A REALTIME SYSTEM FOR HAND GESTURE CONTROLLED OPERATION OF IN-CAR DEVICES

Martin Zobl, Michael Geiger, Björn Schuller, Manfred Lang, Gerhard Rigoll

Institute for Human-Machine Communication, Munich University of Technology, D-80290 München
{martin.zobl, michael.geiger, björn.schuller, manfred.lang, gerhard.rigoll}@ei.tum.de

ABSTRACT

The integration of more and more functionality into the human machine interface (HMI) of vehicles increases the complexity of device handling. Thus optimal use of different human sensory channels is an approach to simplify the interaction with in-car devices. This way the user convenience increases as much as distraction may decrease. In this paper a videobased realtime hand gesture recognition system for in-car use is presented. It was developed in course of extensive usability studies. In combination with a gesture optimized HMI it allows intuitive and effective operation of a variaty of in-car multimedia and infotainment devices with handposes and dynamic hand gestures.

1. INTRODUCTION

[1] gives a comprehensive survey of existing gesture recognition systems. The most important area of application in the past was sign language recognition [2]. Due to fast technical evolution with increasing complexity of the HMI in the last years, applications in the technical domain have become more important. Examples are controlling of the computer dektop environment [3, 4, 5], of presentations [6] and operation of multimedia systems [7]. Especially in the car domain new HMI solutions have been in focus of interest [8, 9] to reduce distraction effects and to simplify the usage. In usability studies, gesture controlled operation of a variety of in-car devices proved to be intuitive, effective [10, 11] and less distracting than haptical user input with knobs and buttons [12]. The presented approach nevertheless is part of a multimodal system. In the following section a short introduction to the whole system is given. Accordingly the single components are presented. At the end, results are discussed and an outlook about future work is given.

2. GESTURE INVENTORY

The used gesture inventory is fitted to the findings in usability studies [10, 11], which makes it suitable to a mean user. It consists of 22 dynamic gestures grouped to twelve gesture classes (including one trash class) and six handposes (including one trash class). The set of gestures consists of dynamic referencing (e.g. ’pointing’), kinemimic (e.g. ’waving to the left/right/up/down’), symbolic (e.g. ’pointing’ for ”engage”) and mimic (e.g. ’lift virtual phone’) gestures. Handposes like for example ’grab’, ’open hand’ or ’relaxed’ were added to the inventory to allow additional functionality inside the user interface. In figure 1, some examples out of the gesture inventory are given.

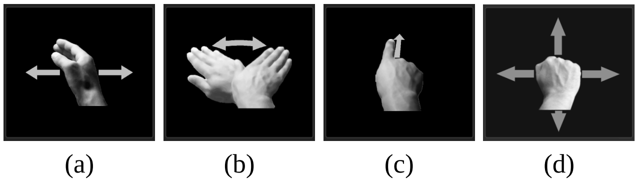
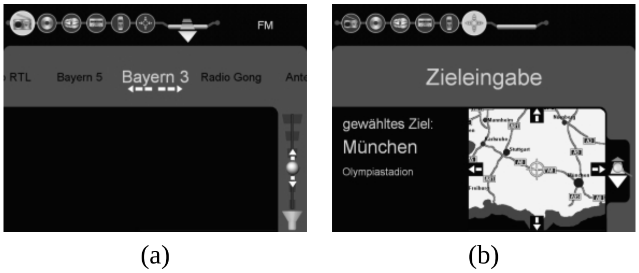

(a) (b) (c) (d)

Fig. 1. Examples out of the gesture inventory with possible directions: ’wave to the left/right’(a) to change to the previous/next function, ’wipe’ (b) to stop a system action, ’point’ (c) to confirm and ’grab’ (d) for direct manipulation of e.g. the volume.



gewähltes Ziel: München

(a) (b)

Fig. 2. GECOM in radio (a) and navigation (b) mode.

3. OVERVIEW

The predescribed gestures are used to control the gesture optimized HMI GECOM (GEsture COntrolled Man machine interface) [13] (see figure 2). It has functionality of a navigation system, multimedia and communication devices. The gesture recognition system consists of different processing stages as shown in figure 3. In the feature extraction stage,

