Gesture Components for Natural Interaction with In-Car Devices

Martin Zobl, Ralf Nieschulz, Michael Geiger, Manfred Lang, and Gerhard Rigoll

Institute for Human-Machine Communication, Munich University of Technology,
D-80290 München, Germany,
zobl, nieschulz, geiger, lang, rigoll@ei.tum.de,
WWW home page: http://www.mmk.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 decreases. In this paper the gesture part of a multimodal system is described. It consists of a gesture optimized user interface, a real time gesture recognition system and an adaptive help system for gesture input. The components were developed in course of extensive usability studies. The so built HMI allows intuitive, effective and assisted operation of infotainment in-car devices, like radio, CD, telephone and navigation system, with handposes and dynamic hand gestures.

1 Introduction

In [1] a comprehensive survey of existing gesture recognition systems is given. 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 and a broad variety of application possibilities, applications in the technical domain have become more important in the last years. Examples are controlling the computer desktop environment [3–5] and presentations [6] as well as operating 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 this environment strict constraints limit the possibilities of user interface design. A drivers primary task always is controlling the car. This task should not be interfered by other tasks, like controlling a HMI. So only short time slots can be used for interaction between the user and the HMI. Additionally feedback possibilities are very limited, because displays are not placed in the primary view of the user. In usability studies, gesture controlled operation of infotainment in-car devices proved to be intuitive, effective [10, 11] and less distracting than haptical user input with knobs and buttons [12].

For this reason the development of a gesture operated HMI is worthwhile. To lower one's inhibition threshold of this new operating type, an automatic,

From: GW 2003, Ed.: A. Camurri, LNCS 2915 Springer Verlag

adaptive help system to provide unobtrusive assistance for gestural operation is a reasonable completion. Regarding human and machine as an overall system, a more stable overall system behavior is achieved casually. Of course the presented components concerning gestures are part of a multimodal system, as some functions like a selection out of long lists are better performed with speech.

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 Overview

In figure 1 the gesture components and their relationship is shown. The user interface (see section 4) is driven by the performed gestures. These are recognized by a gesture recognition system (see section 5). Additionally the associated confidence measures and timestamps of the recognized gestures are sent for use with

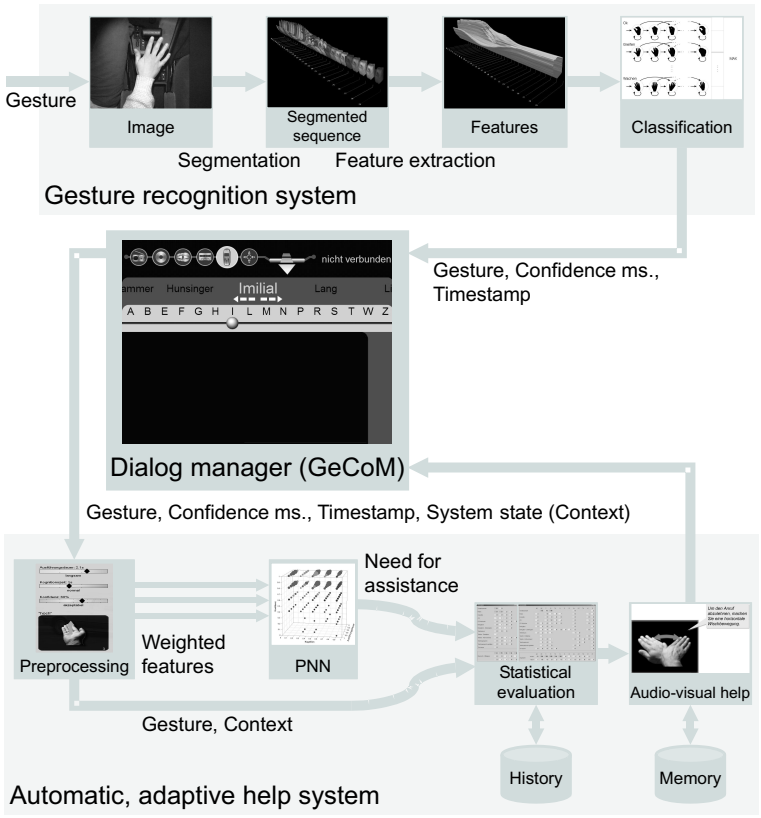
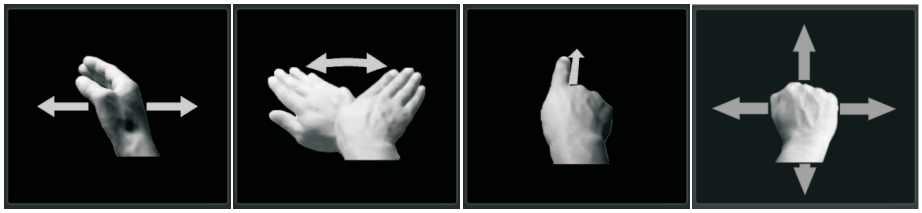


