Hand Gesture Recognition in Real-Time for Automotive Interfaces: A Multimodal Vision-based Approach and Evaluations

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

***Abstract*—In this work, we develop a vision-based system that employs a combined RGB and depth descriptor in order to classify hand gestures. The method is studied for a human-machine interface application in the car. Two interconnected modules are employed: one that detects a hand in the region of interaction and performs user classification, and the second performing the gesture recognition. The feasibility of the system is demonstrated using a challenging RGBD hand gesture data set collected under settings of common illumination variation and occlusion..**

***Index Terms*—Human-machine interaction, hand gesture recognition, driver assistance systems, infotainment, depth cue analysis.**

I. Introduction

Recent years have seen a tremendous growth in novel devices and techniques for human-computer interaction (HCI). These draw upon human-to-human communication modalities in order to introduce a certain intuitiveness and ease to the HCI. In particular, interfaces incorporating hand gestures have gained popularity in many fields of application. In this paper, we are concerned with the automatic visual interpretation of dynamic hand gestures, and study these in a framework of an in-vehicle interface. A real-time, vision-based system is developed, with the goal of robust recognition of hand gestures performed by driver and passenger users. The techniques and analysis extend to many other application fields requiring hand gesture recognition in visually challenging, real-world settings.

Motivation for in-vehicle gestural interfaces: In this paper, we are mainly concerned with developing a vision-based, hand gesture recognition system that can generalize over different users and operating modes, and show robustness under challenging visual settings. In addition to the general study of robust descriptors and fast classification schemes for hand gesture recognition, we are motivated by recent research showing advantages of gestural interfaces over other forms of interaction for certain HCI functionalities.

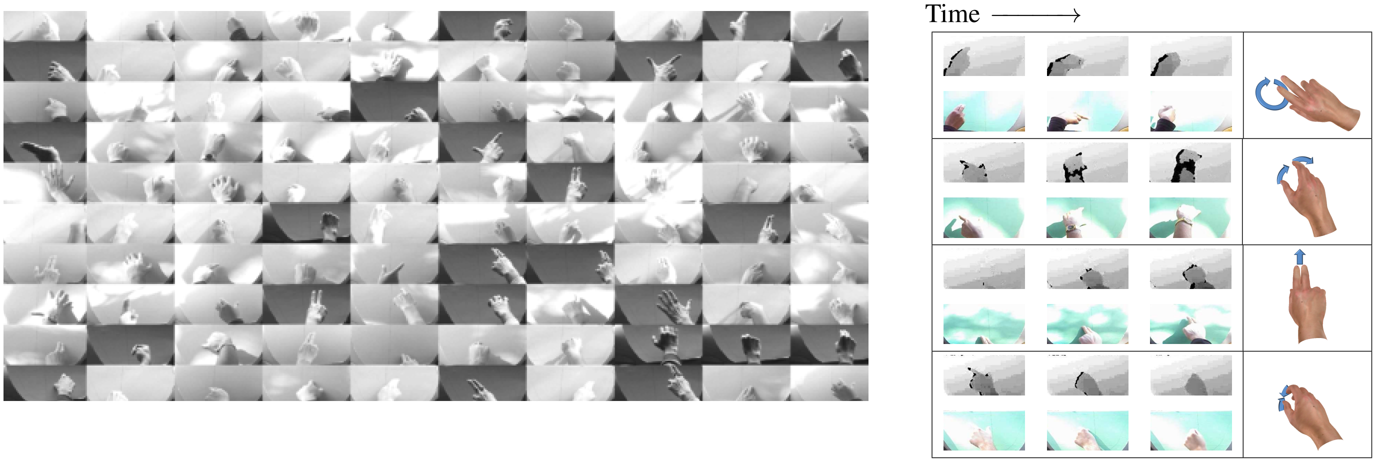
Among tactile, touch, and gestural in-vehicle interfaces, gesture interaction was reported to pose certain advantages over the other two, such as lower visual load, reduced driving errors, and a high level of user acceptability [1]–[3]. The reduction in visual load and non-intrusive nature led many automotive companies to research such HCI [4] in order to alleviate the growing concern of distraction from interfaces with increasingly complex functionality in today's vehicles [5]–[7]. Following a trend in other devices where multi-modal interfaces opened ways to new functionality, efficiency, and comfort for certain users (as opposed to interaction approaches based solely on tangible controllers), we propose an alternative or supplementary solution to the in-vehicle interface. As each modality has its strengths and limitations, we believe a multi-modal interface should be pursued for leveraging advantages from each modality and allowing customization to the user.

Advantages for developing a contact-less vision-based interface solution: The system proposed in this paper may offer several advantages over a contact interface. First, camera input could possibly serve multiple purposes, in addition to the interface. For instance, it allows for analysis of additional hand activities or salient objects inside the car (as in [8]–[12]), important for advanced driver assistance systems. Furthermore, it allows for the determination of the user of the system (driver or passenger), which can be used for further customization. Second, it offers flexibility to where the gestures can be performed, such as close to the wheel region. A gestural interface located above the wheel using a heads up display was reported to have high user acceptability in [2]. In addition to allowing for interface location customization and a non-intrusive interface, the system can lead to further novel applications, such as for use from outside of the vehicle. Third, there may be some potential advantages in terms of cost, as placing a camera in the vehicle involves a relatively easy installation. Just as contact gestural interfaces showed certain advantages compared to conventional interfaces, contact-free interfaces and their effect on driver visual and mental load should be similarly studied. For instance, accurate coordination may be less needed when using a contact-free interface as opposed to when using a touch screen, thereby possibly reducing glances at the interface.

Challenges for a vision-based system: The method must generalize over users and variation in the performance of the gestures. Segmentation of continuous temporal gesture events is also difficult. In particular, gesture recognition in the volatile environment of the vehicle's interior differs significantly from gesture recognition in the constrained environment of an office. Firstly, the algorithm must be robust to varying global illumination changes and shadow artifacts. Secondly, since the camera is mounted behind the front-row seat occupants in our study and gestures are performed away from the sensor, the hand commonly self-occludes itself throughout the performance of the gestures. Precise pose estimation (as in [13], [14]) is difficult and was little studied before in settings of harsh illumination changes and large self-occlusion, yet many approaches rely on such pose information for producing the discriminatory features for gesture classification. Finally, fast computation (ideally real-time) is desirable.

In order to study these challenges, we collected a RGB-

The authors are with the Laboratory for Intelligent and Safe Automobiles (LISA), University of California San Diego, San Diego, CA 92093-0434 USA (e-mail: eohnbar@ucsd.edu; mtrivedi@ucsd.edu).



(a) Large variation in illumination and performance of the gestures.

(b) Example gestures in the dataset.

Fig. 1: Examples of the challenges for a vision-based in-vehicle gesture interface. Illumination artifacts (saturation, high contrast shadows, etc.) throughout the performance of the gestures in the dataset are shown. Gestures are performed away from the sensor, resulting in frequent self-occlusion. The type of gestures varies from coarse hand motion to fine finger motion. (b) The gestures shown are (top to bottom): clockwise O swipe, rotate clockwise, scroll up, pinch\zoom-in.

Depth (RGBD) dataset of 19 gestures, performed 2-3 times by 8 subjects (each subject preformed the set as both driver and passenger) for a total of 886 instances. Examples of gesture samples and the challenging settings are shown in Fig. 1. The dataset collected allows for studying user and orientation invariance, the effects of occlusion, and illumination variability due to the position of the interface in the top part of the center console. Different common spatio-temporal feature extraction methods were tested on the dataset, showing its difficulty (Table IV).

