Proceedings of the 4th International Conference on Automotive User Interfaces and
Interactive Vehicular Applications (AutomotiveUI '12), October 17–19, 2012, Portsmouth, NH, USA

Hand Gesture-based Visual User Interface for Infotainment

Eshed Ohn-Bar
eohnbar@ucsd.edu

Cuong Tran
cutran@ucsd.edu

Mohan Trivedi
mtrivedi@ucsd.edu

Laboratory for Intelligent and Safe Automobiles (LISA)
University of California, San Diego, CA 92093, USA

ABSTRACT

We present a real-time vision-based system that discriminates hand gestures performed by in-vehicle front-row seat occupants for accessing the infotainment system. The hand gesture-based visual user interface may be more natural and intuitive to the user than the current tactile interaction interface. Consequently, it may encourage a gaze-free interaction, which can alleviate driver distraction without limiting the user's infotainment experience. The system uses visible and depth images of the dashboard and center-console area in the vehicle. The first step in the algorithm uses the representation of the image area given by a modified histogram-of-oriented-gradients descriptor and a support vector machine (SVM) to classify whether the driver, passenger, or no one is interacting with the region of interest. The second step extracts gesture characteristics from temporal dynamics of the features derived in the initial step, which are then inputted to a SVM in order to perform gesture classification from a set of six classes of hand gestures. The rate of correct user classification into one of the three classes is 97.9% on average. Average hand gesture classification rates for the driver and passenger using color and depth input are above 94%. These rates were achieved on in-vehicle collected data over varying illumination conditions and human subjects. This approach demonstrates the feasibility of the hand gesture-based in-vehicle visual user interface.

Categories and Subject Descriptors

H.5.2 [**Information Interfaces and Presentation (e.g., HCI)**]: User Interfaces; I.4.9 [**Computing Methodologies**]: Image Processing and Computer Vision—*Applications*

Keywords

Contact-free; hand-gesture recognition; infotainment; kinect; user determination

AutomotiveUI'12, October 17-19, Portsmouth, NH, USA.

1. INTRODUCTION

Today, there is an abundance of new devices that drivers can look at or interact with while driving the vehicle. Such devices provide useful information, entertainment, and connectivity. The potential for such technology is great, as web applications, location-based services, and passive and active safety systems become standard in vehicles. These devices, while providing drivers the capacity for enhanced efficiency and productivity, also form potential problems arising from distraction and inattention. Consequently, there are increasing safety concerns regarding the interaction with devices that may increase visual load and cause the driver to shift his or her gaze from the road [5, 8, 9, 20]. We focus on infotainment-based interactions, in particular, posing an alternative to the current mainstream tactile (buttons) interaction that may require the driver to look away from the road [11]. According to [12], drivers involved in infotainment system use during near crashes exhibit distinct glance behaviors, generally suggesting lower levels of awareness about their driving environment.

Several types of interfaces have been proposed to accommodate the dynamic settings of in-vehicle content delivery [4, 6, 13, 14]. With the objective of alleviating visual load on the driver while maintaining a natural interface, auditory, tactile, and multi-sensory information displays were proposed. Recently, proximity sensing systems have shown promise in increased usability of the in-vehicle interface [15]. As discussed in [15], touch-based interactions may pose distractions because they have a small region of interest and still cause driver's gaze to shift away from the road. Hence, the usability and intuitiveness potential of contact-less interfaces seem advantageous. In particular, a visual user interface holds the potential for a more comprehensive and holistic analysis of driver and environment behavior [17, 16]. Novel vision-based systems have been introduced for driver assistance, such as automatic lane detection or automatic parking [18, 19]. Furthermore, from a hardware perspective, placing a camera or a Kinect is a relatively cheap and easy installation process compared to some of the aforementioned interfaces. A vision-based solution allows for a dynamic region of interest, and the potential of interacting with the interface in a closer proximity to the wheel.

With the aforementioned opportunities for distraction, a common solution has been to limit the functionality of the infotainment system with the goal of making it less distracting to the driver, thereby making the devices less useful to the passenger. A far better solution would incorporate knowledge of the current user accessing the system, as well as an intuitive interactivity interface for alleviating driver distraction and providing optimal information to passengers, a solution explored in [2]. For that purpose, we propose a **Ha**nd **G**esture-based **V**isual **U**ser **I**nterface for infotainment (**HaG VUI**) system that classifies who of the front-row seat occupants is per-

Figure 1: Color and depth image ROI used to determine the hand gesture and the user in the HaG VUI system.