Fig. 1. System Overview

the adaptive help system (see section 6). The help system gets information about the performed gestures with confidences and timestamps as long as information about the state of the user interface. With these features the need for help and type of help is calculated and audio visual help is presented in the user interface when necessary.

3 Gesture Inventory

The used gesture inventory is fitted to the findings in usability studies [10, 11] which makes is suitable to a mean user. In figure 2, examples out of the gesture inventory are shown. There are eleven gesture classes of dynamic hand gestures (some containing several equivalent gestures) and four hand poses. Dynamic hand gestures are used for indirect manipulation (discrete control steps). Hand poses can be applied for different tasks. Two examples out of the handposes are discussed here. With the hand pose 'open' the dynamic gesture recognizer is activated and then is waiting for dynamic gesture input. An activation mechanism is necessary, 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 persons inside the car. The hand pose 'grab' is applied for direct manipulation. This direct manipulation allows the user to control functions that are inconvenient to handle with single dynamic gestures like adjusting the music volume or moving a navigation map in 3D [13].



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

Fig. 2. 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.

4 User Interface

As a result of usability studies, we developed a Gesture Controlled Man-Machine Interface (GeCoM) [12]. It was evaluated in the course of several usability investigations (Wizard-of-Oz methodology) in our driving simulator. By iterative re-design, the interface was optimized for gestural control. The implemented

functional range consists of typical devices of automotive infotainment systems like radio, CD, phone and navigation. The HMI is displayed over a 10" TFT display mounted in the mid console.

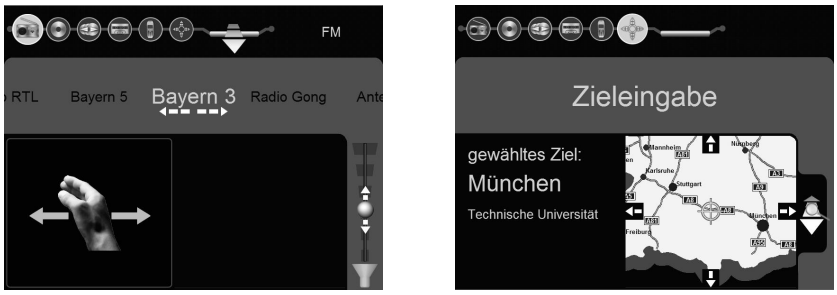
The probably most important task for composition of GeCoM is its the visual representation. Especially when performing kinemimic gestures, the user follows the alignment of the displayed elements without exception. Horizontal elements are exclusively controlled with horizontal movements, whereas vertical structures are controlled with vertical movements (see fig. 3). Beyond it, a strong correlation

(a) (b)

Fig. 3. Vertical (a) and horizontal alignment (b) of menu points with visual presentation for indirect (left) and direct (right) manipulation.

between the user's behavior exists even when no interface is displayed at all. A large number of subjects use for example horizontal gestures to the right in the sense of 'next function' and horizontal gestures to the left in the sense of 'previous function'. Accordingly up and down movements are used to raise or lower a control variable (e.g. volume).

Being aware of this, a horizontal aligned primary interaction structure with selectable menu points and a vertical aligned secondary structure for controlling the volume was implemented (see fig. 4). To reconsider the relation between the gesture and the system reaction, state changes are smoothly animated. The active device is represented by a self-explanatory pictogram. The displayed infor-



(a) (b)

Fig. 4. GeCoM in radio (a) and navigation mode (b). In (a) a help screen is displayed to assist the user in changing the radio station.

mation is reduced to a minimum to allow the user an instantaneous recognition of the system state. The described visual attributes support the user in building a correct system model, which is a precondition for controlling without averting the gaze off the road. In addition, acoustical feedback in form of beeps and speech is given with every system reaction.

5 Gesture Recognition

5.1 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 then multiplied with the gray values of the original image to consider the brightness. The resulting image is 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 $\widetilde{I}[x, y]_n$ can be written as follows.

$$\widetilde{I}_n[x, y] = \begin{cases} 1 \; \forall \; I_n[x, y] \cdot |I_n[x, y] - B_n[x, y]| \geq T \\ 0 \text{ otherwise} \end{cases} \tag{1}$$

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 background and ambient light changes. Figure 5 illustrates the used segmentation cues and their combination.

