing the classification rate of observations (feature vector) of a given class detected as a given class and 2) the average correct classification rate (CCR) which is the CCR averaged over the three classes. Worst performance is that which can be achieved by random guessing, which is 33% classification rate in each element of the confusion matrix. We also present the proportion of time in a minute when the system is detecting in error.

B. Validation of Performance

The confusion matrix was generated for the three classes using fivefold cross validation. The mean and standard deviation of each element in the confusion matrix were found. The standard deviation was always less than 0.5%.

To ensure that the trained results would perform well in real situations, the data set was collected at various times of day (noon, afternoon, twilight, and night) with various individuals (eight male individuals of average build from $5'0''$ to $6'0''$ in height) in both driver and passenger positions. One sequence was captured with a variety of clutter (flashlight, cardboard, paper, mouse pad, tools, and cups) introduced into the ROI to capture the statistics of feature vectors of those instances. A total of 18 video sequences containing 114 886 examples and

TABLE I
ATTRIBUTES SUMMARY OF THE 18 SEQUENCES OF VIDEO DATA USED FOR TRAINING AND TESTING THE CLASSIFIER OF THE VUD SYSTEM. TEST SUBJECTS ARE LABELED A THROUGH H, FOR EIGHT SUBJECTS IN ALL. SEQUENCES 1–4 WERE CAPTURED WITH A STATIONARY VEHICLE; THE REMAINING WERE CAPTURED IN A MOVING VEHICLE. WEATHER CONDITIONS ARE INDICATED INDOOR (I), OVERCAST (O), SUNNY (S), AND AT NIGHT (N)

Grp.	Seq.	Frames	Driver/Passenger	Weather	Time
1st	1	4814	A/*	I	N/A
	2	6000	*/A	I	N/A
	3	6947	A/*	O	6pm
	4	4089	*/A	O	6pm
	5	11,740	A/B	O	7pm
	6	7093	C/A	S	12pm
	7	7699	D/E	S	12pm
Group Total		48,382	# Samples: {31,963, 8650, 7769}		
2nd	8	13,012	A/A	N	9pm
	9	4978	B/G	S	12pm
	10	4202	G/B	S	12pm
	11	6908	C/A	S	12pm
	12	5445	A/C	S	12pm
	13	4845	A/*	S	12pm
	14	5039	H/G	S	4pm
	15	5002	G/H	N	4pm
	16	6961	H/A	N	9pm
	17	3987	A/H	N	9pm
	18	6125	A/H	N	9pm
Group Total		65,604	# Samples: {36,504, 11,529, 17,571}		
Total		114,886	# Samples: {68,467, 20,179, 25,340} {# No One, # Driver, # Passenger}		

TABLE II
SUMMARY OF VUD PERFORMANCE. VARIOUS IMAGE PATCH SIZES WERE USED. THE PERFORMANCE IS DESCRIBED USING THE CONFUSION MATRIX AND AVERAGE PROPORTION OF 1 MIN IN ERROR COMPUTED WITH FIVEFOLD CROSS VALIDATION. (a) EXP 1. (b) EXP 2. (c) EXP 3

(a) Exp 1

	P(predicted|actual)		
	NoOne	Drver	Psngr
NoOne	95.42	2.01	2.57
Drver	1.82	97.63	0.54
Psngr	2.12	0.57	97.31
Mean Correct Rate			96.79%
Mean Time in Error [s/min]			1.926

(b) Exp 2

	P(predicted|actual)		
	NoOne	Drver	Psngr
NoOne	94.98	2.44	2.58
Drver	2.66	96.41	0.93
Psngr	3.02	1.08	95.90
Mean Correct Rate			95.76%
Mean Time in Error [s/min]			2.544

(c) Exp 3

	P(predicted|actual)		
	NoOne	Drver	Psngr
NoOne	93.73	2.77	3.49
Drver	2.40	97.18	0.42
Psngr	2.94	0.59	96.47
Mean Correct Rate			95.79%
Mean Time in Error [s/min]			2.526

63 min of video in various illumination conditions were used for training and testing. The conditions under which the data set was collected are summarized in Table I. The eight subjects in the data capture are denoted as A through H. An asterisk $(^*)$ denotes no occupant for that video sequence. Individuals wore short- and long-sleeve shirts. For most sequences, the data were captured while the vehicle was in motion, driven along a circular route in which the direction of sunlight could shine into the vehicle from every direction at least once. Different times of day yielded different angles of the sun and character of the sun.

