8—26 Gesture Control for use in Automobiles

Suat Akyol, Ulrich Canzler *
Dept. of Technical Computer Science
University of Technology (RWTH)
Aachen, Germany

Klaus Bengler, Wolfgang Hahn †
BMW AG
Munich, Germany

Abstract

Gesture interfaces are gaining relevance for human-machine communication, since it is expected that they make interaction more intuitive. Particularly vision based approaches are widely preferred. This paper describes a novel vision based real-time gesture recognition system, designed for operating in an automotive environment. It is used within an application for retrieving traffic news and e-mails from a message storage. Image processing and pattern matching techniques, specially adapted to the complex environmental conditions, represent the systems basics.

1 Introduction

Vision based gesture recognition is a popular research issue. Recently Pavlovic et al. [8] composed a survey of this topic. The most widespread application is sign language recognition [3][11][5], but it hasn't reached the stage of practical applicability yet. Restrictions regarding user, environment and vocabulary are still too strong. Also gesture control is used with technical applications, like in human-robot interaction [12], for augmented desk interfaces [9][13] or crane control [7]. Although the common motivation of such work is to yield more natural and intuitive and thereby more efficient interfaces, the actual usefulness in practice is not always obvious. Thus it is not surprising that only a few commercial products exist, e.g. the Siemens Virtual Touchscreen (SiVit [10]). It represents an "info kiosk", where the user can reference a spot on a projection space by pointing gestures.

Until now gesture control has not been utilised in automobiles, even though it hypothetically provides a set of advantages: reduced visual and mental distraction compared to conventional interfaces, an at least partial saving of mechanical input devices and more comfortable natural input. Hence we prototypically developed a vision based real-time recognition system, to operate a message storage by gestures in an automotive environment. The demonstrator exists as a laboratory setup and as a mobile version in an experimental vehicle.

The next section describes the technical requirements and some preliminary considerations. Afterwards, the image acquisition and processing schemes are presented. Then the functionality of the chosen application is explained, before the last section gives a final discussion and outlook.

2 Requirements and Setup

The general conditions inside an automobile comprise intense variations of illumination, changing users and non-uniform backgrounds. Besides this, the user acceptance is an important item, why measures like visible illumination, clothing restrictions or extensive calibration requirements can not be tolerated. Accordingly we defined the following criteria as a prerequisite for the given purpose:

- robustness against environmental noise
- invisible illumination
- user independence
- no calibration or training by user
- small and comprehensible set of gestures
- system reaction with minimal latency

Our demonstrator is designed for the recognition of one-hand gestures, since the other must be on the steering wheel. The gesturing area is above the gear shift stick, where a single camera is placed above to record it. This choice was made for two reasons: first, the arm can lie comfortably on the arm rest, and second, this area is not observable for other road users, which is necessary to avoid misunderstandings in public traffic. Figure 1 shows the setup and an image of the camera view.

The camera view in figure 1 represents the typical situation of one hand addressing a gesture command

*Address: Ahornstr.55, 52074 Aachen, Germany.
E-mail: [akyol,canzler]@techinfo.rwth-aachen.de
†Address: FIZ, Knorrstr.147 80788 Munich, Germany.
E-mail: [klaus.bengler,wolfgang.hahn]@bmw.de

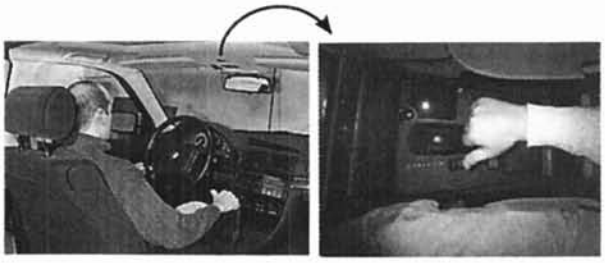

Figure 1: demonstrator setup and camera view

to the system. For use in practice some special cases must be considered in advance, too. The most frequently expected ones are gathered in figure 2. E.g. figure 2a shows a resting hand and a part of the driver's leg, which intrudes into the scene. In (b) there is no hand at all. In both cases the system may not interpret a command, since it is obviously not the drivers intention to do so. Multiple hands acting simultaneously in the gesturing area can occur e.g. if the co-driver reaches into the gesturing area unintentionally (c) or performs gestures to take command of the system (d). Hence it is necessary to apply some relevancy ranking in order to resolve ambiguities, which requires the extraction of motion information.