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

Fig. 5. Cues for hand segmentation: The grabbed image (a), when only thresholded (b) or background subtracted (c). Combination of background subtraction and thresholding (d).

After segmentation, a modified *forearm filter* based on [14] is applied to remove the forearm´s influence on the features. Moment based features like area, center of mass (trajectory) and Hu's moments [15] (handform) are calculated from the segmented image.

5.2 Recognition Performance

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

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

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

Hand Pose Recognition Since the form of the hand is independent of area, position and hand rotation, Hu's moments are used for hand pose description. For classification, the Mahalanobis-Distance between the actual feature vector and every class representing prototype (previously trained) is calculated. To avoid a system reaction on casual hand poses, the distances are smoothed by a sliding window filter. Additionally a trash model is introduced. The reciprocal values (*scores*) of the smoothed distances are finally transformed into confidence measures as described in section 5.4.

Recognition of Dynamic Gestures In dynamic gestures also the form of the hand as well as the relative trajectory data contains relevant information. Not only one vector, but a vector stream here 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 quantization. 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 5.4, too.

5.3 Results

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

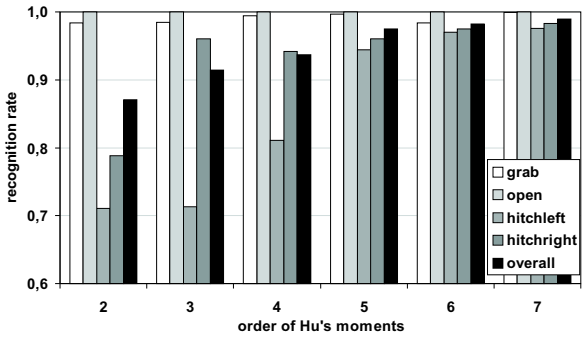


Fig. 6. Recognition rates for four handposes depending on the order of Hu's moments for building the feature vector. Single hand pose results are shown as long as the overall recognition rate.

For the evaluation of the hand pose recognition 500 datasets per class were collected. 350 datasets were used to train the prototypes and 150 datasets to test the recognition performance. In figure 6 the recognition rate over the feature vector dimension is shown. With increasing the accuracy of the hand pose description, the recognition result is increasing. Some poses ('grab', 'open hand') already show very good recognition rates (95%) with low feature dimensions. The other poses (e.g. 'hitchhiker left', 'hitchhiker right') are very similar with respect to the Hu's moments (rotation invariance) and can only be separated in higher feature dimensions. To achieve accurate results using the described methods, the forearm filter proved as a precondition. The so achieved recognition results nearly reach those using pictures with manually removed forearm segments. 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. Since the duration of certain gestures is sometimes as short as seven frames, the HMMs consist of seven states. Given a simple forward structure, not more than seven states can be used. In figure 7 the results of different feature and HMM configurations is given. They already show a very high recognition rate at low feature and codebook dimensions. The low feature dimension means, that a lot of the gestures out of the vocabulary could be recognized using only relative trajectory data and the low codebook dimension, that the used features cluster very well in the feature space.

The recognition results demonstrate, that the described gesture recognition system works very well, for both hand pose and dynamic gesture recognition, when adapted to a single user.

5.4 Confidence Measure

A Maximum-Likelihood decision about the hand pose or dynamic gesture based only on the best match is relatively uncertain. A measure is needed to show how

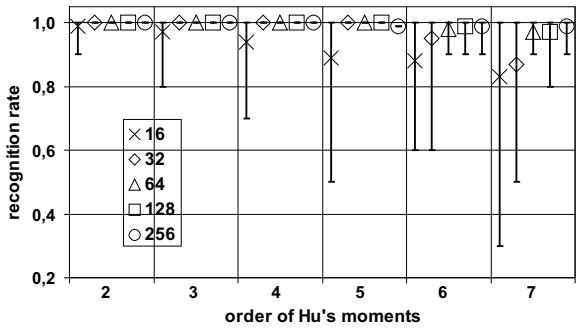


Fig. 7. Average recognition rates for 11 dynamic gesture classes depending on the order of Hu's moments for building the feature vector. The codebook size of the semi continuous HMMs is given as parameter between 16 and 256. The error bars show the rate of the worst and best recognized gesture class.