In this paper, we pursue a no-pose approach for recognition of gestures. A set of common spatio-temporal descriptors [15]–[17] are evaluated in terms of speed and recognition accuracy. Each of the descriptors is compared over the different modalities (RGB and depth) with different classification schemes (kernel choices for a Support Vector Machine classifier [18]) for finding the optimal combination and gaining insights into the strengths and limitations of the different approaches. Finally, the gesture dataset is used to study effects of different training techniques, such as user-specific training and testing, on recognition performance. The results of this study demonstrate the feasibility of an in-vehicle gestural interface using a real-time system based on RGBD cues. The gesture recognition system studied is shown to be suitable for a wide range of functionalities in the car.

II. Related Research Studies

As the quality of RGB and depth output from cameras improve and hardware prices decline, a wide array of applications spurred an interest in gesture recognition in the research community. Relevant literature related to gesture recognition and user interfaces is summarized below.

Video descriptors for spatio-temporal gesture analysis: Recent techniques for extracting spatio-temporal features from video and depth input for the purpose of gesture and activity recognition are surveyed in [19], [20]. Generally, hand gesture recognition methods may extract shape and motion features that represent temporal changes corresponding to the gesture performance, as in [17], [39]. These can be extracted locally using spatio-temporal interest points (as in [22], [23]) or sampled densely. Such features may be hand crafted, as done in this paper, or learned using a convolutional network [24]. Information of pose, although difficult to obtain in our application, is also highly useful for recognition, as demonstrated in [25]–[30].

Hand gesture recognition with RGBD cues: The introduction of high-quality depth sensors at a lower cost, such as the Microsoft Kinect, facilitated the development of many gesture recognition systems. In particular, hand gesture recognition systems were developed with applications in fields of sign language recognition [31]–[34], driver assistance [35], [36], smart environments [37]–[39], video games [40], medical instrumentation [41], [42], and other human-computer interfaces [43]–[45]. Hand gesture recognition systems commonly use depth information for background removal purposes [46]–[49]. [46] proposed using a Finger-Earth Mover's Distance (FEMD) for recognizing static poses. Hand detection is commonly performed using skin analysis [33], [48]. In [33], depth information is used to segment the hand and estimate its orientation using PCA with a refinement step. The classification of static gestures is performed using an average neighborhood margin maximization classifier combined with depth and hand rotation cues. In [34], a nearest neighbor classifier with a dynamic time warping (DTW) measure was used to classify dynamic hand gestures of digits from zero to nine. A Hidden Markov Model (HMM) may also be used [50] for gesture modeling. Minnen *et al.* [51] used features of global image statistics or grid coverage, and a randomized decision forest for depth-based static hand pose recognition. There has been some work in adapting color descriptors to be more effective when applied to depth data. As noted by [52], common RGB based techniques (e.g. spatio-temporal interest points as in Dollár *et al.* [53])

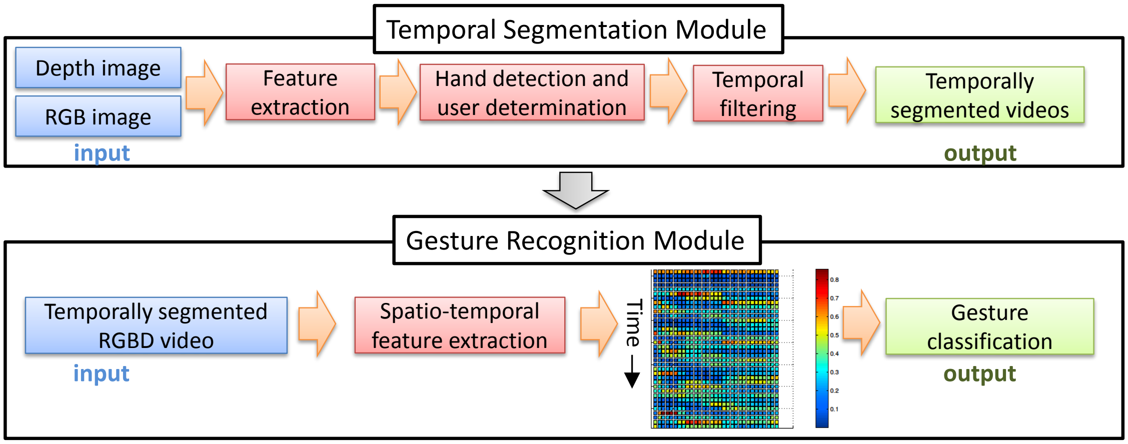

Temporal Segmentation Module
Depth image
RGB image
input
Feature extraction
Hand detection and user determination
Temporal filtering
Temporally segmented videos
output
Gesture Recognition Module
Temporally segmented RGBD video
input
Spatio-temporal feature extraction
Time
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
Gesture classification
output

Fig. 2: Outline of the main components of the system studied in this paper for in-vehicle gesture recognition. First, the hand detection module provides segmentation of gestures and determines the user, which is either the passenger or the driver. This is followed by spatio-temporal feature analysis for performing gesture classification.

may not work well on the output of some depth sensors, and need to be adjusted as in [54].

In this work we focus on approaches that do not involve tracking of hand pose. Each descriptor is applied to the RGB and depth modality separately, and finally these are early-fused together by concatenation. Common spatio-temporal feature extraction methods such as a histogram of 3D oriented gradients (HOG3D) [16], motion boundary descriptors and dense trajectories [17], and other forms of gradient-based spatio-temporal feature extraction techniques [15] will be evaluated on the challenging dataset. For classification, an SVM classifier is employed [18].

Hand gesture interfaces in the car: Finally, we briefly review works with affinity to the vehicle domain. A similar effort to ours was reported in Zobl *et al.* [55], where a CCD camera and NIR LEDs illumination in a simulator were used to perform gesture recognition out of an elaborate gesture inventory of 15 gestures. The gestures used were both static and dynamic. Static gestures may be used to activate the dynamic gesture recognizer. A HMM is employed to perform the dynamic gesture recognition. The inventory is not explicitly mentioned, as well as the speed of the algorithm, and only one subject was used. There also has been some work towards standardization of the in-vehicle gestural interaction space [56]. Althoff *et al.* [57] studied 17 hand gestures and six head gestures using an infrared camera, and a HMM and rule-based classifier. Endres *et al.* [58] used a Theremin device, a contact-less device consisting of two metal antennas. Moving the hand alters the capacity of an oscillating current, generating a signal which is fed to a DTW classifier.

III. Hand Gesture Recognition in the Car

A. Experimental Setup and Dataset

The proposed system uses RGB and depth images in a region of interest (ROI). In our experiments, this ROI was chosen to be the instrument panel (shown in Fig. 1 and Fig. 3). In order to demonstrate the feasibility of the system, we collected a dataset containing 19 hand gestures. The dataset is publicly available at http://cvrr.ucsd.edu/LISA/hand.html. Each gesture was performed about three times by eight subjects. Each subject performed the set two times, once as the driver and once as the passenger. The gestures are all dynamic, as these are common in human-to-human communication and existing gestural interfaces. The size of the RGB and depth maps are both 640 × 480, and the ROI is 115 × 250. Altogether, the dataset contains 886 gesture samples. The main focus of this work is recognition of gestures under illumination artifacts, and not the effects of the interface on driving. Therefore, subjects were requested to drive slowly in a parking lot while performing the gestures, as the gestures were verbally instructed. Subjects 1 and 4 performed the gestures in a stationary vehicle. It was observed that following the initial learning of the gesture set, both passenger and driver carried the gestures more naturally. At times this resulted in the hand partially leaving the pre-defined infotainment ROI, as strokes became large and more flowing. These large and inaccurate movements provided natural variations which were incorporated into the training and testing set.

