Towards Selecting Robust Hand Gestures for Automotive Interfaces

Shalini Gupta, Pavlo Molchanov, Xiaodong Yang, Kihwan Kim, Stephen Tyree and Jan Kautz

***Abstract*— Driver distraction is a serious threat to automotive safety. The visual-manual interfaces in cars are a source of distraction for drivers. Automotive touch-less hand gesture-based user interfaces can help to reduce driver distraction and enhance safety and comfort. The choice of hand gestures in automotive interfaces is central to their success and widespread adoption. In this work we evaluate the recognition accuracy of 25 different gestures for state-of-the-art computer vision-based gesture recognition algorithms and for human observers. We show that some gestures are consistently recognized more accurately than others by both vision-based algorithms and humans. We further identify similarities in the hand gesture recognition abilities of vision-based systems and humans. Lastly, by merging pairs of gestures with high miss-classification rates, we propose ten robust hand gestures for automotive interfaces, which are classified with high and equal accuracy by vision-based algorithms.**

I. INTRODUCTION

Distracted driving has been identified as a serious threat to road safety in many countries around the world [1]. In the United States, driver distraction was involved in 10% of all police-reported crashes in 2013 and resulted in injuries to 424,000 people [2]. Distraction is defined as the diversion of the driver's attention away from the primary task of driving towards a competing activity [3]. It results in the impairment of the driver's situational awareness, decision making and driving performance. There are various forms of distraction including visual, cognitive, physical and auditory. Visual-manual interfaces, *e.g.*, haptic controls and touch screens in cars or cell phones, are a significant source of driver distraction [4].

The research on visual-manual distraction in cars establishes a link between visual attention (*e.g.* eyes off the road) and crash risk [5]. Hand gesture-based user interfaces (UIs) in cars allow drivers to keep their eyes on the road while performing secondary tasks. These UIs can improve safety and comfort in cars especially when coupled with appropriate mechanisms to provide feedback. Additionally, hand gestures performed with hands on/close to the steering wheel can result in low physical distraction. Unlike voice-based interfaces, gesture UIs are less likely to be affected by ambient sounds in vehicles and cause less auditory distraction. Vision-based gesture recognition systems can also be extended to incorporate functionality for driver monitoring.

A recent evaluation of human subjects found that, in cars, users perceived gesture UIs to be more secure, less distracting and slightly more useful than touch screens [6]. Users also reported gesture interfaces to be more desirable and worth buying, but less reliable than touch screens. The choice of hand gestures in automotive UIs is an important design consideration which can greatly influence their success and widespread adoption. Research on automotive gesture UIs has gained significant traction in recent years, but has primarily focused on algorithms for hand gesture recognition in cars [6], [7], [8], [9], [10], [11], [12], [13]. Some work to standardize vehicular gesture interfaces is also underway [14], but little attention has been paid to selecting the best set of gestures for automotive UIs.

The ideal choice of hand gestures for automotive UIs should jointly optimize for safety, ease of use, robustness, and the intended application [15]. In this work we primarily focus on robustness, which in turn depends on the accuracy of automatic gesture recognition systems for each class of hand gesture. As a step towards identifying the most robust hand gestures, we evaluate the recognition accuracy of 25 different hand gestures intended for use in automotive gestural interfaces. We evaluate the recognition accuracy of 5 different state-of-the-art gesture recognition algorithms and 6 human observers. Since the most successful algorithms for hand gesture recognition in cars are based on computer vision technology that uses RGB and/or depth (D) data as input, we consider only vision-based gesture recognition algorithms.

We show that some gestures are consistently recognized more accurately than others by both vision-based systems and humans. This implies the presence of inherent variability in the recognizability of different types of hand gestures. We further identify similarities in how well vision-based systems and humans observers are able to recognize the various kinds of gestures. Lastly, by merging the pairs of gestures with the highest miss-classification rates, we propose ten robust hand gestures for automotive interfaces, which are classified with high and equal accuracy by vision-based algorithms.

The paper is organized as follows. In Section II, we summarize the existing work on vision-based hand gesture recognition algorithms. In Section III we describe our experimental methodology including the dataset and the vision-based algorithms that we use as well as the human study and statistical analysis that we conduct to evaluate the recognition accuracies of the various gesture types. Section IV enumerates the results of our experiments. We conclude in Section V with a summary of our main findings and identify directions for future research.

NVIDIA, 2701 San Tomas Expy, Santa Clara, CA 95050, USA {shalinig, pmolchanov, xiaodongy, kihwank, styree, jkautz}@nvidia.com

II. RELATED WORK

A. Gesture Recognition Systems

Hand gesture recognition using computer-vision techniques in the uncontrolled lighting conditions encountered in a car is challenging. RGB-based techniques for gesture recognition in cars that use special infrared (IR) illuminators and near-IR cameras have been proposed [7], [8], [6]. These previous methods use hand-crafted features, including Hu moments [7], decision rules [8], or contour shape features [6] along with HMM classifiers [7], [8]. Ohn-Bar and Trivedi use RGBD data, histogram of gradients (HOG) features and a support vector machine (SVM) classifier [9], whereas Kopinski et al. [13] extract point feature histograms from depth images, and classify them with a multi-layer perceptron (MLP).

