"Geremin": 2D Microgestures for Drivers Based on Electric Field Sensing

Christoph Endres
German Research Center for Artificial Intelligence
Saarbrücken, Germany
Christoph.Endres@dfki.de

Tim Schwartz
Cluster of Excellence Multimodal Computing and Interaction
Saarland University

Christian Müller
German Research Center for Artificial Intelligence
Saarbrücken, Germany
Christian.Mueller@dfki.de

ABSTRACT

We introduce the "Geremin" approach on in-car 2D microgesture recognition, which belongs to the category of electric field sensing techniques detecting the presence of a human hand near a conductive object (not affected by light and dynamic backgrounds, fast response times). The core component is essentially a modified "Theremin", an early electronic musical instrument named after the Russian inventor Professor Léon Theremin. Gesture recognition is done using a Dynamic Time Warp DTW algorithm. With respect to the application domain, we follow the direction of "selective mapping to theme or function" suggested in the literature. For gesture location, we propose the immediate proximity of the steering wheel, which has the advantage of providing gesture-based interaction without requiring the driver to take off hand(s). The major motivating factor for the proposed approach is reducing installation costs. Although, we use a single-antenna setup for this study, our results indicate that the gain in recognition accuracy justifies the use of two or more.

ACM Classification Keywords

H.5.2 Information interfaces and presentation: User Interfaces, User-centered design

General Terms

Design,Human factors

Author Keywords

automotive, gesture recognition, touch-free interaction, driving safety

INTRODUCTION

In increasing awareness for a safe driver interface, several car manufactures are carrying out research in gesture recognition in collaboration with universities and research institutes [7]. Gestures represent a comfortable addition to existing interaction without the decline in recognition accuracy under noisy conditions that speech still suffers from [1]. However, as pointed out by [7], research has identified other problems: As gesture recognition approaches are mainly camera-based, maintaining accuracy in varying light conditions and with dynamic backgrounds are a major issue. Moreover, since image processing algorithms are computationally expensive, real time operation with standard hardware is not easy to obtain. [11] furthermore points out that in despite of the costs for a respective camera being less than $5, the big hurdle for introducing gesture recognition is justifying these costs.

In a comprehensive overview article, [4] conclude that lasers and capacitive infrared techniques have been referred to in the literature but "no publication on a working system has been identified". The approach presented here belongs to the category of electric field sensing techniques that were initially pioneered by MIT [12]. The technique, which detects the presence of a human hand near a conductive object, has been proven to not being affected by light and dynamic backgrounds while having fast response times.

Aside from creating a reliable recognition system, a challenging task is to design a consistent and easy-to-use HMI that leverages the technology in order to actually create a safer way of interaction. According to [4], gestures are originally not self-revealing and therefore need explanation and visual reminders. At the same time, the authors acknowledge that providing visual reminders would necessarily neutralize any potential safety benefit. Nevertheless, [7] expects numerous automotive applications by 2020. Today, three different application domains for mapping automotive hand gestures have been addressed [4]: 1. Direct mapping of gestures to the complete functionality of in-vehicle devices (e.g. radio, CD, navigation system). Although this approach could lead to a very consistent interface, the authors conclude that too many gestures would be needed and thus many of them would not be natural. 2. Mapping to in-vehicle controls, mimicking each individual control type (push button switch, push and hold button switch, rotary position selector, ...). This

IUI'11, February 13–16, 2011, Palo Alto, California, USA.

type is not recommended either by [4], because creating natural mimic gestures for each control type has substantial limitations. According to the authors, the third category, selective mapping to theme or function, "appears to have the most realistic practical possibilities". [11] provides examples belonging to this category: waving-off incoming calls and using one's index finger with a clockwise/counterclockwise rotation to raise respectively lower the stereo volume. [8] proposes skipping between music titles, albums, radio stations or enabling/disabling audio sources. Selective mapping is the application domain we aim at as well as detailed below.

Figure 1: A sequence of images taken from a video showing a person executing a clockwise circular gesture while driving. The video was taken without a functional gesture recognition system in a standard car in order to illustrate the intended functionality.

Finally, gesture location requires some attention as well. Here, [4] identify 1. (driver) windshield area, 2. central windshield area, and 3. center stack area as the variants investigated in the literature. Our approach belongs to the first category. However, we further restrict the location to the area to the immediate proximity of the steering wheel. In our opinion, this has the advantage of providing gesture-based interaction without requiring the driver to take off hand(s). We postulate a "hands-free" gesture interaction as it is shown in Figure 1.