Fig. 4 shows the illumination variation among videos and subjects. A temporal sum was performed over the number of pixel intensities above a threshold in each gesture video to produce an average intensity score for the video,

$$\text{Intensity Score} = \frac{1}{m \times n \times T} \sum_{t=1:T} |\{(x,y) : I_t(x,y) > 0.95\}| \tag{1}$$

That is, the average number of high intensity pixels over the $m \times n$ images I_t in a video of length T. A large variation in the dataset is observed in Fig. 4, both within the same subject and among subjects.

Interface location: Among previously proposed gestural interfaces, the location of the interface varies significantly. In our study, the gestures were performed by the center console,

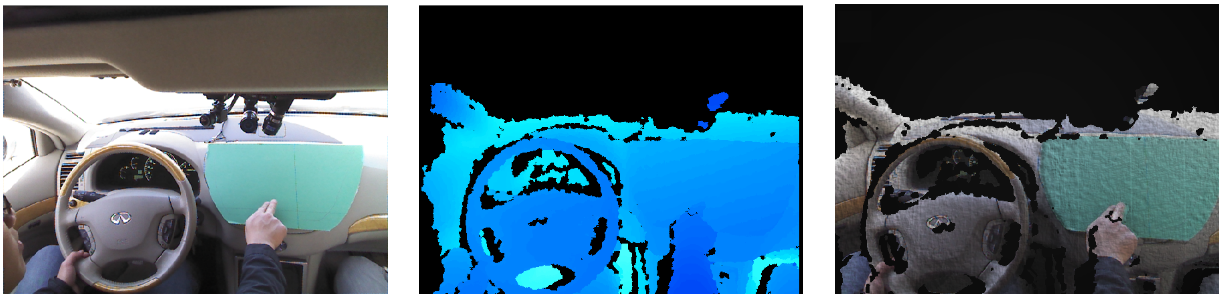


Fig. 3: Camera setup (color, depth, and point cloud) for the in-vehicle vision-based gesture recognition system studied in this work.

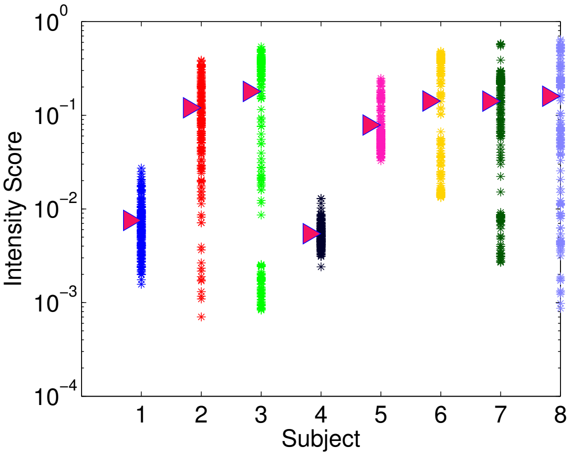


Fig. 4: Illumination variation among different videos and subjects as the average percent of high pixel intensities (see Eqn. 1). Each point corresponds to one gesture sample video. The triangles plot the overall mean for each subject. Videos with little to no illumination variation were taken using subjects 1 and 4.

TABLE I: Attribute summary of the eight recording sequences of video data used for training and testing. Weather conditions are indicated as overcast (O) and sunny (S). Time of capture was done in afternoon and mid-afternoon. Skin-color varies from light (C1) to intermediate (C2) and dark brown/black (C3).

Subject	Gender	Weather	Skin Color
1	M	O	C2
2	M	S	C2
3	M	S	C2
4	M	O	C3
5	M	S	C1
6	M	S	C3
7	F	S	C3
8	M	S	C1
Total Samples: {# Driver, # Passenger} = {450, 436}			

as shown in Fig. 3. We chose a position that would be difficult for a vision-based system due to illumination artifacts and self-occlusion. In future design, the location of the interface should depend on whether the system aims to replace or supplement existing secondary controls, and the type of feedback that will be used.

Gesture inventory: The inventory is as follows. Two-finger swipe gestures: *swipe left*, *swipe right*, *swipe down*, *swipe up*, *swipe V*, *swipe X*, *swipe + (plus)*. The motion in these is mostly performed with the fingers, and not with the hand, as opposed to the scroll where the fingers move with the entire hand in the direction of the scrolling: *scroll left*, *scroll right*, *scroll down*, and *scroll up*. One tap gestures can be done with one or three fingers, *one tap-1* and *one tap-3*. Next we have the *open* and *close*, a fist following a spread open palm or vice-versa. Finally, we use a two finger *pinch* as shown in Fig. 1-bottom, and the *expand* (opposite motion), as well as *rotate counter-clockwise* and *rotate clockwise* (Fig. 1-second row). We note that there were small variations in the performance of some of the gestures; for instance the *swipe X* and *swipe +* can be performed in multiple ways, depending on the starting position of the hand.

Gesture functionality: The 19 gestures are grouped into three subsets with increasing complexity for different in-vehicle applications as shown in Table II. A set of functionalities is proposed for each gesture.

For GS1 (phone), the *open* and *close* gestures are used to answer or end a call. Scrolls provide volume control, and the *swipe* + provides the 'info/settings/bring up menu' button. GS2 involves additional gestures for music control. Swipes provide the 'next' and 'previous' controls. A tap with one finger pauses, and with three fingers allows for a voice search of a song. Finally, the X and V swipes provide feedback and ranking of a song; so that the user can 'like' or 'dis-like' songs. This gesture set contains gestures that can be used for general navigation through other menus if needed. Finally, the more complex GS3 contains refined gestures purposed for picture or navigation control. A one finger tap is used for 'select', the scrolls for moving throughout a map, two finger rotation gestures rotate the view, and *expand* and *pinch* allows for zoom control. *Swipe up* and *swipe down* are used for transition between bird-eye view to street view.

B. Hand Detection and User Determination

Both recognition and temporal segmentation must be addressed. Since recognition was found to be a challenging task on its own, it is the main focus of this paper. In particular, spatio-temporal features are evaluated in terms of speed,

TABLE II: Three subsets of gestures chosen for evaluation of application-specific gesture sets.

Gesture Set 1 (GS1) Phone	Gesture Set 2 (GS2) Music\Menu Control	Gesture Set 3 (GS3) Picture\Navigation
SwipePlus	SwipeX	SwipeUp
SwipeV	SwipeV	SwipeDown
Close	OneTap3	OneTap1
ScrollUp2	ScrollUp2	ScrollUp2
ScrollDown2	ScrollDown2	ScrollDown2
Open	OneTap1	ScrollRight2
	SwipeRight	ScrollLeft2
	SwipeLeft	RotateCntrClwse
		RotateClwse
		Expand\ZoomOut
		Pinch\ZoomIn

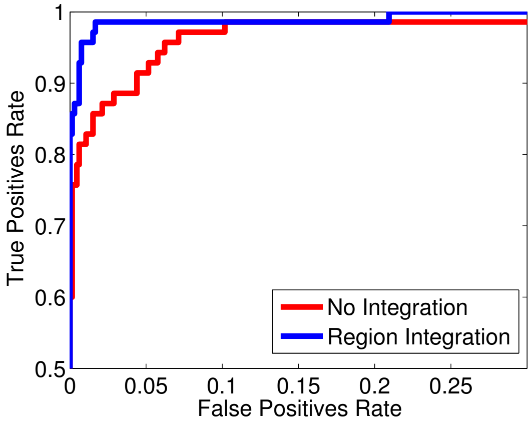


Fig. 5: Driver hand presence detection in the instrument panel region. As the instrument panel region is large with common illumination artifacts, cues from other regions in the scene (such as the wheel region) can increase the robustness of the hand detection in the instrument panel region.

performance, and varying generalization. Although temporal segmentation is a difficult problem as well, in this work we employ a simple segmentation of temporal gestures using a hand presence detector, so that the hand must leave the ROI between different gestures.