Figure 2: special cases in gesturing area

3 Image Acquisition

The detection of the hand pose from a camera recorded image sequence requires sufficient contrast between the hand and the background. Usually, a constant level of illumination is not given under driving conditions. Lightness depends on daytime and weather conditions. So there is a basic necessity for an auxiliary lighting source, which must not influence the inmates or hinder the driver from conducting the car. For this reasons we chose an LED-array emitting near-infrared light (NIR) of 950 nm wavelength. This light is not visible for the human eye but can be detected with a slightly modified CCD-camera. Unfortunately, this choice eliminates any colour information and permits only to measure intensities, which is done with a resolution of 8 bit. An automatic lens aperture provides a smooth intensity level, despite to fluctuation that occur in practice. Additionally, a daylight filter is placed in front of the lens to filter out the visible spectrum, which is not needed in our application. The lighting module as well as the camera are integrated in the car's roof and allow to seize a recording area of approximately 60 cm x 50 cm. Images are digitised by a framegrabber hosted on a standard personal computer.

4 The Processing Scheme

Considering all above mentioned requirements we developed the processing scheme drafted in figure 3. Intermediary processing results are visualised there, too. At the beginning, each image is segmented using simple thresholding with a global threshold. The processing speed of this procedure can not be matched by any other kind of segmentation. Then the boundaries of all connected regions are tracked and labelled with primary features like size, scope, centroid and Hu-moments [4].

Within an image sequence of predefined duration corresponding regions are matched between subsequent images (we will call those regions "objects" henceforth). This is done by tracking the centroids as described in [2] and by assuming that small regions represent negligible noise. Mean values of motion features, like speed or acceleration, are computed over the sequence and spatiotemporal scope properties are logged. This allows to assign the objects to different dynamic classes.

To reduce the calculation outlay of the following processing steps, some objects are eliminated by a pre-selection method, which is based on fuzzy scoring of the observed features. The fuzzy sets have been developed by analysing typical ranges of features for the hands of several test persons. The fuzzy inference labels some objects with the attribute "is unlikely to be a hand" and disregards them.

The remaining objects represent possible candidates for being a hand. Since a hand is always located at the end of the users arm, which is irrelevant and disturbs the recognition, it needs to be isolated from it. Therefore we apply an iterative algorithm for filtering the arm out of a 2D projection of a hand-arm constellation. This procedure has been described in [1]. As a result the user is free to wear clothing with short or long sleeves and to hold his hand at any position in the gesturing area.

After the arm filtering a vector of rotation, translation and scale independent features is composed and passed to the following classification step. The winner of the classification gets a bonus score, if it can positively be assigned to a known reference. This bonus score serves as a bias for the next preselection cycle and prevents the focus of attention from frequent swapping between objects when the situation is ambiguous.

The classifier itself uses maximum likelihood decision. In a training phase several feature vectors

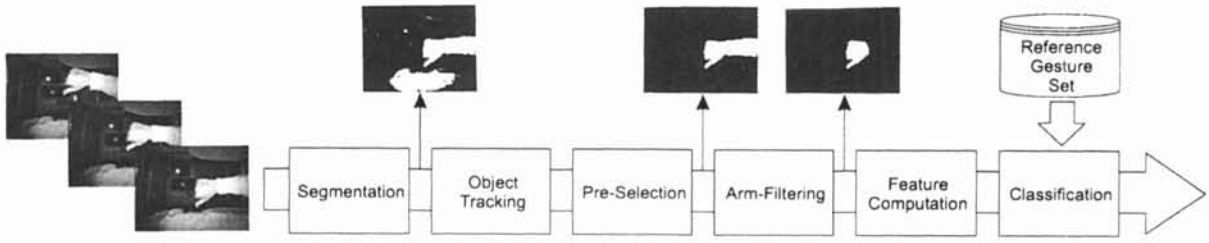

Figure 3: processing scheme

of each reference gesture are presented to the classifier. This allows to estimate a probability distribution function for each dimension (which is a feature). Generally all inter-feature dependencies have to be regarded and modelled by a multivariate distribution function. Under the assumption of stochastically independent features the distribution functions may be modelled as univariate. The recognition phase returns a probability, which describes the similarity between the currently observed object and a known reference. The object with the highest overall probability is the winner of this procedure.