Besides hand-crafted features, recent works learn feature representations from deep neutral networks for gesture and action classification. Neverova et al. [10] employ convolutional neural networks to combine color and depth inputs from hand regions with upper-body skeletons to recognize sign language gestures. Molchanov et al. [11], [12] apply three-dimensional (3D) convolutional neural networks (CNN) to segmented gesture video sequences along with space and time video augmentation techniques to avoid overfitting. They report state-of-the-art performance, in both accuracy and speed, on a benchmark in-car hand gesture recognition dataset [9]. Independently of vision-based techniques, hand recognition systems that use micro-Doppler signatures of electromagnetic signals have also been developed [11], [16], [17], [18], but they result in lower accuracy.

Techniques that were originally geared towards video classification and action recognition can also be applied to hand gesture recognition. Improved dense trajectories (iDT) [19] rely on image gradients and optical flow, use hand-crafted descriptors based on appearance and motion cues, and are considered state-of-the-art for action classification. Super normal vectors (SNV) [20] perform action recognition on depth videos using a novel scheme for aggregating normals, yielding high accuracy. The two-stream CNN architecture [21] uses two separate CNNs for spatial (video frame) and temporal (optical flow) streams that are combined through late fusion. Finally, the convolutional 3D (C3D) method [22] performs classification by analyzing individual short video clips and classifies their averaged responses via an SVM classifier, which leads to state-of-the-art accuracy.

B. Gesture Types

Despite the many systems that have been proposed for in-car gesture recognition, no standard has emerged for which types of gestures to use. Most published work focuses on recognition with little attention to choosing a set of gestures to improve recognition accuracy. The early work by Zobl et al. [23] reports that there is a "limited gesture vocabulary with a high inter- and intra-individual conformity for a variety of applications", but unfortunately does not specify that set of gestures. The more recent work by Riener et al. [14] finds that, when subjects are allowed to freely choose their own set of gestures, there is a low variability within a subjects' gestures (subjects consistently use the same gesture for the same functionality), but variability is high between subjects (subjects use different gestures for the same functionality), which is perhaps not surprising. We set out to answer whether particular gestures are more amenable to automatic recognition. While this is related to the aforementioned studies, our work looks at gesture recognition from the point of view of a system designer.

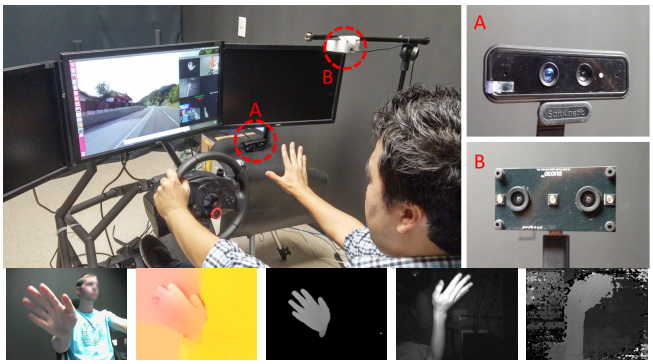

Fig. 1. (Top left) The driving simulator for data collection with the main monitor displaying simulated driving scenes and a user interface for prompting gestures, (A) a SoftKinetic depth camera (DS325) for recording depth and RGB data, and (B) a DUO 3D camera capturing stereo IR data. Both sensors capture 320×240 pixels at 30 frames per second. (Bottom row) Examples of each recorded modality, from left: RGB, optical flow, depth, IR-left, and IR-disparity.

III. METHOD

A. Dataset

We employ a dataset containing 1532 videos of hand gesture from 25 different gesture classes [24]. The data were captured indoors with a car simulator under both bright and dim artificial lighting environments, as shown in Fig. 1. It includes gestures performed by 20 different subjects. Each subject used their right hand to perform the gestures while controlling the steering wheel with their left hand. During a recording session, each subject repeated each of the 25 different gestures three times in a random order. The simulator had a user interface that prompted subjects to perform particular hand gestures with an audio instruction and a sample video clip of the gesture to perform. Despite the uniformity of gesture prompts, we observe natural inter- and intra-subject variability in gesture performance. This natural variability is preserved in the data set.

The gesture classes include moving the hand left, right, up or down; moving two fingers left, right, up or down; "clicking" with the index finger; calling someone (beckoning with the hand); opening or shaking the hand; showing the index finger, two fingers or three fingers; pushing the hand up, down, out or in; rotating two fingers clockwise (CW) or counter-clockwise (CCW); pushing forward with two fingers; closing the hand twice; and showing the symbols for "thumb up" or "OK". (See the examples shown in Fig. 2.) These gestures were partly adopted from existing commercial systems [25] and from popular automotive gesture recognition