The first module in the system performs hand detection in a chosen ROI. The classification may be binary, detecting whether a hand or not is present in the ROI, or multiclass for user determination, as in [59]. In the latter case, a three class classification performs recognition of the user: 1) no one; 2) driver; or 3) passenger. This is done with a simplified version of the histogram of oriented (HOG) algorithm [60] which will be described below and an SVM classifier. For clarity and reproducibility, we detail the implementation of the visual features extraction used in this work.

HOG spatial feature extraction: Let $I(x,y)$ be an $m \times n$ signal. The discrete derivatives G_x and G_y are approximated using a $1D$ centered first difference $[-1,0,1]$ to obtain the magnitude, G, and quantized orientation angles into B bins, Θ. The image is split into $M \times N$ blocks. We found that overlapping the blocks produces improved results, and throughout the experiments a 50% overlap between the cells is used. Let G^s, Θ^s denote a cell for $s \in \{1,\dots,M \cdot N\}$, so that the q^{th} bin for $q \in \{1,\dots,B\}$ in the histogram descriptor for the cell is

$$h^s(q) = \sum_{x,y} G^s_{x,y} \cdot \mathbf{1}[\Theta^s(x,y) = \theta] \tag{2}$$

where $\theta \in \{-\pi + \frac{2\pi}{B} : \frac{2\pi}{B} : \pi\}$ and $\mathbf{1}$ is the indicator function. The local histogram is normalized using an L2-normalization: $h^s \rightarrow h^s/\sqrt{\|(h^s)\|_2 + \epsilon}$. Finally, the descriptor at frame t is the concatenation of the histograms from the cells

$$h_t = [h^1, \dots, h^{M \cdot N}]. \tag{3}$$

For additional details and analysis on this part of the algorithm we refer the reader to [59].

Region integration for improved hand detection: The specific setup of location and size of the ROI can have a significant impact on the illumination variation and background noise in the ROI. Because the location of the ROI in our setup produces common illumination artifacts, we found that using visual information from other ROIs in the scene improves hand detection performance under ambiguous and challenging settings [61]. For instance, features extracted from the wheel, gear shift, and side hand-rest regions were shown to increase detection accuracy for the driver's hand in the ROI (Fig. 5).

C. Spatio-Temporal Descriptors from RGB and Depth Video

The first module described in the previous section produces a video sequence, which then requires spatio-temporal feature extraction for the classification of the gesture instance. We consider four approaches, each is applied to the RGB and depth video independently. These are compared in Table III in terms of extraction time and dimensionality. In calculation of extraction time, we time feature extraction for each video, divide by the number of frames, and average over the videos in the dataset. Given a set of video frames, we choose a descriptor function, $\phi : \mathbb{R}^m \times \mathbb{R}^n \times \mathbb{R}^T \rightarrow \mathbb{R}^d$, for producing the d dimensional feature vector for gesture classification.

HOG: A straightforward temporal descriptor is produced by choosing a vectorization operator on the spatial descriptors in each frame, h_t, $t \in \{1,\dots,T\}$. In this case, the video is first resized to $T = 20$ frames by linear interpolation so that the descriptor is fixed in size.

$$\phi(I_1,\dots,I_T) = \big[h_1,\dots,h_T\big] \tag{4}$$

The pipeline for this algorithm contains three parameters, namely M, N, and B. We use $B = 8$ orientation bins in all of the experiments, and fix $M = N$, so that only one parameter can be varied, as shown in Fig. 6.

HOG^2: Another choice of ϕ is motivated by [15], [62]. In this case, the spatial descriptors are collected over time to form a 2D array (visualized in Fig. 2) of size $T \times (M \cdot N \cdot B)$. Changes in the feature vector correspond to changes in the shape and location of the hand. Consequently, the spatial HOG algorithm described in Section III-B is applied again using a $M^1 \times N^1$ grid of cells and B^1 angle quantization bins to extract a compact temporal descriptor of size $M^1 \cdot N^1 \cdot B^1$. The approach is termed HOG^2, since it involves applying the same algorithm twice (once in the spatial domain, and then again on

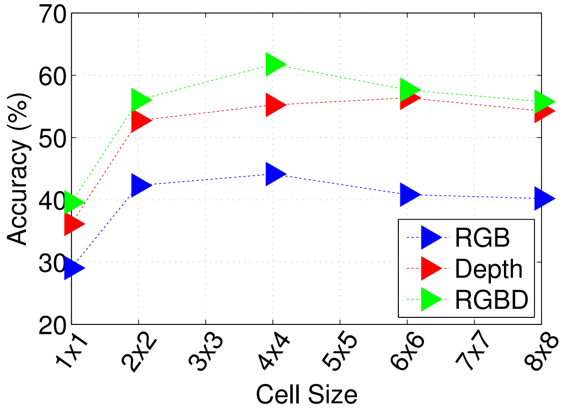


Fig. 6: Varying the cell size parameters in the HOG-based gesture recognition algorithm with a linear SVM for a RGB, depth, and RGB+Depth descriptors. A fixed 8 bin orientation histogram is used. Results are shown on the entire 19 gestures dataset using leave-one-subject-out cross-validation (cross-subject test settings).

those histograms over time). In this case, $\phi : \mathbb{R}^m \times \mathbb{R}^n \times \mathbb{R}^T \rightarrow \mathbb{R}^{M^1 \cdot N^1 \cdot B^1}$

$$\phi(I_1, \ldots, I_T) = \text{ HOG } (\begin{bmatrix} h_1 \\ \vdots \\ h_T \end{bmatrix}) \tag{5}$$

As in [15], we also use the mean of each of the spatial HOG features over time in the feature set. Generally, the dimensionality of HOG^2 is much lower than the corresponding temporal HOG concatenation. There are three additional parameters for the second operation of HOG, but we fix those to be the same as in the spatial HOG feature extraction so that $M^1 = M$, $N^1 = N$, $B^1 = B$.

HOG-PCA: Alternatively, we can reduce the dimensionality of the concatenated histograms descriptor (HOG) using Principal Component Analysis (PCA). In this case, we pre-compute the eigenspace using the training samples, and at test time project the temporal HOG concatenation feature using the eigenspace to derive a compact feature vector. Studying this operation is useful mainly for comparison with HOG^2.

HOG3D (Kläser *et al.* [16]): A spatio-temporal extension of HOG, where 3D gradient orientations are binned using convex regular polyhedrons in order to produce the final histogram descriptor. The operation is performed on a dense grid, and a codebook is produced using k-means. In our experiments, we optimize k over $k \in \{500, 1000, 2000, 3000, 4000\}$. k-means is run five times and the best results are reported.

DTM (Heng *et al.* [17]): The dense trajectories and motion boundary descriptor uses optical flow to extract dense trajectories, around which shape (HOG) and motion (histograms of optical flow) descriptors are extracted. Trajectory shape descriptors encode local motion patterns, and motion boundary histograms (MBH) are extracted along the x and y directions. Similarly to HOG3D, we follow the author original implementation with a dense sampling grid and a codebook produced by k-means.

TABLE III: Comparison of average extraction time per frame in milliseconds for each descriptor and for *one modality* - RGB or depth. Note that extracting RGBD cues from both modalities will require about twice the time. Experiments were done in C++ on a Linux 64-bit system with 8GB RAM and Intel Core i7 950 @ 3.07 GHz x 8. Asterisk * prefix - requires codebook construction.