To avoid false classification the winning objects probability must exceed a pre-defined threshold. Besides, if the object shows too little motion, then the classification result is verified by a second classifier, which is based on boundary shape correlation. This measure makes recognition harder for objects with low motion and is motivated by the statement of Kittler et al. [6], who argue it is very unlikely to have two different classifiers to yield the same error at the same time. Finally, the overall result of the classification triggers the corresponding functionality of the application, provided that it is stable for a minimum number of frames.

At present, the processing rate equals the PAL-framerate of 25 fps, when a resolution of 192 x 144 pixels and a Pentium-II 333 MHz machine are used. Tests with a reference set of 20 gestures yield a recognition rate of 90,3 %. Even 98,1 % are reached with the set of 6 gestures, that is utilised in the application (see figure 4). Although the correlation classifiers performance is not as accurate as the maximum likelihood method, the combination of both showed to be reasonable, since error rates below 1 % are obtained for both gesture sets.

5 The Application

The number of information systems in automobiles is permanently increasing. Examples for this are traffic message memories in car radios that allow the driver to receive the messages on demand and other kinds of telematic systems. Gesture control is a new possibility to operate them.

Our application is a gesture controlled storage for acoustic messages. Incoming messages of different categories like traffic, email, answering machine etc. are stored. The driver can control message playback by performing gestures and gets speech output via car speakers and textual information via a display. The control concept is derived from usual radios, because it is generally known and thus intuitively understandable. Figure 4 gives an overview of the utilised gestures and the corresponding functions.

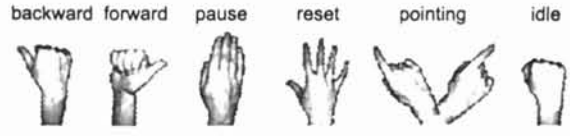

Figure 4: gestures and functions

Archived entries are played back sequentially after activation. It is possible to skip messages by performing the 'forward' respectively the 'backward' gesture, whereby short execution skips one message and longer execution skips several messages. Alternatively the user may use the 'pointing' gesture (left or right). Playback mode can be interrupted by the 'pause' gesture and resumed by the 'forward' gesture, while the 'reset' gesture quits it. The 'idle' gesture represents the typical resting position of the hand. Therefore it is used as a garbage model, which means it is recognised but has no function.

Because interaction with the system is supposed to be possible without eye contact, an orientation aid is given via speech output by announcing the position number of the message. Additional graphical information on a display is for short control glances and for familiarising inexperienced users with the system. Figure 5 shows the graphical output. It has a caption and a small window aside it to indicate the current hand position in the gesturing area. Below is a text field containing short message info. The icon bar at the bottom shows the gestures and functions, that are available according to the current state. E.g. if the playback is paused, the icon for the 'pause' gesture is not visible. In case a known gesture has been recognised, it's symbol is shortly highlighted to affirm the input.