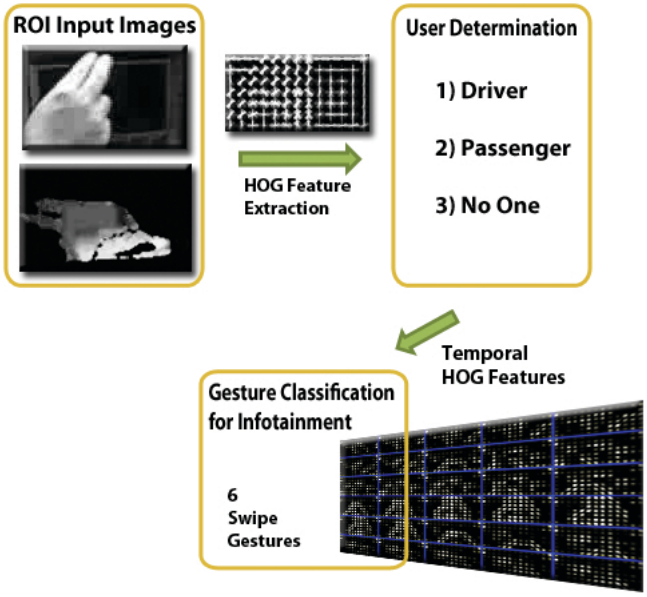

Figure 2: Hand Gesture-based Visual User Interface for Infotainment system (HaG VUI).

forming a gesture to interact with the infotainment system, and the type of hand gesture that is being performed. By tailoring the information to the user engaged with it-the driver, passenger, or no one, the system may improve vehicle safety. In particular, the customization could allow the infotainment settings to better adapt to the current user, thereby possibly resulting in a reduced visual load to the driver. Furthermore, intuitive gesture interaction may encourage the user to interact with the infotainment system in a gaze-free manner.

2. HAND GESTURE-BASED VISUAL USER INTERFACE FOR INFOTAINMENT

The HaG VUI system (figure 2) determines the user whose hand is accessing the infotainment device through the classification of vision and depth input in a region of interest (ROI) located in the center console by the gear shift (figure 1). Our system takes color and depth images from a Kinect sensor viewing the center-console area. The challenge of developing an in-vehicle vision-based system lies in the need for a classification algorithm that is robust to the various operating modes of the vehicle. Therefore, classification performance should be maintained through appearance and illumination variations. Since the depth input from the Kinect was shown to be less reliable in direct lightning conditions, the position of the sensor in the experiments was cruical. We attached the Kinect to the vehicle's roof behind the front row occupants facing the gearshift where the hand gestures were performed.

The initial user determination part of our system builds on the work of Cheng *et al*. [2] where a system for user discrimination based on a modified histogram-of-oriented-gradients (HOG) image descriptor [3] representation of image and near-infrared input is presented. The modified HOG descriptor is derived by taking the gradient of the image patch and dividing the gradient image into smaller cells. Within each cell, we quantize the angles (orientations) of the gradient vectors and bin them to form a histogram. The concatenation of the resulting histogram from each cell forms the final HOG descriptor for the image patch. We use a 2×2 grid of cells with 8 histogram bins resulting in 32-dimensional ($2 \times 2 \times 8$) feature vector. A support vector machine (SVM) [1] is used to classify the current user in the ROI. The user is defined as 1) driver; 2) passenger; and 3) no one. This allows for the tailoring of the infotainment system to the current user in order to provide more informative or less distracting driver assistance. Other methods have been devised for such purpose [7, 10] but the proposed method in Cheng *et al*. is significantly efficient and robust. The success of the method inspired an extension into an improved system of occupants-vehicle interaction through gestures. Our approach addresses the user determination and gesture recognition problem using direct classification of images of the infotainment control region. No tracking is required, and we assume the vehicle interior as a constant background.