From: ICME 2003, IEEE

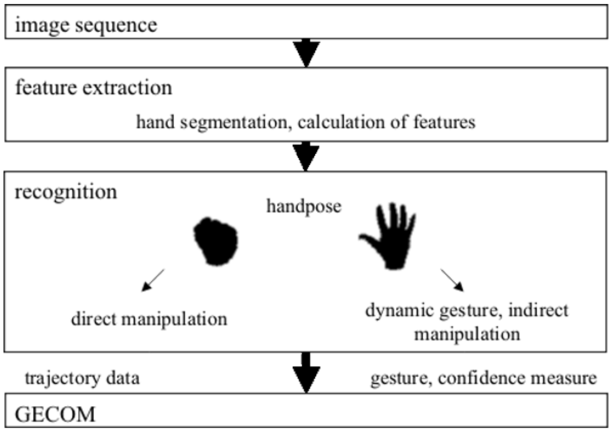


image sequence
feature extraction
hand segmentation, calculation of features
recognition
handpose
direct manipulation
dynamic gesture, indirect manipulation
trajectory data
gesture, confidence measure
GECOM

Fig. 3. Overview: Gesture recognition system

the hand is segmented from the background and gesture describing features are calculated (see section 4). With these features handpose recognition is performed (see section 5.1). Depending on the recognized pose there are two possible modes - direct and indirect manipulation.

3.1. Direct Manipulation Mode

As long as the handpose ‘grab‘ is performed, continuous trajectory data (see table 1) of the hand is sent to GECOM. This direct manipulation is used to control functions that are inconvenient to use with single dynamic gestures like adjusting the music volume or moving a navigation map in 3D [13].

3.2. Indirect Manipulation Mode

The mode for indirect manipulation is activated by the handpose ‘open hand‘. The next recognized dynamic gesture (see section 5.2) is sent to GECOM. Each dynamic gesture input is used to execute one command in GECOM. A continuous spotting approach would fail, because some of the gestures out of the inventory are as common (e.g. ‘to the left‘, ‘to the right‘) that they could be used casually by the driver while talking to other people inside the car.

Indirect manipulation can be achieved with special handposes, too. ‘Pointing‘ for example selects the displayed item.

4. FEATURE EXTRACTION

For image acquisition, a standard CCD camera is mounted at the roof with its field of vision centered to the mid console. This is the area where most gestures were performed by test subjects in usability studies. As proposed in [9] the camera is equipped with a daylight filter and the scene is illuminated by NIR LEDs (950nm) to achieve independence from ambient light as well as to prevent the driver from being disturbed. Fields are grabbed with 25fps at a resolution of 384*144 to avoid *frame comb* that would destroy the features in case of fast hand movements . For spatial segmentation, it is assumed that the hand is a large object that does not belong to the background and is very bright because of the NIR illumination. Thus, on the original image background subtraction is performed to remove pixels not belonging to the hand. The so processed image is multiplied with the gray values of the original image to consider the brightness. The resulting image is then thresholded at an adaptive level T. At time step n, with the original image $I_n[x, y]$ and the background image $B_n[x, y]$, the hand mask $\widetilde{I}_n[x, y]$ can be written as

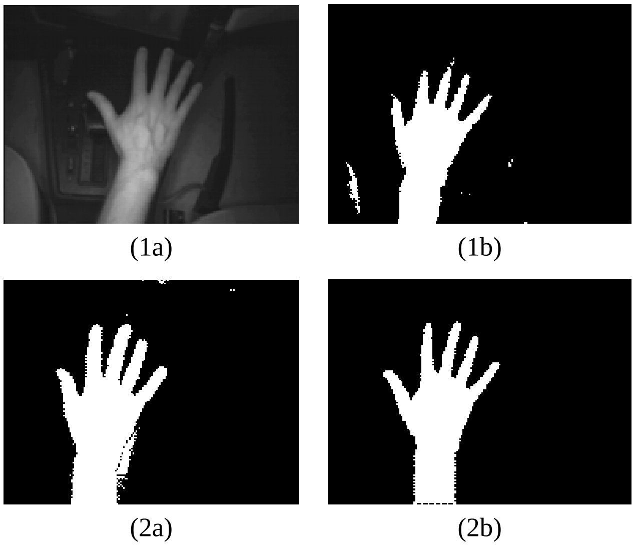


Fig. 4. Steps in hand segmentation: grabbed image (1a), adaptive thresholding (1b), background subtraction (2a), combination of background subtraction and thresholding (2b)

$$\widetilde{I}_n = \begin{cases} 1 & , I_n \cdot |I_n - B_n| \geq T \\ 0 & , else \end{cases} \tag{1}$$

Small objects are removed with cleaning operators. The background image is updated in a sliding mean window with every region that does not belong to the hand, to adapt to changing background. Figure 4 illustrates the used segmentation steps. After segmentation, a *forearm filter* [14] is applied to remove the forearm´s influence on the features. Moment based features like area, center of mass and Hu moments [15] are calculated from the segmented image.