Each frame of the video is manually annotated with the category to which it belongs. Namely, each frame may show either no one, the driver, or the passenger is placing their hand over the infotainment control area. There are a total of 68 467, 20 179, and 25 340 unique frames collected for each of the three classes, respectively.

C. System Parameter Optimization

Three feature types were analyzed to validate our choice of image patch dimensions. Intuitively, the forearm is a good indicator of whose hand is accessing the infotainment controls. A rectangular image ROI of size 140 × 80, as depicted in Table II(a), appears to capture both the hand and the forearm compactly. The other two image patches consisted of a square image patch of sizes 80 × 80 and 140 × 140, centered around the hand, as shown in Table II(b) and (c).

The multiclass SVM classifier with the RBF kernel parameters (the slack parameter C and the RBF kernel width γ) is tuned with a grid search of values for the optimal tuple (C, γ) that yields the highest average correct detection rate (mean of the diagonal elements of the normalized confusion matrix). A subset of the complete data set (5000 per class) was used in the grid search. The optimal values are $C = 25$ and $\gamma = 100$.

The results of the three patch sizes are summarized in Table II. The differences in percentage points were subtle, but the rectangular image patch produced the best results of the three with 96.79%. To give a better sense of the performance as a function of time, the duration of time in error was calculated. The amount of time when the system was in error in 1 min is calculated by the average percentage of time in error multiplied by 60 s.

The average number of seconds in error in 1 min for the three types of features were 1.926, 2.544, and 2.526 s, respectively. There is an improvement in confining the image patch to a rectangular area of the aisle by 0.5 s.

Despite the care taken during annotation to ensure that a consistent strategy was used for when a hand is in transition to and from the image patch region, the percentage of the hand remaining in the image patch is usually used to determine whether the hand is still in the region. There may still exist some inconsistencies in the transition frames; some may be annotated as hand still in the region when in fact it is not, and *vice versa*. Furthermore, in transition frames, the hands are often blurred.

To determine how much of the inconsistency adversely influenced the performance, we examine the proportion of errors that exist in the transition region. A transition region is defined as $\pm L$ seconds surrounding the point of transition in the annotation file. The width of the transition region is $2L$. As expected, 50% of the errors occur within 0.5 s of the transition. The frame rate is 30 Hz. This means that the most confusing frames are when the hands enter and exit the image patch region. The proportion of errors tapers off as the transition

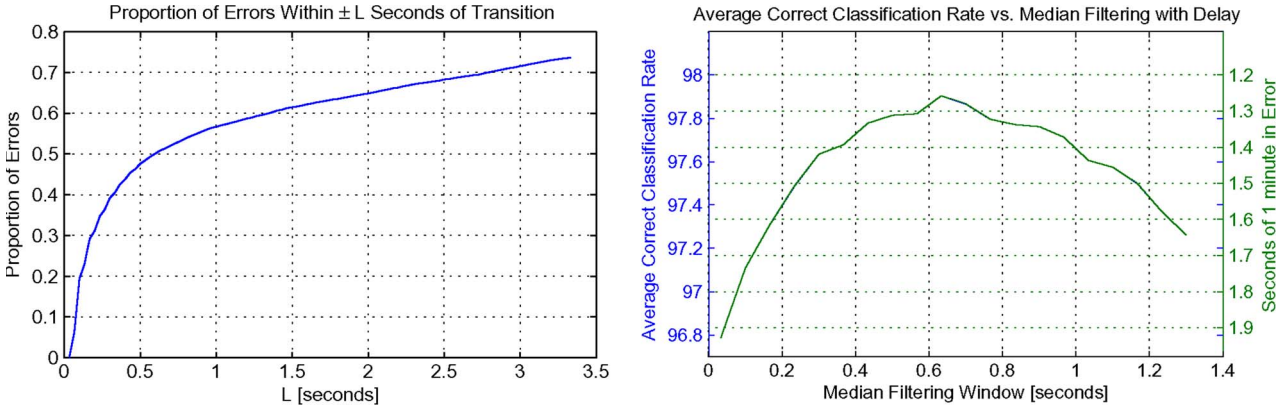

Fig. 3. Errors in the VUD system are examined to determine the proportion of errors in the transition times, which are when the hands of the occupants enter or exit the image patch region. Average correct classification rate versus median filtering with delay. Applying a smoothing median filter yields better correct classification rates with an optimal window of 0.63 s in width. The corresponding delay is 0.315 s. The frame rate is 30 Hz.

window increases, as shown in Fig. 3. The implication is that 50% of the errors can be avoided by utilizing a delay before deciding on the presence of a hand.