Once a user is detected, the second stage of the system extracts gesture characteristics from temporal dynamics of the initial step. In this stage of the HaG VUI, we collect a 32-dimensional HOG feature vector which is being computed at each frame by the initial stage of the system. Then, once the user's hand leaves the ROI, we form a 'histogram image' composed of the collection of these feature vectors. Changes in the feature vector over time correspond to changes of the location and shape of the user's hand in sub-regions of the ROI. Since we are interested in recognizing local temporal patterns expressed in the spatial descriptors over the frames, we perform a second HOG feature extraction on this temporal sequence of feature vectors. In our experiments, this temporal HOG descriptor proved to be powerful in describing motion patterns. Secondly, the average over time for each entry in the spatial 32-dimensional feature vector is computed and used as a feature as well. Hence we derive a 64-dimensional feature vector, which is then used with a SVM to classify the gesture into one of six hand gestures (figure 3). The outlined procedure is performed both on the color and the depth input from the Kinect. A RBF kernel multiclass SVM is used with parameter values of $C = 15$ and $\gamma = 0.15$ for features derived from RGB input, and $C = 15$, $\gamma = 0.05$ for features derived from depth input.

Seq	# Samples	Weather	Time	Testbed
1	130	S	2pm	LISA-Q
2	177	S	3pm	LISA-Q
3	166	O	5pm	LISA-Q
4	140	O	5pm	LISA-Q
Total	613	{# Driver, # Passenger} = {299, 314}		

Table 1: Attributes summary of the four recording sequences of video data used for training and testing the hand gesture classifier in the HaG VUI system. Data was collected in two different moving vehicles using eight subjects. Weather conditions are indicated overcast (O) and sunny (S).

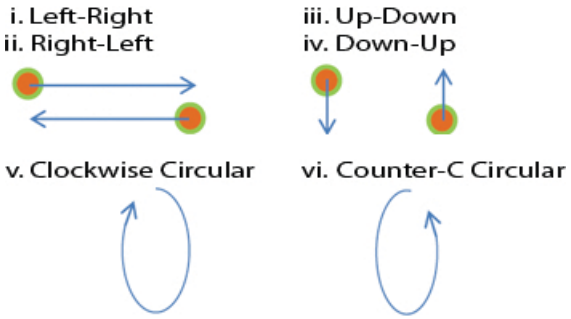

Figure 3: Gesture set comprising of 6 swipe gestures. The orange circle marks the starting position of each of the gestures. Circular swipes had an arbitrary initial position.

3. EXPERIMENTAL VALIDATION IN LISA TESTBEDS

In order to best exemplify the feasibility of the HaG VUI system in real-world application, we chose a gesture set (figure 3 that we thought would be suitable to perform while driving. Instructions regarding the performance of the gestures were given verbally. This resulted in large variations in the execution of the gestures. For instance, some users performed the hand movement with their entire hand as one object moving along the swipe gesture, while other users performed the same gesture almost entirely with a finger or two, and minimal wrist movement. Even simple gestures such as `Left-Right Swipes` included variations in the time it took to perform the gesture, as well as the location and trajectory in which the hand enters and exits the ROI. To verify the robustness of the HaG VUI, the data set was collected at various times of the day in a moving vehicle with eight subjects, providing four video recordings with a total of 613 gesture samples (table 1). The subjects were members of the LISA team. Individuals wore short and long-sleeve shirts and were of varying nationalities, skin colors, and gender. The participants were verbally instructed on how to perform the swipe gestures. They were told to perform each gesture whenever they felt comfortable, with the requirement that the hand must leave the ROI before a new gesture is performed. The gestures were performed in sequences of 10 each, with the driver and passenger alternating turns. Videos were recorded on a circular route in which the direction of sunlight could shine into the vehicle from multiple directions in one video. As mentioned in [2], the most difficult time for the initial stage of the system to differentiate between the three classes of users is at transition moments when a hand enters the ROI. Hence, we utilize a delay of 0.3 seconds before extracting features for the gesture classification stage of the system.

A Kinect camera was chosen primarily for the depth input of the camera; at night, the front-row area can still be captured without illumination using the infrared-based depth information, although this was not tested in our experiment. Image pixel resolution was 640×480 and 320×240 for color and depth input, respectively, at 30 fps. The system first extracts a rectangular image patch sized 80×120 as depicted in figure 1. Feature extraction and classification takes approximately 60 ms on an Intel Pentium D 3.2-GHz PC.