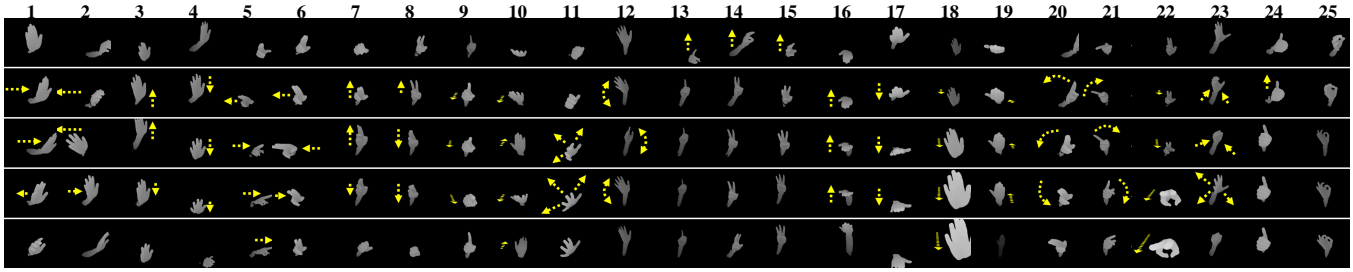

Fig. 2. Each column shows a different gesture class ($1-25$). The gestures included (from left to right): moving the hand left, right, up, or down; moving two fingers left, right, up, or down; clicking with the index finger; calling someone (beckoning with the hand); opening and shaking the hand; showing the index finger, two fingers or three fingers; pushing the hand up, down, out or in; rotating two fingers clockwise or counter-clockwise; pushing forward with two fingers; closing the hand twice; and showing "thumb up" or "OK". The top and bottom rows show the starting and ending depth frames, respectively, of the "nucleus" phase for each gesture, and yellow arrows indicate the motion performed in intermediate frames.

research datasets [9], [11]. The intended application of these gestures is, *e.g.*, to manipulate the visual-manual interfaces in cars including those for controlling the radio, fan, thermostat, and navigation systems. For example a gesture involving a horizontal hand/finger motion to the left/right could be used to move the music player to a previous/next audio track, respectively.

The data were collected with RGBD (Fig. 1A) and IR (Fig. 1B) sensors that each covered different views of the hand. A SoftKinetic DS325 sensor acquired front-view RGBD videos and a top-mounted DUO 3D sensor recorded a pair of stereo-IR streams. In addition, dense optical flow [26] was computed from the color stream and the IR disparity map was computed from the IR-stereo pair [27]. The data were randomly split by subject into training (70%) and testing (30%) sets, resulting in 1050 training and 482 test gesture videos.

B. Vision-based Algorithms

In previous work [12], we proposed a CNN-based algorithm for hand gesture recognition, which outperformed all other algorithms on the VIVA 2015 challenge's in-car hand gesture recognition dataset [9]. We recently extended the CNN-based multi-modal algorithm to use a recurrent three-dimensional convolutional neural network (R3DCNN), which led to a significant improvement in accuracy over the CNN-based method [24]. In the same work, we employed the connectionist temporal classification cost function [28] for training the R3DCNN classifier to support early detection in unsegmented video streams with zero or negative lag.

We compared the performance of the proposed R3DCNN algorithm to several state-of-the-art gesture and video action recognition methods: HOG+HOG^2 [9], improved dense trajectories (iDT) [19], super normal vector (SNV) [20], two-stream CNNs [21], and convolutional 3D (C3D) [22]; and to the performance of human subjects. For the details of how we implemented these algorithms, we refer interested readers to Molchanov et al. [24].

We observed that the multi-modal version of our R3DCNN algorithm with depth, color, optical flow, IR disparity and IR image data outperformed all vision-based algorithms with an accuracy of 83.8%. The other top performing algorithms, in order to decreasing accuracy, were: C3D with depth (accuracy of 78.8%), iDT with MBH (76.8%), iDT with color and optical flow (73.4%) and SNV with depth (70.7%) [24].

These top-performing algorithms cover a broad array of successful gesture recognition techniques ranging from those that use hand-crafted features to those with purely data-driven approaches for feature extraction. They also represent many different pattern classification techniques. These top-performing algorithms are more likely to be representative of the true performance of vision-based algorithms, whereas the poorly-performing methods are likely to yield noisy, uncorrelated classification results. Hence, to identify the set of gestures that are classified most accurately by vision-based algorithms, we used these five top-performing vision-based gesture recognition algorithms for all further analysis.

C. Annotation by Humans

We also conducted a subjective user study to evaluate the performance of human subjects at recognizing dynamic hand gestures in video clips [24]. We designed a custom graphical user interface for the task (Fig. 3), wherein we displayed the RGB video clips of gestures acquired with the front-view SoftKinetic camera (Fig. 1A). Six human subjects manually labeled, in one continuous session, all 482 gesture videos from the test partition of our dataset. We re-scaled all gestures temporally to 80 frames and presented them at 30 frames per second. Clips were presented in a different random order to each subject.

Before the experiment, all annotators were required to familiarize themselves with the 25 gesture classes by reviewing example video clips of each. While labeling the gestures, the annotators viewed each gesture video clip only once. Additionally, a black frame was presented in the gesture viewing window after the clip had played and while the annotators decided on its label. This allowed the subject to view the gesture for only its actual duration and not beyond, and prevented the use of any identifying information that may be contained in the last frame of some gestures (*e.g.*, "thumb up"). Newly presented gestures were marked as belonging to the "none" category so as not to bias the annotator's decision, but subjects had to assign one label from the 25 gestures classes to every gesture before proceeding. There was no time limit on how long the subjects could take to select a label. Additionally, the annotators could review an example video clip of any gesture type at any time during the labeling session.