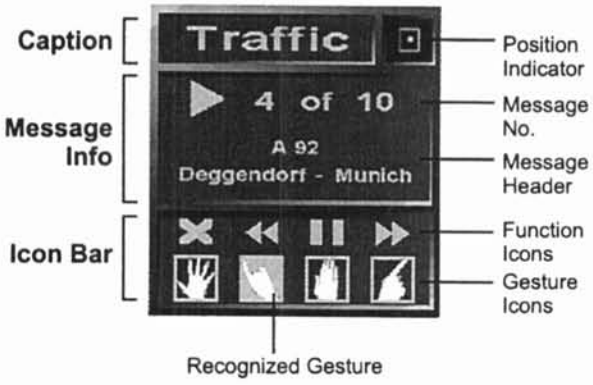

Figure 5: visual output

6 Discussion and Outlook

The described recognition system proves to be well suited to the given task. Tests with an experimental vehicle yield a permanent user independent functionality despite to varying lighting conditions like day and night or tunnel passing. Practically no recognition errors are encountered, even if the co-driver reaches his hand into the gesturing area. Recognition is not reliable any more if the segmentation is erroneous. Reasons can be direct sunlight or overlapping objects. Furthermore, the interior material of the car may not have NIR-reflecting characteristics, otherwise the contrast between hand and background would certainly be insufficient. The portrayed problems could be solved by using distance images, which can be acquired by a laser scanner or structured light technique.

The interaction with the system is rated as very natural and intuitive by the subjects. This originates from combining the chosen features, that are robust against intense variations of gesture performance, with the immediate system reaction as a result of the high processing speed. Recognition of dynamic gestures instead of static hand postures would even increase the impression of naturalness. Therefore we prepared a classifier based on Hidden Markov Models, since their suitability to this task has been reported in different works [3][7][11].

Finally we can state, that the developed gesture recognition system convincingly documents the usefulness of gesture control in cars. The next step towards practical use is going to be a study regarding user acceptance and efficiency. Then the gesture recogniser will be combined with a speech recogniser and mechanical input devices to create a novel multimodal input concept. The dedicated application is a multimedia car information system where each modality will be used according to it's suitability.

References

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

[2] C. Cedras, M. Shah. Motion-Based Recognition: A Survey. Image and Vision Computing, vol. 13, pp. 135-145, 1995.

[3] H. Hienz, K.-F. Kraiss, B. Bauer. Continuous Sign Language Recognition using Hidden Markov Models. 2nd Int. Conference on Multimodal Interfaces, Hong Kong, pp.IV10-IV15, 1999.

[4] M.-K. Hu. Visual Pattern Recognition by Moment Invariants. IRE Transactions on Information Theory, vol. IT8, pp. 179187, Feb. 1962.

[5] K. Imagawa, S. Lu., S. Igi. Color-Based Hands Tracking System for Sign Language Recognition. 3rd Int. Conference on Automatic Face and Gesture Recognition, Japan, pp. 462-467, 1998.

[6] J. Kittler, M. Hatef, R.P.W. Duin, J. Matas. On Combining Classifiers. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.20, No.3, pp. 226-239, 1998.

[7] S. Müller, S. Eickeler, G. Rigoll. Crane Gesture Recognition Using Pseudo 3-D Hidden Markov Models. 4th Int. Conf. on Automatic Face and Gesture Recognition, France , pp. 398-402, 2000.

[8] V. Pavlovic, R. Sharma, T. Huang. Visual Interpretation of Hand Gestures for Human-Computer Interaction: A Review. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 19, no. 7, pp. 677-695, 1997.

[9] Y. Sato, Y. Kobayashi. Fast Tracking of Hands and Fingertips in Infrared Images for Augmented Desk Interface. 4th Int. Conference on Automatic Face and Gesture Recognition, France, pp. 462-467, 2000.

[10] Fa. Siemens. Product Information for Siemens Virtual Touchscreen-SiVit: http://www.atd.siemens.de/td_electronic/produkte/sivit/sivit.htm

[11] T. Starner, A. Pentland. Visual Recognition of American Sign Language Using Hidden Markov Models. International Workshop on Automatic Face and Gesture Recognition, Switzerland, pp. 189-194, 1995.

[12] J. Triesch, C.v.d. Malsburg. A Gesture Interface for Human-Robot-interaction. 3rd International Conference on Automatic Face and Gesture Recognition, Japan, pp. 546-551, 1998.

[13] Y. Zhu, H. Ren, G. Xu, X. Lin. Toward Realtime Human-Computer Intertaction with Continuous Dynamic Hand Gestures. 4th Int. Conference on Automatic Face and Gesture Recognition, France, pp. 544-549, 2000.