Descriptor	Extraction Time (in ms)	Dimensionality
HOG	2.8	2560
HOG^2	2.83	256
HOG-PCA	3.25	256
DTM[17]	54	2000*
HOG3D[16]	372	1000*

We emphasize that in our implementation, only HOG3D and DTM require codebook construction with k-means. In these, a video sequence is represented as a bag of local spatio-temporal features. k-means is used to produce the codebook by which to quantize features, and each video is represented as a frequency histogram of the visual words (assignment to visual words is performed using the Euclidean distance). The other techniques involve a global descriptor computed over the entire image patch. Furthermore, we experimented with a range of descriptors, such as the Cuboids [53] and HON4D [52], but even after parameter optimization these did not show improvement over the aforementioned baselines.

D. *Classifier Choice*

SVM [18] is used in the experiments due to its popularity in the action recognition literature with varying types of descriptors [16], [17]. In SVM classification, a Mercer similarity or kernel function needs to be defined. We study the following three kernel choices. Given two data points, $\mathbf{x}_i, \mathbf{x}_j \in \mathbb{R}^d$, the linear kernel is given as,

$$K_{LIN}(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \mathbf{x}_j \tag{6}$$

an RBF-χ^2 kernel,

$$K_{\chi^2}(\mathbf{x}_i, \mathbf{x}_j) = exp(-\frac{1}{2C}\sum_{k=1:d} \frac{(x_{ik} - x_{jk})^2}{x_{ik} + x_{jk}}) \tag{7}$$

where C is the mean value of the χ^2 distances over the training samples, and a histogram intersection kernel (HIK),

$$K_{HI}(\mathbf{x}_i, \mathbf{x}_j) = \sum_{k=1:d} \min(x_{ik}, x_{jk}) \tag{8}$$

IV. Experimental Evaluation and Discussion

Spatio-temporal descriptor analysis: The descriptors mentioned in Section III were compared with the three kernels in Table IV. The results are shown for the entire 19 gesture dataset with leave-one-subject-out cross validation (cross-subject test settings). As shown in Fig. 6, a 4×4 cell size in computing the HOG-based descriptors was shown to work well. The temporal HOG descriptor shows best results across modalities and kernels. Although lower performing descriptors benefit significantly from the non-linear kernels, the benefits for HOG are small. Overall, the DTM and HOG3D baselines

TABLE IV: Comparison of gesture classification results using the different spatio-temporal feature extraction methods on the entire 19 gesture dataset in a leave-one-subject-out cross validation (cross-subject test settings). Average and standard deviation over the 8 folds are shown. In bold are the best results for each modality and for each kernel for the SVM classifier; Linear, RBF-χ^2, and histogram intersection kernel. The best result overall is prefixed by an asterisk.

	K_{LIN}	K_{χ^2}	K_{HI}
Descriptor \ Modality	RGB (%)		
DTM	41.3 ± 14.1	47.0 ± 12.3	47.7 ± 12.0
HOG3D	35.8 ± 9.5	39.1 ± 8.5	37.8 ± 6.4
HOG	44.1 ± 11.8	46.5 ± 15.9	47.3 ± 14.1
HOG-PCA	44.3 ± 8.8	38.0 ± 9.2	42.1 ± 11.6
HOG^2	33.1 ± 8.9	35.4 ± 9.1	34.9 ± 8.6
HOG+HOG-PCA	45.4 ± 12.7	47.2 ± 14.3	49.0 ± 14.1
HOG+HOG^2	$\mathbf{47.9 \pm 13.8}$	$\mathbf{50.8 \pm 17.2}$	$\mathbf{52.3 \pm 16.2}$
Descriptor \ Modality	Depth (%)		
DTM	37.1 ± 9.5	40.8 ± 9.9	43.2 ± 11.8
HOG3D	40.6 ± 7.8	43.0 ± 11.4	44.2 ± 8.6
HOG	55.2 ± 13.9	57.0 ± 17.0	57.4 ± 15.6
HOG-PCA	49.1 ± 11.9	48.7 ± 13.7	48.8 ± 13.4
HOG^2	46.9 ± 12.8	49.6 ± 14.4	49.0 ± 14.7
HOG+HOG-PCA	55.9 ± 13.6	57.1 ± 16.7	57.8 ± 15.7
HOG+HOG^2	$\mathbf{57.5 \pm 14.6}$	$\mathbf{57.6 \pm 17.9}$	$\mathbf{58.6 \pm 15.8}$
Descriptor \ Modality	RGB+Depth (%)		
DTM	47.8 ± 13.2	51.5 ± 15.3	54.0 ± 14.8
HOG3D	36.7 ± 8.5	41.3 ± 9.0	44.6 ± 9.7
HOG	61.8 ± 15.7	62.1 ± 15.5	62.2 ± 16.8
HOG-PCA	56.2 ± 12.4	56.5 ± 13.7	57.3 ± 13.2
HOG^2	49.6 ± 14.6	52.3 ± 13.5	52.3 ± 14.5
HOG+HOG-PCA	62.2 ± 15.8	62.5 ± 16.0	62.1 ± 16.1
HOG+HOG^2	$\mathbf{63.3 \pm 15.3}$	$*\mathbf{64.5 \pm 16.9}$	$\mathbf{63.1 \pm 16.7}$

are outperformed by the rest, possibly since these are densely sampled over the ROI yet background information does not contain useful information for the recognition (unlike other action recognition scenarios).

Inspecting the different HOG descriptors studied in this work, we observe that although the HOG^2 shows comparable results to DTM and HOG3D, it is outperformed by the HOG scheme. Interestingly, it appears to contain complementary information to the HOG scheme when combined, more so than when using the HOG-PCA scheme (although the two descriptors have the same dimensionality). This is the main reason for which HOG-PCA was studied in this work, and not for improving the results over HOG. Because HIK SVM with the HOG+HOG^2 descriptor showed good results, it is used in the remaining experiments.

Evaluation on gesture subsets: As mentioned in Section III-A, a 19 gesture dataset may not be suitable for the application of an automotive interface. A set of three subsets was chosen and experiments were done using three testing methods, with results shown in Table V. The three test settings are as follows: *1/3-Subject*: a 3-fold cross validation where each time a third of the samples from each subject are reserved for training and the rest for testing. *2/3-Subject*: Similarly to 1/3-Subject, but two thirds of the samples are reserved for training from each subject and the remaining third for testing. *Cross-subject*: leave-one-subject-out cross validation. Results are done over 8 subjects and averaged.

The purpose of such a study is mostly in evaluating the generalization of the proposed algorithm, as well as the effect of user-specific training. The confusion matrix for each gesture subset using 2/3-Subject test settings are shown in Fig. 8. Table V reveals a lower accuracy on the challenging cross-subject testing, as expected. The reason is that within the 8 subjects there were large variations in the execution of each gesture.

Basic interface with a mode switch: Equipped with insight on the least ambiguous gestures from Fig. 8, we study a final gesture subset (Fig. 7) that provides a basic gesture interaction at high recognition accuracy (shown in Table VI). We propose to use one of the gestures, such as a one tap with three fingers (OneTap3) in order to navigate among functionality modes while keeping the same gesture set.