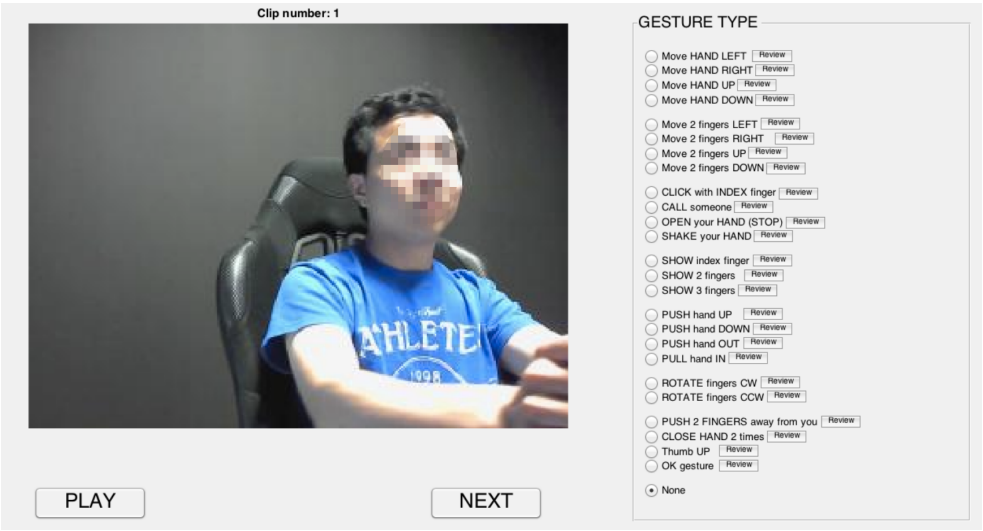

Fig. 3. The graphical user interface used by human subjects to label video clips of hand gestures.

D. Recognition Rates of Gestures

We wish to ascertain if among the set of 25 gestures there are some that are consistently recognized more accurately than others by multiple top-performing vision-based gesture recognition algorithms. In order to evaluate this hypothesis, we compute the confusion matrices for each of the top five vision based classifiers: R3DCNN, C3D, iDT with MBH, iDT with color and optical flow, and SNV for the 482 gestures in the test partition of our dataset. From the confusion matrices, we further compute the $sensitivity_{i,j}$, $precision_{i,j}$,, and the $F1_{i,j}$ scores for each of the $n = 5$ vision-based classifiers and for each of the $k = 25$ gesture types. We employ the standard definitions for these metrics as:

$$sensitivity_{i,j} = \frac{TP_{i,j}}{TP_{i,j} + FN_{i,j}}, \tag{1}$$

$$precision_{i,j} = \frac{TP_{i,j}}{TP_{i,j} + FP_{i,j}}, \tag{2}$$

$$F1_{i,j} = \frac{2TP_{i,j}}{2TP_{i,j} + FN_{i,j} + FP_{i,j}}, \tag{3}$$

where $TP_{i,j}$, $FP_{i,j}$ and $FN_{i,j}$ are the numbers of true positive, false positive, and false negative cases for the classifier i and the gesture type j, respectively. Note that while $sensitivity$ and $precision$ measure the different types of classification errors performed by a system, the $F1$ score, which is their harmonic mean, is an overall indicator of the classification accuracy of a system.

We accumulate all $sensitivity$, $precision$ and $F1$ scores into 5×25 sized matrices with the classifier and gesture effects represented along the rows and columns, respectively. Similarly, we also compute the 6×25 $sensitivity$, $precision$, and $F1$ score matrices for the 6 human annotators.

The distributions for the $sensitivity$, $precision$ and $F1$ measures are not Gaussian. Hence we perform the non-parametric Friedman test [29] on all the six $sensitivity$, $precision$ and $F1$ matrices to ascertain the presence of consistent differences in the recognition rates of different gesture types. The test provides the mean ranks of the column effects, or in our case, gesture types and a probability value (p) for whether the ranks of the column effects are significantly similar or not. A low value ($p < 0.05$) indicates insufficient evidence to accept the hypothesis that the ranks of the column effects are similar and vice-versa. On observing significant differences in the recognition accuracies of a set of gestures, we rank them in decreasing order of their mean ranks for $F1$ scores.

We further identify pairs of gestures with significantly different recognition rates for vision-based algorithms and humans subjects. To do so, we conduct the multiple comparison test [30] using a nonparametric version of the balanced two-way ANOVA on the ranks of the gestures. The multiple comparison test computes standard (95%) confidence intervals for the ranks of each of the 25 gestures. It further conducts statistical comparisons between all possible pairs of the 25 gestures (called "pairwise comparisons") to determine if they have significantly different ranks. An example of such "pairwise comparisons" is shown in Fig. 4 (left), where the ranks of the 25 gestures in terms of their $F1$ scores for vision-based algorithms, are compared.

E. Comparison to Humans' Performance

We also wish to determine if the gesture recognition rates of the various gestures for vision-based algorithms are correlated to the recognition rates of human annotators. We compute the Spearman's correlation coefficient (r) [31] between the mean $F1$ ranks of vision-based algorithms and the mean $F1$ ranks of human annotators for the 25 gesture types. Further, we compute the correlation coefficient

between the average $F1$ scores (for the various gestures) of vision-based algorithms and human subjects.

F. Merging Gestures

Lastly, we propose a methodology to identify a smaller set of gestures, which are recognized with high and roughly equal accuracy. The procedure we adopt progressively merges pairs of gestures that have the largest misclassification errors.

We begin by identifying the lowest ranked gesture in terms of $F1$ scores. For example, for vision-based systems, this gesture is number 4 "move hand down" from Fig. 4 (right). We pair this gesture with another gesture that results in the highest observed false positive/negative error rate for the lowest ranked gesture, provided the error rate is greater than a pre-specified threshold (0.10 for vision-based algorithms and 0.025 for human annotators). For example, from the fourth row of Table I, which contains the average confusion matrix of the vision-based systems, the gesture 17 "push hand down" results in the highest false negative rate for the gesture 4 "move hand down" with a value of 0.36. Since this miss-classification error is greater than 0.1 we merge the gestures 4 and 17 into one class. If the lowest ranked gesture's highest miss-classification rate is less than the selected threshold value, we select the next lowest ranked gesture (*e.g.*, gesture 8 "move 2 fingers down" from Fig. 4 (right)) to merge with another gesture.