4. EXPERIMENTAL EVALUATION AND DISCUSSION

Classification of the user into three classes of driver, passenger, or no one, has an average accuracy rate of 97.9%. The performance of the gesture classification step is shown in tables 2 and 3 for the color and depth input. The performance metrics were all calculated using a fivefold cross validation. Table 2 shows the accuracy in terms of a confusion matrix for the two input modalities. The mean performance using color image features is 98.41% and 97.60% for driver and passenger, respectively. Features derived from depth images were shown to result in a lower overall correct classification rate of the hand gestures, with 94.24% and 95.74% for the driver and passenger, respectively.

There are several reasons for the misclassifications seen in table 2. First, variations in the performance of the gesture, such as in the trajectory of entering and leaving the ROI, may lead to incorrect classification. For instance, a driver may perform a `Left-Right Swipe` and withdraw the hand through the ROI in a certain trajectory resembling a performances of a `Clockwise Swipe` gesture. Interestingly, a subject who wore a long-sleeve shirt and performed a `Left-Right Swipe` was the reason of such misclassification. The circular swipes contain large variations in gesture performance as these don't have a general initial position, and may involve more self-occlusion. Table 2b shows the classification results using depth input, where we can see how out of the six gestures, four contain instances of misclassification as a `Clockwise Swipe` gesture.

Confusion between gesture classes with opposite movement patterns is also apparent, such as between `Left-Right Swipe` and a `Right-Left Swipe`. For instance, in a `Left-Right Swipe` gesture performed by the passenger, a large portion of the gesture may involve moving the hand towards the initial position on the left side of the ROI. Then, a a quick movement to the right combined with exiting from the ROI is followed. As this may produce a similar temporal descriptor to a gesture that differ purely in direction, the `Right-Left Swipe` gesture, appropriate measures to eliminate these intra-class misclassifications need to be devised. In particularly, we would like to separate entering or exiting segments from the actual gesture performed. One possibility for doing so is using depth information (when it's reliable) to distinguish these segments by the location of the movement in 3-D space. The hope was that a depth-based input could alleviate such situations, as the motion of a hand exiting the ROI is not entirely performed in the same plane where the gesture was performed. Unfortunately, due to the close range and varying lightning condition, the depth didn't seem to provide such an advantage. Using depth-based features was shown to produce slightly lower rate of correct classification, mostly due to noise in the input data. The hand gestures `Clockwise Circular Swipe` and `Counter-Clockwise Circular Swipe` have motion patterns involving all the subregions of the ROI, thereby are captured well in the temporal histograms for these sub-cells, and produce high rates of correct classification. On the other hand, the other swipes are affected by the

Actual \ *Result*	RghtLft	LftRght	UpDwn	DwnUp	CntrClk	Clkwse
RghtLft	**96.56**	1.22	2.22	0.00	0.00	0.00
LftRght	3.57	**95.32**	0.00	0.00	0.00	1.11
UpDwn	0.00	0.00	**98.25**	0.89	0.00	0.86
DwnUp	0.00	0.00	0.00	**100.00**	0.00	0.00
CntrClk	0.00	0.00	0.91	0.00	**99.09**	0.00
Clkwse	0.00	0.00	0.85	0.00	0.88	**98.28**

(a) HaG VUI using a temporal descriptor derived from **RGB input**.
Mean correct classification rate **%98.01**.

Actual \ *Result*	RghtLft	LftRght	UpDwn	DwnUp	CntrClk	Clkwse
RghtLft	**93.12**	4.66	0.00	1.22	0.00	1.00
LftRght	8.23	**89.98**	0.00	0.00	1.79	0.00
UpDwn	0.00	0.00	**93.87**	2.62	0.89	2.62
DwnUp	0.00	0.00	1.92	**94.82**	2.17	1.09
CntrClk	0.00	0.00	0.00	0.00	**98.18**	1.82
Clkwse	0.88	0.00	0.00	0.00	0.85	**98.28**

(b) HaG VUI using a temporal descriptor derived from **depth input**.
Mean correct classification rate **%95.00**.