V. Concluding Remarks

In this paper, we studied the feasibility of an in-vehicle, vision-based gesture recognition system. Although our work is concerned with gesture recognition in naturalistic settings and not the psychological aspects of the interface, our experimental design attempted to accompany other successful existing gestural interaction interfaces. Following a trend in other devices where multi-modal interfaces opened ways to new functionality and efficiency, with additional comfort for some users, we sought a similar solution to the in-vehicle interface. As each interaction modality has its strengths and limitations, we believe a multi-modal interface should be pursued for leveraging advantages from each modality and allowing customization to the user. Each should be designed and studied carefully in order to avoid a high mental workload in remembering or performing gestures, provide appropriate feedback, and maximize intuitiveness.

In an attempt to propose a complete system, first a hand detection and user determination step was used, followed

TABLE V: Recognition accuracy and standard deviation over cross-validation using different evaluation methods discussed in Section IV. Increasing the number of user-specific samples results in improved recognition. RGB+Depth is the two descriptors concatenated and a HIK SVM. The overall category is the mean over the column for each modality, for showing the best modality settings and the effects of the test settings.

	1/3-Subject	2/3-Subject	Cross-Subject
	RGB (%)		
GS1	95.0 ± 1.1	96.5 ± 2.7	75.5 ± 16.7
GS2	91.0 ± 1.7	94.7 ± 1.3	63.8 ± 16.6
GS3	91.4 ± 2.0	94.6 ± 1.7	56.2 ± 14.7
Overall	92.5 ± 1.6	95.3 ± 1.9	65.2 ± 16.0

	Depth (%)		
GS1	92.7 ± 0.3	94.1 ± 1.6	80.9 ± 12.4
GS2	90.5 ± 1.5	93.6 ± 1.9	72.6 ± 19.4
GS3	87.0 ± 2.1	90.3 ± 2.1	67.3 ± 16.0
Overall	90.1 ± 1.3	92.3 ± 1.9	73.6 ± 15.9

	RGB+Depth (%)		
GS1	95.6 ± 1.1	96.5 ± 1.6	82.4 ± 15.1
GS2	92.9 ± 1.8	96.1 ± 1.2	73.8 ± 13.7
GS3	93.2 ± 1.9	96.0 ± 2.2	72.0 ± 15.6
Overall	**93.9 ± 1.6**	**96.2 ± 1.7**	**76.1 ± 14.8**

by a real-time spatio-temporal descriptor and gesture classification scheme. Out of a set of 19 gestures, four subsets were constructed for different interactivity applications. A careful evaluation of different temporal descriptors showed the challenging nature of the dataset, with RGBD fusion proving to be beneficial for recognition.

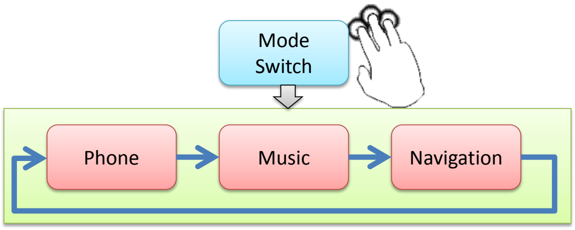
Future extensions should further analyze the role of each of the spatio-temporal descriptors in increasing illumination-, occlusion-, and subject-invariance of the system. Temporal segmentation of gestures without requiring the hand to leave the ROI may result in a more comfortable interface to use. The studied RGBD feature set might be useful for studying naturalistic hand gestures [8], [9]. The location of the ROI can be studied in order to determine optimal natural interactivity and the gesture subset can be further refined and evaluated. Since incorporating samples of a subject in training resulted in a significantly higher recognition performance for that subject in testing, online learning could further improve classification rates. Finally, we hope that the evaluation in this work and the public dataset will inspire the development of new and improved hand gesture recognition techniques.

VI. Acknowledgments

This research was performed at the UCSD Computer Vision and Robotics Research and LISA: Laboratory for Intelligent and Safe Automobiles. The authors would like to thank Dr. Cuong Tran and Mr. Ashish Tawari for their helpful advice. We thank our colleagues for their help in the data collection process, and the reviewers and editor for their helpful comments.

References

[1] J. M. Gregers, M. B. Skov, and N. G. Thomassen, "You can touch, but you can't look: interacting with in-vehicle systems," in *SIGCHI Conf. Human Factors in Computing Systems*, 2008.



Gesture Set 4 (GS4)
OneTap3
SwipeV
ScrollUp2
ScrollDown2
ScrollRight2
ScrollLeft2

Fig. 7: Equipped with the analysis of the previously proposed gesture subsets, a final gesture set composed of less ambiguous gestures is defined and studied. The subset is designed for basic interaction, with one of the gestures used to switch between different functionality modes.

TABLE VI: Recognition accuracy using RGB+Depth and a HIK SVM on Gesture Set 4.

	1/3-Subject	2/3-Subject	Cross-Subject
	RGB+Depth (%)		
GS4	98.4 ± 0.5	99.7 ± 0.6	92.8 ± 8.8

[2] M. Alpern and K. Minardo, "Developing a car gesture interface for use as a secondary task," in *CHI Human factors in computing systems*, 2003.

[3] F. Parada-Loira, E. González-Agulla, and J. L. Alba-Castro, "Hand gestures to control infotainment equipment in cars," in *IEEE Intell. Veh. Symp.*, 2014.

[4] C. A. Pickering, K. J. Burnham, and M. J. Richardson, "A research study of hand gesture recognition technologies and applications for human vehicle interaction," in *Institution of Engineering and Technology Conference on Automative Electronics*, 2007.

[5] W. J. Horrey, "Assessing the effects of in-vehicle tasks on driving performance," *Ergonomics in Design*, no. 19, pp. 4–7, 2011.

[6] G. Jahn, J. F. Krems, and C. Gelau, "Skill aquisition while operating in-vehicle information systems: Interface design determines the level of safety-relevant distractions," *Human Factors and Ergonomics Society*, no. 51, pp. 136–151, 2009.

[7] 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.

[8] E. Ohn-Bar, A. Tawari, S. Martin, and M. M. Trivedi, "Predicting driver maneuvers by learning holistic features," in *IEEE Intell. Veh. Symp.*, 2014.

[9] E. Ohn-Bar, S. Martin, and M. M. Trivedi, "Driver hand activity analysis in naturalistic driving studies: challenges, algorithms, and experimental studies," *Journal of Electronic Imaging*, vol. 22, no. 4, 2013.