In light of this, median filtering and a delay to the responses were used, and the average correct detection rate was measured. The rates for the various median filter window widths were used, and the resulting plot is shown in Fig. 3. A window of 0.63 s in width (19 samples) gave the best average correct detection rate of 97.9%, or 1.257 s of 1 min in error, which is an improvement of 2/3 s worth of errors.

D. Robustness

The VUD system's invariance to translation in the x- and y-axis directions was also analyzed. These qualities influence the flexibility in camera placement during the installation process. The image patch ROI is specified as shown in Fig. 4. This ROI was shifted to various positions in the image ± 30 px along the x and y-directions. The range of translations is depicted in Fig. 4(a). The effect of those translations on the average correct classification rate is shown as a heat image in Fig. 4(b). The average correct classification rates are collected from test-set frames in sequence 5 (see Table I).

The results show that using the rectangular ROI, the performance of the VUD remains above 85% at ± 5 px deviation from the original location and above 80% at ± 10 px. The 10-pixel deviation is to approximately half the length of the occupant's finger.

V. Discussion and Concluding Remarks

We presented a VUD system. The system consisted of a visible and near-infrared imaging device observing the front-row seat area in the vehicle. Using the HOG image descriptor to describe the area over the controls, an SVM was shown to obtain 96.8% average correct classification rate over the three classes: 1) driver; 2) passenger; or 3) no one is accessing the infotainment controls. The rate is improved to 97.9% average correct classification rate with median filtering. With an offset of 10 pixels in any direction, the rate could still be maintained at better than 80%.

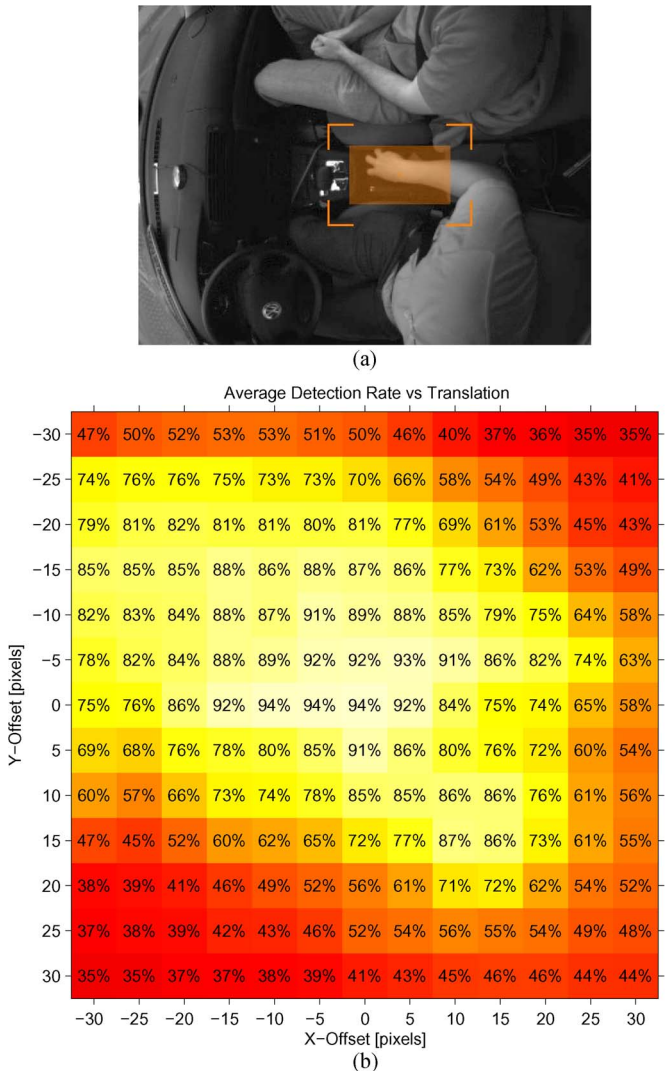

Fig. 4. VUD translation invariance evaluation. (a) Image ROI used to determine the user. The larger rectangle marks the boundary of all the translations of the image patch in measuring the translation invariance of the VUD system. (b) The corresponding average classification rate for each translation of the image patch is shown in the heat image to the right. The image and the graph can be used as reference in positioning the camera and image patch location during the installation of the VUD system.

The system is intended to improve the safety and comfort of the vehicle by enabling the vehicle to determine which occupant is accessing the vehicle's infotainment controls, which is often characterized as one of the more distracting elements in a vehicle. It is a safety device in the sense that the vehicle would know whether there is a potential of the driver to be distracted in a critical situation, taking one more step toward a fully automatic driver's situational awareness-estimation system [14]–[17]. It is a comfort device in the sense that the passenger can still be allowed to access the infotainment controls to aide the driver in navigational and convenience needs.