	$\sqrt{A}$	ΔA	C	ΔC	HU
handpose recognition	-	-	-	-	+
dynamic gesture recog.	-	+	-	+	+
direct manipulation	+	-	+	-	-

Table 1. Features used for the different tasks. A: area, C: center of mass, HU: Hu Moments.

5. RECOGNITION

A feature vector is formed for every image. It consists of features that are necessary for the respective task (see Table 1).

5.1. Handpose Recognition

Since the handform is independent of area, position and hand rotation, Hu moments are used for handpose description. For classification, the Mahalanobis-Distance between the actual feature vector and every class representing prototype is calculated. To avoid a system reaction on casual handposes, the distances are smoothed by a sliding window filter. The reciprocal values (*scores*) of the smoothed distances are finally transformed into confidence measures as described in section 5.3.

5.2. Recognition of Dynamic Gestures

In dynamic gestures the handform as well as the relative trajectory contains relevant information. Not only one vector, but a vector stream has to be classified. In the first stage, the vectors containing the gesture are cut out from the vector stream with a *movement detection* that uses the derivatives of the movement features (area, center of mass). In the second stage the cut feature vector sequence is fed to Hidden Markov Models (HMMs) [16]. Here semi-continuous HMMs are used because of their low quantity of parameters and smooth vector quantisation. The viterbi search through the models delivers a score for every model (representing a gesture) given a feature vector sequence. These scores are transformed into confidence measures as described in section 5.3.

5.3. Confidence Measure

A Maximum-Likelihood decision about the handpose or dynamic gesture based only on the best match is relatively uncertain. A measure is needed to show how safe the decision for the best match is - regarding the output of every model given a vector or a vector sequence. Further this measure should spread between zero and one to resemble a probability. With the number of existing gesture classes N and class i delivering the best match, the measure

$$c_i = \frac{score_i}{\sum_{j \in N} score_j} \qquad (2)$$

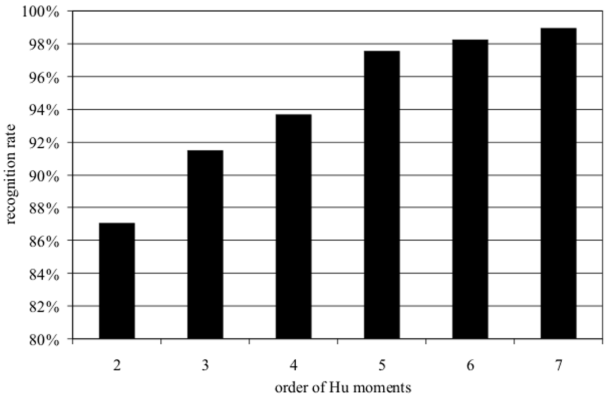


Fig. 5. Recognition rates for handpose recognition over the order of used Hu moments for building the feature vector

fits to our demands and delivers good results with the described classifiers. When the best score is high and the other scores are low then $c \rightarrow 1$. When every score is equal then $c = \frac{1}{N}$. Now for every class a threshold is defined above which the recognition is accepted. Below this threshold it is rejected.

6. RESULTS

The recognition results are preliminary results for offline recognition with datasets from one person.

For the evaluation of the handpose recognition 500 datasets per class were collected. 250 datasets were used to train the prototypes and 150 datasets to test the recognition performance. In figure 5 the recognition rate over the feature vector dimension is shown. With increasing the accuracy of the handpose description, the recognition result is increasing. Some poses (e.g. ‘grab‘, ‘open hand‘, ‘pointing‘) already show very good recognition rates (95%) with low feature dimensions. Other poses (e.g. ‘hitchhiker left‘, ‘hitchhiker right‘) are very similar with respect to the Hu moments (rotation invariance!) and can only be seperated in a higher dimensional feature space. To achieve accurate results using the described methods, the forearm filter proved as a precondition. The so achieved recognition results nearly reach those using hand segmented pictures. The evaluation of the dynamic gesture recognition system was done with 20 datasets per gesture. 13 sets were used to train the HMMs and seven to test the recognition. The HMMs consist of seven states, because the duration of some gestures is sometimes as short as seven frames. Given a simple forward structure, not more than seven states can be used. Since the recognition rate is 100% when using the first two Hu moments in addition to the relative trajectory and a codebook consisting of 128 prototypes, no figure is included.

The results show, that the gesture recognition works very well for both handpose and dynamic gesture recognition when it is adaptet to a single user. False rejection and acceptance levels have not been tested so far with the presented confidence measure, because of the lack of multiple user data. Future work will be an online evaluation with different subjects while controlling GECOM to get an overall result. A gesture controlled HMI should only be part of a multimodal HMI in which the user is allowed to control every functionality with the optimal modality (haptics, speech, gestures). The so build HMI will enable the user to handle complex multimedia systems like in-car devices in an intuitive and effective way while driving a car.