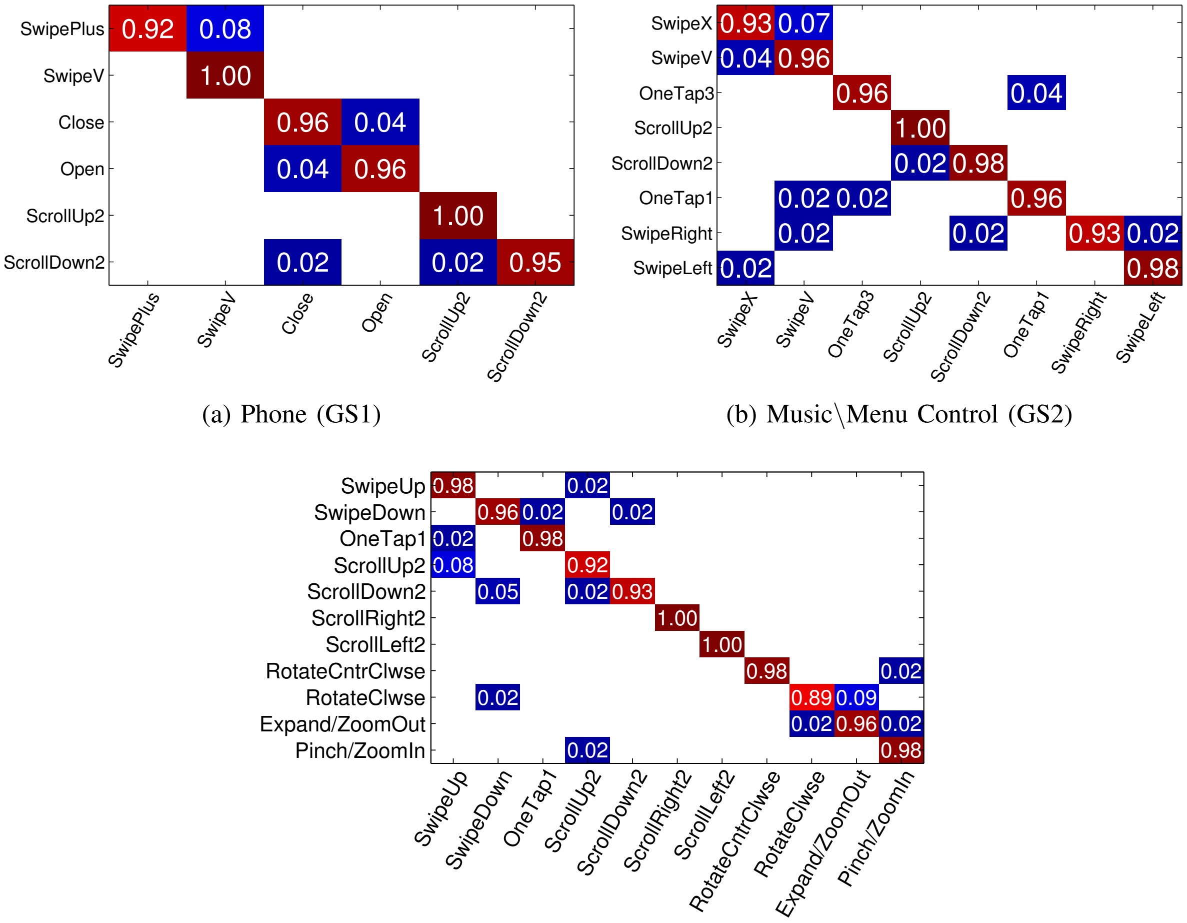
[10] C. Tran and M. M. Trivedi, "Vision for driver assistance: Looking at people in a vehicle," in *Visual Analysis of Humans*, 2011, pp. 597–614.

[11] 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.

[12] E. Ohn-Bar, A. Tawari, S. Martin, and M. M. Trivedi, "Vision on wheels: Looking at driver, vehicle, and surround for on-road maneuver analysis," in *IEEE Conf. Computer Vision and Pattern Recognition Workshops-Mobile Vision*, 2014.

[13] I. Oikonomidis, N. Kyriazis, and A. A. Argyros, "Efficient model-based 3D tracking of hand articulations using Kinect," in *British Machine Vision Conf.*, 2011.

[14] C. Qian, X. Sun, Y. Wei, X. Tang, and J. Sun, "Realtime and robust hand tracking from depth," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2014.



(a) Phone (GS1)

(b) Music\Menu Control (GS2)

(c) Picture\Navigation (GS3)

Fig. 8: Results for the three gesture subsets for different in-vehicle applications using 2/3-Subject test settings, where 2/3 of the samples are used for training and the rest for testing in a 3-fold cross validation. A RGB+Depth combined descriptor was used. Average correct classification rates are shown in Table V.

[15] E. Ohn-Bar and M. M. Trivedi, "Joint angles similarities and HOG^2 for action recognition," in *IEEE Conf. Computer Vision and Pattern Recognition Workshops-Human Activity Understanding from 3D Data*, 2013.

[16] A. Kläser, M. Marszalek, and C. Schmid, "A spatio-temporal descriptor based on 3D-gradients," in *British Machine Vision Conf.*, 2008.

[17] W. Heng, A. Kläser, C. Schmid, and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *Intl. Journal of Computer Vision*, vol. 103, no. 1, pp. 60–79, May. 2013.

[18] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Intell. Syst. and Tech.*, vol. 2, pp. 27:1–27:27, 2011.

[19] M. B. Holte, C. Tran, M. M. Trivedi, and T. B. Moeslund, "Human pose estimation and activity recognition from multi-view videos: Comparative explorations of recent developments," *IEEE J. Sel. Topics Signal Process.*, vol. 6, no. 5, pp. 538–552, 2012.

[20] R. Poppe, "A survey on vision-based human action recognition," *Image and Vision Computing*, vol. 28, no. 6, pp. 976–990, 2010.

[21] S. Sathyanarayana, G. Littlewort, and M. Bartlett, "Hand gestures for intelligent tutoring systems: Dataset, techniques evaluation," in *IEEE Intl. Conf. Computer Vision Workshops*, 2013.

[22] S. Hadfield and R. Bowden, "Hollywood 3D: Recognizing actions in 3D natural scenes," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2013.

[23] Y. Zhu, W. Chen, and G. Guo, "Evaluating spatiotemporal interest point features for depth-based action recognition," *Image and Vision Computing*, vol. 32, no. 8, pp. 453–464, 2014.

[24] N. Neverova, C. Wolf, G. Paci, G. Sommavilla, G. W. Taylor, and F. Nebout, "A multi-scale approach to gesture detection and recognition," in *IEEE Intl. Conf. Computer Vision Workshops*, 2013.

[25] A. Erol, G. Bebis, M. Nicolescu, R. D. Boyle, and X. Twombly, "Vision-based hand pose estimation: A review," *Computer Vision and Image Understanding*, vol. 108, no. 12, pp. 52–73, 2007.

[26] C. Keskin, F. Kirac, Y. Kara, and L. Akarun, "Real time hand pose estimation using depth sensors," in *IEEE Intl. Conf. Computer Vision Workshops*, 2011.

[27] G. Evangelidis, G. Singh, and R. Horaud, "Skeletal quads: Human action recognition using joint quadruples," in *IEEE Intl. Conf. Pattern Recognition*, 2014.

[28] R. Vemulapalli, F. Arrate, and R. Chellappa, "Human action recognition by representing 3d human skeletons as points in a lie group," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2014.

[29] J. Wang, Z. Liu, and Y. Wu, "Learning actionlet ensemble for 3d human action recognition," in *Human Action Recognition with Depth Cameras*, ser. SpringerBriefs in Computer Science. Springer International Publishing, 2014, pp. 11–40.

[30] I. Kapsouras and N. Nikolaidis, "Action recognition on motion capture data using a dynemes and forward differences representation," *Journal of Visual Communication and Image Representation*, 2014.

[31] N. Pugeault and R. Bowden, "Spelling it out: Real-time ASL fingerspelling recognition," in *IEEE Intl. Conf. Computer Vision Workshops*, 2011.

[32] C. Helen, H. Brian, and B. Richard, "Sign language recognition," in *Visual Analysis of Humans*, 2011, pp. 539–562.

[33] D. Uebersax, J. Gall, M. V. den Bergh, and L. V. Gool, "Real-time sign language letter and word recognition from depth data," in *IEEE Intl. Conf. Computer Vision Workshops*, 2011.

[34] P. Doliotis, A. Stefan, C. McMurrough, D. Eckhard, and V. Athitsos, "Comparing gesture recognition accuracy using color and depth in-

formation,” in *ACM Conf. Pervasive Technologies Related to Assistive Environments*, 2011.

[35] E. Ohn-Bar, S. Martin, A. Tawari, and M. M. Trivedi, “Head, eye, and hand patterns for driver activity recognition,” in *IEEE Intl. Conf. Pattern Recognition*, 2014.

[36] E. Ohn-Bar and M. M. Trivedi, “The power is in your hands: 3D analysis of hand gestures in naturalistic video,” in *IEEE Conf. Computer Vision and Pattern Recognition Workshops-Analysis and Modeling of Faces and Gestures*, 2013 (best paper award).