After merging the pair of highly miss-classified gestures, we recompute the average confusion matrix for the new sets of gestures and re-analyze the differences in their recognition rates using the Friedman test. If statistical differences in the recognition rates of the new sets of gestures exist, we repeat the procedure and eliminate the gesture sets with the lowest ranked $F1$ score from the current set. We terminate the procedure when no statistical differences between the recognition rates of the sets of gestures are present or when all gesture sets' mis-classification rates are less than the preselected threshold.

IV. RESULTS

A. Recognition Rate of Gestures

The Friedman test revealed statistically significant differences in the recognition rates of the 25 hand gestures for vision-based algorithms. The tests of all the three classification performance metrics ($sensitivity$, $precision$ and $F1$ scores for vision-based algorithms), resulted in probability values $p \ll 0.01$. Pairwise comparisons of the gestures' ranks with respect to their $F1$ scores are shown in Fig. 4 (left). The gestures, sorted (from top to bottom) in increasing order of their mean ranks for $F1$ scores, are shown in Fig. 4 (right). The average of the confusion matrices for the 5 vision-based algorithms is shown in Table I.

Observe that the vision-based algorithms recognize the "shake hand" gesture most accurately. On the other end of the spectrum, they classify the "move hand down" gesture least accurately. Additionally, with the help of the statistical comparisons between pairs of gestures three groups of gestures with high, medium and low recognition rates, respectively, can be loosely identified. These groups are indicated by the colors yellow, blue and red, respectively, in Fig. 4. In this categorization, all gestures in the "low" category (in red) are significantly worse than the most accurate gesture (12); and no gestures in the "high" category (yellow) are significantly dissimilar from the most accurate gesture (12) in terms of their recognition rates.

Similarly, the Friedman tests of the $sensitivity$, $precision$ and $F1$ scores for human annotators also revealed significant differences between the recognition rates of the 25 gestures. All probability values for the Friedman tests were low ($p \ll 0.01$). The statistical pairwise comparisons of the gestures for human annotators and their ranks are shown in Fig. 5 (left). Similar to vision-based algorithms, human subjects also recognized the "shake hand" gesture most accurately, and the "move hand down" gesture least accurately. The groups of gestures with high, medium and low recognition accuracy for human subjects are also indicated in Fig. 5 (right) with the colors yellow, blue and red, respectively.

B. Vision-based Algorithms vs. Humans

A plot of the average $F1$ scores (for the 25 gestures) for the vision-based algorithms versus the average $F1$ scores of the human subjects is shown in Fig. 6. Each gesture class is represented as a single dot. The $F1$ scores of vision-based algorithms were moderately, but significantly correlated to the those of human subjects with a correlation coefficient of $r = 0.487$ and a probability value of $p = 0.014$. Similarly, the mean ranks of the gestures, for vision-based algorithms were moderately correlated to those of human subjects ($r = 0.493$, $p = 0.012$). Here, we computed the ranks from the $F1$ score matrices of the vision-based algorithms and humans, respectively.

It is also interesting to note from Fig. 4 (right) and Fig. 5 (right) that the "shake hand", "thumb up", "close hand two times", and "open hand" gestures were among the most accurately classified gestures for both humans and vision-based algorithms; and the gestures "move hand down" and "show two fingers" were among the least accurately classified gestures.

These results demonstrate abilities of vision-based algorithms to perform closely to human annotators. Moreover, they indicate agreement in correctly recognizing certain types of hand gestures versus others. It suggests the presence of inherent differences between the various gestures which render them less or more easily recognizable. The variability in the recognizability of different gesture classes could be attributed to the range of variability in how humans perform them.

C. Mis-classification Errors

A number of gestures in our vocabulary are consistently misclassified by vision-based algorithms (Fig. 4). The major sources of error can be observed from the average confusion matrix of the vision-based algorithms (Table I). Notably the