7. REFERENCES

[1] V. Pavlovic, R. Sharma, and T. Huang, "Visual interpretation of hand gestures for human-computer interaction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19.7, pp. 677–695, 1997.

[2] H. Hienz, K. Kraiss, and B. Bauer, "Continuous sign language recognition using hidden markov models," in *Proceedings, 2nd Int. Conference on Multimodal Interfaces, Hong Kong, China, 1999*, 1999, pp. IV10–IV15.

[3] F. Althoff, G. McGlaun, B. Schuller, P. Morguet, and M. Lang, "Using multimodal interaction to navigate in arbitrary virtual vrml worlds," in *Proceedings, PUI 2001 Workshop on Perceptive User Interfaces, Orlando, Florida, USA, November 15-16, 2001*. Association for Computing Machinery, 2001, ACM Digital Library: www.acm.org/uist/uist2001. CD-ROM.

[4] Y. Sato and Y. Kobayashi, "Fast tracking of hands and fingertips in infrared images for augmented desk interface," in *Proceedings, 4th Int. Conference on Automatic Face and Gesture Recognition, Grenoble, France, 2000*, 2000, pp. 462–467.

[5] P. Morguet and M. Lang, "Comparison of approaches to continuous hand gesture recognition for a visual dialog system," in *Proceedings, ICASP 1999 Int. Conceference on Acoustics and Signal Processing, Phoenix, Arizona, USA, March 15-19, 1999*. IEEE, 1999, pp. 3549–3552.

[6] C. Hardenberg and F. Bérard, "Bare-hand human-computer interaction," in *Proceedings, PUI 2001 Workshop on Perceptive User Interfaces, Orlando, Florida, USA, November 15-16, 2001*. Association for Computing Machinery, 2001, ACM Digital Library: www.acm.org/uist/uist2001. CD-ROM.

[7] "Jestertek Inc. Homepage," www.jestertek.com.

[8] E. Klarreich, "No more fumbling in the car," in *nature, Glasgow, Great Britain, November, 2001*. British Association for the Advancement of Science, 2001, Nature News Service.

[9] S. Akyol, U. Canzler, K. Bengler, and W. Hahn, "Gesture control for use in automobiles," in *Proceedings, MVA 2000 Workshop on Machine Vision Applications, Tokyo, Japan, November 28-30, 2000*. IAPR, 2000, pp. 28–30, ISBN 4-901122-00-2.

[10] M. Zobl, M. Geiger, P. Morguet, R. Nieschulz, and M. Lang, "Gesture-based control of in-car devices," in *VDI-Berichte 1678: USEWARE 2002 Mensch-Maschine-Kommunikation/Design, GMA Fachtagung USEWARE 2002, Darmstadt, Germany, June 11-12, 2002*, Düsseldorf, 2002, VDI, pp. 305–309, VDI-Verlag.

[11] M. Zobl, M. Geiger, K. Bengler, and M. Lang, "A usability study on hand gesture controlled operation of in-car devices," in *Abridged Proceedings, HCI 2001 9th Int. Conference on Human Machine Interaction, New Orleans, Louisiana, USA, August 5-10, 2001*, New Jersey, 2001, pp. 166–168, Lawrence Erlbaum Ass.

[12] M. Geiger, M. Zobl, K. Bengler, and M. Lang, "Intermodal differences in distraction effects while controlling automotive user interfaces," in *Proceedings Vol. 1: Usability Evaluation and Interface Design , HCI 2001 9th Int. Conference on Human Machine Interaction, New Orleans, Louisiana, USA, August 5-10, 2001*, New Jersey, 2001, pp. 263–267, Lawrence Erlbaum Ass.

[13] M. Geiger, R. Nieschulz, M. Zobl, and M. Lang, "Gesture-based control concept for in-car devicess," in *VDI-Berichte 1678: USEWARE 2002 Mensch-Maschine-Kommunikation/Design, GMA Fachtagung USEWARE 2002, Darmstadt, Germany, June 11-12, 2002*, Düsseldorf, 2002, VDI, pp. 299–303, VDI-Verlag.

[14] U. Broekl-Fox, *Untersuchung neuer, gestenbasierter Verfahren für die 3D-Interaktion. PhD thesis*, Shaker Publishing, 1995.

[15] M. Hu, "Visual pattern recognition by moment invariants," *IRE Transactions on Information Theory*, vol. IT8, pp. 179–187, 1962.

[16] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, pp. 257–286, 1989.