[37] R.-D. Vatavu, “User-defined gestures for free-hand tv control,” in *European conference on Interactive TV and Video*, 2012.

[38] G. Panger, “Kinect in the kitchen: testing depth camera interactions in practical home environments,” in *ACM Human Factors in Computing Systems*, 2012.

[39] J. P. Wachs, M. Kölsch, H. Stern, and Y. Edan, “Vision-based hand-gesture applications,” *ACM Commun.*, 2011.

[40] C. Kirmizibayrak, N. Radeva, M. Wakid, J. Philbeck, J. Sibert, and J. Hahn, “Evaluation of gesture based interfaces for medical volume visualization tasks,” in *Virtual Reality Continuum and Its Applications in Industry*, 2011.

[41] L. Gallo, A. Placitelli, and M. Ciampi, “Controller-free exploration of medical image data: Experiencing the Kinect,” in *Computer-Based Medical Systems*, 2011.

[42] E. Saba, E. Larson, and S. Patel, “Dante vision: In-air and touch gesture sensing for natural surface interaction with combined depth and thermal cameras,” in *IEEE Conf. Emerging Signal Process. Applicat.*, 2012.

[43] C.-Y. Kao and C.-S. Fahn, “A human-machine interaction technique: Hand gesture recognition based on hidden markov models with trajectory of hand motion,” *Procedia Engineering*, vol. 15, pp. 3739–3743, 2011.

[44] E. Ozcelik and G. Sengul, “Gesture-based interaction for learning: time to make the dream a reality,” *British Journal of Educational Technology*, vol. 43, no. 3, pp. 86–89, 2012.

[45] Z. Ren, J. Meng, and J. Yuan, “Depth camera based hand gesture recognition and its applications in human-computer-interaction,” in *IEEE Conf. Inform., Commun. and Signal Process.*, 2011.

[46] J. M. Teixeira, B. Reis, S. Macedo, and J. Kelner, “Open/closed hand classification using Kinect data,” in *IEEE Symp. on Virtual and Augmented Reality*, 2012.

[47] M. V. den Bergh, D. Carton, R. D. Nijs, N. Mitsou, C. Landsiedel, K. Kuehnlenz, D. Wollherr, L. V. Gool, and M. Buss, “Real-time 3D hand gesture interaction with a robot for understanding directions from humans,” in *IEEE RO-MAN*, Aug. 2011.

[48] G. Bailly, R. Walter, J. Mller, T. Ning, and E. Lecolinet, “Comparing free hand menu techniques for distant displays using linear, marking and finger-count menus,” in *Human-Computer Interaction-INTERACT*, 2011, vol. 6947, pp. 248–262.

[49] C. Keskin, A. Cemgil, and L. Akarun, “DTW based clustering to improve hand gesture recognition,” in *Human Behavior Unterstanding*, 2011, vol. 7065, pp. 72–81.

[50] D. Minnen and Z. Zafrulla, “Towards robust cross-user hand tracking and shape recognition,” in *IEEE Intl. Conf. Computer Vision Workshops*, 2011.

[51] O. Oreifej and Z. Liu, “HON4D: Histogram of oriented 4D normals for activity recognition from depth sequences,” in *IEEE Conf. Computer Vision and Pattern Recognition*, 2013.

[52] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, “Behavior recognition via sparse spatio-temporal features,” in *IEEE Intl. Conf. Computer Vision Workshops*, 2005.

[53] L. Xia and J. Aggarwal, “Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera,” in *IEEE Conf. Computer Vision and Pattern Recognition*, 2013.

[54] M. Zobl, R. Nieschulz, M. Geiger, M. Lang, and G. Rigoll, “Gesture components for natural interaction with in-car devices,” in *Gesture-Based Communication in Human-Computer Interaction*, 2004, vol. 2915, pp. 367–368.

[55] A. Riener, A. Ferscha, F. Bachmair, P. Hagmüller, A. Lemme, D. Muttenthaler, D. Pühringer, H. Rogner, A. Tappe, and F. Weger, “Standardization of the in-car gesture interaction space,” in *ACM Automotive User Interfaces and Interactive Vehicular Applications*, 2013.

[56] F. Althoff, R. Lindl, and L. Walchshaeusl, “Robust multimodal hand and head gesture recognition for controlling automotive infotainment systems,” in *VDI-Tagung: Der Fahrer im 21*, 2005.

[57] C. Endres, T. Schwartz, and C. A. Müller, “Geremin”: 2D microgestures for drivers based on electric field sensing,” in *ACM Conf. Intell. User Interfaces*, 2011.

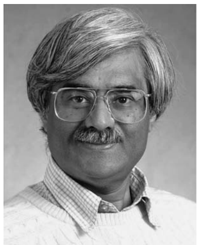
[58] S. Y. Cheng and M. M. Trivedi, “Vision-based infotainment user determination by hand recognition for driver assistance,” *IEEE Trans. Intell. Transp. Syst.*, vol. 11, no. 3, pp. 759–764, Sep. 2010.

[59] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *IEEE Conf. Computer Vision and Pattern Recognition*, 2005.

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

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

Eshed Ohn-Bar (S′14) received his M.S. degree in electrical engineering in 2013 from the University of California, San Diego (UCSD). He is currently working towards a Ph.D. degree with specialization in signal and image processing. His research interests include computer vision, intelligent vehicles, and driver assistance systems.



Mohan Manubhai Trivedi (S′14) received the B.E. (with honors) degree in electronics from Birla Institute of Technology and Science, Pilani, India, in 1974 and the M.S. and Ph.D. degrees in electrical engineering from Utah State University, Logan, UT, USA, in 1976 and 1979, respectively. He is currently a Professor of electrical and computer engineering and he is the Founding Director of the Computer Vision and Robotics Research Laboratory, University of California San Diego (UCSD), La Jolla, CA, USA. He has also established the Laboratory for Intelligent and Safe Automobiles, Computer Vision and Robotics Research Laboratory, UCSD, where he and his team are currently pursuing research in machine and human perception, machine learning, human-centered multimodal interfaces, intelligent transportation, driver assistance, active safety systems and Naturalistic Driving Study (NDS) analytics. His team has played key roles in several major research initiatives. These include developing an autonomous robotic team for Shinkansen tracks, a human-centered vehicle collision avoidance system, a vision-based passenger protection system for smart airbag deployment, and lane/turn/merge intent prediction modules for advanced driver assistance. He regularly serves as a Consultant to industry and government agencies in the United States, Europe, and Asia. He has given over 70 Keynote/Plenary talks at major conferences. Prof. Trivedi is a Fellow of the International Association of Pattern Recognition (for contributions to vision systems for situational awareness and human-centered vehicle safety) and the Society for Optical Engineering (for contributions to the field of optical engineering). He received the IEEE Intelligent Transportation Systems Societys highest honor, Outstanding Research Award in 2013, the Pioneer Award (Technical Activities) and the Meritorious Service Award of the IEEE Computer Society, and the Distinguished Alumni Award from Utah State University, Logan, UT, USA. He is a co-author of a number of papers winning Best Papers awards. Two of his students were awarded Best Dissertation Awards by the IEEE ITS Society (Dr. Shinko Cheng 2008 and Dr. Brendan Morris 2010) and his advisee Dr. Anup Doshis dissertation judged among the five finalists in the 2011 by the Western (USA and Canada) Association of Graduate Schools. He serves on the Board of Governors of IEEE ITS Society and on the Editorial advisory board of the IEEE Trans on Intelligent Transportation Systems.