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

$$c_i = \frac{score_i}{\sum_{j \in N} score_j} \tag{2}$$

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 $c \to 1$. When every score is equal then $c = \frac{1}{N}$. Now for every gesture class a threshold is defined above which the recognition is accepted. Below this threshold it is rejected. This becomes necessary because some gestures are relative similar to others, while other gestures are totally different. This would lead to an always low confidence measure for similar gestures. False rejection and acceptance levels have not been tested so far with the presented confidence measure, because of the lack of multiple user data.

6 Adaptive Help System

The adaptive help system is implemented in a 2-stage methodology. The first stage is a neural network based classification that determines if a user needs help while performing a certain task. This stage works automatically if desired, which is the default mode. Alternatively a user is able to request a help information manually. In the second stage a postprocessing based on statistics determines which help this user actually needs in the given context (q.v. [17, 18]).

6.1 Need of Assistance

For every gesture the user performs, the following data is send from the gesture recognition system to the adaptive help system, in order to infer the user's *need of assistance*: the gesture type (e.g. 'right') the appropriate confidence measure as described in section 5.4 and the gestures start and ending time. This input is then preprocessed to adapt it to a user's gestural behavior. In short, all features are weighted with memory functions regarding all past gestures of this person. The preprocessed feature vector now provides information about the user adapted quality of a gesture by its weighted confidence measure, as well as the user adapted execution duration and cognition time (q.v. [18]).

The feature vector is then used as input to a neural network, which supplies the statement, whether a user needs assistance, or not, at that time. The neural network is built as a probabilistic neural network (PNN) based on radial basis functions (RBFs). About 2000 gestures out of a test series about gestural operation of in-car devices (q.v. sec. 4) were used as training corpus. First, the optimal positions of the neurons of the PNN in its feature space are determined, applying a LVQ-algorithm (linear vector quantization) to the training material. That way an effective, lean and powerful neural network design is obtained. Build on that the PNN is trained to classify a users *need of assistance* with the mentioned training corpus.

The performance of this classifier was tested using about another 1000 gestures from the study. With further consideration of the operating context the recognition rate of the neural network is 96% and the error rate 4%. As can be seen in figure 8 the recognition rate of the case, that the user actually doesn't need help, is 98%, while the rate of the other case, where the user needs assistance, is only 81%.

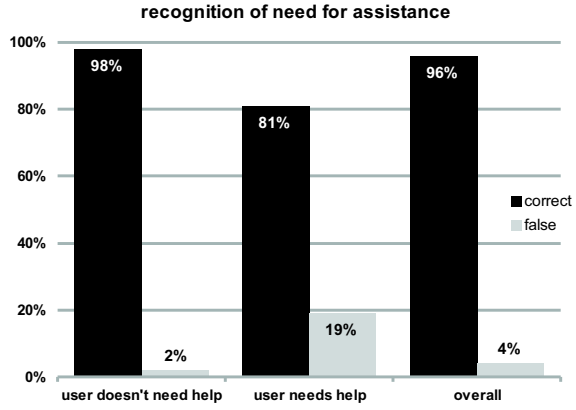


Fig. 8. Recognition rate of the probabilistic neural network used in the first stage of the help system

6.2 Help Content

If *need of assistance* is detected in the first stage of the help system or if a user demands for help explicitly, the help content is determined in the second stage. Therefore the current context of the HMI and a users operation history is taken into account. The following information is gathered: which gestures have been used in what context at what time in a correct respectively false manner by this user so far, and which help content has been provided to this user in what context at what time so far. Out of this data a weight is calculated for every separate help content and for every help type (that is to represent cognitive coherences between related help contents) using memory functions.

A bayesian network, which contains the description of the whole help corpus, is continuously adapted by means of these weights. After the adaptation of the network, it is able to infer the statistically most probable help content. The help content thus calculated is then audio-visually presented to the user via GeCoM (v. figure 4) [17, 18]. If the provided information strikes the user as insufficient or wrong, he can request further assistance.

As results of a usability study regarding gestural operation of in-car devices combined with the presented help system show (v. figure 9), the system had to provide 1.35 help contents on average for satisfaction of the users. This is a significant enhancement compared to conventional online help systems. Usually one time-consuming has to search through a more or less extensive menu structure, which is possibly shortened by the current context. The study has also shown, that even if a user does not require any help actually, the provided information is useful most of the time nevertheless.