ENVISIONED HMI CONCEPT

As indicated above, we envision a selective mapping of a limited set of "microgestures" performed in the immediate area of the steering wheel without taking a hand off. In particular, we focus the following task: raising (+) or lowering (-) the status of a certain object, which can be windows (up, down), seat heating (warmer, cooler), climate control (warmer, cooler), volume (higher, lower), etc. We are aware of the fact that other forms of interaction exist that ultimately have to be integrated into the gesture concept. However, we believe that +,- is a reasonable starting point, because this kind of manipulation is very common in the car. Therefore, we chose the following gestures for the study presented here: CIRCLE CW, RIGHT, UP for + and CIRCLE CCW, LEFT, DOWN for -. It is important to notice that aside from this broad outline of the concept, we do not elaborate on the human-centered part, which is orthogonal the goal of this paper: exploring the basic technical constraints (recognition rates) of the gesture recognition component. Questions like which of the suggested gestures are preferred by the users (if any at all) or how much driver distraction is accompanied with our "microgestures" in comparison to other modes (touch, eye-gaze, speech, turn-and-push dial), etc. are subject of future research.

THE "GEREMIN" APPROACH

With the "Geremin" system for 2D-gestures as it is proposed here, a gesture (for example a circular gesture as depicted in Figure 1) is translated by the an electric field sensing component, which is essentially a modified "Theremin" (also called "aetherophone"), an early electronic musical instrument named after the Russian inventor Professor Léon Theremin, who patented the device in 1928 [2]. The Theremin is controlled without contact from the player and consists of two metal antennas. Moving the hand towards or away from an antenna alters the capacity of an oscillating circuit. Hereby, hand and antenna form the two plates (conductors) of a capacitor. With the original instrument, one of the player's hands controls the frequency (pitch) and the other the amplitude (volume). In trying to get to the bottom of minimizing installation costs, we used a single-antenna setup for this study. Note that the gestures themselves are nevertheless two-dimensional. Our results revealed, however, that the gain in recognition accuracy justifies the use of two (or more) antennas. The sound generated is fed into a signal processing component that translates the pitch curve into a vector of numbers (whose number of dimensions correspond to the number of antennas used). We use the tool Praat [3] to do this. The feature vector is finally used as input for the GESTURE RECOGNITION component.

For gesture recognition we use the classification framework described in [5] . The framework provides several common classifiers for learning and recognizing signals of an arbitrary number of dimensions, such as Multilayer Perceptron Neural Networks (MLP) , Support Vector Machines (SVM) and Multi-Dimensional Dynamic Time Warp (DTW) [10]. For the present study, we used DTW because it outperformed the other two algorithms in a series of pretests. DTW measures similarity between two sequences which may vary in time or speed. It has been applied to video, audio, and graphics. A well known application is automatic speech recognition, to cope with different speaking rates.

RELATED WORK

The present study is consistent with the literature in terms of gesture set, number of subjects, and evaluation procedure. All systems below use a Hidden Markov Model (HMM) for classification, which is a very common approach with vision-based systems [2]. [6] introduce a vision-based drawing tool for augmented desk interaction using a regular camera in conjunction with an infrared camera, which tracks the finger temperature. The 2D gesture set comprises CIRCLE, SQUARE, and TRIANGLE, clockwise (cw) and counterclockwise (ccw). The system was evaluated with seven subjects. An average accuracy of 92.5 % and 97.5 % was obtained in the single-finger respectively multi-finger condition. [9] describe a WiiMote-based interaction component designed to be used for a variety of applications. The gesture set comprises SQUARE CW, CIRCLE CW, QUARTER CIRCLE CW, Z-SHAPE, and a TENNIS move. Training and testing

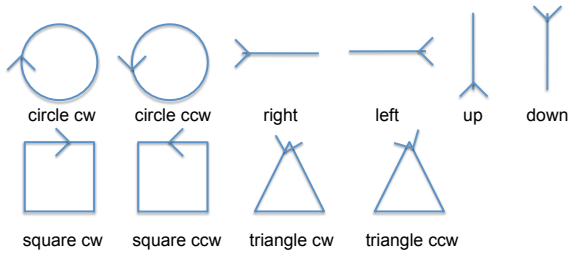

Figure 2: Top row: a set of (target) gestures believed to be useful in the automotive context. Bottom row: additional gestures used in the study to evaluate the recognition accuracy of target gestures.

was done on the same data (self-training). Results on a study with six subjects (one woman, five men) revealed an accuracy of 89.5 % on the five-class problem. [1] suggest a combination of head and hand gestures to operate selected in-car devices. The set of hand gestures comprises 17 different gestures. In order to improve the recognition rates, the probability of gestures which are not in the current system context are lowered. An evaluation was performed with six subjects revealing an average accuracy of 86 % on the entire set.