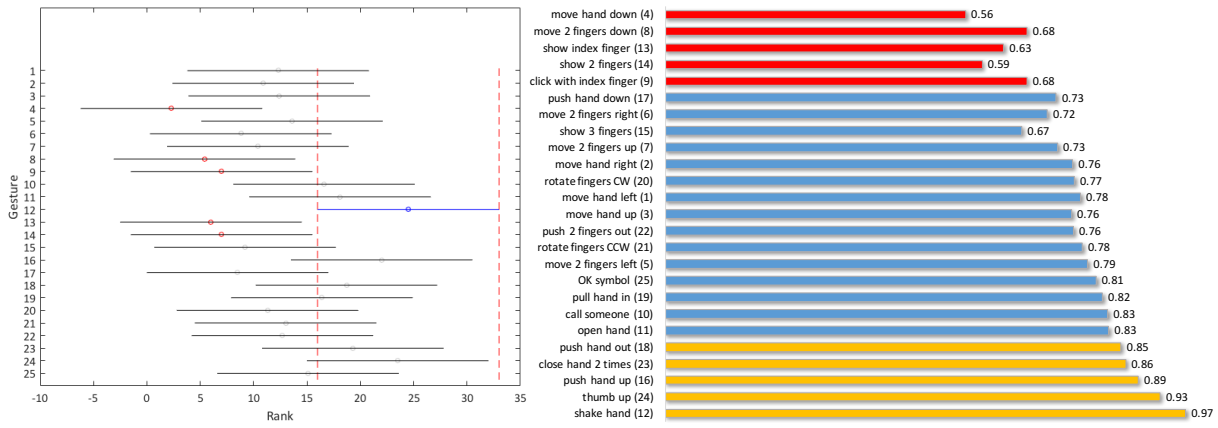

Fig. 4. **Ranking of gestures for vision-based algorithms** (Left) Shows the ranks of the 25 gestures (Fig. 2) and the standard errors of the ranks with horizontal lines. The ranks were computed from the $F1$ scores. The vertical lines represent the 95% confidence interval for the highest ranked gesture (12). All gestures with estimated (central) rank value within the 12th gesture's confidence interval are not regarded as being different from it. (Right) Shows the gestures sorted in increasing order of their mean $F1$ ranks from top to bottom. The gesture type and its number (in parenthesis) is listed to the left; and the average $F1$ score is listed to the right of each bar. Observe the set of gestures that are consistently recognized most (yellow) and least (red) accurately by vision-based algorithms.

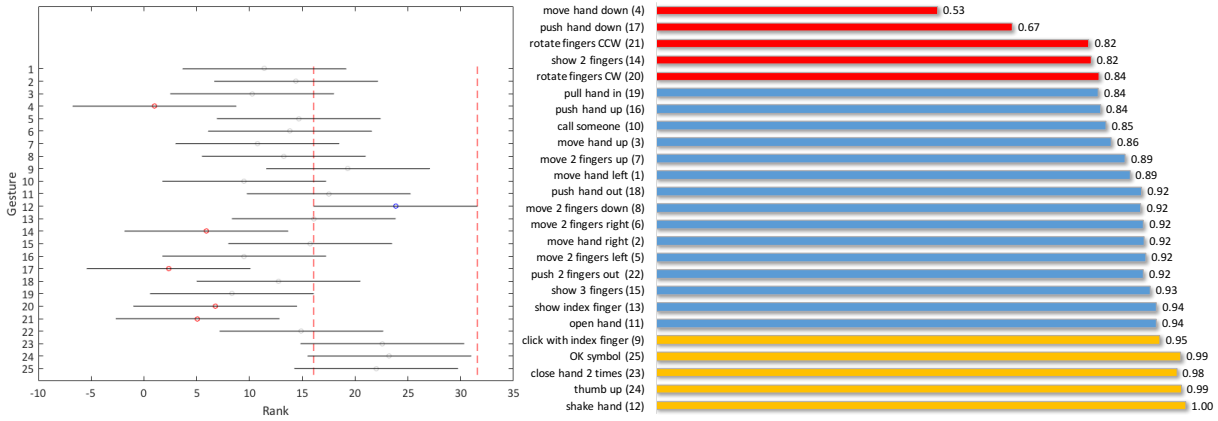

Fig. 5. **Ranking of gestures for human subjects** (Left) Shows the ranks of the 25 gestures (Fig. 2) and the standard errors of the ranks with horizontal lines. The ranks were computed from the $F1$ scores. The vertical lines represent the 95% confidence interval for the highest ranked gesture (12). All gestures with estimated (central) rank value within the 12th gesture's confidence interval are not regarded as being different from it. (Right) Shows the gestures sorted in increasing order of their mean $F1$ ranks from top to bottom. The gesture type and its number (in parenthesis) is listed to the left; and the average $F1$ score is listed to the right of each bar. Observe the set of gestures that are consistently recognized most (yellow) and least (red) accurately by human subjects.

"move hand down" gesture was most often miss-classified as the "push hand down" gesture and vice-verse. Both gestures involve similar global hand motions and differ primarily in the shape of the hand. Similarly the motion of fingers versus that of the whole hand along the horizontal direction was often confused. This suggests that the vision-based algorithms may be emphasizing motion cues over cues related to the shape of the moving object. The low resolution of the sensor used in our analyses may also be a contributing factor for this observation.

Dynamic hand gestures contain three overlapping temporal phases of "preparation" to bring the hand from the resting to the starting position; "nucleus", which contains the primary motion for the gesture; and "retraction" where the subject moves the hand back to the resting position. We observed a trend, wherein, vision-based algorithms confused between gestures which involved the same general type of motion (*e.g.*, horizontal, vertical, or rotation), but in opposite directions, *e.g.*, up/down, left/right, CW/CCW. These types of gestures often contain motions in different directions during their different phases. Hence it is plausible that vision-based algorithms rely more heavily on the presence of certain types of sub-motions as opposed to their evolution over time.

D. Merged Gestures

We obtained 10 sets of gestures by progressively merging pairs of gestures with the highest observed miss-classification rates for vision-based classifiers, as described in Section III-F. The final sets are shown in Table II and are listed in no particular order. Several of the sets of merged gestures can be described with a single gesture listed in the third column of Table II. Note that the gestures listed in the third

TABLE I

THE AVERAGE CONFUSION MATRIX OF VISION-BASED ALGORITHMS FOR THE 25 GESTURES. LARGE MISS CLASSIFICATION ERRORS (≥ 0.10) ARE INDICTED IN COLOR AND ZERO VALUES AND NOT INCLUDED.