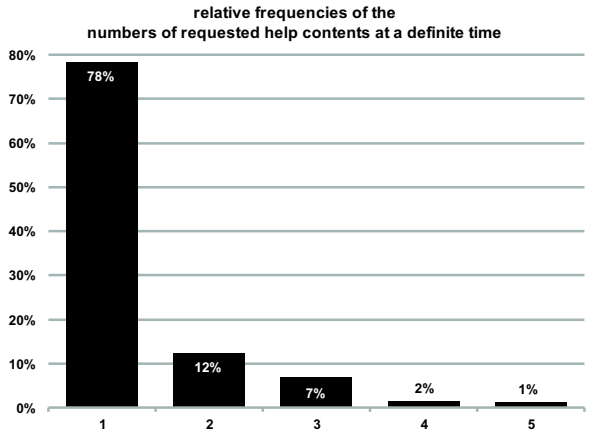


Fig. 9. Performance of the second stage of the help system: users had to request 1.35 help contents on average to achieve appropriate help

7 Summary and Outlook

In this paper a system of gesture components for a natural interaction with in-car devices was presented. It was shown, that consequent user centered redesign coupled with implementation of new techniques enhance the operation of the HMI. By the use of an integrated adaptive help system the whole gesture operated HMI obviously gains convenience. In particular the learning procedure is significantly accelerated. Moreover an earlier overcome of a persons inhibition threshold in using gestures is achieved.

Future work will be an online evaluation of the system with different subjects to get an overall result. Nevertheless, gestural operation should be part of a multimodal system in which the user is allowed to control every functionality with the optimal or familiar 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.

References

1. Pavlovic, V., Sharma, R., Huang, T.: Visual interpretation of hand gestures for human-computer interaction. IEEE Transactions on Pattern Analysis and Machine Intelligence **19.7** (1997) 677–695
2. Hienz, H., Kraiss, K., Bauer, B.: Continuous sign language recognition using hidden markov models. In: Proceedings, 2nd Int. Conference on Multimodal Interfaces, Hong Kong, China, 1999. (1999) IV10–IV15
3. Althoff, F., McGlaun, G., Schuller, B., Morguet, P., Lang, M.: 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, ACM Digital Library: www.acm.org/uist/uist2001. CD-ROM (2001)
4. Sato, Y., Kobayashi, Y.: 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) 462–467
5. Morguet, P., Lang, M.: 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) 3549–3552
6. Hardenberg, C., Bérard, F.: 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, ACM Digital Library: www.acm.org/uist/uist2001. CD-ROM (2001)
7. : Jestertek Inc. Homepage. (www.jestertek.com)
8. Klarreich, E.: No more fumbling in the car. In: nature, Glasgow, Great Britain, November, 2001, British Association for the Advancement of Science, Nature News Service (2001)
9. Akyol, S., Canzler, U., Bengler, K., Hahn, W.: Gesture control for use in automobiles. In: Proceedings, MVA 2000 Workshop on Machine Vision Applications, Tokyo, Japan, November 28-30, 2000, IAPR, ISBN 4-901122-00-2 (2000) 28–30

10. Zobl, M., Geiger, M., Morguet, P., Nieschulz, R., Lang, M.: 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, VDI, VDI-Verlag (2002) 305–309
11. Zobl, M., Geiger, M., Bengler, K., Lang, M.: 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, Lawrence Erlbaum Ass. (2001) 166–168
12. Geiger, M., Zobl, M., Bengler, K., Lang, M.: 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, Lawrence Erlbaum Ass. (2001) 263–267
13. Geiger, M., Nieschulz, R., Zobl, M., Lang, M.: 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, VDI, VDI-Verlag (2002) 299–303
14. Broekl-Fox, U.: Untersuchung neuer, gestenbasierter Verfahren für die 3D-Interaktion. PhD thesis. Shaker Publishing (1995)
15. Hu, M.: Visual pattern recognition by moment invariants. IRE Transactions on Information Theory **IT8** (1962) 179–187
16. Rabiner, L.R.: A tutorial on hidden markov models and selected applications in speech recognition. Proceedings of the IEEE **77** (1989) 257–286
17. Nieschulz, R., Geiger, M., Zobl, M., Lang, M.: Need for assistance in automotive gestural interaction. In: VDI-Berichte 1678: USEWARE 2002 Mensch-Maschine-Kommunikation/Design, GMA Fachtagung USEWARE 2002, Darmstadt, Germany, June 11-12, 2002, Düsseldorf, VDI, VDI-Verlag (2002) 293–297
18. Nieschulz, R., Geiger, M., Bengler, K., Lang, M.: An automatic, adaptive help system to support gestural operation of an automotive mmi. 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, Lawrence Erlbaum Ass. (2001) 272–276