EXPERIMENT

The initial set of 2D-gestures used in the present study is depicted in Figure 2. The upper row shows examples of gestures, which we believe are useful in the actual application scenario. The lower row of Figure 2 represents a set of additional gestures with a certain geometric complexity. Particularly, we investigate the discriminability of the target gestures CIRCLE CW and CIRCLE CCW from the respective square-shaped variants (SQUARE CW and SQUARE CCW). Also, we are interested in the question, how much the recognition of linear gestures suffer from having only one dimension input signals.

The "Geremin" was mounted behind a (fixed) steering wheel of a car. Two separate data sets were created using a horizontal respectively a vertical antenna. Seven subjects (five men, two women) were asked to perform the gesture set twice (training/testing) holding the steering wheel with both hands like it is shown in Figure 1. From the initial set of gestures detailed above, variants were created by modifying execution speed and accuracy. Subjects were instructed to gesture a) slow and accurate, b) fast and accurate, and c) fast and inaccurate resulting in a total of 840 cases (7 subjects * 10 gestures * 3 qualities * 2 sets * 2 antenna alignments). The starting point was fixed according to the arrows in Figure 2.

RESULTS

The following facts are helpful for interpreting the results detailled below: With the 10-class problem at hand, chance level is at 10 %. Training and test were done on a single gesture (first respectively second gesture of each subject). No parameter tuning or adaption of the classification algorithm was made.

avg 64,28	circlecw	circleccw	right	left	up	down	squarecw	squareccw	triangle cw	triangle ccw
circlecw	**28,57**	28,57					14,29		28,57	
circleccw	28,57	**71,43**								
right			**71,43**	14,29		14,29				
left				**100,00**						
up				14,29	**71,43**	14,29				
down				14,29	14,29	**71,43**				
squarecw							**57,14**		42,86	
squareccw		14,29						**85,71**		
triangle cw				14,29			42,86		**42,86**	
triangle ccw		42,86						14,29		**42,86**

Table 1: Confusion matrix showing the recognition accuracy in percent for a one-dimensional "Geremin" (vertical antenna) in the slow + accurate condition. Rows: actual gestures (ground truth); columns: hypothesized gestures (classifier output); diagonal: correctly classified gestures; upper left corner: average accuracy (recall). Zero cells were emptied out. Recall values for fast + accurate and fast + inaccurate are 51.4 % respectively 57.1 %. See text for interpretation.

Table 1 provides a confusion matrix showing the recognition accuracy in percent for a one-antena "Geremin". The antenna was installed vertically behind the steering wheel. The numbers are results in the slow and accurate condition. Rows represent ground truth, i.e. actual gestures, while columns contain hypothesized gestures (output of the classifier). Hence, the diagonal represents the correctly classified cases. LEFT was correctly recognized in 100 % of the cases which makes sense since LEFT is a single horizontal line that can be caught up by a vertical antenna fairly good. RIGHT is confused with left and down in some of the cases. We attribute this to the fact that gestures were made with the right index finger holding the steering wheel. In order to perform a RIGHT gesture, subjects needed to first bend the finger and then stretch it to the left. As a result, the movement tilted a little bit and therefore was not strictly horizontal. CIRCLE CW and CIRCLE CCW were confused with each other – which is also not surprising given that we have only one signal dimension.

Results obtained with an horizontal antenna are along the lines of the observations above. Here, UP achieves the highest accuracy while DOWN suffers from similar drawbacks as we have seen with RIGHT. The average accuracy in the horizontal-only condition is lower (48 % for slow + accurate). We attribute this to the fact that the location/alignment of the antenna was suboptimal. The reader is invited to interpret the remainder of the numbers in Table 1 along these lines, which, altogether, appeal to our intuition. The average recall values for fast + accurate and fast + inaccurate are 51.4 % respectively 57.1 % (vertical antenna alignment). There is no (significant) difference between the accurate and inaccurate conditions. It seems to be the case that performing gestures faster impedes accuracy.