Truth	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
1	.72	**.10**	.03		.05	.04				.02			.01								.01		.02		
2	**.10**	.71			.02	**.11**				.02	.02											.01		.01	
3	.01		.69	**.10**	.01		.07	.01	.01						.01	.03	.02	.03					.02		
4	.02		.01	.50		.01	.02	.04		.01							**.36**	.01					.02		
5					.87	.01		.01					.01	.02	.01					.01	.06				
6		.01			**.19**	.68									.01					.04	.01	.06			
7			.02	.03	.01		.76	.04					.01	.02	.04	.01				.03		.02		.01	
8				.02		.01	**.18**	.65	.02					.01	.04		.02			.02	.01	.01			
9							.03	.06	.63		.02	.01	**.15**	.01	.02	.01						.01	.05		.01
10										.85					.01				**.11**	.02				.01	
11					.01	.01					.91	.01		.04	.02										
12			.02									.98													
13									.08		.03		.65	**.14**	.01							.02		.05	.02
14								.01	.05		.02		.07	.66	**.16**							.03			
15								.01	.01		.03		.03	**.19**	.71			.01							.01
16			.03				.01			.03				.01		.91	.01								
17			.01	**.10**				.01									.86	.01						.01	
18		.02	.01	.03		.01											.06	.82		.01		.02		.01	
19				.02			.01			**.11**				.01		.08			.77						
20					.01			.06	.02	.01										.86	.03		.01		
21					.01				.02											**.22**	.73		.01		
22								.02	.07				.01	**.10**	.02					.01		.74		.01	.01
23					.02						**.11**	.01									.01		.85		
24										.01			.03											.96	
25		.01				.03					.07			.02	.07			.04				.02			.73

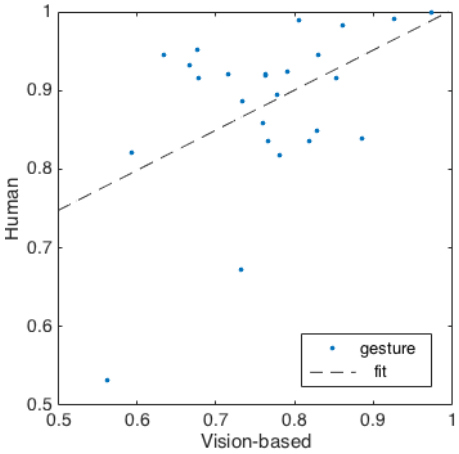

Fig. 6. The average $F1$ scores of vision-based algorithms for the 25 gestures plotted against the corresponding scores for human annotators. Each dot represents a different gesture. The best-fit line to the observed data is also plotted.

TABLE II

MERGED GESTURES

Number	Merged Gestures	Equivalent Gesture
1	move hand left move hand right, move 2 fingers left, move 2 fingers right	Move horizontally
2	move hand up, move hand down, push hand down, push hand out	Move hand vertically
3	move 2 fingers up, move 2 fingers down	Move 2 fingers vertically
4	click with index finger, show index finger, show 2 fingers, show 3 fingers, push 2 fingers out, OK symbol	Show fingers
5	call someone, pull hand in	Beckon
6	open hand, close hand 2 times	Open/close hand
7	shake hand	shake hand
8	push hand up	push hand up
9	rotate fingers CW, rotate fingers CCW	rotate
10	thumb up	thumb up

column of Table II are much less specific than the original 25 gestures and allow subjects more flexibility in choosing how to perform them.

The merged gestures had an average $F1$ score of 0.89. By comparison, the original 25 gestures had a much lower average $F1$ score of 0.77. In addition, these 10 gestures obtained via merging also had a higher average $F1$ score than that of the individual 10 gestures from the original set with the highest ranked $F1$ scores (Fig. 4 (right)). The latter set of gestures has an $F1$ score of 0.86. Also, while the gestures in the merged set (Table II) had statistically indistinguishable recognition rates ($p = 0.08$), the high $F1$ ranked individual gestures did not ($p = 0.003$). We observed an identical trend on comparing the 10 merged gestures for human observers versus their 10 highest ranked gestures in terms of $F1$ scores (Fig. 5 (right)). The $F1$ scores for the original 25 gestures, 10 gestures after merging, and the 10 gestures with the highest ranked $F1$ scores for human annotators were 0.88, 0.99 and 0.96, respectively. Furthermore, there was more parity in the recognition rates of the 10 gestures resulting from merging ($p = 0.007$) versus those resulting from selecting the 10 gestures with the highest ranked $F1$ scores ($p = 8.51e^{-5}$). We also re-trained the best performing vision-based R3DCNN classifier with data from the depth

channel only with the 10 (a) merged gestures in Table II and (b) highest ranked gestures in Fig. 4 (right). Owing to reduction in the number of gesture classes both sets of gestures resulted in higher accuracies of 93.9% and 93.2%, respectively, with the merged gestures resulting in slightly higher accuracy.

Lastly, we compared the memorability of the 10 merged gestures with that of the 10 gestures with the highest ranked $F1$ scores for vision-based classifiers. We use the accuracy of human observers at correctly recognizing a particular set of gestures as a proxy for the memorability of that set. We observed that the set of merged gestures were more memorable: humans were able to correctly recognize with an accuracy of 96.62%, as opposed to the 10 high ranked gestures, which humans recognize correctly with an accuracy of 92.73%. Hence, we recommend the final set of 10 gestures listed in third column of Table II for automotive interfaces because of their high and equivalent recognition rates, and high memorability for human observers.