Table 2: Summary of the HaG VUI performance using the two modalities. Gesture classification accuracy is computed with five-fold cross validation. The results shown are the averaged results for the driver and passenger hand gestures classification performance. Each row sums up to 100%.

noise as the illumination variation introduces artifacts that are not a part of the gesture being performed.

It is important to note that in the data collection process segments of the hand gestures performed by the subjects were partially out of the ROI chosen. These examples are essential, as we are interested in an intuitive means of communicating with the infotainment center. For instance, a portion of the hand performing `Clockwise Circular Swipe` may exit the top of the ROI in the middle of the motion and return to the ROI when completing the swipe. While this results in a more challenging classification task, it exemplifies the strengths of a vision-based interactivity system which allows for the incorporation of information from the arm for inferring the correct gesture. Also, a dynamic ROI can also be utilized. These inaccuracies in the gesture performance are only natural, as a careful performance of gestures requires an increased cognitive and perhaps visual attentiveness. Since the instructions regarding the performance of the gestures were given verbally, the performance of the hand gestures varied significantly. For instance, some subjects chose to perform the swipes with their entire hand moving as one object along the swipe gesture, while others performed the same gesture almost entirely with a finger or two, yet with minimal wrist movement (and possibly self-occlusion). While the feasibility of the system as an intuitive interface was not investigated in our work, we did observe how both the passenger and driver became increasingly natural in their carrying out of the gestures (thereby extending beyond the ROI while performing larger strokes at times). These large and inaccurate movements provided natural variations which we incorporated into our training and testing set because they exemplify the potential of the HaG VUI system.

	Precision	Recall
RGB	97.97	98.13
Depth	95.24	94.93

$$\text{Recall} = \frac{|\{Actual\} \cap \{Predicted\}|}{|\{Actual\}|}$$

$$\text{Precision} = \frac{|\{Actual\} \cap \{Predicted\}|}{|\{Predicted\}|}$$

Table 3: Summary of the HaG VUI performance using precision and recall rates.

5. CONCLUDING REMARKS

The result of this work suggests that a reliable vision-based system in the volatile environment of the vehicle interior is feasible. The high accuracy results for the gesture recognition stage of the system are encouraging, yet in order to further evaluate the robustness of the system under different driving conditions, larger scale data collection including more users, driving routes, and different camera positions should be performed. The usability of the proposed system may also be extended, such as to coupling it with head-up displays (HUDs) which projects information directly into the field of view of the driver. Additional efforts must be made to study the safety and benefits of occupants-vehicle interactions using the system. The gesture set must then be evaluated and redefined to optimally accommodate the vehicle occupants. Additionally, the interface was studied for instances of gestures in which the hand left the ROI entirely before another gesture was performed or another user's hand entered the ROI. The best scheme for multiple gesture instances classification must be devised. Furthermore, the optimal ROI location in the vehicle should be evaluated as well. The system performance should be analyzed with respect to camera jitter, as vibrations may shift the ROI and affect the gesture classification performance. A system component for re-positioning and aligning of the ROI in real-time according to specific dashboard keypoints to correct for vibrations or shifting in camera position may be desired. For such purpose, a dynamic ROI could utilize the Kinect's motor. Finally, while our work provides separate analysis for the two modalities, an efficient incorporation of RGB and depth towards a more robust classification scheme could be devised. As the proposed gesture recognition methodology is only one possible approach in many to extract spatio-temporal features for the purpose of gesture classification, a more extensive comparison between gesture recognition methodologies on the dataset is an important future step.

6. ACKNOWLEDGMENTS

The authors would like to thanks the members of the Laboratory for Intelligent and Safe Automobiles for their help in data collection, and the reviewers for their valuable comments.

References

[1] CHANG, C.-C., AND LIN, C.-J. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology 2* (2011), 27:1–27:27. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[2] CHENG, S. Y., AND TRIVEDI, M. M. Vision-based infotainment user determination by hand recognition for driver assistance. *IEEE Transactions on Intelligent Transportation Systems 11*, 3 (September 2010), 759–764.

[3] DALAL, N., AND TRIGGS, B. Histograms of oriented gradients for human detection. In *Proc. IEEE CVPR* (2005), vol. 1, pp. 886–893.