In order to get an outlook on a two-dimensional setup, we fused the horizontal and vertical dimensions. We are

avg 65,71	circlecw	circleccw	right	left	up	down	squarecw	squareccw	trianglecw	triangleccw
circlecw	28,57				14,29	14,29			42,86	
circleccw		57,14			14,29				14,29	14,29
right			42,86	14,29	14,29			28,57		
left				71,43	14,29				14,29	
up					100,00					
down				14,29	14,29	57,14				14,29
squarecw	14,29						57,14	14,29	14,29	
squareccw								100,00		
trianglecw							14,29		85,71	
triangleccw								28,57	14,29	57,14

Table 2: Recognition accuracy (%) a two-dimensional case where vertical antenna and horizontal antenna were recorded separately and fused afterwards. Slow + accurate condition. See text for interpretation.

aware of the fact that the expected recognition accuracy is not as high as with gestures simultaneously recorded using two antennas. This is due to the fact that 1. obviously the two parts do not stem from the same movement and 2. data is not temporally aligned. Table 2 shows the respective confusion matrix. Although the average recognition accuracy is only marginally better, which we believe is due to the above mentioned reasons, the picture of individual confusions changed: 1. CIRCLE CW and CIRCLE CCW are not confused with each other any more; 2. the characteristic pattern of LEFT and UP disappeared. The main confusion is to be observed among the more sophisticated shapes (with respect to gesture recognition), namely SQUARE and TRIANGLE. Also, it is apparent that UP has a high recall but a low precision value, i.e. other gestures are often confounded with UP.

DISCUSSION

The results presented here are promising given that installation costs, one of the major hurdles for introducing new technology in the automotive industry, have been driven down extremely in this approach by using only one antenna. Please note that for two-dimensional gesture recognition, the sensors (here antennas) do not have to be aligned to each other in a way that the one represents the x-axis and the other the y-axis. Basically, there is a large degree of freedom as to where the antennas are installed exactly. Besides experimenting with two or more antennas recorded simultaneously, future work will have to reconsider the training and machine learning part. We believe that one reason (besides the one-antenna setup) why recognition accuracy was not as high as in other studies is that training has only been done on a single gesture. Thus, the model was not able to generalize sufficiently. Also, a Hidden-Markov approach will be considered in future experiments as it was the preferred learning algorithms in other studies. By modeling a series of states and transition probabilities, HMMs are able to catch up the temporal aspects of the gestures better than DTW, which is essentially a similarity measure of final objects. Also, reliable gesture recognition alone does not guarantee a better HMI for the driver. There has to be a consistent mapping of a relatively small set of gestures onto functions. This aspect will also be part of our future work.

Acknowledgements. This work was funded by BMBF (grant number 01IW08004). We would like to thank Robert Neßelrath for providing his classification API.

REFERENCES

1. 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. Jahrhundert*, Braunschweig, Germany, November 22-23 2005.

2. M. Billinghurst. Gesture Based Interaction. In B. Buxton, editor, *Haptic Input*. Cambridge University Press, 2009.

3. P. Boersma. PRAAT, a system for doing phonetics by computer. *Glot International*, 9(5):341–345, 2001.

4. M. J. R. Carl A. Pickering, Keith J. Burnham. A Research Study of Hand Gesture Recognition Technologies and Applications for Human Vehicle Interaction. In *3rd Conf. on Automotive Electronics*, June 2007.

5. R. Nesselrath and J. Alexandersson. A 3D Gesture Recognition System for Multimodal Dialog Systems. In *6th IJCAI Worksh. on Knowledge and Reasoning in Practical Dialogue Systems*, pages 46–51, Passadena, CA, July 12 2009.

6. K. Oka, Y. Sato, and H. Koike. Real-Time Fingertip Tracking and Gesture Recognition. *IEEE Computer Graphics and Applications*, 22:64–71, 2002.

7. C. Pickering. Gesture Recognition Could Improve Automotive Safety. *Asian Engineer - Automotive-Design*, December 2006.

8. U. Reissner. Gestures and Speech in Cars. In *Electronic proceedings of Joint Advanced Student School (JASS)*, March 2007.

9. T. Schlömer, B. Poppinga, N. Henze, and S. Boll. Gesture recognition with a Wii controller. In *TEI '08: Proc. of the 2nd International Conference on Tangible and Embedded Interaction*, pages 11–14, New York, NY, USA, 2008. ACM.

10. G. ten Holt, M. Reinders, and E. Hendriks. Multi-Dimensional DTW for Gesture Recognition. In *13th Conference Advanced School for Computing and Imaging*, 2007.

11. K. Whitfield. Gesture Interface for Automotive CONTROL: Beyond Digital Expletives. *Automotive Design and Production Magazine*, July 2003.

12. T. Zimmerman, J. R. Smith, J. A. Paradiso, D. Allport, and N. Gershenfeld. Applying Electric Field Sensing to Human-Computer Interfaces. In *Proceedings of CHI 1995*, pages 280–287. Press, 1995.