V. CONCLUSIONS

The accuracy of gesture recognition systems is a critical factor in determining the success and widespread adoption of gesture UIs in cars. In this work we focused on the robustness of gestures. We show that there is consistent variability in how accurately different types of hand gestures are recognized by various vision-based algorithms and human subjects. We further identify similarities in the hand gesture recognition abilities of vision-based systems and humans. Lastly, by means of a methodology that we propose for merging pairs of error-prone gesture classes, we identify ten sets of hand gestures, which are accurately and equally classified by vision-based algorithms. To the best of our knowledge, ours is the first work to address the important question of selecting robust gestures for automotive UIs from a system design perspective.

In the future we plan to incorporate the human factors of safety and driver distraction [15] as well as intuitiveness and ease of use in the criteria for selecting hand gestures for automotive UIs. In addition, we plan to incorporate a larger corpus of gesture data containing more gestures per class, in order to increase the statistical confidence of the results and to improve the training of the statistical pattern classifier.

REFERENCES

[1] WHO, "Mobile phone use: A growing problem of driver distraction," World Health Organization, Tech. Rep., 2011.

[2] NHTSA, "Traffic safety facts: A research note," National Highway Traffic Safety Administration, Tech. Rep., 2013.

[3] M. Regan, "Driver distraction: Reflections on the past, present and future," *Distracted driving. Sydney, NSW: Australasian College of Road Safety*, pp. 29–73, 2007.

[4] J. C. Stutts, A. A. Association, *et al.*, *The role of driver distraction in traffic crashes*. AAA Foundation for Traffic Safety Washington, DC, 2001.

[5] S. G. Klauer, T. A. Dingus, V. L. Neale, J. D. Sudweeks, and D. J. Ramsey, "The impact of driver inattention on near-crash/crash risk: An analysis using the 100-car naturalistic driving study data," Tech. Rep., 2006.

[6] F. Parada-Loira, E. Gonzalez-Agulla, and J. Alba-Castro, "Hand gestures to control infotainment equipment in cars," in *IEEE Intelligent Vehicles Symposium*, 2014, pp. 1–6.

[7] 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*. Springer, 2004, pp. 448–459.

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

[9] E. Ohn-Bar and M. Trivedi, "Hand gesture recognition in real time for automotive interfaces: A multimodal vision-based approach and evaluations," *IEEE Trans. on Intelligent Transportation Systems*, vol. 15, no. 6, pp. 1–10, 2014.

[10] N. Neverova, C. Wolf, G. W. Taylor, and F. Nebout, "Multi-scale deep learning for gesture detection and localization," in *ECCVW*, 2014.

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

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

[13] T. Kopinski, S. Magand, A. Gepperth, and U. Handmann, "A light-weight real-time applicable hand gesture recognition system for automotive applications," in *IEEE Intelligent Vehicles Symposium*, 2015, pp. 336–342.

[14] 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 *Int. Conf. on Automotive User Interfaces and Interactive Vehicular Applications*, 2013, pp. 14–21.

[15] N. H. T. S. Administration *et al.*, "Visual-manual nhtsa driver distraction guidelines for in-vehicle electronic devices (docket no. nhtsa-2010-0053)," *Washington, DC: US Department of Transportation National Highway Traffic Safety Administration (NHTSA)*, 2013.

[16] S. Gupta, D. Morris, S. Patel, and D. Tan, "Soundwave: using the Doppler effect to sense gestures," in *CHI*, 2012, pp. 1911–1914.

[17] K. Kalgaonkar and B. Raj, "One-handed gesture recognition using ultrasonic Doppler sonar," in *ICASSP*, 2009, pp. 1889–1892.

[18] S. Dura-Bernal, G. Garreau, C. Andreou, A. Andreou, J. Georgiou, T. Wennekers, and S. Denham, "Human action categorization using ultrasound micro-doppler signatures," in *Human Behavior Understanding*. Springer, 2011, pp. 18–28.

[19] H. Wang, D. Oneata, J. Verbeek, and C. Schmid, "A robust and efficient video representation for action recognition," *International Journal of Computer Vision*, 2015.

[20] X. Yang and Y. Tian, "Super normal vector for activity recognition using depth sequences," in *CVPR*, 2014.

[21] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition," in *NIPS*, 2014.

[22] D. Tran, L. D. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *ICCV*, 2015.

[23] M. Zobl, M. Geiger, K. Bengler, and M. Lang, "A usability study on hand gesture controlled operation of in-car devices," in *Poster Proceedings HCII 2001*, 2001.

[24] P. Molchanov, X. Yang, S. Gupta, K. Kim, S. Tyree, and J. Kautz, "Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural networks," in *IEEE CVPR*, 2016.

[25] Bayerische Motoren Werke AG, "Gesture control interface in BMW 7 Series https://www.bmw.com/." [Online]. Available: https://bmw.com/

[26] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *Scand. Conf. on Im. Anal.*, 2003.

[27] Code Laboratories Inc., "Duo3D SDK https://duo3d.com/." [Online]. Available: https://duo3d.com/

[28] A. Graves, S. Fernández, F. J. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *ICML*, 2006, pp. 369–376.

[29] R. V. Hogg and J. Ledolter, *Engineering statistics*. Macmillan Pub Co, 1987.

[30] Y. Hochberg and A. C. Tamhane, "Multiple comparison procedures," 2009.

[31] R. A. Fisher, *Statistical methods for research workers*. Genesis Publishing Pvt Ltd, 1925.