[4] DOSHI, A., CHENG, S. Y., AND TRIVEDI, M. M. A novel active heads-up display for driver assistance. *IEEE Transactions on Systems, Man, and Cybernetics, Part B 39*, 1 (2009), 85–93.

[5] DREWS, F. A., YAZDANI, H., GODFREY, C. N., COOPER, J. M., AND STRAYER, D. L. Text messaging during simulated driving. *The Journal of the Human Factors and Ergonomics Society*, 51 (January 2011), 762–770.

[6] ECKER, R., BROY, V., HERTZSCHUCH, K., AND BUTZ, A. Visual cues supporting direct touch gesture interaction with in-vehicle information systems. In *Proceedings of the 2nd International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (New York, NY, USA, 2010), AutomotiveUI '10, ACM, pp. 80–87.

[7] HARTER, J. J. E., SCHARENBROCH, G. K., FULTZ, W. W., GRIFFIN, D. P., AND WITT, G. J. User discrimination control of vehicle infotainment system, u.s. patent 6 668 221.

[8] HORREY, W. J. Assesing the effects of in-vehicle tasks on driving performance. *Ergonomics in Design*, 19 (December 2011), 4–7.

[9] JAHN, G., KREMS, J. F., AND GELAU, C. Skill aquisition while operating in-vehicle information systems: Interface design determines the level of safety-relevant distractions. *Human Factors and Ergonomics Society*, 51 (June 2009), 136–151.

[10] KOLSCH, M., AND TURK, M. Analysis of rotational robustness of hand detection with viola jones method. In *ICPR* (2004), pp. 107–110.

[11] LEE, J. D., ROBERTS, S. C., HOFFMAN, J. D., AND ANGELL, L. S. Scrolling and driving: How an mp3 player and its aftermarket controller affect driving performance and visual behavior. *The Journal of the Human Factors and Ergonomics Society* (January 2012).

[12] PEREZ, M. A. Safety implications of infotainment system use in naturalistic driving. *Work: A Journal of Prevention, Assessment and Rehabilitation 41* (2012), 4200–4204.

[13] PITTS, M. J., WILLIAMS, M. A., WELLINGS, T., AND ATTRIDGE, A. Assessing subjective response to haptic feedback in automotive touchscreens. In *Proceedings of the 1st International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (New York, NY, USA, 2009), AutomotiveUI '09, ACM, pp. 11–18.

[14] RICHTER, H., ECKER, R., DEISLER, C., AND BUTZ, A. Haptouch and the 2+1 state model: potentials of haptic feedback on touch based in-vehicle information systems. In *Proceedings of the 2nd International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (New York, NY, USA, 2010), AutomotiveUI '10, ACM, pp. 72–79.

[15] RIENER, A., AND WINTERSBERGER, P. Natural, intuitive finger-based input as a means of in-vehicle information system operation. In *Automotive User Interfaces and Interactive Vehicular Applications* (December 2011), pp. 159–166.

[16] TRAN, C., AND TRIVEDI, M. Driver assistance for "keeping hands on the wheel and eyes on the road". In *Vehicular Electronics and Safety (ICVES), 2009 IEEE International Conference on* (nov. 2009), pp. 97 –101.

[17] TRAN, C., AND TRIVEDI, M. M. Vision for driver assistance: Looking at people in a vehicle. In *Visual Analysis of Humans*, T. B. Moeslund, A. Hilton, V. Kruger, and L. Sigal, Eds. Springer London, 2011, pp. 597–614.

[18] TRIVEDI, M. M., AND CHENG, S. Y. Holistic sensing and active displays for intelligent driver support systems. *Computer 40*, 5 (May 2007), 60–68.

[19] TRIVEDI, M. M., GANDHI, T., AND MCCALL, J. Looking-in and looking-out of a vehicle: Computer-vision-based enhanced vehicle safety. *IEEE Trans. Intell. Transp. Syst. 8*, 1 (Mar 2007), 108–120.

[20] WILLIAMSON, A. R., YOUNG, K. L., NAVARRO, J., AND LENNE, M. G. Music selection using a touch screen interface: effect of auditory and visual feedback on driving and usability. *International Journal of Vehicle Design 57*, 4 (2011), 391–404.