For future work, upgrading the algorithm to affine invariance can be investigated further. To increase the affine invariance of the image descriptors, the image ROI can be repositioned (recalibrated) when the controls are visible on a regular basis to correct for any vibrations of the camera over time. A scheme as simple as template matching of the gradient images with the stored image may be used to align the ROI to the location that produces the best classification rates. It would also be interesting to undertake a comparative study to evaluate the modified HOG features with other suitable features.

Finally, one aspect of the VUD that was not studied in detail is the ability of the system to correctly determine the user when both the passenger's and driver's hands are in the ROI. Anecdotally, the system was observed to be able to determine the correct user even after the other occupant's hand occluded nearly 50% of the ROI by either reaching for the dome light or for the radio controls. This indicates that the feature descriptor is probably rich enough to classify between the three cases in these situations, and the (future) task is to collect and train on sufficient instances where occlusion occurs.

ACKNOWLEDGMENT

The authors would like to thank the kind assistance from Dr. A. Stoschek and J. Camhi, the members of the Computer Vision and Robotics Research Laboratory for their hand in data collection, and the reviewers for their valuable comments.

REFERENCES

[1] J. D. Lee, "Technology and teen drivers," *J. Saf. Res.*, vol. 38, no. 2, pp. 203–213, 2007.

[2] M. Jerbi, S.-M. Senouci, R. Meraihi, and Y. Ghamri-Doudane, "An improved vehicular ad hoc routing protocol for city environments," in *Proc. IEEE Int. Conf. Commun.*, Jun. 2007, pp. 3972–3979.

[3] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE CVPR*, 2005, vol. I, pp. 886–893.

[4] E. Murphy-Chutorian, A. Doshi, and M. M. Trivedi, "Head pose estimation for driver assistance systems: A robust algorithm and experimental evaluation," in *Proc. IEEE Int. Conf. Intell. Transp. Syst.*, Sep. 2007, pp. 709–714.

[5] P. B.-L. Chou, J. Lai, A. Levas, and P. A. Moskowitz, "System for controlling vehicle information user interfaces," U.S. Patent 6 181 996, Jan. 30, 2001.

[6] J. J. E. Harter, G. K. Scharenbroch, W. W. Fultz, D. P. Griffin, and G. J. Witt, "User discrimination control of vehicle infotainment system," U.S. Patent 6 668 221, Dec. 23, 2003.

[7] M. Kölsch and M. Turk, "Analysis of rotational robustness of hand detection with Viola & Jones' method," in *Proc. ICPR*, 2004, pp. 107–110.

[8] S. Y. Cheng, S. Park, and M. M. Trivedi, "Multi-spectral and multi-perspective video arrays for driver body tracking and activity analysis," *Comput. Vis. Image Underst.*, vol. 106, no. 2/3, pp. 245–257, May/Jun. 2007. DOI: 10.1016/j.cviu.2006.08.010.

[9] Q. Yuan, S. Sclaroff, and V. Athitsos, "Automatic 2d hand tracking in video sequences," in *Proc. IEEE Workshop Appl. Comput. Vis.*, Jan. 2005, pp. 250–256.

[10] V. I. Pavlovic, R. Sharma, and T. S. Huang, "Visual interpretation of hand gestures for human-computer interaction: A review," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 677–695, Jul. 1997.

[11] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[12] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 10, pp. 1615–1630, Oct. 2005.

[13] T. Serre, L. Wolf, S. Bileschi, M. R. Esenhuber, and T. Poggio, "Robust object recognition with cortex-like mechanisms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 3, pp. 411–426, Mar. 2007.

[14] M. M. Trivedi, T. Gandhi, and J. McCall, "Looking-in and looking-out of a vehicle: Computer-vision-based enhanced vehicle safety," *IEEE Trans. Intell. Transp. Syst.*, vol. 8, no. 1, pp. 108–120, Mar. 2007.

[15] A. Doshi and M. M. Trivedi, "On the roles of eye gaze and head pose in predicting driver's intent to change lanes," *IEEE Trans. Intell. Transp. Syst.*, vol. 10, no. 3, pp. 453–462, Sep. 2009.

[16] M. M. Trivedi and S. Y. Cheng, "Holistic sensing and active displays for intelligent driver support systems," *Computer*, vol. 40, no. 5, pp. 60–68, May 2007.

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